UNIVERSITÄT
BIELEFELD
Faculty of Business Administration
and Economics

Dissertation

Strength out of Weakness: Harnessing Information Gained from the Pair Structure of Composite Marginal Likelihood Estimation

Sebastian Büscher

Submitted in fulfilment of the requirements for the academic degree of
Doctor rerum politicarum (Dr. rer. pol.)
to the Faculty of Business Administration and Economics,
Bielefeld University

20 December 2024

Supervisor

Prof. Dr. Dietmar Bauer

Examiners

1. Prof. Dr. Dietmar Bauer
2. Prof. Dr. Roland Langrock

Acknowledgements

> "It's a dangerous business, [...] going out of your door. You step into the road, and if you don't keep your feet, there's no knowing where you might be swept off to."
>
> *J.R.R. Tolkien (1954)*

I would like to express my profound gratitude to several individuals who have provided invaluable support throughout my academic journey and without whom the completion of this thesis would not have been possible.

I would like to begin by expressing my profound gratitude to my principal advisor, Dietmar Bauer. His decision to take a chance on me, despite not knowing me or my work beforehand, marked the beginning of a journey that was to be filled with his academic guidance and encouragement. I am immensely grateful for the freedom he afforded me to explore my own ideas and for the counsel he provided, guiding my endeavour of their pursuit, which was enriched by his experience and vast academic knowledge. I am also appreciative of him entrusting me with the responsibility of leading the Regression Analysis lecture, which significantly contributed to my growth as an educator. This has, together with the invaluable experiences I gained from presenting at international conferences under his mentorship, speaking in front of leading academics in the field, greatly enriched my academic and personal growth and improved my communication skills. Working with him at the university has been a genuinely rewarding experience that extends beyond the academic realm.

Furthermore, I am grateful to Roland Langrock, who acted as my second examiner. His constructive criticism during the discussions following numerous presentations at Young Researchers Workshops has been instrumental in improving both the quality of my work and my presentation skills. His insights have revealed new perspectives on my research, the models I utilise, and their applications.

Financial support was gratefully received by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 356500581.

I would like to thank Manuel and Lennart for being exceptional office mates, colleagues, and intellectual sparring partners. Their companionship at various conferences and the stimulating discussions we had at the office have been invaluable to my work. I would be remiss if I did not extend this appreciation to the fellows from the extended W9 group who were instrumental in creating a lively and inspiring environment that helped make working here a truly enjoyable experience.

I am especially indebted to my parents for their continued support throughout my academic endeavours, and since well before this journey started. They enabled and encouraged me to follow my pursuit of knowledge far beyond the borders of Bielefeld, an experience that not only expanded my intellectual horizons but helped me grow as a human being.

Lastly, I would like to express my deepest gratitude to my partner, Kaja, for her unwavering support and patience, for always having my back and for always offering her honest opinion, providing me with an outside perspective when I got lost in the mist of details.

Thank you all for your support and contributions to my journey.

Contents

Acknowledgements iii

List of Abbreviations vii

List of Figures ix

List of Algorithms xi

List of Tables xiii

Overview of Research Papers and Statement of Contribution xv

1 Introduction 1

2 Visual Guidance for Model Specification: Introducing Score Plots for Discrete Choice Models 5

2.1 Introduction . . . 6
2.2 Score Contributions . . . 8
2.2.1 Scores and Residuals in Linear Regression . . . 8
2.2.2 Score Contributions of Discrete Choice Models . . . 10
2.3 Score Plots . . . 14
2.4 Estimation . . . 18
2.5 Synthetic Examples . . . 19
2.5.1 Non-Linear Utility - Quadratic Utility . . . 20
2.5.2 Non-Linear Utility - Steps and Bends . . . 23
2.5.3 Missing Variables . . . 32
2.5.4 Structural Breaks in Time - Parameter Shift . . . 35
2.5.5 Structural Breaks in Time - Parameter Split . . . 38
2.5.6 Auto-Regressive Error Structure . . . 40
2.6 Real-World Example . . . 42
2.6.1 Data Set Description . . . 42
2.6.2 Initial Model . . . 43
2.6.3 Score Plots and Model Selection . . . 44
2.6.4 Final Model and Model Comparison . . . 48
2.7 Conclusion . . . 50
Appendix 2.A . . . 52
Appendix 2.B . . . 54

3 Using Lagrange Multiplier Type Tests to Detect Structural Intra-Person Heterogeneity in CML Estimation in Panel Data Sets 61
3.1 Introduction . . . 62
3.2 CML pair gradient contributions . . . 68
3.3 Simulation evaluation of test properties . . . 75
3.3.1 Finite sample distribution of test statistic under the null hypothesis . . . 76
3.3.2 Detecting autoregressive errors . . . 76
3.3.3 Detecting structural breaks in the model . . . 78
3.4 Simulation Results . . . 79
3.4.1 Distribution under the null hypothesis . . . 79
3.4.2 Rejections rates under violation of the null hypothesis . . . 80
3.5 Conclusion . . . 83
Appendix 3.A . . . 85
Appendix 3.B . . . 86

4 Weighting strategies for pairwise CML estimation in case of unbalanced panels and unaccounted AR structure of the errors 97
4.1 Introduction . . . 98
4.2 Weighting in case of unbalanced panel data . . . 102
4.2.1 Two-Step Optimal Group-Weight CML Estimator . . . 103
4.2.2 Finite-sample simulation study . . . 110
4.3 Weighting in case of unaccounted autoregressive error structure . . . 114
4.3.1 Variance-covariance structure of autocorrelated errors . . . 115
4.3.2 Deriving the asymptotic bias in misspecified cases . . . 117
4.3.3 Finite-sample simulation . . . 119
4.3.4 Discussion . . . 122
4.4 Conclusion . . . 124
Appendix 4.A . . . 126
Appendix 4.B . . . 127
Appendix 4.C . . . 132
Appendix 4.D . . . 133

5 Conclusion and Outlook 135
5.1 Conclusion . . . 135
5.2 Outlook . . . 137

Bibliography 139

List of Abbreviations

ASC alternative specific constant

bME bivariate Mendell-Elston

CDF cumulative distribution function

CLAIC composite likelihood Akaike Information Criterion

CLBIC composite likelihood Bayesian information criterion

CLRT composite likelihood ratio test

CML composite marginal likelihood

DCM discrete choice model

DGP data generating process

DM decision-maker

DQ designed quadrature

FGLS feasible generalised least square

GLS generalised least square

iid independent, identically distributed

LM Lagrange multiplier

MACML maximum approximate composite marginal likelihood

ME Mendell-Elston

ML maximum likelihood

MNP multinomial probit

MNL multinomial logit

MMNL mixed multinomial logit

MSL maximum simulated likelihood

MSM method of simulated moments

MSS method of simulated scores

MVNCDF multivariate normal cumulative distribution function

PDF probability density function

QMC quasi-Monte Carlo

RUM random utility model

SGQ sparse grid quadrature

SJ Solow-Joe

TVBS two-variate bivariate screening

List of Figures

2.1 Illustrative score plots for a linear regression model. 10
2.2 Structure of score plots with illustrative quadratic data. 17
2.3 Score plots of data from a DGP with a quadratic utility function a model with linear utility function. 21
2.4 Visualisation of the similarities between score plots for a linear regression model and the diagonally sliced score plots for a discrete choice model with a quadratic utility function in the DGP. 23
2.5 Score plots of data from a DGP with a utility function exhibiting a non-linear dependency w.r.t. the covariate x. 26
2.6 Visualisation of the similarities between score plots for a linear regression model and the diagonally sliced score plots for a discrete choice model with a step function in the DGP. 27
2.7 Visualisation of the similarities between score plots for a linear regression model and the diagonally sliced score plots for a discrete choice model with a ReLU function in the DGP. 28
2.8 Visualisation of the similarities between score plots for a linear regression model and the diagonally sliced score plots for a discrete choice model with an absolute value function in the DGP. 31
2.9 Score plots of data which include variables x_1, x_2 and x_3 in the DGP, whilst solely including x_a and x_b in the estimated model. 34
2.10 Score plots of data with a shift in parameter values between panel waves. 37
2.11 Score plots of data with a split in parameter values in the population. 39
2.12 Score plots of data with an auto regressive error process in the DGP. . 41
2.13 Case study: Score plots for initial model +ASC. 44
2.14 Case study: Score plots for model + comfort level zero. 45
2.15 Case study: Score plots for model + polynomial dependence on `price`. 46
2.16 Case study: Score plots for model + dummy variable for more expensive alternative. 47
2.17 Case study: Score plots of the score contributions depending on choice id. 48
2.18 Additional score plots of data from a DGP with a quadratic utility function a model with linear utility function. 54
2.19 Additional score plots of data from a DGP with a utility function exhibiting a non-linear dependency w.r.t. the covariate x. 55
2.20 Heatmap score plots of data which include variables x_1 and x_2 in the DGP and in the model. 56

2.21 Additional score plots of data which include variables x_1, x_2 and x_3 in the DGP. . . . 57
2.22 Additional heatmap score plots of data with a shift in parameter values between panel waves. . . . 58
2.23 Additional heatmap score plots of data with a split in parameter values in the population. . . . 58
2.24 Additional heatmap score plots of data with an auto regressive error process in the DGP. . . . 59

3.1 CDFs of the test statistics for the joint test and the tests for the individual parameter components. . . . 93
3.2 Rejection rates of tests for different CML-pair groupings, autoregressive error structure . . . 94
3.3 Rejection rates of tests for different CML-pair groupings, shift in β_1 parameter . . . 95
3.4 Rejection rates of tests for different CML-pair groupings, split in β_1 parameter . . . 96

4.1 Boxplots of optimal weights compared to heuristic weights. . . . 112
4.2 Distribution of relative trace of variance-covariance matrix $\hat{V}_\theta$ of the weighted models. . . . 112
4.3 Visualisation of the asymptotic bias in β, ω^2, and σ^2, respectively, for the four different composite marginal likelihood (CML) pair structures. . . . 120
4.4 Visualisation of the four different CML pair structures applied for the estimation process. . . . 121
4.5 Percentage of successfully estimated models by autocorrelation coefficient ρ and CML pair-type. . . . 121
4.6 Distribution of the mean absolute deviation of the estimated β parameters from the true parameters. . . . 122
4.7 Distribution of mean estimated correlation $\tilde{\rho}$. . . . 122

List of Algorithms

4.1 Algorithm to calculate two-step optimal group-weight CML estimator 132
4.2 Algorithm to calculate two-step optimal group-weight CML estimator with parametric variation 132

List of Tables

2.1 Estimated parameters for the covariate effect from a model with an omitted covariate, compared to expected mediation effect. 34
2.2 Case study: Model selection criteria for different models. 47
2.3 Case study: Comparison of estimated parameters for first and final model. 49

3.1 Shares of componentwise tests with largest test statistic, autoregressive error structure, *FirstLast* grouping 87
3.2 Shares of componentwise tests with largest test statistic, autoregressive error structure, *NearFar* grouping 88
3.3 Shares of componentwise tests with largest test statistic, shift in β_1 parameter, *FirstLast* grouping 89
3.4 Shares of componentwise tests with largest test statistic, shift in β_1 parameter, *NearFar* grouping 90
3.5 Shares of componentwise tests with largest test statistic, split in β_1 parameter, *FirstLast* grouping 91
3.6 Shares of componentwise tests with largest test statistic, split in β_1 parameter, *NearFar* grouping 92

4.1 Ratio of the asymptotic variance for optimal weights versus unweighted case. 127
4.2 Share of `nlm()` exit codes by employed weighting scheme. 133
4.3 Overview of distribution of quotients between the traces of differently weighted models. 133
4.4 Distribution of average ratios between the optimal group weights depending on different initial weights. 134
4.5 Distribution of ratios between the traces of the variance-covariance matrix of the estimator depending on different initial weights. 134
4.6 Overview of distribution of relative mean squared l_2 distance of predicted choice probabilities to true choice probabilities. 134

Overview of Research Papers and Statement of Contribution

This dissertation is based on three research papers written during my tenure as a research associate at Bielefeld University. The thesis comprises an introductory chapter that provides background information on discrete choice modelling, composite marginal likelihood estimation and the R package `Rprobit`, which was developed in part during my tenure as a research associate in collaboration with Dietmar Bauer, Manuel Batram and Lennart Oelschläger (Chapter 1). This is followed by three research papers on composite marginal likelihood estimation (Chapters 2-4). The papers are presented in the order in which a practitioner would integrate their respective contributions and developments into a choice modelling workflow, instead of them being presented in their chronological order. The dissertation then concludes with a summary and an outlook on further research topics. The research papers are presented in their current state of publication and are written to be self-sufficient. Consequently, there may be some overlap in the contents of the papers and differences in notation between the three papers.
A short synopsis of the research papers, along with their current status and statements of contribution of the authors, will be provided here.

Paper 1: Visual Guidance for Model Specification: Introducing Score Plots for Discrete Choice Models (Chapter 2)

This paper was authored by Sebastian Büscher. It is available as a preprint on SSRN. [Büscher, 2024]

This paper introduces score plots, a novel diagnostic tool in the field of discrete choice modelling, able to pose as a visual guide during the model specification process. These score plots utilise score contributions from composite marginal likelihood (CML) pairs and are, in concept and interpretation, akin to residual plots commonly used in linear regression modelling.
The paper derives the conditional expectation of the score contribution, thereby establishing the theoretical foundation of score plots while drawing conceptual parallels to residuals and scores in the linear regression framework. The versatility, practical applicability and straightforward interpretation of these plots were demonstrated through the analysis of several synthetic data sets and a real-world case study. In the presented studies, using the plots, it was possible to successfully identify

(1) non-linearities in the utility function with respect to covariate values, (2) missing covariates, (3) structural breaks over time, and (4) dynamic error processes.
Given the often subjective nature of the model specification process in discrete choice modelling, these score plots offer the potential to enhance the model selection process by offering visual guidance to practitioners.

Sebastian Büscher was solely responsible for all aspects of the paper in its current form, including its initiation, conceptualisation, conducting literature research, developing the methodology and visualisations, carrying out the simulation studies and real-world case study, as well as authoring the paper.

Paper 2: Using Lagrange multiplier type tests to detect structural intra-person heterogeneity in composite marginal likelihood estimation in panel data sets (Chapter 3)

This paper was authored by Sebastian Büscher and Dietmar Bauer. It is under review at the *Journal of Choice Modelling* since August 2024 and is available as a preprint on SSRN.
[Büscher and Bauer, 2024]

This paper contributes to the field of discrete choice modelling by introducing a Lagrange multiplier type test for pooling groups of observational pairs in probit models. The test facilitates the detection of structural breaks, intra-personal heterogeneity, and dynamic effects, such as autoregressive errors, without the necessity of implementing or estimating a dynamic, unrestricted model.
The test is applicable to models estimated using composite marginal likelihood (CML) methods, leveraging the availability of score information at the level of pairs of observations, thereby enabling tests for intra-personal and time-dependent effects to be conducted. This contrasts with maximum likelihood estimation, where score information is only available at the level of individuals.
This paper presents a derivation of the asymptotic distribution of the test statistic under the null hypothesis. The distribution is validated for the finite sample setting through the use of simulation studies. Similar simulations were conducted to analyse the empirical size and power of the test in the context of various violations of the null hypothesis. Furthermore, component-wise versions of the tests can be employed to identify the parameters of the model most affected by the violation, thereby enhancing their practical applicability.

Sebastian Büscher initiated the paper, conceptualised it, authored the original draft, and conducted the simulation studies. All other efforts, including developing the methodology, conducting literature research, performing data analyses and visualisation, and editing the paper, were collaboratively undertaken by Sebastian Büscher and Dietmar Bauer.

Paper 3: Weighting strategies for pairwise composite marginal likelihood estimation in case of unbalanced panels and unaccounted autoregressive structure of the errors (Chapter 4)

This paper was authored by Sebastian Büscher and Dietmar Bauer. It is published in *Transportation Research Part B: Methodological, 181*.
[Büscher and Bauer, 2024]

This paper has a focus on leveraging the power weights in the pairwise CML function and contributes in two ways to the choice modelling literature.
Firstly, it introduces a two-stage optimal group-weights CML estimator for discrete choice probit models with unbalanced panel data. This estimator is derived from asymptotic properties of the weighted CML estimator for probit models. It utilises the gradient and Hessian contributions from individuals with varying numbers of observations and aims to minimise the variance of the estimator via the targeted utilisation of the power weights.
Secondly, the paper explores weighting schemes to address unaccounted autoregressive error structures in discrete choice probit models. It demonstrates that emphasising pairs of distant observations in CML margins via the power weights can mitigate bias introduced to the estimator in the presence of model misspecification, such as when errors follow an autoregressive process, which is not represented in the model.
Both contributions are founded upon rigorous theoretical derivations with respect to their asymptotic properties before demonstrating their practical usage in simulation studies.

Sebastian Büscher initiated the paper, conceptualised it, authored the original draft, and conducted the simulation studies. All other efforts, including developing the methodology, conducting literature research, performing data analyses and visualisation, and editing the paper, were collaboratively undertaken by Sebastian Büscher and Dietmar Bauer.

CHAPTER

1 Introduction

"We demand rigidly defined areas of doubt and uncertainty!"

Douglas Adams (1979)

Discrete choice models (DCMs) have become fundamental tools for analysing human decision-making since the pioneering works by Thurstone [1927] on probit models, Luce [1959] on logit models, and, perhaps most influentially, McFadden [1974] on the random utility model (RUM) framework. These models find extensive application across various fields, including transportation [Ben-Akiva and Lerman, 1985; Bhat and Koppelman, 2003; Büscher et al., 2019], marketing [Chintagunta, 1992; Albert and Chib, 1993; Kim et al., 2020], environmental economics [Louviere et al., 2000], experimental psychology [Johnson and Bruce, 1997; Berkowitsch et al., 2014], and behavioural neuroscience [Haghani et al., 2021]. In transportation science, discrete choice models are pivotal in predicting travel behaviour [Büscher et al., 2019] and assessing mode choice characteristics [Ben-Akiva and Morikawa, 1990], often informing high-stakes infrastructure investments [Short and Kopp, 2005].

DCMs are frequently based on the RUM framework, which underpins prominent models such as multinomial probit (MNP) and multinomial logit (MNL) models. Although probit models [Thurstone, 1927] preceded logit models [Luce, 1959], the MNL model has gained widespread popularity in the scientific community, in parts due to its closed-form solution of the choice probabilities. In contrast, when working

with MNP models, "computational difficulties arise when this model is applied to problems with more than a few alternatives, mainly because the calculation of choice probabilities involves evaluating multiple integrals" [The Royal Swedish Academy of Sciences, 2000]. The importance of discrete choice models gained widespread recognition when Daniel McFadden was awarded the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel in 2000 "for his development of theory and methods for analyzing discrete choice" [Nobel Prize Outreach, 2000].

While the MNL model's popularity is partly due to its numerical simplicity, the flexibility of MNP (and mixed multinomial logit (MMNL)) models has driven innovations in estimation methodologies. Traditionally[1], MNP and MMNL models have been estimated using simulation techniques such as maximum simulated likelihood (MSL), method of simulated moments (MSM) or method of simulated scores (MSS) [Train, 2009], as pioneered by Lerman and Manski [1981] and McFadden [1989]. These methods replace the exact probabilities required to calculate the likelihood, moments, or scores, respectively, with simulated probabilities derived from random draws from the underlying distributions to approximate the cumulative distribution function (CDF). However, these simulation methods encounter computational difficulties as the number of choice alternatives and occasions increases [Bhat, 2014]. Advances in quasi-Monte Carlo (QMC) methods [Bhat, 2003; Hess et al., 2006; Dick et al., 2016], sparse grid quadrature (SGQ) methods [Heiss and Winschel, 2008], and, more recently, designed quadrature (DQ) methods [Ryu and Boyd, 2015; Keshavarzzadeh et al., 2018; Bansal et al., 2022] have significantly reduced the computational time in MSL estimation, but they remain affected by the curse of dimensionality.

To address the computational challenges associated with multivariate normal cumulative distribution function (MVNCDF) evaluations in MNP models, Varin [2008] proposed the use of composite marginal likelihoods (CMLs). In contrast to maximum likelihood (ML) estimation, CML estimation replaces the likelihood function with a pseudo-likelihood where the log-probability of all observations of a decision-maker (DM) is replaced by the weighted sum of the log-probabilities of subsets (margins) of observations. This approach reduces the dimensionality of the MVNCDFs required for calculating choice probabilities, thereby substantially reducing the computational burden. The formulation of the CML allows the margins to represent any subset of choice observations from a DM. Most commonly, pairs of observations are selected as margins, sometimes combined with single observations [Cox and Reid, 2004], although higher order margins such as triplets are occasionally considered [Engler et al., 2005].

[1] Other estimation frameworks include non-parametric estimation [see, for example, Bauer et al., 2022] or Bayesian estimation, for which Oelschläger and Bauer [2024] developed the `RprobitB` package and have investigated within this framework preference classification [Oelschläger and Bauer, 2023] and the estimation of latent class mixed MNP models.

To further enhance computational efficiency, Bhat [2011] combined CML estimation with a deterministic approximation of the MVNCDFs, introducing the maximum approximate composite marginal likelihood (MACML) approach. The original MACML formulation employs pairwise CML combined with an approximation of the MVNCDFs based on the work of Solow [1990] and Joe [1995], hence known as the Solow-Joe (SJ) approximation. Subsequent works have proposed other approximations, including the Mendell-Elston (ME) approximation [Mendell and Elston, 1974] and improvements such as bivariate Mendell-Elston (bME) [Trinh and Genz, 2015] or two-variate bivariate screening (TVBS) proposed by Bhat [2018], among other new variants.

In this dissertation, the MACML approach utilising pairwise margins is employed for the estimation of all choice models presented. In particular, the R package `Rprobit` was employed for the computations, which implements the MACML approach in R with CML function evaluations performed in C++ for enhanced computational efficiency. The development of this codebase by Dietmar Bauer, Manuel Batram, Sebastian Büscher[2], and Lennart Oelschläger was pivotal for the findings presented in this thesis. In turn, the work presented here has contributed to the advancement of the R package. With regard to MACML and the `Rprobit` package, Batram and Bauer [2016, 2019] examined the asymptotic properties of the resulting estimator in terms of consistency and asymptotic normality. Additionally, they discussed model selection criteria and results on model averaging in the MACML context [Batram and Bauer, 2017].

Notwithstanding the progress made in discrete choice models and CML estimation, a number of issues remain unresolved. Firstly, the model selection and utility specification process is often considered to be subjective, resulting in highly variable modelling outcomes that are dependent on the decisions of the researcher [Paz et al., 2019; van Cranenburgh et al., 2022; Rodrigues et al., 2022; Nova et al., 2024]. This dependence of modelling results on the subjective decisions of researchers is a phenomenon known as "researcher degrees of freedom" in the field of psychology [Simmons et al., 2011; Gelman, Andrew and Loken, 2013]. Implementing methods to guide model specification and employing computationally efficient statistical tests for model discrimination may serve to reduce this variability, thereby enhancing the credibility of model-based policy recommendations [Short and Kopp, 2005]. Secondly, the utilisation of CML estimation in place of ML estimation results in an increase of the estimator variance, given that ML estimation is statistically efficient, whereas CML is not. A reduction in estimator variance would therefore facilitate the

[2]Contributions to the `Rprobit` package that are not directly part of this thesis include, but are not limited to, the implementation of the analytical second derivative of the SJ approximation, the implementation of the TVBS approximation and its analytical derivative, parallelisation of the code, automatic normalisation procedures, and various contributions to the general package structure.

application of CML estimation. Thirdly, the implementation of dynamic error processes and complex dependencies between observations is challenging, often leading researchers to prefer simpler models that introduce correlation through mixed effects. However, the use of misspecified models can introduce a bias in the estimated parameters. Therefore, the development of estimation techniques that are capable of mitigating the effects of misspecification would be highly beneficial.

This thesis addresses these issues by exploring the strategic utilisation of the power weights of CML pairs and the informational value of scores, extending their utility beyond model estimation and the estimation of the Fisher information matrix. The dependence between scores and covariates has been identified as a valuable area for investigation in other model types and estimation methods, including partitioning strategies for random forests [Schlosser et al., 2019]. While ML estimation provides score information at the individual level, CML estimation enables the extraction of more granular score information at the level of observational pairs. This allows for a detailed analysis of the manner in which score contributions depend on covariate values and other factors, offering insights into the estimated model and the data generating process (DGP). The utilisation of the power weights, on the other hand, allows for the usage of higher weights for groups of individuals and observations with a low estimated variance, and for emphasising pairs of observations that are less affected by a potential misspecification of the model.

In the following chapters, we will (a) develop diagnostic plots based on the score contribution of pairwise CML margins (Chapter 2), (b) develop Lagrange multiplier-type tests for pooling of pairs of observations (Chapter 3), (c) develop a two-step optimal group-weight estimator for unbalanced panel data (Chapter 4), and (d) investigate weighting strategies to mitigate the effects of unaccounted autocorrelation of the errors (Chapter 4).

These contributions to the field of choice modelling can be integrated into the modelling workflow to guide the model specification process, detect model misspecifications related to dynamic error processes, reduce estimation bias in such cases, and decrease estimator variance in unbalanced panel data scenarios.

CHAPTER

Visual Guidance for Model Specification: Introducing Score Plots for Discrete Choice Models

2

Abstract

Discrete choice models (DCMs) are widely employed to analyse decision-making processes across a range of disciplines, with a particular prevalence in transportation planning, where they provide the basis for predictions that inform high-stakes infrastructure investments and policy decisions. However, the often subjective nature of the utility function specification process introduces variability in model outcomes, thereby undermining the reliability of forecasts and potentially leading to discrepancies between data-driven recommendations and stakeholder decisions.

This paper addresses these challenges by introducing *score plots*, a diagnostic visualisation tool for models estimated using composite marginal likelihood (CML) methods. Drawing conceptual parallels to residual plots in linear regression, score plots employ pairwise score contributions to detect model misspecifications, thereby providing a guided approach to enhance model fit and consequently reduce subjectivity in the model selection process. While the requirement to use CML estimation may appear restrictive, this is only necessary for the model specification phase.

For the final model, maximum likelihood (ML) estimation may be used to harness its statistical efficiency.

The potential of score plots to identify issues such as omitted non-linearities, missing variables, structural breaks, and dynamic error processes is demonstrated through their application to both synthetic data sets and a real-world transportation case study. The integration of score plots into DCM workflows enables analysts to more effectively navigate the intricacies of model specification, thereby enhancing the reliability of predictions and improving the transparency of the model specification process, thus reinforcing the credibility of data-driven recommendations.

Keywords: discrete choice models, diagnostic plots, score plots, modelling workflow, probit modelling, composite marginal likelihood

2.1 Introduction

Since the seminal formulation of probit models by Thurstone [1927], logit models by Luce [1959], and the influential works by McFadden [1974] on the random utility model (RUM) framework, discrete choice models (DCMs) have become pivotal tools for analysing human decision-making across a range of fields, including transportation [Ben-Akiva and Lerman, 1985; Bhat and Koppelman, 2003; Büscher et al., 2019], marketing [Chintagunta, 1992; Albert and Chib, 1993; Kim et al., 2020], environmental economics [Louviere et al., 2000], experimental psychology [Johnson and Bruce, 1997; Berkowitsch et al., 2014], and behavioural neuroscience [Haghani et al., 2021]. However, a primary challenge in applied choice modelling is the specification of the utility function, which frequently involves iterative trial-and-error adjustments guided by domain expertise and subjective decisions regarding covariate inclusion [Paz et al., 2019; van Cranenburgh et al., 2022; Rodrigues et al., 2022]. This subjective nature of model specification often results in DCMs being seen as "partly science and partly art"[Páez and Boisjoly, 2022].

The impact of subjective modelling choices has been empirically demonstrated in a recent study by Nova et al. [2024] by utilizing a serious game setting. In their study, substantial variability in results among practitioners analysing the same data set was observed, which they attribute in parts to differences in the modelling workflow. This phenomenon is echoed in other fields, such as neuroimaging, where Botvinik-Nezer et al. [2020] documented similar variability in a crowd science experiment, and in psychology, where the concept of "researcher degrees of freedom" is well established and viewed critically [Simmons et al., 2011; Gelman, Andrew and Loken, 2013]. Reliable and objective model specification is especially crucial in

fields such as transportation, where DCMs inform long-term predictions of traffic patterns for high-stakes infrastructure investments. High variance in model predictions and ambiguity in the model selection process can rise suspicion with regards to the reliability of planning outcomes, leading to discrepancies between the data driven expert recommendations, informed by DCMs, and the policy decisions taken by stakeholders [Short and Kopp, 2005], highlighting the necessity of reliant models and a transparent model selection process.

Within DCMs, variability in model outcomes is particularly influenced by the specification of the utility function, where the exponential growth in possible combinations of covariates renders exhaustive model searches infeasible as data sets grow larger [Paz et al., 2019; van Cranenburgh et al., 2022; Nova et al., 2024]. One frequently underexplored area in DCM utility specification is the incorporation of non-linear transformations, which significantly expand the space of possible models [Mariel et al., 2021; Nova et al., 2024], though the value of including quadratic dependency of the utility function on covariate values was already recognised by Adamowicz et al. [1998]. While some methods for (semi-)automatic variable selection exist [see, for example, Paz et al., 2019; Ortelli et al., 2021; Rodrigues et al., 2022], visual diagnostic tools, widely valued in linear regression modelling in the form of residual plots [compare Fahrmeir et al., 2013, Chap. 3, Wooldridge, 2015, Chap. 6, and Wickham et al., 2023, Chap. 10], remain underutilized in DCM. This is evidenced by the scarceness of emphasis on visualisation tools and diagnostic plots in the reference literature for DCMs [see, for example, Ben-Akiva and Lerman, 1985; Louviere et al., 2000; Train, 2009; Hess and Daly, 2014] compared to the literature on linear regression. Both, Páez and Boisjoly [2022] and Nova et al. [2024], highlight that visual tools can help analysts explore data patterns, diagnose model fit, and potentially reduce variability in model outcomes and improve model fit by informing the model selection process.

A notable exception is the work of Mariel et al. [2021], who introduce a visualisation tool to detect non-linearities in the utility function with respect to covariate values. Their approach involves dummy coding numerical covariates into attribute levels and plotting the corresponding estimated parameters to identify non-linear effects. However, this method requires a substantial modification of the estimated model, necessitating the inclusion of multiple additional parameters for each numerical covariate, which may potentially influence parameter estimation.

In this paper, we introduce an alternative to residual plots – *score plots* – designed to provide diagnostic insights across a broad range of models estimated using CML estimation. By focusing on pairwise score contributions, score plots offer a diagnostic method applicable to a wide range of models estimated through gradient-based methods, with an emphasis on multinomial probit (MNP) models in our data

examples. The value of utilising information obtained from analysing the dependency between scores and covariate values has also been recognised for other types of models and modelling means, such as for partitioning strategies for random forests [Schlosser et al., 2019]. Whilst pairwise CML gradient information is required for the construction of the plots proposed in this paper, this is only required during the model specification phase and, if feasible, maximum likelihood estimation can be used for the estimation of the final model to profit from its statistical efficiency.

The structure of the paper is as follows: Section 2.2 establishes foundational concepts and statistical properties of score contributions, first in the linear regression framework whilst drawing parallels to residual plots, secondly in the DCM setting; based on the established statistical properties of the score contributions, Section 2.3 introduces score plots, including recommendations for data preprocessing; Section 2.5 applies score plots to detect model misspecifications in synthetic data examples; Section 2.6 demonstrates the method on a real-world transportation data set; and Section 2.7 discusses the results and broader implications for model diagnostics, gives recommendations for integrating score plots into DCM workflows, and concludes with possible further avenues of research with respect to score plots.

2.2 Score Contributions

In this section, we will revisit the concept of scores for the classical normal regression model, with the aim of providing a more intuitive understanding and relating scores to residuals, a concept with which the majority of researchers are more familiar. In the second part of this section, we will then proceed to derive the properties of the scores of DCMs, conditional on the values of the covariates.

2.2.1 Scores and Residuals in Linear Regression

The classical normal regression model [see, for example, Fahrmeir et al., 2013] can be expressed in the form

$$y = X\beta + \varepsilon, \tag{2.1}$$

where y is the dependent variable, $X \in \mathbb{R}^{N \times J}$ is the regressor matrix, $\beta \in \mathbb{R}^J$ is the parameter vector, and the errors follow a normal distribution $\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$. This

leads to the log-likelihood function

$$ll_{ML}(\beta, \sigma^2 \mid X, y) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i'\beta)^2, \tag{2.2}$$

and the corresponding score function w.r.t. the β parameters

$$s^{(j)}(\beta, \sigma^2 \mid X, y) := \frac{\partial}{\partial \beta^{(j)}} ll_{ML}(\beta, \sigma^2 \mid X, y) = \frac{1}{2\sigma^2}\sum_{i=1}^{N} x_i^{(j)}(y_i - x_i'\beta), \tag{2.3}$$

with $s(\beta, \sigma^2 \mid X, y) = (s^{(1)}(\beta, \sigma^2 \mid X, y), \ldots, s^{(J)}(\beta, \sigma^2 \mid X, y))'$. The score contribution of observation i is thus given by

$$s_i^{(j)}(\beta, \sigma^2 \mid X, y) = \frac{1}{2\sigma^2} x_i^{(j)}(y_i - x_i'\beta) = \frac{1}{2\sigma^2} x_i^{(j)}\hat{\varepsilon}_i. \tag{2.4}$$

Therefore, the score contribution of observation i reflects the residual $\hat{\varepsilon}_i = y - x_i'\hat{\beta}$ multiplied by the covariate vector x_i, and thus the score direction w.r.t. the intercept parameter $\beta^{(0)}$ is equivalent to the residual $\hat{\varepsilon}_i$.

For a correctly specified model, where the assumptions of a classical normal regression model [Wooldridge, 2015; Fahrmeir et al., 2013] are met, the conditional expectation of the scores, given the covariates, is zero:

$$\mathbb{E}\left(s_i^{(j)} \mid X, y\right) = \mathbb{E}\left(\frac{1}{2\sigma^2} x_i^{(j)}\hat{\varepsilon}_i \mid X, y\right) = \frac{1}{2\sigma^2} x_i^{(j)}\mathbb{E}\left(\hat{\varepsilon}_i \mid X, y\right) = 0. \tag{2.5}$$

This follows from the assumption of zero conditional expectation of the errors, $\mathbb{E}(\varepsilon_i \mid X, y) = 0$. However, if this assumption is violated, for example in the case of a quadratic dependence of y on x, this result does not hold. In such cases, a plot of the scores conditional on covariate values can be used to detect this violation.

To illustrate this, the plots in Figure 2.1 display the score contributions of a linear model

$$y_n = \beta^{(0)} + \beta^{(1)} x_n + \varepsilon_n, \tag{2.6}$$

estimated on 50 observations from the data generating process (DGP)

$$y_n = \beta^{(0)} + \beta^{(1)} x_n + \beta^{(2)} x_n^2 + \varepsilon_n, \tag{2.7}$$

with $\varepsilon_n \sim \mathcal{N}(0, 1)$, $\beta^{(0)} = 1$, $\beta^{(1)} = 1$, and $\beta^{(2)} = 0$ for two panels on the left of the figure and $\beta^{(2)} = 1$ for the panels on the right.

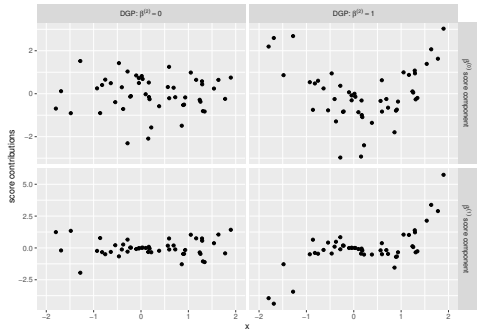


Figure 2.1: Illustrative score plots for a linear regression model. The estimated models include only a linear dependency of x with parameter $\beta^{(1)}$. The data from two different DGPs is presented, differing in $\beta^{(2)}$, the effect of x^2, as indicated above the panels.

2.2.2 Score Contributions of Discrete Choice Models

In the formulation of DCMs, the RUM framework is frequently utilised, wherein the assumption is made that to each alternative $j \in \{1, \ldots, J\}$ and choice occasion $n \in \{1, \ldots, N\}$, a latent random utility

$$U_{n,j} = V_{n,j} + \varepsilon_{n,j}, \tag{2.8}$$

is assigned, where $\varepsilon_{n,j}$ denotes the random component of the utility and $V_{n,j}$ denotes the deterministic component. In the classical approach, a linear form is assumed for the deterministic utility

$$V_{n,j} = X_{n,j}\beta, \tag{2.9}$$

with $\beta \in \mathbb{R}^{K\times 1}$, and $X_{n,j} \in \mathbb{R}^{1\times K}$. In the RUM framework, it is then assumed that the observed choice y_n represents the alternative with the highest random utility, and therefore $y_n = \arg\max_j U_{n,j}$. The distribution of the random errors $\varepsilon_n = (\varepsilon_{n,1}, \ldots, \varepsilon_{n,J})'$ plays a crucial role in the formulation of DCMs, as distributional assumptions give rise to different model families, such as the multinomial logit (MNL), or the MNP [see, for example, Train, 2009]. In the context of this paper, the focus is on MNP models, which imply the assumption that the errors are jointly normally distributed.

Cross-Sectional Discrete Choice Models

In the case of cross-sectional data, or panel data where independence between the observations of one decision-maker (DM) can be assumed, the errors $\varepsilon_{n,j}$ can be correlated across alternatives j but are assumed to be independent across observations n. Let $f(\varepsilon_{n,j};\sigma_\varepsilon)$ now denote the joint probability density function (PDF) of the errors $\varepsilon_{n,j}$, parametrised by $\sigma_\varepsilon \in \mathbb{R}^{K_\varepsilon \times 1}$, then the choice probability of alternative i being chosen by DM n is

$$\begin{aligned} P_{n,i}(\theta) &= \Pr(U_{n,i}(\beta) > U_{n,j}(\beta)\ \forall j \neq i;\sigma_\varepsilon) \\ &= \Pr(\varepsilon_{n,j} - \varepsilon_{n,i} < V_{n,i}(\beta) - V_{n,j}(\beta)\ \forall j \neq i;\sigma_\varepsilon) \\ &= \int_{\varepsilon_n} I(\underbrace{\varepsilon_{n,j} - \varepsilon_{n,i}}_{\tilde{\varepsilon}_{n,j,i}} < \underbrace{V_{n,i}(\beta) - V_{n,j}(\beta)}_{\tilde{V}_{n,j,i}(\beta)}\ \forall j \neq i) f(\varepsilon_n;\sigma_\varepsilon)\, d\varepsilon_n \\ &= \int_{\tilde{\varepsilon}_{n,i}} I(\tilde{\varepsilon}_{n,j,i} < \tilde{V}_{n,j,i}(\beta)\ \forall j \neq i) \tilde{f}(\tilde{\varepsilon}_{n,i};\sigma_\varepsilon)\, d\tilde{\varepsilon}_{n,i}, \end{aligned} \tag{2.10}$$

where $\tilde{f}$ denotes the PDF of the error differences $\tilde{\varepsilon}_{n,i} = (\tilde{\varepsilon}_{n,1,i},\ldots,\tilde{\varepsilon}_{n,i-1,i}, \tilde{\varepsilon}_{n,i+1,i},\ldots,\tilde{\varepsilon}_{n,J,i})' = (\varepsilon_{n,1} - \varepsilon_{n,i},\ldots,\varepsilon_{n,i-1} - \varepsilon_{n,i},\ \varepsilon_{n,i+1} - \varepsilon_{n,i},\ldots,\varepsilon_{n,J} - \varepsilon_{n,i})'$, and $\theta = (\beta,\sigma_\varepsilon)' \in \mathbb{R}^{(K+K_\varepsilon)\times 1}$ the joint parameter vector parametrising the model.

Assuming independence of the errors between observations, the log-likelihood function can then be written as

$$\begin{aligned} ll_{ML}(\theta;X,y) &= \log\left(\prod_n \prod_i P_{n,i}(\theta)^{I(y_n=i)}\right) \\ &= \sum_n \sum_i I(y_n = i)\log(P_{n,i}(\theta)) = \sum_n \log(P_{n,y_n}(\theta)), \end{aligned} \tag{2.11}$$

where $P_{n,y_n}(\theta)$ denotes the predicted probability of choice occasion n, depending on the parameter vector θ.

The score w.r.t. the k-th model parameter $\theta^{(k)}$ can thus be expressed as

$$\frac{\partial}{\partial\theta^{(k)}} ll_{ML}(\theta;X,y) = \sum_n \frac{\partial}{\partial\theta^{(k)}} \log\left(P_{n,y_n}(\theta)\right) = \sum_n \frac{\frac{\partial}{\partial\theta^{(k)}} P_{n,y_n}(\theta)}{P_{n,y_n}(\theta)}. \tag{2.12}$$

Definition 2.1. *Let $P_{n,y_n}(\theta)$ be the predicted choice probability of the chosen alternative $y_n \in \{1, \dots, J\}$ of choice occasion n, depending on the parameter vector $\theta \in \mathbb{R}^{(K+K_\varepsilon)}$. Then, the score contribution of observation n in the score direction k is defined as*

$$s_n^{(k)}(\theta; X_n, y_n) := \frac{\frac{\partial}{\partial \theta^{(k)}} P_{n,y_n}(\theta)}{P_{n,y_n}(\theta)}. \tag{2.13}$$

For the expected scores, conditional on the covariate values X_n, we get the following result.

Lemma 2.1. *Let θ_0 be the true parameter vector governing the DGP, and let $\hat{\theta}$ be an estimator of θ_0 such that $P_{n,i}(\hat{\theta})$ is a consistent estimator of $P_{n,i}(\theta_0)$. Furthermore, assume the probabilities for all observed choice occasions n and all choice alternatives i to be bounded from below ($P_{n,i}(\theta) > \delta \; \forall n, i$ for some $\delta > 0$) and that the gradients of the probabilities w.r.t. θ are bounded ($\left|\frac{\partial}{\partial \theta^{(k)}} P_{n,i}(\theta)\right| < M$ for some $0 < M < \infty$). Then the score contributions conditional expected value will converge to zero as the number of observations N goes to infinity, so*

$$\mathbb{E}\left(s_n^{(k)}(\hat{\theta}; X_n, y_n) \mid X_n\right) \overset{N\to\infty}{\longrightarrow} 0. \tag{2.14}$$

Please refer to Appendix 2.A for the proof.

Panel Discrete Choice Models

For panel data sets, where a number of T choice occasion are recorded for each DM n[1], the errors $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \dots, \varepsilon_{n,t,J})'$ can be correlated across observations t from one DM by being drawn from a joint distribution. An alternative approach is to assume that the errors $\varepsilon_{n,t}$ are independent across observations and correlation between observations of one decision maker can then be introduced by the usage of random parameters. This assumes that for each DM n a latent random parameter β_n is drawn from a given distribution [see, for example, Train, 2009]. In either case, the joint probability of all observations $y_n = (y_{n,1}, \dots, y_{n,T})'$ of one DM cannot be expressed as the product of the individual choice probabilities. Consequently, the joint probability of all observation of one DM is given by

$$P_{n,y_n}(\theta) = \Pr(U_{n,1,y_{n,1}} > U_{n,1,j} \; \forall j \neq y_{n,1}, \dots, U_{n,T,y_{n,T}} > U_{n,T,j} \; \forall j \neq y_{n,T}; \theta). \tag{2.15}$$

[1] In general, each DM n might face a different number of choice occasions T_n, for sake of readability we will, however, w.l.o.g. assume a balanced panel data set with $T_n = T \forall n$.

After normalisation for utility differences, the dimension of the integral required to solve for this joint probability is $T(J-1)$, compared to $J-1$ for the cross-sectional case, increasing in turn the computational burden significantly. The resulting log-likelihood function is then

$$ll_{ML}(\theta; X, y) = \sum_n \log\left(P_{n,y_n}(\theta)\right). \tag{2.16}$$

An alternative to ML estimation is the usage of pairwise CML estimation or its advancements [see, for example Varin, 2008; Varin et al., 2011; Bhat, 2011]. In this approach, the joint probability of all observations from one DM is replaced by the (power weighted) product of all possible pairs of observations $(a, b) \in \{1, \ldots, T\}^2$ of that DM. The resulting CML log-likelihood can then be expressed as

$$\begin{aligned} ll_{CML}(\theta; X, y) &= \sum_n \sum_{1 \le a < b \le T} w_{n,a,b} \log\Big(\Pr(U_{n,a,y_{n,a}} > U_{n,a,j}\ \forall j \neq y_{n,a}, \\ &\qquad\qquad U_{n,b,y_{n,b}} > U_{n,b,j}\ \forall j \neq y_{n,b}); \theta\Big) \\ &= \sum_n \sum_{1 \le a < b \le T} w_{n,a,b} \log(P_{n,a,b}(\theta)), \end{aligned} \tag{2.17}$$

where $w_{n,a,b}$ denotes the power weight of the pair of observations (a, b) of DM n and $P_{n,a,b}$, defined by the second equation, is the joint probability of the pair of choice observations (a, b) of DM n. This reduces the dimension of the integral required to be solved to $2(J-1)$, of which at most $T(T-1)/2$ have to be solved for each DM, depending on the power weights $w_{n,a,b}$. This can result in a notable reduction in the computational burden, although it comes at the loss of the statistical efficiency of the ML estimator.[2] Nevertheless, the increase in computational speed can be considerable enough to make the CML estimator a preferred option over ML estimation, despite the advancements in computational power that have occurred over the past decades [see, for example Delporte et al., 2025].

Similarly to the score contribution of individual observations in the cross-sectional case (Definition 2.1), we can also define the score contributions of pairs of observations for panel data models estimated using a pairwise CML approach.

Definition 2.2. *Let $P_{n,a,b}(\theta)$ be the predicted choice probability of the chosen alternatives $(y_{n,a}, y_{n,b}) \in \{1, \ldots, J\}^2$ of a pair of choice occasions $(a, b) \in \{1, \ldots, T\}^2$ of DM n, depending on the parameter vector $\theta \in \mathbb{R}^{(K+K_\varepsilon)}$. Then, the score contribution of the pair of*

[2] Büscher and Bauer [2024] have shown, however, how variance optimal power weights can be used to partially mitigate this loss in efficiency for unbalanced panel data sets.

observations (a, b) of DM n in the score direction k is defined as

$$s^{(k)}_{n,a,b}(\theta; X_n, y_n) = \frac{\frac{\partial}{\partial \theta^{(k)}} P_{n,a,b}(\theta)}{P_{n,a,b}(\theta)}. \tag{2.18}$$

By interpreting any pair of choices as a single choice between J^2 choice alternatives, the results of Lemma 2.1 can be directly expanded to the score contributions from a pair of observations.

Lemma 2.2. *Let θ_0 be the true parameter vector governing the DGP, and let $\hat{\theta}$ be an estimator of θ_0 such that $P_{n,a,b}(\hat{\theta})$ is a consistent estimator of $P_{n,a,b}(\theta_0)$. Furthermore, assume the probabilities for all observed pairs of choice occasions (a, b) for all DMs n and all pairs of choice alternatives (i, i') to be bounded from below ($\Pr(y_{n,a} = i, y_{n,b} = i'; \theta, X_{n,a,b}) > \delta \ \forall n, i, i'$ for some $\delta > 0$) and that the gradients of the probabilities w.r.t. θ are bounded ($\left|\frac{\partial}{\partial \theta^{(k)}} P_{n,a,b}(\theta)\right| < M$ for some $0 < M < \infty$). Then the score contributions conditional expected value converges to zero as the number of observations TN goes to infinity, so*

$$\mathbb{E}\left(s^{(k)}_{n,a,b}(\theta; X_n, y_n) \mid X_n\right) \overset{TN \to \infty}{\longrightarrow} 0. \tag{2.19}$$

The proof is a direct consequence of the results presented in Lemma 2.1.

2.3 Score Plots

In light of the conditional zero expectation of the score contributions (Lemmas 2.1 and 2.2), we will now proceed to construct diagnostic plots based on the score contributions of pairs of observations, which we will refer to as *score plots*.

Given that panel data models with correlation between observations estimated using ML estimation offer only score information at the level of DMs, we will focus on models estimated using pairwise CML estimation, as these offer score information from pairs of observations. This allows us to plot covariate values on the x- and y-axes and utilise a third visual channel [see Healy, 2018] to add information on the score contributions to the plot. In particular, we will be mapping information on the score contributions to the colour of elements within the plot.

In order to enhance the proposed plots, we will introduce some transformations of the raw score data before plotting it. In this section, we will provide an overview of the concepts underlying these transformations, before defining them in greater detail in Definition 2.3.

To enhance the interpretability of the score information, we will standardise the score contributions of the pairs of observations for each score direction separately. This will be achieved by subtracting the average score and then dividing by the empirical standard deviation of the score contributions. The resulting standardises scores will be referred to as *relative scores* (rel. score). The application of this standardisation procedure results in the unification of the different score directions on a single scale. A value of one for the relative score indicates a deviation of plus one standard deviation from the average score contribution of that score direction.

Given our interest in the expected score contribution conditional on variable values from covariates or the temporal position of the observations, we will calculate summary statistics of the score contributions. In the case of discrete variables, this is a straightforward process for any combination of variable values x_a and x_b, where x_a denotes the variable value of the first observation and x_b that of the second observation of the pair (a, b). In the case of continuous variables, binning will be employed to discretise the variable values prior to the calculation of the summary statistics. In either case, it is preferable to have observations of different choice alternatives being chosen in each data bin used for the plots, as this allows for the averaging of score contributions across different choice alternatives. As summary statistics of the score contributions employed in the score plots, we will mainly focus on two: (1) The mean relative scores ($\overline{\text{rel. score}}$) across all observation pairs within one data bin. This summary statistic provides insight into the magnitude and direction of any potential bias in the expected scores. (2) A summary statistic, designated as "score stat", calculated as the mean relative score multiplied by the square root of the number of pairs within the respective data bin. In case of zero conditional expected score contribution and due to the scaling of the relative scores by the empirical standard deviation of the scores, the score stat would have zero conditional expectation and unit standard deviation. The selection of an appropriate colour scale for the plots enables the interpretation of the plots in a manner analogous to statistical tests. In our cases, a colour scale has been employed which visualises whether the average relative score within a data bin diverges from zero by a margin exceeding two standard deviations. It should be noted, however, that this does not constitute a statistical test, given that the score contributions cannot be assumed to exhibit constant variance across different covariate values.

Definition 2.3. *For given score information* $s^{(k)}_{n,a,b}(\theta; X_n, y_n)$ *from a DCM estimated using a pairwise CML estimator, define the* mean score*, the* empirical standard deviation of the scores*, and the* relative score *of the score direction* k *as*

$$\overline{s^{(k)}}(\theta; X, y) := \frac{\sum_n \sum_{1 \le a < b \le T} s^{(k)}_{n,a,b}(\theta; X_n, y_n)}{\sum_n \sum_{1 \le a < b \le T} 1}, \tag{2.20}$$

$$\widehat{\mathrm{sd}}(s^{(k)}(\theta; X, y)) := \frac{\sum_n \sum_{1 \le a < b \le T} \left(s^{(k)}_{n,a,b}(\theta; X_n, y_n) - \overline{s^{(k)}}(\theta; X, y) \right)^2}{(\sum_n \sum_{1 \le a < b \le T} 1) - 1}, \tag{2.21}$$

$$\text{rel. score}^{(k)}_{n,a,b}(\theta) := \frac{s^{(k)}_{n,a,b}(\theta; X_n, y_n) - \overline{s^{(k)}}(\theta; X, y)}{\widehat{\mathrm{sd}}(s^{(k)}(\theta; X, y))}. \tag{2.22}$$

For knots $\kappa_0, \dots, \kappa_P \in \mathbb{R}$ *define the boxcar functions* $\Pi_p(x)$ *as*

$$\Pi_p(x) := \begin{cases} 1 & x \in [\kappa_{p-1}, \kappa_p) \\ 0 & \textit{else} \end{cases}. \tag{2.23}$$

Using the boxcar functions $\Pi_p(x)$*, define the* mean relative score *and the* score statistic *for rectangle data bins, defined by the knots* $\kappa_0, \dots, \kappa_P$*, as*

$$\overline{\text{rel. score}}^{(k)}_{p,q}(\theta) := \frac{\sum_{a,b} \Pi_p(x_a) \Pi_q(x_b) \, \text{rel. score}^{(k)}_{n,a,b}(\theta)}{\sum_{a,b} \Pi_p(x_a) \Pi_q(x_b)}, \tag{2.24}$$

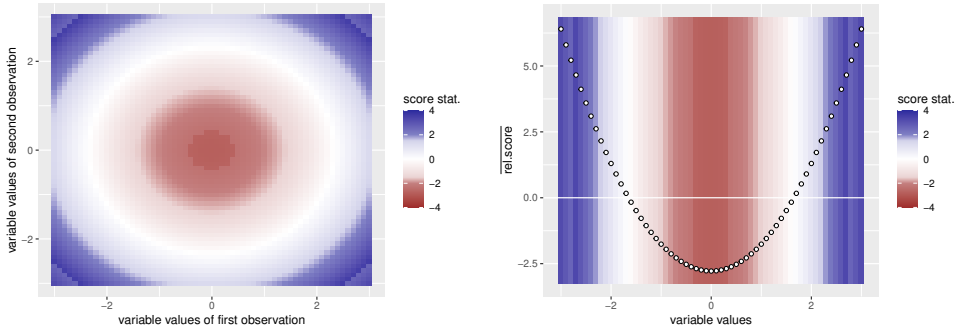
$$\text{score stat}^{(k)}_{p,q}(\theta) := \overline{\text{rel. score}}^{(k)}_{p,q}(\theta) \sqrt{\sum_{a,b} \Pi_p(x_a) \Pi_q(x_b)}. \tag{2.25}$$

The mean relative score $\overline{\text{rel. score}}_{p,q}$ can thus be seen as an estimator of the expected score for a pair of observations (a, b) with $x_a \in [\kappa_{p-1}, \kappa_p)$ and $x_b \in [\kappa_{q-1}, \kappa_q)$.

With these concepts in mind, we propose two different types of plots, *heatmap score plots* and *sliced score plots*, which we describe below and whose general structures are shown in Figures 2.2a and 2.2b.

Heatmap Score Plot

This plot (example shown in Figure 2.2a) uses a tile plot. The value x_a of the independent variable x for the first observation a of a pair of observations (a, b) is mapped onto the x-axis, while the corresponding value x_b of the second observation b is mapped on the y-axis. The score stat is then mapped onto the colour of the tiles. The positions of the tiles are defined by the data bins and thus by a grid spanned by the knots $\kappa_0, \dots, \kappa_P$. The colours help to indicate



(a) Heatmap score plot

(b) Sliced score plot

Figure 2.2: Structure of score plots with illustrative quadratic data.

whether the score contributions within the corresponding tiles are, on average, substantially different from zero.

Sliced Score Plot
This plot (example shown in Figure 2.2b) uses a scatter plot with a coloured background. As the name suggests, a slice, or cross-section, through the heatmap score plot is taken, where either the bin for the first (or second) observation is kept fixed – reflecting a horizontal (vertical) slice through the heatmap score plot – or only observations where the variable values x_a and x_b for both observations are in the same data bin ($\Pi_p(x_a) = \Pi_p(x_b)\ \forall p$) – reflecting a slice along the diagonal of the heatmap score plot. The values of the varying variable (x_a or x_b) are then mapped onto the x-axis and the $\overline{\text{rel. score}}$ values onto the y-axis. The background is then coloured according to the score stat values of the corresponding data bins, with a horizontal white line at $\overline{\text{rel. score}} = 0$ for further visual guidance. In this way, the scatter plot helps to identify possible shapes of score dependencies on variable values, while the background colours are identical to those in the heatmap score plot, thus indicating a possible substantial difference from zero.

For the independent variable(s) of the plots, we suggest using either (1) the covariate values of the observations and producing a plot for each variable under consideration, (2) the predicted deterministic part of the utilities U for different choice alternatives[3], or (3) the time point at which the choice occasions were observed. Using the observation time as independent plot variable can be particularly useful for

[3]In the synthetic examples utilised, the usage of the predicted deterministic part of the utilities as an independent plot variable was found to be advantageous solely for models of minimal complexity.

performing model diagnostics regarding dynamic error processes or time-dependent behaviour of the model.

Remark 2.1. *Since the calculation of a CML estimator usually assumes symmetry in the observations, only one of the pairs (a, b) and (b, a) is included in the log-CML function. For the plots, both pairs should be included so that a 'mirrored' data set is added to the original data set, where the roles of the first and second observations in each pair are reversed. Thus, for each pair (a, b) represented in the data set of score contributions, another pair (a', b') is added to the data set with $a' = b$, $b' = a$, $x_{a'} = x_b$, and $x_{b'} = x_a$ for all variables x.*

Remark 2.2. *Two further issues that may arise for the score plots, and should therefore be considered, are the following: (1) Since the score contributions of the observations depend inversely on the predicted probabilities of the observations (see Definition 2.2), in case of small predicted choice probabilities small numerical or statistical errors in the predicted choice probabilities can lead to large errors in the expected score contributions. (2) Since we are interested in the mean score contributions across a range of different observed choice alternatives, data bins in which a single alternative dominates the observed choices can prevent an unbiased estimation of the expected score contributions. In order to offset both effects, it is proposed that data bins should be excluded in which one alternative dominates the observed choices. One option would be to remove data bins from the plot where one alternative has more than $p_{max} >> 1/J$ of the share of observed choices. Another option would be to only include data bins where each alternative makes up at least $p_{min} << 1/J$ of the share of observed choices. The choice of the thresholds p_{min} and p_{max} depends on data availability. In this paper we have used $p_{min} = 0.05$.*

Remark 2.3. *As an alternative to data binning, one could also use kernel smoothing [see, for example, Hastie et al., 2009, Chapter 6] to obtain a continuous approximation of the conditional expectation of the scores as a function of the covariate values, for example using a Nadaraya–Watson kernel-weighted average to estimate the expected relative score $\overline{\text{rel. score}}(x_a, x_b)$ for a point (x_a, x_b) and multiplying this by the square root of the sum of the kernel weights used for this estimate to obtain the* score stat *for that point.*

2.4 Estimation

The following sections present a series of data examples, for which we have employed the use of MNP models with independent normal distributions for the distribution of possible random effects. To estimate the models, we employed pairwise CML estimation in lieu of ML estimation. In particular, we utilised a variant of maximum approximate composite marginal likelihood (MACML), as intro-

duced by Bhat [2011], with a Solow-Joe approximation [Solow, 1990; Joe, 1995] of the multivariate normal cumulative distribution functions (MVNCDFs). The numerical procedures were implemented in `R` [Version 4.2.2, R Core Team, 2021] with the calculation of the CML function written in `C++11` to enhance computational performance and integrated in `R` via the `Rcpp` package by Eddelbuettel and Francois [2011]. The negative CML function is minimised using the `R` function `nlm`, using analytic gradients, and the variance–covariance matrix of the estimator is calculated using the analytic Hessian matrix of the CML function. The `R` and `C++` code employed for the estimation process is bundled into an `R` package named `Rprobit` and is accessible on github via `https://github.com/dbauer72/Rprobit`.

Remark 2.4. *It is noteworthy that to generate score plots utilising software designed for ML estimation, which has to be capable of providing score information at the DM level, the original data set can be transformed into a data set in which each new DM represents exactly one pair of observations from a DM in the original data set. This enables pairwise CML estimation. Following the model selection and model diagnosis phase, the practitioner can revert to the original data set and estimate the final model using ML estimation, thereby capitalising on the statistical efficiency of the ML estimator.*

2.5 Synthetic Examples

In this section, we will employ synthetic data examples and estimate DCMs with either a misspecified utility function or a misspecified error specification. The two types of score plots proposed in Section 2.3, namely *heatmap score plots* and *sliced score plots*, were generated for the misspecified models, with the objective of detecting indications of the misspecifications. While the heatmap score plots were generated for each model and considered independent plot variable, the sliced score plots were only produced when the corresponding heatmap score plot exhibited discernible structures, indicating potential model misspecifications, to facilitate further investigation into the nature of the misspecification. For the sliced score plots, we employed a 'diagonal' slice with $\Pi_p(x_a) = \Pi_p(x_b)\ \forall p$ whenever covariates were used, and 'horizontal' slices for plots with the observation time as independent plot variable.

Binary probit models were employed as illustrative examples to demonstrate the functionality of the proposed plots. All data sets were simulated to represent a balanced panel data set, comprising $N = 1\,000$ DMs and $T = 30$ observations per DM. These observations were allocated in two panel waves, the first comprising $\tau_1 = 1$, $\tau_2 = 2, \ldots, \tau_{15} = 15$, and the second $\tau_{16} = 366, \ldots, \tau_{30} = 380$. In this context, τ_t indicates the time point at which choice occasion t was observed.

As independent variables employed for the construction of the score plots we considered (a) the covariate values of the observations, (b) the predicted utility of the second choice alternative, and (c) the temporal position of the observations. In the case of the plots for which the temporal positions constituted the independent variable, an adequate number of observations for each discrete combination of time points was available (one for each of the 1 000 DMs), obviating the necessity for binning. For the continuous variables, data binning was employed in accordance with the methodology outlined in Remark 2.2, with $\kappa_0 = -\infty$, $\kappa_{101} = +\infty$, κ_1 set to the empirical 1% quantile of the respective data, κ_{100} to the 99% quantile, and the remaining knots selected to be distributed equidistant between κ_1 and κ_{100}, resulting in 100 data bins for each variable, and thus a heatmap score plot comprising 10 000 tiles.

In order to present the results in a concise manner, we will refrain from including every plot for which no patterns were expected to be visible, as they were using independent plot variables that were not connected to the misspecification of the respective model, in this section. Instead, we will include the majority of these plots in 2.7.

2.5.1 Non-Linear Utility - Quadratic Utility

In certain instances, utilities may exhibit non-linear behaviour w.r.t. specific covariate values. To reflect this, we have implemented a quadratic utility function w.r.t. a covariate x into the DGP. Subsequently, we have estimated two distinct models: one that is correctly specified and another that considers only a linear effect w.r.t. the covariate x.

Data Generating Process and the Misspecified Model

In this scenario, the DGP under consideration is

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{2.26}$$

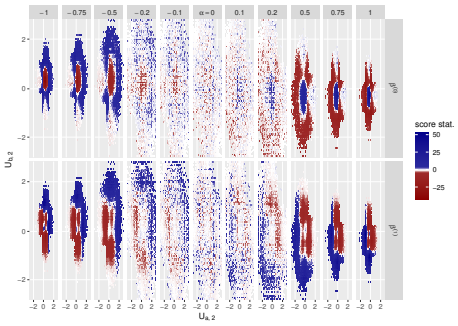
$$U_{n,t,2} = \beta^{(0)} + \beta^{(1)} x_{n,t} + \beta^{(2)} x_{n,t}^2 + \gamma_n + \varepsilon_{n,t,2}, \tag{2.27}$$

with $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}\right)$ and $\gamma_n \sim \mathcal{N}(0, 0.25)$. In the DGP, the values $\beta^{(0)} = -\alpha$, $\beta^{(1)} = 1$, and $\beta^{(2)} = \alpha$ were employed. The covariate values were drawn independent, identically distributed (iid) from a standard normal distribution ($x_{n,t} \sim \mathcal{N}(0,1)$). The data was utilised to estimate $\hat{\theta} = (\hat{\beta}^{(0)}, \hat{\beta}^{(1)}, \hat{\beta}^{(2)})'$ for two models: one correctly specified model and one model with $\hat{\beta}^{(2)} = 0$ fixed.

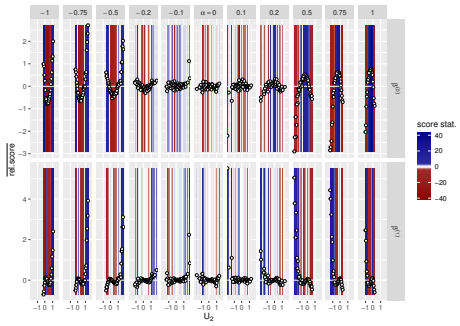
In order to investigate the effect of different degrees of quadratic effects, a number of data sets with varying values of the parameter α were generated.[4]

Plots and Discussion

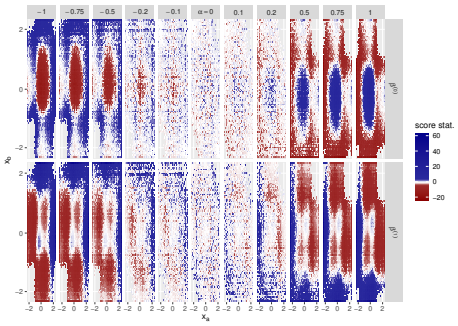
The plots resulting from two models described above are presented in Figures 2.3 and 2.18. Inspection of the score plots with either the covariate x or the predicted util-



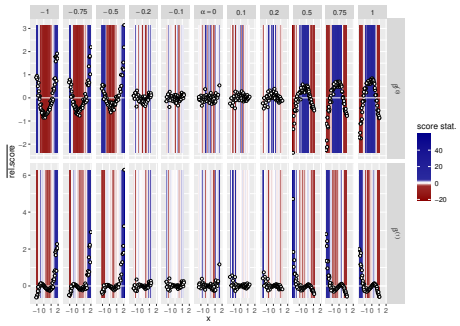
(a) Heatmap score plot with predicted expected utility of second alternative on the axis, misspecified model



(b) Sliced score plot with predicted expected utility of second alternative on the x-axis, misspecified model



(c) Heatmap score plot with covariate x on the axis, misspecified model



(d) Sliced score plot with covariate x on the x-axis, misspecified model

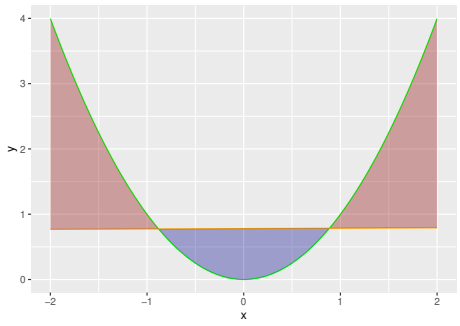
Figure 2.3: Score plots of data from a DGP with a utility function quadratic in x and a model with linear dependency. Each sub-figure contains a grid of plots, with each column representing a different data set and a different value of α, representing the effect of x^2 on the utility, as indicated above the columns. Each row represents the score direction w.r.t. a different model parameter, indicated on the right of the rows.

[4]The values used for α were $\{-1, -0.75, -0.5, -0.2, -0.1, 0, 0.1, 0.2, 0.5, 0.75, 1\}$.

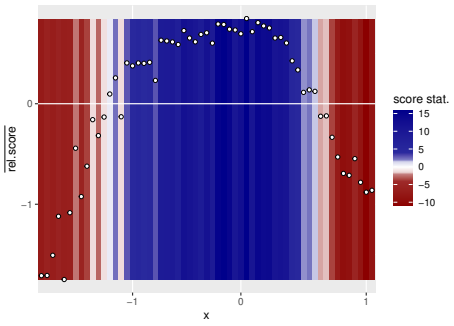
ity of the second alternative as the independent plot variable reveals clear patterns in the score contributions of the misspecified models (Figure 2.3). As the predicted utility is, in this case, a linear function of the covariate x, the patterns are equivalent for these plots and will therefore only be discussed for those depending on the covariate x. The facets within the plots referring to the models with $\alpha = 0$, i.e. no quadratic effect of the quadratic variable x^2, show no discernible patterns, which is to be expected given that in this case the model is correctly specified. The patterns observed in the heatmap score plots (Figure 2.3c) of the score direction related to the alternative specific constant (ASC) parameter $\beta^{(0)}$ show, for a negative value of the quadratic parameter α in the DGP, predominantly negative values for x_a and x_b near zero, and predominantly negative values are otherwise observed. Similarly, when the quadratic parameter α is positive in the DGP, the same pattern is observed, but with reversed signs. This pattern is also clearly visible when the score contributions to the score direction related to the ASC parameter $\beta^{(0)}$ in the sliced score plots in Figure 2.3d are inspected. Here, a quadratic dependency of the score contributions on the covariate values seems evident.

In order to elucidate this observed pattern, we have visualised the expected score contributions of a linear regression model next to the sliced score plots for the data from the model with $\alpha = 1$ in Figure 2.4. The plot of the score direction related to the parameter $\beta^{(1)}$ of the covariate effect exhibits a more intricate pattern, which can nevertheless be more readily interpreted when compared to the expected score contributions for the score direction of the slope parameter in a linear regression model. For $x_a > 0$ and $x_b > 0$, the observed pattern for the $\beta^{(1)}$ score direction is identical to that observed for the $\beta^{(0)}$ score direction. Similarly, for $x_a < 0$ and $x_b < 0$ the same pattern is evident, albeit with a reversed sign. In the remaining quadrants of the plot, the pattern is more complex and can be interpreted as a smooth continuation of the aforementioned pattern. Upon observing the form of the deterministic part of the utility, as specified in (2.26), and applying the chain rule to the score contributions specified in Definition 2.2, it becomes evident that the score contributions w.r.t. the score direction related to the effect $\beta^{(k)}$ of a covariate x_k are, among other factors, scaled by the value of the covariate x_k. This facilitates comprehension of the observed pattern, as the same case, though in a more straightforward manner, is also true for the score contributions in a linear regression model, as illustrated in Figure 2.4c and discussed in Section 2.2.1.

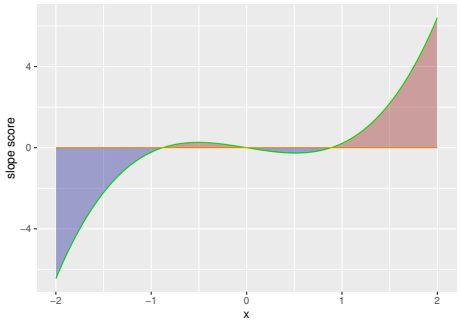
The plots in Figure 2.18 demonstrate that for a correctly specified model, all these patterns are absent and there is no dependency of the score contributions on covariate values or predicted utilities. In any of the models, plotting the score contributions by time reveals no discernible patterns, as expected, given that neither the observations nor the covariate values depend on the time of the observations.



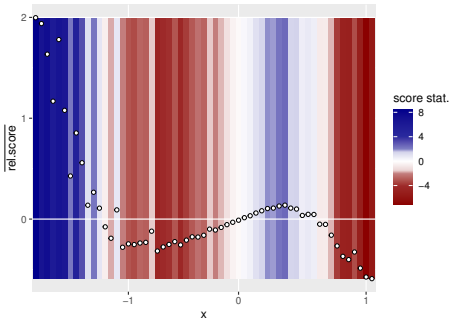
(a) Visualisation of the residual function (coloured areas) of a linear model (orange line) fitted to data from a quadratic function (green line).



(b) Diagonally sliced score plot for the score direction related to the ASC $\beta^{(0)}$.



(c) Visualisation of the score function of the slope parameter (coloured areas and green line) fitted to data from a quadratic function.



(d) Diagonally sliced score plot for the score direction related to the effect $\beta^{(1)}$ of the covariate x.

Figure 2.4: Visualisation of the similarities between score plots for a linear regression model (left) and the diagonally sliced score plots for a discrete choice model with a quadratic dependency on x with effect size $\alpha = 1$ in the utility of the DGP and a linear dependency in the model (right).

2.5.2 Non-Linear Utility - Steps and Bends

In some cases, utilities may exhibit non-smooth behaviour w.r.t. some covariate values. In order to reflect this, we implemented different non-linear and non-smooth transformations of the covariate x in the DGP and estimated a model once with the correct non-linear transformation of the covariate and a misspecified version with the original variable x in the model. Furthermore, an additional continuous covariate z

was added to the model to more accurately reflect the characteristics of a real-world data set.

Data Generating Process and the Misspecified Model

In this case, the DGP is given by

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{2.28}$$

$$U_{n,t,2} = \beta^{(0)} + \beta^{(1)} f_\circ(x_{n,t}) + \beta^{(2)} z_{n,t} + \gamma_n + \varepsilon_{n,t,2}, \tag{2.29}$$

with $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}\right)$ and $\gamma_n \sim \mathcal{N}(0, 0.25)$. In the DGP, the values $\beta^{(0)} = 0$ and $\beta^{(1)} = \beta^{(2)} = 1$ were employed. The covariate values were drawn iid from a standard normal distribution, so $x_{n,t} \sim \mathcal{N}(0,1)$ and $z_{n,t} \sim \mathcal{N}(0,1)$.

Subsequently, the data was employed to estimate two models: one correctly specified model with $f_\circ(x_{n,t})$ as a covariate, and one misspecified model with the untransformed variable $x_{n,t}$. The parameter vector being estimated was $\theta = (\beta_0, \beta_1, \beta_2)'$, while the remaining parameters were held fixed at their true parameter values.

The following non-smooth data transformation functions were considered and implemented:

$$f_{\text{step}}(x) = I(x \geq 0), \tag{2.30}$$

$$f_{\text{ReLU}}(x) = \max(x, 0), \tag{2.31}$$

$$f_{\text{wedge}}(x) = |x|. \tag{2.32}$$

Here, ReLU stands for Rectified Linear Unit, as used in deep learning [for an early example, see Householder, 1941]. It should be noted that a second covariate was added to introduce additional noise into the model, in comparison to the models presented in Section 2.5.1. This reflects a more realistic setting.

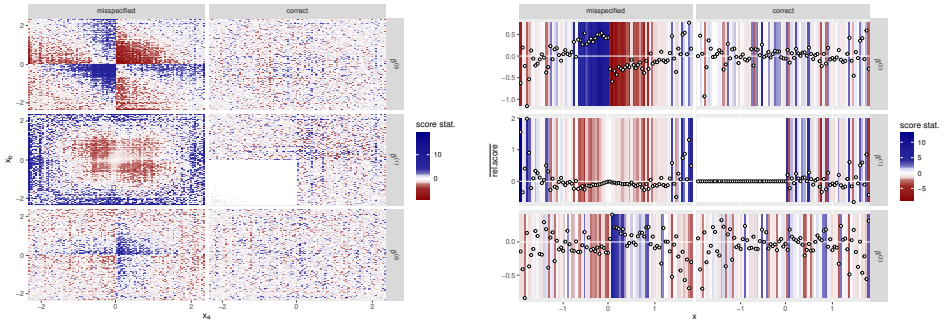
Plots and Discussion

The plots generated from the score data of the aforementioned models are presented in Figures 2.5 and 2.19. Upon examination of the plots with the untransformed covariate x as the independent plot variable, it becomes evident that distinct patterns in the score contributions contingent on x are discernible in the misspecified models. However, these patterns disappear once the correctly specified model is estimated, which incorporates the transformed covariate $f_\circ(x)$. This is the case for all three considered non-smooth transformations and all score directions, though the effect is most pronounced for the score directions related to the ASC parameter $\beta^{(0)}$ and the parameter for the effect of x, $\beta^{(1)}$.

In the case of a step function transformation of the covariate ($f_{\text{step}}(x)$, eq. (2.30)) the heatmap score plot (Figure 2.5a) of the score direction relating to the ASC parameter $\beta^{(0)}$ is clearly segmented into the four quadrants, with the origin in the centre. In quadrants II and IV (upper left and lower right), it can be observed that the lower left corners have predominantly negative score contributions, whilst the upper left corner has predominantly positive score contributions. In the quadrant with both observations having negative values for x (quadrant III, lower left), the observations in proximity to the origin exhibit predominantly positive score contributions, whereas the pairs with large negative x values demonstrate slightly negative score contributions. The pairs of observations in quadrant I (upper right), where $x_a > 0$ and $x_b > 0$, exhibit predominantly negative score contributions for observations with covariate values close to zero. This pattern diminishes as the covariate values increase.

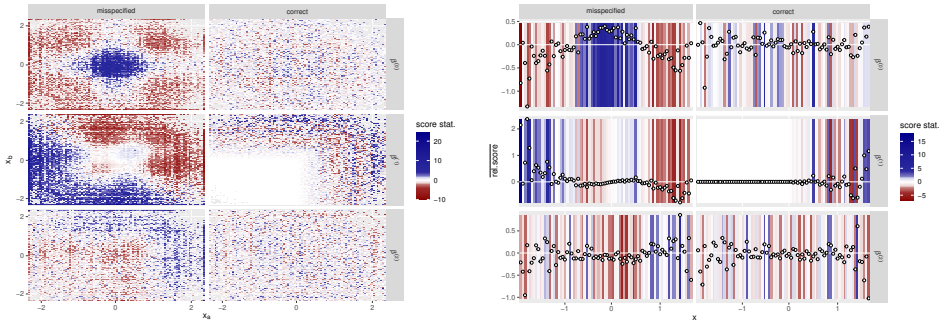
This partitioning into the four distinct quadrants is to be expected, given that the discrete step separates the deterministic part of the utility function of the DGP into those four quadrants. Additionally, the pattern observed within the quadrants, where there appear to be approximately equi-utility regions perpendicular to the diagonal ($\Pi_p(x_a) = \Pi_p(x_b)$), with monotonically rising score values for rising values of the covariates, is consistent with expectations. This pattern is also clearly discernible in the diagonally sliced score plot presented in Figure 2.5b. When this plot is related to the expected scores in a linear regression setting, a remarkable resemblance is revealed (for comparison, see Figures 2.6a and 2.6b).

The pattern observed in the heatmap score plot for the score direction relating to the $\beta^{(1)}$ parameter for the misspecified model can be most accurately described as a four-leaved clover centred at the origin, with each leaf situated within one of the quadrants of the plot, where the clover represents an area of predominantly negative score contributions, surrounded by an area of predominantly positive score contributions. The sliced score plot is again useful for interpreting this pattern, particularly in comparison to the expected scores for the slope parameter in a linear regression model (compare Figures 2.6c and 2.6d). As with the previous case, the



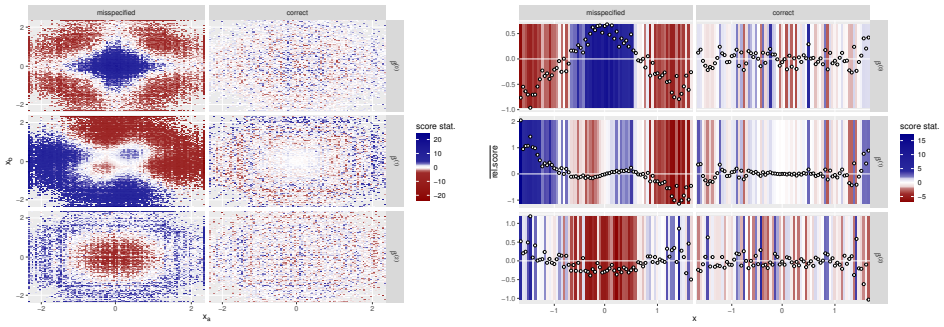
(a) Heatmap score plot, $f_{\text{step}}(x)$

(b) Sliced score plot, $f_{\text{step}}(x)$



(c) Heatmap score plot, $f_{\text{ReLU}}(x)$

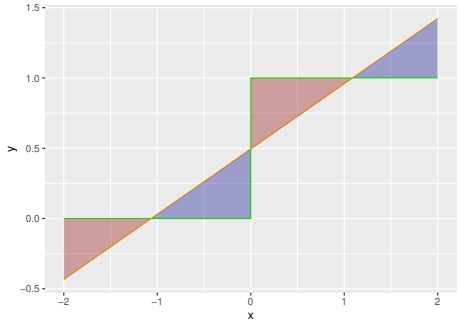
(d) Sliced score plot, $f_{\text{ReLU}}(x)$



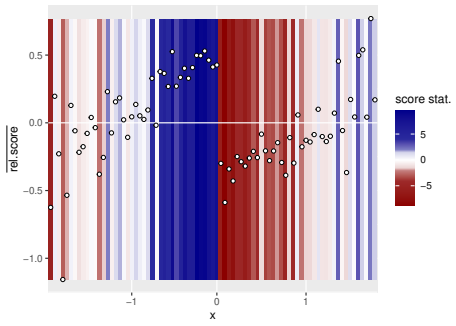
(e) Heatmap score plot, $f_{\text{wedge}}(x)$

(f) Sliced score plot, $f_{\text{wedge}}(x)$

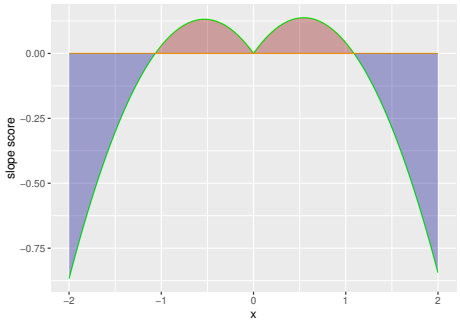
Figure 2.5: Score plots of data from a DGP with a utility function exhibiting a non-linear dependency $f_\circ(x)$ w.r.t. the covariate x. The covariate values of x are used as the independent plot variable. Within each sub-figure, the left column of plots represents the misspecified model, whereas the right column represents the correctly specified model.



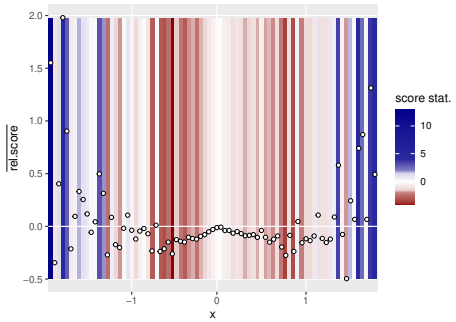
(a) Visualisation of residual function (coloured areas) of a linear model (orange line) fitted to data from a step function (green line).



(b) Diagonally sliced score plot for the score direction related to the ASC $\beta^{(0)}$.



(c) Visualisation of score function of the slope parameter (coloured areas and green line) fitted to data from a step function.

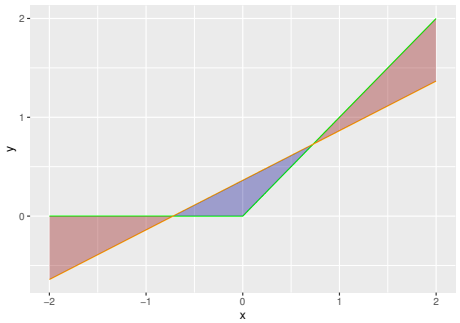


(d) Diagonally sliced score plot for the score direction related to the effect $\beta^{(1)}$ of covariate x.

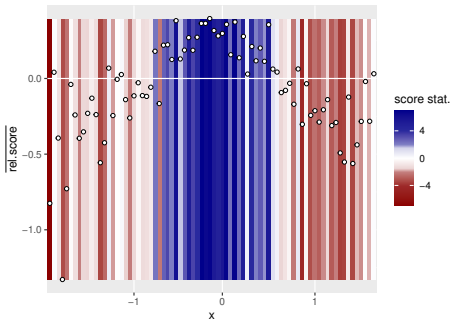
Figure 2.6: Visualisation of the similarities between score plots for a linear regression model (left) and the diagonally sliced score plots for a discrete choice model with a step function in the DGP and linear dependency in the model (right).

observed pattern in the sliced score plot closely resembles the expected scores in a linear model with a comparable misspecification.

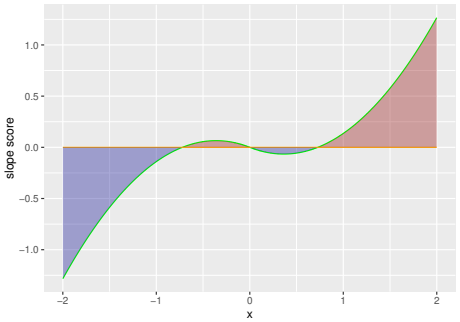
Additionally, the plot for the score direction related to the effect of z (parameter $\beta^{(2)}$) exhibits discernible patterns, particularly the segmentation into the four segments is visible. This may initially appear counterintuitive, given that the effect size of z should be estimated in an unbiased manner as the misspecification in x is independent of z and that x and z are not correlated. However, as the scale of the model



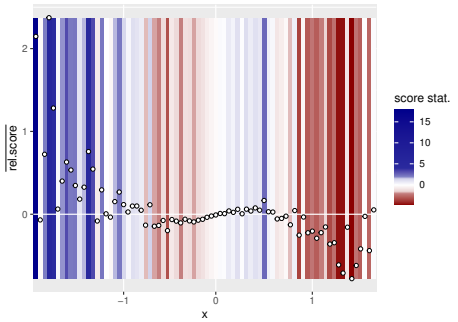
(a) Visualisation of residual function (coloured areas) of a linear model (orange line) fitted to data from a ReLU function (green line).



(b) Diagonally sliced score plot for the score direction related to the ASC $\beta^{(0)}$.



(c) Visualisation of score function of the slope parameter (coloured areas and green line) fitted to data from a ReLU function.



(d) Diagonally sliced score plot for the score direction related to the effect $\beta^{(1)}$ of covariate x.

Figure 2.7: Visualisation of the similarities between score plots for a linear regression model (left) and the diagonally sliced score plots for a discrete choice model with a ReLU function in the DGP and linear dependency in the model (right).

has been fixed by setting the variance of the error terms ε to 0.25, the effect size of the covariates must be interpreted in relation to the scale of the model. The misspecification of the model can result in different error variances being optimal for log-CML maximisation for different values of x. Since these are fixed, this is compensated for by adjusting the scale of the model, which affects all parameters of the model. We will keep this effect in mind, as it can be observed in subsequent models and score plots.

The score plots for the models estimated on data with a ReLU transformation $f_{\text{ReLU}}(x) = \max(x, 0)$ in the DGP (Figures 2.5c and 2.5d) and those estimated on data with an absolute value function $f_{\text{wedge}}(x) = |x|$ in the DGP (Figures 2.5e and 2.5f) exhibit strikingly similar patterns, with those from the model with the absolute value function in the DGP being more pronounced. As both functions are piecewise linear with a change in slope at $x = 0$, these similarities are to be expected.

With regard to the score direction related to the ASC parameter $\beta^{(0)}$, we observe predominantly positive score contributions in the vicinity of the origin, followed by increasingly negative score contributions as we move away from the centre. In regions where either the x value of the first observation or that of the second observation is close to zero (near the axes), the score contributions are also close to zero, when sufficiently far from the origin.

The sliced score plots are once more useful for further investigating the dependency of the score contributions upon the covariate values. We will, again, employ the analogy to the linear regression case and direct the readers' attention to the visualisation in Figures 2.7a and 2.8a, which we will compare to the diagonally sliced score plots, as shown in Figures 2.7b and 2.8b. The similarities in the resulting plots are, once more, self-evident.

The analogies between the expected scores in the linear regression case facilitate the interpretation and understanding of the patterns visible in the heatmap score plots w.r.t. the score direction of the parameter $\beta^{(1)}$, the effect of the covariate x. The scatter plots in the sliced score plots (Figures 2.7d and 2.8d) show the same dependencies of the score contributions on the covariate values as the expected score contributions for the linear regression model do (Figures 2.7c and 2.8c).

Furthermore, patterns for the score direction w.r.t. $\beta^{(2)}$, the effect of the covariate z, can be observed, albeit with diminished intensity in comparison to the other score directions. As with the case of the model estimated on data with a step function in the DGP, this may be attributed to different variances of the errors being optimal to maximise the log-CML function for different values of x, due to the misspecification in the covariate x. The error variances are, however, kept fixed to fix the scale of the model, and consequently all other parameters would require rescaling in order to reflect a change in the variance of the errors. The pattern observed in the heatmap score plot for the score direction relating to the $\beta^{(1)}$ parameter for the misspecified model can be most accurately described as a four-leaved clover centred at the origin, with each leaf situated within one of the quadrants of the plot, where the clover represents an area of predominantly negative score contributions, surrounded by an area of predominantly positive score contributions. The sliced score plot is again useful for interpreting this pattern, particularly in comparison to the expected scores for the slope parameter in a linear regression model (compare Figures 2.6c and 2.6d). As

with the previous case, the observed pattern in the sliced score plot closely resembles the expected scores in a linear model with a comparable misspecification.

Additionally, the plot for the score direction related to the effect of z (parameter $\beta^{(2)}$) exhibits discernible patterns, particularly the segmentation into the four segments is visible. This may initially appear counterintuitive, given that the effect size of z should be estimated in an unbiased manner as the misspecification in x is independent of z and that x and z are not correlated. However, as the scale of the model has been fixed by setting the variance of the error terms ε to 0.25, the effect size of the covariates must be interpreted in relation to the scale of the model. The misspecification of the model can result in different error variances being optimal for log-CML maximisation for different values of x. Since these are fixed, this is compensated for by adjusting the scale of the model, which affects all parameters of the model. We will keep this effect in mind, as it can be observed in subsequent models and score plots.

The score plots for the models estimated on data with a ReLU transformation $f_{\text{ReLU}}(x) = \max(x, 0)$ in the DGP (Figures 2.5c and 2.5d) and those estimated on data with an absolute value function $f_{\text{wedge}}(x) = |x|$ in the DGP (Figures 2.5e and 2.5f) exhibit strikingly similar patterns, with those from the model with the absolute value function in the DGP being more pronounced. As both functions are piecewise linear with a change in slope at $x = 0$, these similarities are to be expected.

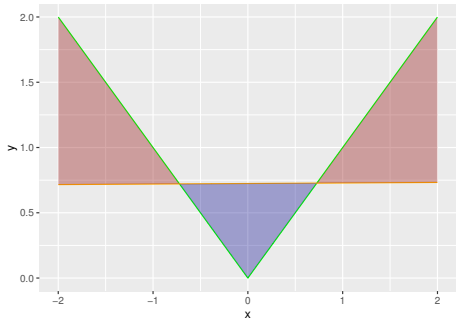
With regard to the score direction related to the ASC parameter $\beta^{(0)}$, we observe predominantly positive score contributions in the vicinity of the origin, followed by increasingly negative score contributions as we move away from the centre. In regions where either the x value of the first observation or that of the second observation is close to zero (near the axes), the score contributions are also close to zero, when sufficiently far from the origin.

The sliced score plots are once more useful for further investigating the dependency of the score contributions upon the covariate values. We will, again, employ the analogy to the linear regression case and direct the readers' attention to the visualisation in Figures 2.7a and 2.8a, which we will compare to the diagonally sliced score plots, as shown in Figures 2.7b and 2.8b. The similarities in the resulting plots are, once more, self-evident.

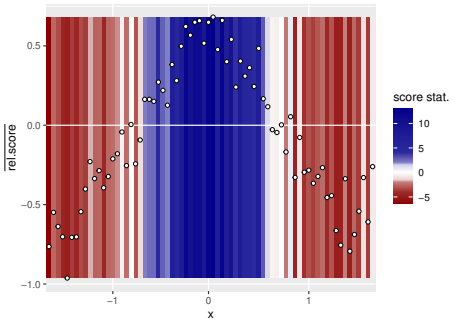
The analogies between the expected scores in the linear regression case facilitate the interpretation and understanding of the patterns visible in the heatmap score plots w.r.t. the score direction of the parameter $\beta^{(1)}$, the effect of the covariate x. The scatter plots in the sliced score plots (Figures 2.7d and 2.8d) show the same dependencies of the score contributions on the covariate values as the expected score contributions for the linear regression model do (Figures 2.7c and 2.8c).

Furthermore, patterns for the score direction w.r.t. $\beta^{(2)}$, the effect of the covariate

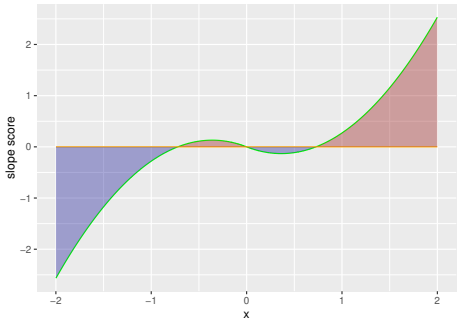
z, can be observed, albeit with diminished intensity in comparison to the other score directions. As with the case of the model estimated on data with a step function in the DGP, this may be attributed to different variances of the errors being optimal to maximise the log-CML function for different values of x, due to the misspecification in the covariate x. The error variances are, however, kept fixed to fix the scale of the model, and consequently all other parameters would require rescaling in order to reflect a change in the variance of the errors.



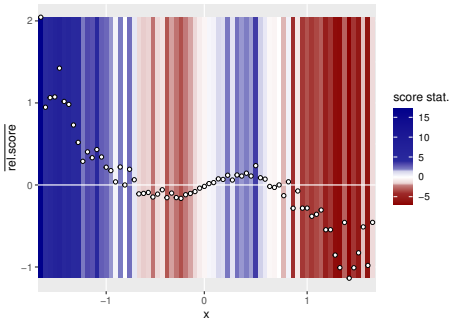
(a) Visualisation of residual function (coloured areas) of a linear model (orange line) fitted to data from a absolute value function (green line).



(b) Diagonally sliced score plot for the score direction related to the ASC $\beta^{(0)}$.



(c) Visualisation of score function of the slope parameter (coloured areas and green line) fitted to data from a absolute value function.



(d) Diagonally sliced score plot for the score direction related to the effect $\beta^{(1)}$ of the covariate x.

Figure 2.8: Visualisation of the similarities between score plots for a linear regression model (left) and the diagonally sliced score plots for a discrete choice model with an absolute value function in the DGP and linear dependency in the model (right).

Lastly, it is to note that none of the plots in which time is the independent plot variable demonstrate any discernible patterns (see Figures 2.19c, 2.19f and 2.19i). This is to be expected, given that the DGP is not dependent on time. In the case of the plots in which the covariate z or the predicted utility of the second alternative is the independent plot variable (Figure 2.19) no strong discernible patterns are visible. The sole exceptions are for the plots with data from a DGP with a ReLU function in the utility, where a slight tendency of negative score contributions for the score direction related to the parameter $\beta^{(1)}$ of the effect of covariate x for positive values of the predicted utility can be observed (and positive score values for other values of the predicted utility), see Figure 2.19e. A similar pattern with reversed signs emerges for the same score direction but with covariate z as the independent plot variable. With regard to the plot in which the predicted utility is the independent plot variable, this can be attributed to the fact that the predicted utility is correlated with the covariate x and thus also with the misspecification of the model. It should be noted, however, that the inclusion of additional covariates in the model results in a reduction in the effectiveness of score plots with the predicted utility as the independent plot variable, in comparison to the results presented in Section 2.5.1.

In all cases, any discernible pattern in the score plots for the misspecified model is obliterated when the transformed variable $f_\circ(x)$ in lieu of the original variable x is included in the utility function, so when the correctly specified model is estimated.

In conclusion, the cases presented demonstrate the potential of using score plots to detect non-linearities with respect to covariates. Furthermore, by comparing the sliced score plots for the score direction related to the ASC parameters and the residuals from linear regression models, insights can be gained into the shape of the non-linearity in the utility function.

2.5.3 Missing Variables

Diagnosis plots may also be employed to assess variables that have not (yet) been included in the model, with a view to determining whether their inclusion could prove advantageous. In order to ascertain the suitability of the proposed score plots for this purpose, we simulated data from different DGPs and estimated models in which one of the covariates was omitted.

Data Generating Process and the Misspecified Model

In this case, the DGP is specified as

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{2.33}$$

$$U_{n,t,2} = \beta^{(0)} + \beta^{(1)} x_{1,n,t} + \beta^{(2)} x_{2,n,t} + \beta^{(3)} x_{3,n,t} + \gamma_n + \varepsilon_{n,t,2}, \tag{2.34}$$

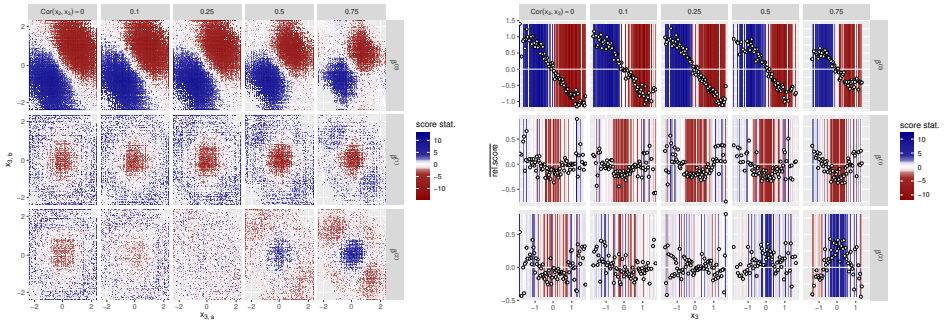
with $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}\right)$ and $\gamma_n \sim \mathcal{N}(0, 0.25)$. The parameters of the DGP were set to $\beta^{(0)} = 0$ and $\beta^{(1)} = \beta^{(2)} = 1$. When x_3 was included in the DGP, the corresponding parameter was set to $\beta^{(3)} = 1$. The values of the covariate x_1 were drawn iid from a standard normal distribution, so $x_{1,n,t} \sim \mathcal{N}(0, 1)$. In order to reflect the possibility of correlated covariates, the values of x_2 and x_3 were drawn iid across decision makers and choice occasions from a joint normal distribution, such that $(x_{2,n,t}, x_{3,n,t})' \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}\right)$.

A restricted model was then estimated for each data set, wherein $\theta = (\beta^{(0)}, \beta^{(1)}, \beta^{(2)})'$ was estimated with $\beta^{(3)}$ fixed at $\beta^{(3)} = 0$. For the data sets where x_3 was also included in the DGP, a second model was estimated, wherein the corresponding parameter $\beta^{(3)}$ was also included in the set of estimated parameters.

Plots and Discussion

The score plots for the models estimated on data from a DGP comprising solely the variables x_1 and x_2 are presented in Figure 2.20. They exhibit no discernible patterns, as would be anticipated given that the model is correctly specified and the missing covariates exert no influence on the model. This is also the case for the plot with x_3 as the independent plot variable, despite the fact that x_3 is correlated with the covariate x_2 for four out of the five simulated data sets.

For the data sets generated from DGPs that include the covariates x_1, x_2 and x_3 in the utility function and solely x_1 and x_2 in the estimated model, discernible patterns for the score plots with x_3 as the independent plot variable are evident, as illustrated in Figure 2.9. In the heatmap score plots w.r.t. the score direction corresponding to the ASC $\beta^{(0)}$ (Figure 2.9a), the lower left quadrant exhibits predominantly positive score contributions, whereas the upper right quadrant displays predominantly negative score contributions. This pattern is most pronounced for data sets with $\text{Cor}(x_b, x_c) = 0$ and becomes less pronounced as the correlation coefficient increases. This is to be expected, as the inclusion of x_2 in the model results in a mediation effect, whereby part of the effect of x_3 is picked up by an increased estimated effect of the correlated covariate x_2. The estimated coefficient $\beta^{(2)}$ of the



(a) Heatmap score plot, misspecified model

(b) Sliced score plot, misspecified model

Figure 2.9: Score plots of data which include variables x_1, x_2 and x_3 in the DGP, whilst solely including x_a and x_b in the estimated model. The values displayed above the plot grids indicate the covariance between x_2 and x_3. The values of the covariate x_3 are used as the independent plot variable.

$\text{Cor}(x_2, x_3)$	0	0.1	0.25	0.5	0.75
$\hat{\beta}^{(2)} / \hat{\beta}^{(1)}$	1.003 466	1.102 746	1.252 645	1.507 06	1.739 261
$\beta_0^{(2)} + \text{Cor}(x_2, x_3)\beta_0^{(3)}$	1	1.1	1.25	1.5	1.75

Table 2.1: Estimated parameters for the effect of covariate x_2, relative to the effect of covariate x_1 ($\hat{\beta}^{(2)} / \hat{\beta}^{(1)}$) from a model with omitted covariate x_3, compared to the expected result via mediation due to an omitted variable x_3 with true effect size $\beta_0^{(3)}$.

effect of x_2 would therefore be biased. This effect was exactly what was observed in the estimated parameters, presented in Table 2.1, for which we re-scaled the models such that $\beta^{(1)} = 1$. As a result, the estimated parameter $\hat{\beta}^{(2)}$ closely matched $\beta_0^{(2)} + \text{Cor}(x_2, x_3)\beta_0^{(3)}$. This finding is consistent with expectations and aligns with what would be observed in a residual plot in a linear regression setting, where such an effect is also referred to as mediation [Baron and Kenny, 1986]. The sliced score plots in Figure 2.9b demonstrate a discernible linear dependency of the score contributions for the score direction of the ASC parameter upon the covariate values of x_3, reinforcing the parallels with residuals in a linear regression setting.

The plots of the score directions related to the effects of the covariates x_1 and x_2 (parameters $\beta^{(1)}$ and $\beta^{(2)}$, respectively) demonstrate that in the absence of a correlation between the left-out variable x_3 and the covariate x_2, there are predominantly negative score contributions for x_3 close to zero for both observations, and positive

score contributions for covariate values with a greater absolute value. As the correlation between the left-out covariate x_3 and the included covariate x_2 increases, the pattern in the plots for the score direction of x_2 initially diminishes towards a correlation of $\text{Cor}(x_2, x_3) = 0.25$. Subsequently, it resurfaces with a reversed sign for larger correlations. As with the previous cases, for the score direction related to $\beta^{(1)}$, this can be attributed to the scale of the model being fixed by setting the error variance to 0.25. For observations where a higher estimated error variance would increase the log-CML value of the observation, this would be compensated by a change in the scale of the model, consequently affecting the magnitude of all parameters. In addition, the correlation between x_2 and the omitted covariate x_3 exerts and influence on the score direction related to $\beta^{(2)}$.

As in the other cases, score plots with the temporal position of the observations as the independent plot variable demonstrate an absence of discernible patterns (Figures 2.21b and 2.21e). This is also true for plots with the predicted utility as the independent plot variable, as it is a linear combination of variables which are correctly included in the model (Figures 2.21a and 2.21d). The inclusion of the missing covariate x_3 into the model also results in the absence of any pattern previously observed when the variable x_3 was not included in the model (Figure 2.21c).

It can thus be reasoned that the proposed score plots can assist in the identification of missing variables for a model, obviating the necessity of the estimation of new models for each variable under consideration. Instead, score plots with the considered variables as independent plot variables can be generated. Subsequently, only those variables for which discernible patterns are visible are then included in a potential model. Thereafter, classical model selection criteria can be applied to make a final decision regarding the inclusion or exclusion of the variable under consideration.

2.5.4 Structural Breaks in Time - Parameter Shift

The data examples presented in this section illustrate a change in the effect of a covariate between two panel waves. This could be, for example, due to a change in policies or a change in the disposition towards an attribute in the population occurring between panel waves.[5]

[5]The data used for this example was also used in Büscher and Bauer [2024].

Data Generating Process and the Misspecified Model

In this scenario, the DGP under consideration is

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{2.35}$$

$$U_{n,t,2} = \beta^{(0)} + \beta^{(1)}_{n,t} x_{n,t} + \gamma_n + \varepsilon_{n,t,2}, \tag{2.36}$$

with $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_\varepsilon^2 \end{pmatrix}\right)$ and $\gamma_n \sim \mathcal{N}(0, \sigma_\gamma^2)$. For the DGP, the values $\beta^{(0)} = 1$, $\sigma_\gamma = 1$, and $\sigma_\varepsilon = 1$ were used. In order to simulate a change in attitude towards an attribute, a discrete change was implemented in the parameter $\beta^{(1)}_{n,t}$ between the two panel waves, with identical parameter values for all DMs, specifically

$$\beta^{(1)}_{n,t} = \begin{cases} 1 - \alpha, & \tau_t < 365, \\ 1 + \alpha, & \tau_t \geq 365, \end{cases} \tag{2.37}$$

where, τ_t indicates the time point at which choice occasion t was observed, as mentioned at the beginning of the section. The covariate values were drawn iid from a standard normal distribution, so $x_{n,t} \sim \mathcal{N}(0, 1)$.

For estimation, the misspecified model

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{2.38}$$

$$U_{n,t,2} = \beta^{(0)} + \beta^{(1)} x_{n,t} + \gamma_n + \varepsilon_{n,t,2}, \tag{2.39}$$

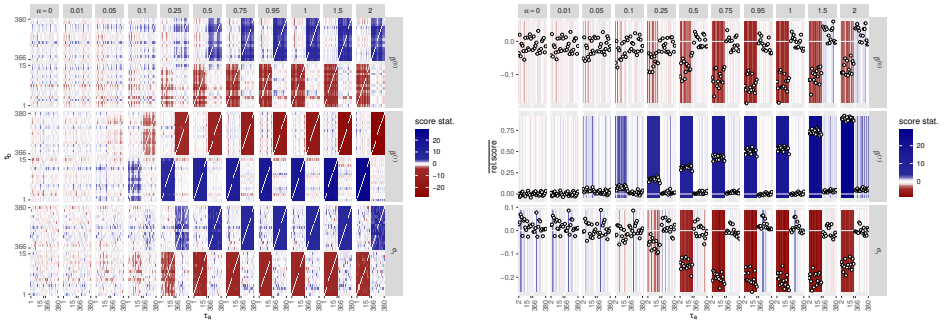
was employed, with $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_\varepsilon^2 \end{pmatrix}\right)$ and $\gamma_n \sim \mathcal{N}(0, \sigma_\gamma^2)$. For identification purposes, the value of σ_ε was fixed at 1, while $\theta = (\beta^{(0)}, \beta^{(1)}, \sigma_\gamma)'$ was estimated. In this instance, a joint value of $\beta^{(1)}_{n,t} = \beta^{(1)}$ was assumed for all individuals and time points.

In order to evaluate the impact of the magnitude of the parameter shift between the panel waves, a series of data sets with varying α values were generated.[6]

Plots and Discussion

The score plots for the data sets considered in this section are presented in Figures 2.10 and 2.22. Upon examination of the heatmap score plots with time as the independent plot variable (Figure 2.10a), a discernible pattern emerges in the

[6]The values used for α were $\{0, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95, 1, 1.5, 2\}$.



(a) Heatmap score plot

(b) Sliced score plot, $\tau_b = 1$

Figure 2.10: Score plots of data with a shift in parameter values between panel waves in the DGP with temporal position of the observations as the independent plot variable.

average score contributions depending on the temporal positions of the associated observations. The pattern is most prominent for the score directions relating to the parameter $\beta^{(1)}$, where pairs with both observations in the first panel wave have, on average, predominantly positive score contributions, while pairs with both observations in the second panel wave have, on average, negative score contributions. Pairs comprising observations from both panel waves (one observation from each wave) do not exhibit a discernible pattern. This is consistent with our expectations, as we estimate a shared $\beta^{(1)}$ for the entire panel data set, whereas in the DGP, we have a smaller value for the first panel wave and a larger one for the second panel wave. For pairs with observations from both waves, the estimate will, on average, be unbiased. In contrast, for pairs from the same panel wave, the estimated parameter will be either too small or too large for both observations, which is then reflected in the score contribution of the pair. The observed patterns become discernible at $\alpha = 0.05$ and intensify as α increases.

Similarly, the same overall pattern, though with reversed signs, can be observed for the score directions related to $\beta^{(0)}$ and σ_γ, indicating, as in the other cases, that the effect of the misspecification is most prominent for the score directions related to the misspecification, though also has an effect on the other score directions. This is further emphasised in the sliced score plot, in which the slice is selected such that the temporal position of the second observation is fixed at 1 (Figure 2.10b). This plot provides a more effective visual representation of the impact of the misspecification, demonstrating that the effect is most pronounced for the score direction associated with $\beta^{(1)}$, the parameter subject to the misspecification.

The plots in which the covariate x or the predicted utility is used as the independent plot variable do not display any discernible patterns, as illustrated in Figure 2.22.

This demonstrates the feasibility of identifying parameter shifts over time through the utilisation of these diagnostic plots. As the pattern is most pronounced for the score direction related to $\beta^{(1)}$, it is also possible to ascertain which part of the model is the source of the misspecification.

2.5.5 Structural Breaks in Time - Parameter Split

As in Section 2.5.4, the data examples in this section represent a change in the effect of a covariate between the panel waves. However, in this example, the population is divided into two equal groups, designated P_1 and P_2. In the first panel wave, the entire population has a joint parameter vector governing the DGP, whilst in the second panel wave, the parameter that governs the effect of a covariate is increased for one half of the population, while it is decreased for the other half. This could reflect a divergence in the sentiments held by the population regarding a particular attribute between two panel waves, such as attitudes towards green energy, public transportation, electric vehicles, or immigration.[7]

Data Generating Process and the Misspecified Model

The DGP and model for this case are designed analogously to those in Section 2.5.4, with the exception of the definition of the parameter $\beta^{(1)}_{n,t}$ in the DGP, which in this case is defined as

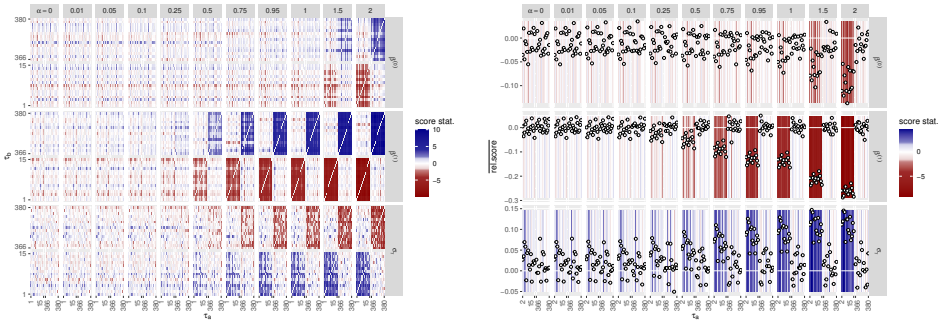
$$\beta^{(1)} n, t = \begin{cases} 1, & \tau_t < 365, \\ 1 + \alpha, & \tau_t \geq 365 \text{ and } n \in P_1, \\ 1 - \alpha, & \tau_t \geq 365 \text{ and } n \in P_2. \end{cases} \tag{2.40}$$

As in Section 2.5.4, the impact of varying the magnitude of the parameter split was investigated by utilising a range of α values were used in the generation of synthetic data sets.[8]

[7] The data used for this example was also used in Büscher and Bauer [2024].
[8] The values used for α were $\{0, 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.95, 1, 1.5, 2\}$.

Plots and Discussion

The score plots for the models estimated on data with a split in the parameter value in the DGP between panel waves are presented in Figures 2.11 and 2.23. Upon exam-



(a) Heatmap score plot

(b) Sliced score plot, $\tau_b = 1$

Figure 2.11: Score plots of data with a split in parameter values in the population of DMs in the second panel wave in the DGP with temporal position of the observations as the independent plot variable.

ination of the heatmap score plot with the temporal position of the choice occasions serving as the independent plot variable (Figure 2.11a), it becomes evident that there is a discernible correlation between the average score contributions and the temporal positions of the observations. The pattern is most prominent for the score directions relating to the parameter $\beta^{(1)}$, whereby pairs with both observations in the first panel wave have, on average, predominantly negative score contributions, while pairs with both observations in the second panel wave have, on average, positive score contributions. Pairs comprising observations from both panel waves (one observation from each wave) do not exhibit a discernible pattern. Similar patterns can be observed for the score directions related to $\beta^{(0)}$ and σ_γ, though they are less pronounced, and for the score direction relating to σ_γ, with reversed signs. The relating sliced score plot, in which the slices are selected such that the temporal position of the second observation in each pair is equal to 1, is shown in Figure 2.11b. This plot also highlights the aforementioned dependency, and demonstrates that the effect is strongest for the score direction related to the parameter $\beta^{(1)}$.

The plots in which the covariate x or the predicted utility is used as the independent plot variable do not exhibit any discernible patterns, as illustrated in Figure 2.23.

This illustrates the feasibility of identifying parameter shifts over time through the utilisation of these diagnostic plots. As the pattern is most pronounced for the score direction related to $\beta^{(1)}$, it is even possible to obtain evidence regarding the

source of the misspecification within the model. A comparison of the plot for the parameter split (Figure 2.11) with those for the parameter shift (Figure 2.10) from Section 2.5.4 reveals strong similarities. This is to be expected, given that the misspecifications are similar between the two cases. However, it also makes it challenging to identify the specific misspecification present in the model from the plots alone.

2.5.6 Auto-Regressive Error Structure

In this example, binary choice probit data sets are generated with an auto-regressive error structure, with each data set characterised by a different auto-correlation coefficient. Subsequently, a probit model was estimated for each data set, which does not account for the auto-regressive error structure[9]. There are numerous potential sources of such error processes in real-world data sets. One illustrative example is the omission of variables from the model that are correlated over time, for instance, the effect of the weather on transportation mode choice.

Data Generating Process and the Misspecified Model

In this scenario, the DGP is specified as

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{2.41}$$

$$U_{n,t,2} = \beta^{(0)} + \beta^{(1)} x_{n,t} + \varepsilon_{n,t,2}, \tag{2.42}$$

$$\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' = \rho \varepsilon_{n,t-1} + \tilde{\varepsilon}_{n,t}, \tag{2.43}$$

with $\tilde{\varepsilon}_{n,t} = (\tilde{\varepsilon}_{n,t,1}, \tilde{\varepsilon}_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 1-\rho^2 & 0 \\ 0 & 1-\rho^2 \end{pmatrix}\right)$, $\beta^{(0)} = 1$, $\beta^{(1)} = 1$, and varying values of ρ between -0.99 and 0.99 for different simulation setups[10]. The regressors $x_{n,t}$ are drawn iid from a standard normal distribution.

The estimation is then based on a misspecified model, wherein the correlation between observations from one decision maker is assumed to be from a mixed parameter for the ASC instead of from the auto-regressive error process. The model is thus specified as

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{2.44}$$

$$U_{n,t,2} = \beta^{(0)} + \beta^{(1)} x_{n,t} + \gamma_n + \varepsilon_{n,t,2}, \tag{2.45}$$

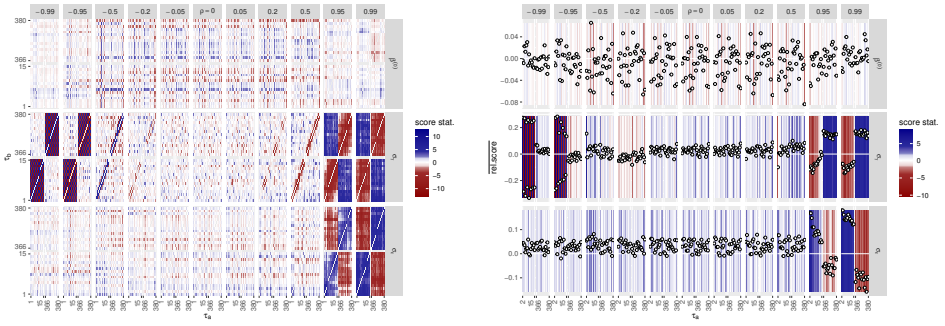
[9] The data used for this example was also used in Büscher and Bauer [2024].

[10] Values used for ρ were $\{-0.99, -0.95, -0.5, -0.2, -0.05, 0, 0.05, 0.2, 0.5, 0.95, 0.99\}$

with $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_\varepsilon^2 \end{pmatrix}\right)$ and $\gamma_n \sim \mathcal{N}(0, \sigma_\gamma^2)$. So, in contrast to the DGP, in the model we assume independent errors across different observations. The parameter $\beta^{(1)}$ is set to 1 to ensure identification and the parameter vector $\theta = (\beta^{(0)}, \sigma_\gamma, \sigma_\varepsilon)'$ is estimated.

Plots and Discussion

The score plots for the models estimated on data with an auto-regressive error structure are presented in Figures 2.12 and 2.24. Upon examination of the heatmap score



(a) Heatmap score plot

(b) Sliced score plot, $\tau_b = 1$

Figure 2.12: Score plots of data with an auto regressive error process in the DGP with temporal position of the observations as the independent plot variable.

plots with the covariate x or the predicted expected utility of the second alternative as the independent plot variable, as illustrated in Figure 2.24, no discernible patterns emerge, as would be expected.

In contrast, the score plots with time on the independent axes (Figure 2.12) demonstrates clear patterns in the score directions related to the parameters σ_γ and σ_ε. For the score direction relating to σ_γ, a discernible pattern emerges when $|\rho| \geq 0.2$. For positive auto-correlation coefficients, observational pairs with a small temporal distance between the choice occasions exhibit a negative score, whereas observations with larger temporal distances predominantly demonstrate positive score contributions. This indicates that the covariance between observations is dependent on their temporal distance, which is true for the DGP considered here.

In the case of negative auto-correlation coefficients, an even more detailed pattern can be observed in the score contributions related to σ_γ. While no discernible pattern emerges for pairs of observations seperated by a considerable temporal distance, a pronounced tendency is observed in the score contributions for pairs of observations

with small temporal distance. Notably, the sign of these contributions exhibits an alternating pattern, depending on whether the temporal distance is an even or odd number of steps. Given that two observations with an auto-regressive error structure with a negative auto-correlation coefficient ρ and odd temporal distance exhibit a negative correlation ($\rho^{2k+1} < 0$), while they have a positive correlation when at an even distance ($\rho^{2k} > 0$), the observed alternating pattern can be interpreted as evidence of such a correlation structure. Although the complete pattern is only discernible for what might be considered extreme negative correlation coefficients, $\rho \leq -0.95$, first indications are already apparent for the case with $\rho = -0.2$. In this instance, pairs of observations separated by one temporal step predominantly exhibit negative score contributions, whereas those separated by two temporal steps display cases of predominantly positive score contributions. In the case with $\rho = -0.5$, the pattern is more pronounced, although the signs of the scores are reversed.

An examination of the sliced score plot (Figure 2.12b), in which the slices were selected such that $\tau_b = 1$, reveals that the observed patterns begin to diminish within the first panel wave as the temporal distance between observations increases. However, the patterns for observational pairs across panel waves remain consistent.

For the score direction relating to σ_ε, a discernible pattern emerges only for extreme positive auto-correlation coefficients with $\rho \geq 0.95$. No patterns are visible for the score direction relating to $\beta^{(0)}$, which is to be expected given that the parameter $\beta^{(0)}$ relates to the effect of the covariate, which is independent of the auto-regressive error structure.

In conclusion, this shows that with these plots, it is not only possible to find indications of auto-correlated error structures in discrete choice models, it is further possible to find indications towards the nature of this process (positive or negative correlation).

2.6 Real-World Example

In order to demonstrate the practical applicability of the proposed score plots, we applied them to enhance model selection for an established real-world data set, the "`Train`" data set as included in the `R` package `mlogit` [Croissant, 2020].

2.6.1 Data Set Description

The stated preference data set was collected by the Hague Consulting Group in 1987 as a within-mode experiment, in which respondents were presented with two dif-

ferent rail options and were asked to choose between them. The objective of the experiment was to evaluate the relative importance of four rail service attributes: fare price, journey time, number of rail-to-rail transfers, and the comfort level. This data set was selected for the following reasons: (1) it has been extensively studied and is therefore well known within the transportation and choice modelling community [see, for example, Ben-Akiva and Morikawa, 1990; Bradley and Daly, 1991; Ben-Akiva et al., 1993; Meijer and Rouwendal, 2006]; and (2) it is publicly accessible to practitioners and researchers, allowing for the replication of our findings, given that it is included as a data example in the `R` package `mlogit` [Croissant, 2020].

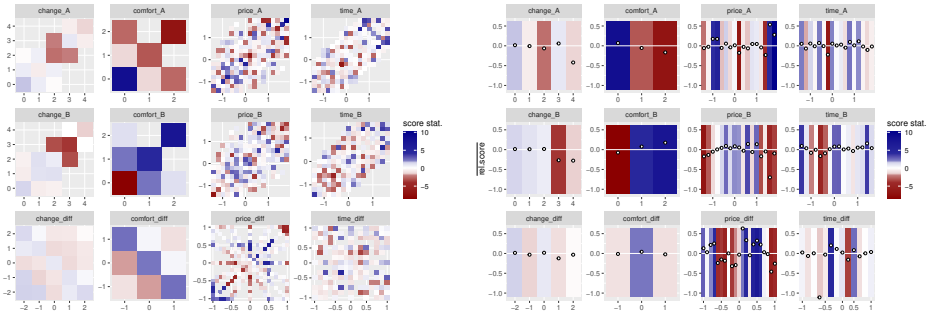
The data set, as included in the `mlogit` package, comprises 235 individuals, with between 5 and 19 stated choice per individual (12.5 on average), and a total of 2929 observations of stated preferences between two train alternatives. For the two alternatives (denoted `A` and `B`), the following variables were shown: fare price (`price`), travel time (`time`), number of train-to-train transfers (`change`), and a measure of comfort level (`comfort`). No sociodemographic information on the decision makers is available.

2.6.2 Initial Model

In line with the model used by Meijer and Rouwendal [2006], the initial model included all available covariates in a linear form in the utility function, with a fixed `price` effect coefficient for all individuals and the effect coefficients for `time`, `change`, and `comfort` treated as random between individuals, with a diagonal covariance matrix. The resulting model is

$$\begin{aligned} U_{j,n,t} = \beta_{\texttt{price}}\texttt{price}_{j,n,t} + \gamma_{\texttt{comfort},n}\texttt{comfort}_{j,n,t} + \gamma_{\texttt{change},n}\texttt{change}_{j,n,t} \\ + \gamma_{\texttt{time},n}\texttt{time}_{j,n,t} + \varepsilon_{j,n,t}, \end{aligned} \tag{2.46}$$

for $j \in \{\texttt{A}, \texttt{B}\}$ with $\varepsilon_{j,n,t} \sim \mathcal{N}(0, 0.5^2)$, $\gamma_{\texttt{var},n} \sim \mathcal{N}(\beta_{\texttt{var}}, \sigma^2_{\texttt{var}})$, for all $\texttt{var} \in \{\texttt{comfort}, \texttt{change}, \texttt{time}\}$. Accordingly, the parameter vector that was required to be estimated was $\theta = (\beta_{\texttt{price}}, \beta_{\texttt{comfort}}, \beta_{\texttt{change}}, \beta_{\texttt{time}}, \sigma_{\texttt{comfort}}, \sigma_{\texttt{change}}, \sigma_{\texttt{time}})'$. The index n denotes the individual and t the choice occasion. In order to facilitate numerical stability, the `time` and `price` covariates were standardised by subtracting their respective overall means and dividing by their overall standard deviations. As `comfort` reflects a categorical variable and `change` presented with only five distinct values, these were not standardised.



(a) Heatmap plot

(b) Diagonally sliced plot with $\Pi_p(x_a) = \Pi_p(x_b)\ \forall p$

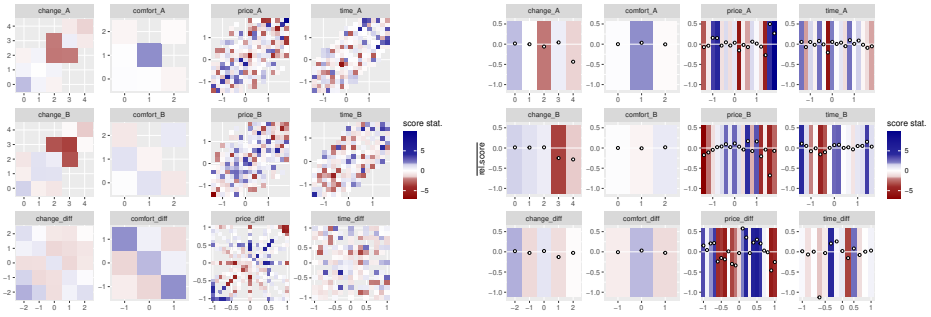
Figure 2.13: Score plots of the score contributions related to the ASC parameter of alternative `B` for the initial model with added ASC. Each panel employs a different independent plot variable, as indicated above the respective panels.

2.6.3 Score Plots and Model Selection

In the following, we will attempt to progressively enhance the original model, and in each step utilise the score plots to decide which change to the model we will implement. As we proceed, a record will be kept of the log-CML value (ll_{CML}) of each model, along with the composite likelihood Akaike Information Criterion (CLAIC) and the composite likelihood Bayesian information criterion (CLBIC) [Gao and Song, 2010], as presented in Table 2.2. This will allow for the observation of the progress made during the model selection stage.

In order to most effectively identify non-linear effects w.r.t. the covariate values in the utility function we first add an ASC for the alternative `B`. This allows the score direction related to the ASC parameter $\beta_{\text{B}}^{(0)}$ to be calculated for each pair of observations. Subsequently, the heatmap score plots and the diagonally sliced score plots were generated, incorporating the covariate values of alternatives `A` and `B`, as well as their differences, as independent plot variables. For the initial model, as defined in (2.46), with an additional ASC parameter $\beta_{\text{B}}^{(0)}$, these plots are presented in Figure 2.13.

A dependence of the mean score on the differences in comfort level can be observed when the score plots with the `comfort_diff` as the independent plot variable are examined in Figure 2.13. Upon examination of the scores depending on the comfort levels `comfort_A` and scores by `comfort_B` of the two alternatives, it becomes evident that the mean scores for pairs of observations with `comfort_A` $= 0$ or `comfort_B` $= 0$ for both observations exhibit a different sign compared to those with



(a) Heatmap plot

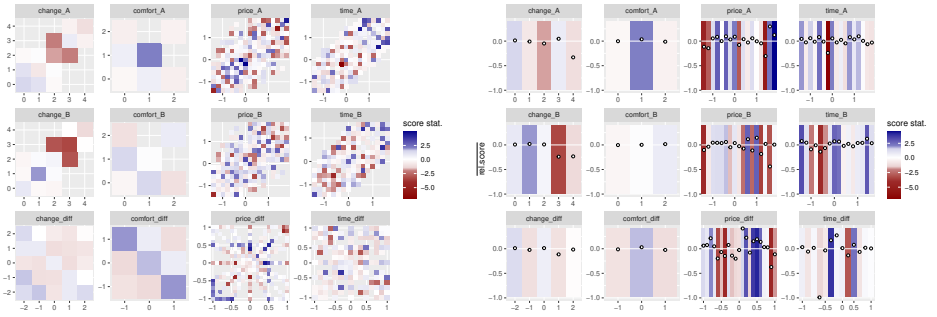
(b) Diagonally sliced plot with $\Pi_p(x_a) = \Pi_p(x_b)\ \forall p$

Figure 2.14: Score plots of the score contributions related to the ASC parameter of alternative `B` for the model with added dummy variable for the comfort level zero. Each panel employs a different independent plot variable, as indicated above the respective panels.

other comfort levels. It was therefore decided that a dummy variable for the comfort level zero should be included in the next model. The resulting score plots from this model are presented in Figure 2.14.

The initial observation made with regard to the plots presented in Figure 2.14 is that the dependency of the mean scores on the comfort level has diminished. Upon examination of the impact of price differences, particularly in the diagonally sliced score plot (see Figure 2.14b), it becomes evident that positive score contributions are observed for price differences approaching -1, while negative mean score contributions are evident for price differences between -0.6 and -0.1. This is followed by a resurgence of positive score contributions up to a price difference of approximately $+0.6$, and then again negative average score contributions for price differences approaching $+1$ (measured in standard deviations of ticket prices). It was therefore decided to add a non-linear dependency of the utility on the `price` into the utility function by including the `price` variable as a polynomial function of order three. The resulting score plots from this model can be seen in Figure 2.15.

The inclusion of a polynomial dependence of order three of the utility on the price does not entirely eliminate the dependence of the score contributions on the price differences. A pattern of predominantly positive score values for positive price differences and negative score contributions for negative price differences, provided that the price differences are not too large, remains evident. Accordingly, a dummy variable, $I(\Delta\texttt{price} > 0)$, was added to the model to indicate whether the alternative is the more expensive one. The resulting score plots from this model are presented



(a) Heatmap plot

(b) Diagonally sliced plot with $\Pi_p(x_a) = \Pi_p(x_b)\ \forall p$

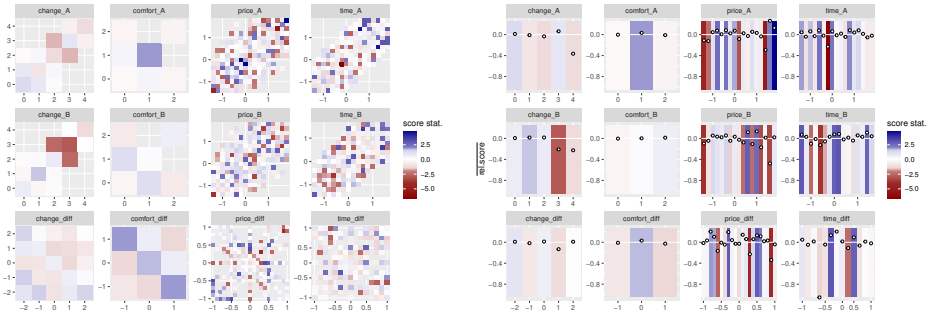
Figure 2.15: Score plots of the score contributions related to the ASC parameter of alternative `B` for the model with added polynomial dependence of order three of the `price`. Each panel employs a different independent plot variable, as indicated above the respective panels.

in Figure 2.16.

The data still exhibits cumulations of predominantly positive and negative score values depending on the price, though no discernible patterns are evident. It is important to exercise caution when interpreting the patterns observed in the plots for the covariate values of alternative `A` or `B` (upper two rows in Figure 2.16), given that the data originates from an unlabelled stated choice experiment. This is because, in the absence of a discernible pattern in the difference of the covariates between the two alternatives, estimating a different covariate effect depending on the alternative would elude meaningful interpretation. Further iterations of this process may be conducted to improve the model, for instance, by investigating the correlation between the score contributions and the number of train-to-train transfers (`changes`). We have elected to refrain from pursuing this matter further, for the sake of keeping the paper concise.

Table 2.2 provides a summary of the composite marginal log-likelihood, the CLAIC, and the CLBIC for the models that were considered during the model selection phase. This illustrates that by reducing the dependency of the score contributions on the covariate values, as evidenced by the plots in Figures 2.13-2.16, as a consequence all three metrics were enhanced in each step of the model selection process conducted here.

As an additional diagnostic step, the mean score contributions the different score directions, as a function of the `choiceid` (the identifier for the choice occasion), are presented in Figure 2.17. In the case of the score directions associated with the para-



(a) Heatmap plot

(b) Diagonally sliced plot with $\Pi_p(x_a) = \Pi_p(x_b)\ \forall p$

Figure 2.16: Score plots of the score contributions related to the ASC parameter of alternative `B` for the model with added dummy variable for if the alternative is the more expensive one Δ`price` > 0. Each panel employs a different independent plot variable, as indicated above the respective panels.

model	ll_{CML}	CLAIC	CLBIC
initial + ASC	−3 407.989	6 831.977	6 879.836
+ $I($`comfort` $= 0)$	−3 350.933	6 719.865	6 773.707
+ `price`^{2} + `price`^{3}	−3 262.426	6 546.851	6 612.658
+ $I(\Delta$`price` $> 0)$	−3 237.505	6 499.009	6 570.798

Table 2.2: Model selection criteria for different models with each model containing the previous as a restricted version.

meter of `price`^{2}, a discernible pattern emerges. Here, the score contributions of pairs of observations where one of the observations has a `choiceid` of 11 or below exhibit predominantly positive score contributions, whereas other pairs of observations demonstrate predominantly negative score contributions. A Lagrange multiplier type test was conducted to test the null hypothesis that the parameter vector θ is identical for observations with `choiceid` ≤ 11 and those with `choiceid` > 11. The resulting test statistic is $LM = 1.952\,604$. As the test statistic asymptotically follows an F-distribution with 12 and 223 degrees of freedom, the critical value of the test statistic at a significance level of 5% is $c_\alpha = 1.795\,782$. Consequently, the null hypothesis is rejected at a 5% significance level [for details on the test applied here, please refer to Büscher and Bauer, 2024]. Such a change of parameters within a stated choice experiment after a number of choices could be indicative of survey fatigue [see, for example, Porter et al., 2004], as it might indicate a change in the

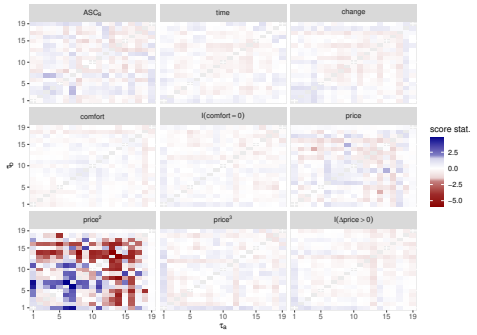


Figure 2.17: Score plots of the score contributions depending on choice id for the model with added dummy variable for if the alternative is the more expensive one $\Delta\texttt{price} > 0$. The covariate to which the score direction relates to is indicated above each panel of the plot.

choice criteria over time. To confirm this hypothesis would require further investigation, which is beyond the scope of this paper.

2.6.4 Final Model and Model Comparison

In order to create the final model, the ASC for alternative `B` was removed, as the alternatives in the data set are unlabelled, and therefore an ASC would be devoid of meaningful interpretation. The final model is therefore represented by

$$\begin{aligned} U_{j,n,t} = {} & \beta_{\texttt{price}}\texttt{price}_{j,n,t} + \beta_{\texttt{price2}}\texttt{price}^2_{j,n,t} \\ & + \beta_{\texttt{price3}}\texttt{price}^3_{j,n,t} + \beta_{\Delta\texttt{price}} I(\Delta\texttt{price}_{j,n,t} > 0) \\ & + \gamma_{\texttt{comfort},n}\texttt{comfort}_{j,n,t} + \beta_{\texttt{comfort0}} I(\texttt{comfort}_{j,n,t} = 0) \\ & + \gamma_{\texttt{change},n}\texttt{change}_{j,n,t} + \gamma_{\texttt{time},n}\texttt{time}_{j,n,t} + \varepsilon_{j,n,t}, \end{aligned} \tag{2.47}$$

for $j \in \{\texttt{A},\texttt{B}\}$, with $\Delta\texttt{price}_{\texttt{A},n,t} = (\texttt{price}_{\texttt{A},n,t} - \texttt{price}_{\texttt{B},n,t})$, $\Delta\texttt{price}_{\texttt{B},n,t} = (\texttt{price}_{\texttt{B},n,t} - \texttt{price}_{\texttt{A},n,t})$, $\varepsilon_{j,n,t} \sim \mathcal{N}(0, 0.5^2)$, and $\gamma_{\texttt{var},n} \sim \mathcal{N}(\beta_{\texttt{var}}, \sigma^2_{\texttt{var}})$, for all $\texttt{var} \in \{\texttt{comfort}, \texttt{change}, \texttt{time}\}$. A comparison of the estimated parameters and their estimated standard deviations of the initial model with the four original covariates implemented linearly in the utility function and the final model is presented in Table 2.3. In this table, the reported means are the $\beta_{\texttt{var}}$ parameters, while the variances are the estimated variances $\sigma^2_{\texttt{var}}$ of the random effects.

	initial model		final model	
	β	σ	β	σ
price	−1.674053		−1.344249	
	(0.163971)		(0.240417)	
price^2			0.358791	
			(0.098793)	
price^3			−0.054903	
			(0.024921)	
$I(\Delta\texttt{price} > 0)$			−0.522602	
			(0.140342)	
comfort	−0.898898	0.995239	−1.645860	1.019973
	(0.091785)	(0.109312)	(0.193808)	(0.127949)
$I(\texttt{comfort} = 0)$			−0.818266	
			(0.186253)	
change	−0.316850	0.658973	−0.445346	0.850831
	(0.070152)	(0.129402)	(0.082560)	(0.133965)
time	−0.795230	1.038829	−1.077024	1.226195
	(0.090155)	(0.128812)	(0.118042)	(0.151149)
ll_{CML}	−3408.651		−3237.822	
CLAIC	6831.301		6497.645	
CLBIC	6873.178		6563.451	
CV(%correct)	0.690798		0.717730	
	(0.002937)		(0.001830)	
CV($\bar{ll}$)	−0.584478		−0.554956	
	(0.002586)		(0.001135)	

Table 2.3: Comparison of estimated parameters for first and final model of the case study. Standard errors of the estimates are in parentheses below the respective estimate values.

Additionally, the table presents the composite marginal log-likelihood $ll_{\texttt{CML}}$ of both models, their CLAIC and CLBIC, and two cross-validation metrics, as recommended by Parady et al. [2021]. In order to obtain five holdout values for both cross-validation metrics, the respondents were partitioned into five equally sized groups of 47 individuals. For each set of individuals the models were estimated on the observations of the remaining 188 individuals. The mean predicted log-likelihood ($\bar{ll}$) and the percentage of correctly predicted observations (%correct) in the remaining sample were then calculated. The mean of these holdout values is then presented in Table 2.3 as the cross-validation value of the measures, accompanied by their empirical standard deviations in parentheses below. All estimated parameters in both

models are statistically significant at the 5% significance level[11],and the final model represents an improvement over the initial model w.r.t. all considered model comparison metrics, including the two cross-validation measures. Therefore, by utilising the score plots as guides for model selection, we were able to enhance the initial model considerably and discern a non-linear dependency within the data, as well as an indication of survey fatigue, which might otherwise have remained undetected. This practical example demonstrates the utility of score plots in guiding model selection and diagnosing model issues in real-world data sets.

2.7 Conclusion

In this paper, we have demonstrated how the often complex workflows in discrete choice modelling, which can lead to substantial variations in modelling outcomes between researchers, can be guided by the utilisation of the proposed score plots. These plots, inspired by residual plots in regression modelling, provide a flexible and intuitive diagnostic tool, informed by statistical properties of the score contributions, as discussed in Sections 2.2 and 2.3.

The synthetic data examples presented in Section 2.5 highlight the versatility of the proposed CML score plots as diagnostic tools. Specifically, they enable practitioners to: (1) identify and characterise non-linearities in covariate effects within the utility functions; (2) detect structural breaks in the DGP and locate them in terms of their temporal position and the model component most affected; (3) uncover temporal dependencies in the covariance between observations, such as those introduced by an auto-regressive error structure; and (4) guide the model selection process by identifying missing covariates.

In Section 2.6, a real-world case study demonstrates the practical utility of these plots for model specification and diagnosis, particularly in guiding the inclusion of non-linear transformations of covariates and in detecting potential signs of survey fatigue, as indicated by a structural break in the model. The score plots are computationally efficient and interpretable, reducing the need to estimate a large number

[11]Büscher and Bauer [2024] have demonstrated that variance-optimal CML weights can be used for models estimated on unbalanced panel data, with the potential to reduce the variance of the estimated coefficients. This approach may serve to counteract the loss of statistical efficiency associated with CML estimation in comparison to ML estimation. This, however, introduces complications in the comparison between different models, as the log-CML function is effectively altered by the weights, and thus ll_{CML}, CLAIC, CLBIC and CV($\bar{ll}$) are not suitable for model comparison between models estimated using different power weights. Accordingly, we have elected to refrain from utilising these optimal power weights in the present case study.

of models to explore various combinations of covariates and their transformations.

By incorporating these plots into the workflow, researchers can reduce uncertainty in model selection and feature engineering, significantly accelerating the process (as shown in the examples in Sections 2.5.1, 2.5.2, 2.5.3 and 2.6). Additionally, during model diagnosis, score plots provide insights into potential violations of model assumptions such as homoscedasticity and error independence (example in Section 2.5.6), or evidence of changes in the DGP over time (examples in Sections 2.5.4, 2.5.5 and 2.6).

We propose a revised modelling workflow for employing score plots in practice: (1) During model selection, include ASCs for all but the reference choice alternative to enable ASC-related score plots. (2) Use score plots for the ASC score directions to guide the inclusion of additional covariates in the model, with the covariate of interest as the independent plot variable. (3) Check for non-linearities w.r.t. a covariate x by using score plots for the ASC score directions with the covariate x as the independent plot variable. (4) Post model specification, optionally exclude ASCs lacking meaningful interpretation from the final model, i.e. in the case of unlabelled alternatives. (5) Use the time of the observations as independent plot variable to uncover dynamic model properties, such as auto-regressive error processes or time-dependent model behaviour. (6) At each step, complement the score plots with appropriate statistical tests or model selection criteria to validate and justify the decisions made during the model selection process.

The use of score plots offers substantial benefits by providing guidance to practitioners to enhance the specification of utility functions, explore covariate non-linearities, and thus potentially reduce the variability of modelling outcomes across researchers. While pairwise CML score information is required for the score plots during the model specification and diagnosis phases, maximum likelihood estimation can be used for final model estimation to take advantage of its statistical efficiency.

Further investigation of score plots would be facilitated by (1) applying score plots to a diverse set of real-world data sets to provide additional evidence of their applicability in practical statistical modelling; (2) applying score plots to different types of models other than MNP models, such as mixed multinomial logit (MMNL) models or models outside the DCM context; and (3) investigating plots based on single observations estimated using CML or ML methods.[12] This may facilitate the wider usage of score plots by practitioners.

[12]Initial efforts in this direction were conducted by us, wherein the assumption of independence between all observations was made to calculate score contributions for individual observations. However, this approach, did not yield results as promising as those presented in this paper, and thus requires further investigation.

CRediT author statement

Sebastian Büscher: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original draft, Writing - Review & Editing, Visualization.

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 356500581 which is gratefully acknowledged. The author would also like to extend their gratitude to Dietmar Bauer, Manuel Batram and Lennart Oelschläger, who contributed to the codebase used for the calculations, and Kaja Balzereit for proofreading the original manuscript.

Declarations of interest

Declarations of interest: none

The funding agency Deutsche Forschungsgemeinschaft (DFG) had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Appendix 2.A

Lemma 2.1. *Let θ_0 be the true parameter vector governing the DGP, and let $\hat{\theta}$ be an estimator of θ_0 such that $P_{n,i}(\hat{\theta})$ is a consistent estimator of $P_{n,i}(\theta_0)$. Furthermore, assume the probabilities for all observed choice occasions n and all choice alternatives i to be bounded from below ($P_{n,i}(\theta) > \delta \; \forall n, i$ for some $\delta > 0$) and that the gradients of the probabilities w.r.t. θ are bounded ($\left|\frac{\partial}{\partial \theta^{(k)}} P_{n,i}(\theta)\right| < M$ for some $0 < M < \infty$). Then the score contributions conditional expected value will converge to zero as the number of observations N goes to*

infinity, so

$$\mathbb{E}\left(s_n^{(k)}(\hat{\theta}; X_n, y_n) \mid X_n\right) \overset{N\to\infty}{\longrightarrow} 0. \tag{2.14}$$

Proof of Lemma 2.1. Using

(a) that the sum of the choice probabilities over all alternatives has to equal one and thus the derivative of the sum is equal to zero ($\sum_{i=1}^{J} \frac{\partial}{\partial \theta^{(k)}} P_{n,i}(\hat{\theta}) = 0$) and

(b) that the expected value of the score is the sum of the conditional expectations of the score, conditional on the chosen alternative, times their choice probabilities,

(c) the dominated convergence theorem [Çinlar, 2011], which, applied to sequence of random variables $\{X_n\}$ states that if (1) $\exists X$ such that $X_n \xrightarrow{P} X$ and (2) $\exists Y$ with $\mathbb{E}(|Y|) < \infty$ such that $|X_n| \leq Y \; \forall n \in \mathbb{N}$, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$, and

(d) $\sum_{i=1}^{J} \frac{\frac{\partial}{\partial \theta^{(k)}} P_{n,i}(\hat{\theta})}{P_{n,i}(\hat{\theta})} P_{n,i}(\theta_0) < \sum_{i=1}^{J} \frac{M}{\delta} P_{n,i}(\theta_0) = \frac{M}{\delta} < \infty$,

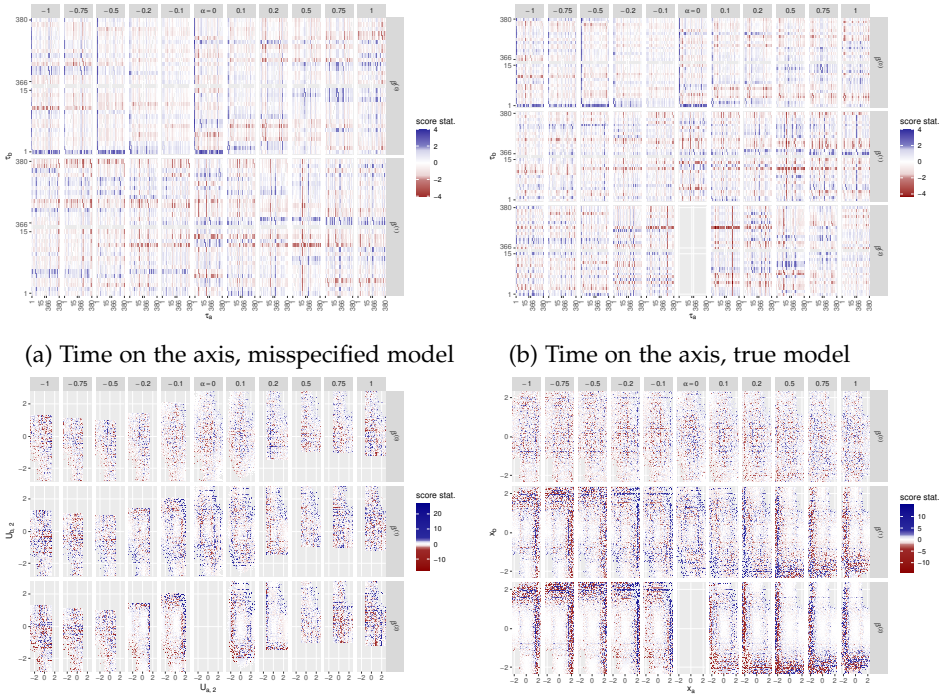
we get for a consistent estimator of $P_{n,i}(\theta_0)$

$$\begin{aligned}
\lim_{N\to\infty} \mathbb{E}\left(s_n^{(k)}(\hat{\theta}; X_n, y_n) \mid X_n\right) &= \lim_{N\to\infty} \mathbb{E}\left(\sum_{i=1}^{J} \frac{\frac{\partial}{\partial \theta^{(k)}} P_{n,i}(\hat{\theta})}{P_{n,i}(\hat{\theta})} P_{n,i}(\theta_0) \mid X_n\right) \\
&= \lim_{N\to\infty} \mathbb{E}\left(\sum_{i=1}^{J} \frac{\partial}{\partial \theta^{(k)}} P_{n,i}(\hat{\theta}) \frac{P_{n,i}(\theta_0)}{P_{n,i}(\hat{\theta})} \mid X_n\right) \\
&= \mathbb{E}\left(\lim_{N\to\infty} \sum_{i=1}^{J} \frac{\partial}{\partial \theta^{(k)}} P_{n,i}(\hat{\theta}) \underbrace{\frac{P_{n,i}(\theta_0)}{P_{n,i}(\hat{\theta})}}_{\overset{N\to\infty}{\to} 1} \mid X_n\right) \\
&= \mathbb{E}\left(\lim_{N\to\infty} \underbrace{\sum_{i=1}^{J} \frac{\partial}{\partial \theta^{(k)}} P_{n,i}(\hat{\theta})}_{=0} \mid X_n\right) \\
&= 0.
\end{aligned} \tag{2.48}$$

□

Appendix 2.B

This appendix includes the additional plots from Section 2.5, which did not show any clear patterns and for which the nature of the corresponding model and DGP would not suggest any patterns in the score contributions of pairs of observations to surface. For sake of completeness and clarity, we have included these plots in this appendix.



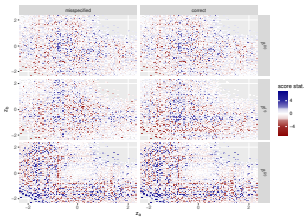
(a) Time on the axis, misspecified model

(b) Time on the axis, true model

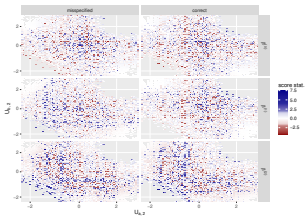
(c) Heatmap score plot with predicted expected utility of second alternative on the axis, correctly specified model

(d) Heatmap score plot with covariate x on the axis, correctly specified model

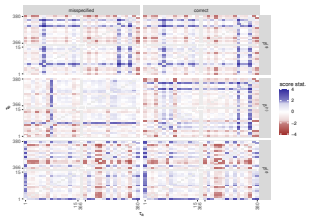
Figure 2.18: Additional score plots of data from a DGP with a utility function quadratic in x and a model with linear dependency. Each sub-figure contains a grid of plots, with each column representing a different data set and a different value of α, representing the effect of x^2 on the utility, as indicated above the columns. Each row represents the score direction w.r.t. a different model parameter, indicated on the right of the rows.



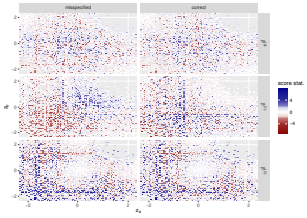
(a) Heatmap score plot, $f_{\text{step}}(x)$, covariate z on the axis



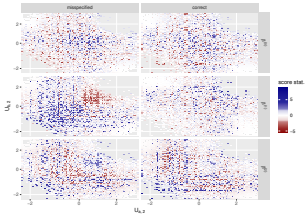
(b) Heatmap score plot, $f_{\text{step}}(x)$, predicted expected utility of second alternative on the axis



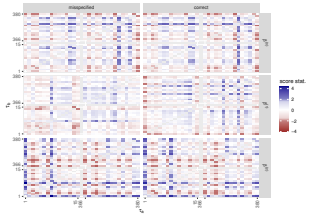
(c) Heatmap score plot, $f_{\text{step}}(x)$, time on the axis



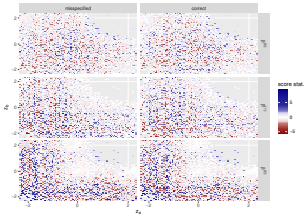
(d) Heatmap score plot, $f_{\text{ReLU}}(x)$, covariate z on the axis



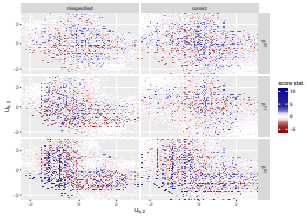
(e) Heatmap score plot, $f_{\text{ReLU}}(x)$, predicted expected utility of second alternative on the axis



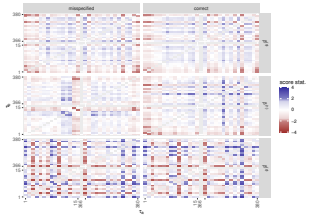
(f) Heatmap score plot, $f_{\text{ReLU}}(x)$, time on the axis



(g) Heatmap score plot, $f_{\text{wedge}}(x)$, covariate z on the axis

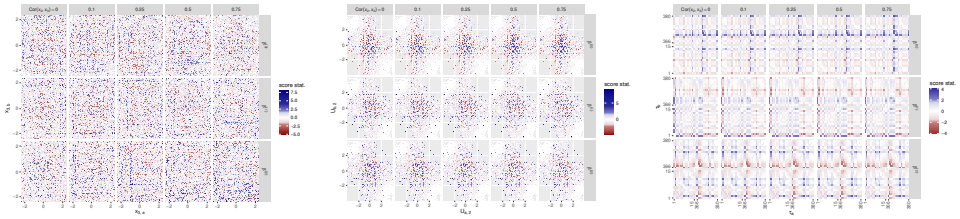


(h) Heatmap score plot, $f_{\text{wedge}}(x)$, predicted expected utility of second alternative on the axis



(i) Heatmap score plot, $f_{\text{wedge}}(x)$, time on the axis

Figure 2.19: Additional score plots of data from a DGP with a utility function with a non-linear dependency $f_{\circ}(x)$ on x. Within each sub-figure the left column of plots represent the misspecified model, whilst the right one represents the correctly specified model.

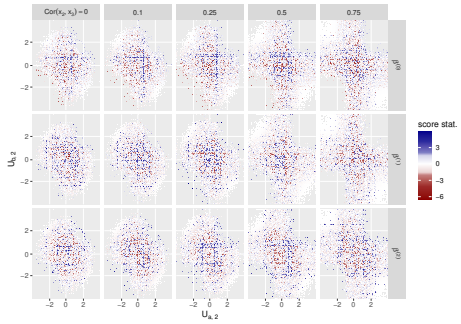


(a) Covariate x_3 on the axes

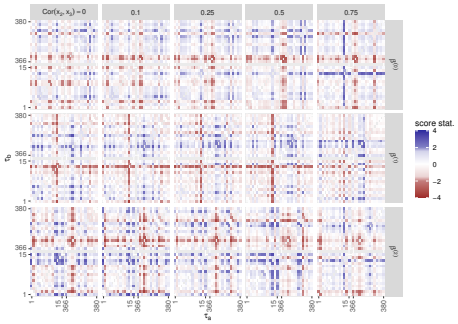
(b) Predicted expected utility of second alternative on the axes

(c) Time on the axis

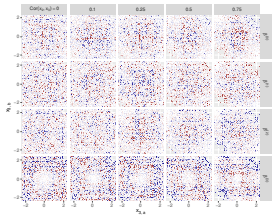
Figure 2.20: Heatmap score plots of data which include variables x_1 and x_2 in the DGP and in the model. The values displayed above the plot grids indicate the covariance between x_2 and x_3.



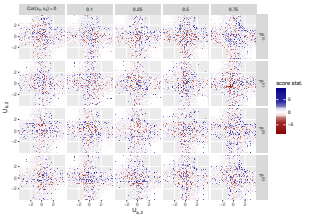
(a) Predicted expected utility of second alternative on the axis, misspecified model



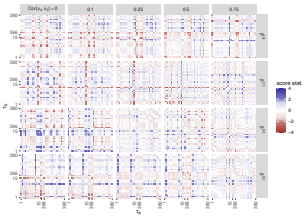
(b) Time on the axis, misspecified model



(c) Covariate x_3 on the axis, correctly specified model

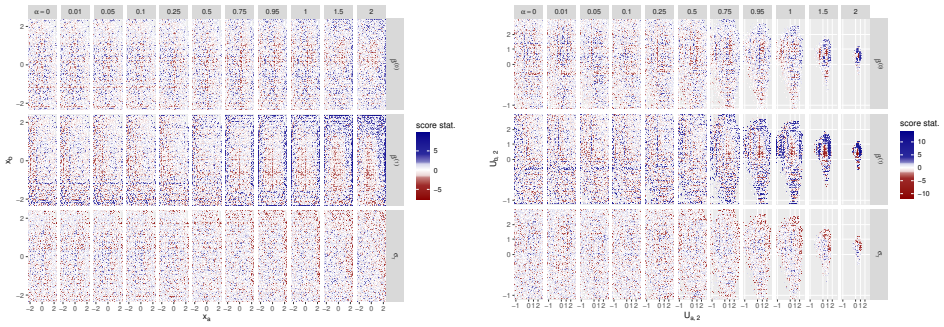


(d) Predicted expected utility of second alternative on the axis, correctly specified model



(e) Time on the axis, correctly specified model

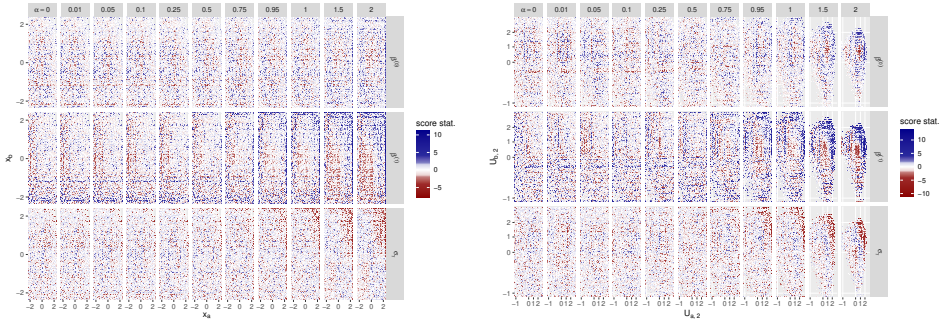
Figure 2.21: Additional score plots of data which include variables x_1, x_2 and x_3 in the DGP. Misspecified models (top two plots) only contain the variables x_1 and x_2, the correctly specified models (bottom two plots) also include x_3. The values displayed above the plot grids indicate the covariance between x_2 and x_3.



(a) Covariate x on the axis

(b) Predicted expected utility of second alternative on the axis

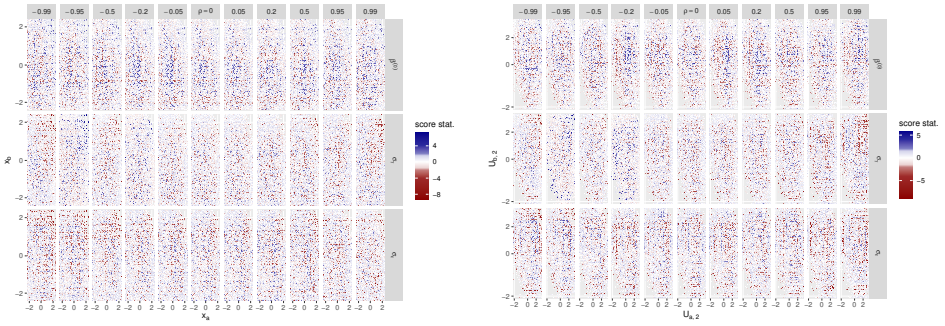
Figure 2.22: Additional heatmap score plots of data with a shift in parameter values between panel waves in the DGP.



(a) Covariate x on the axis

(b) Predicted expected utility of second alternative on the axis

Figure 2.23: Additional heatmap score plots of data with a split in parameter values in the population of DMs in second panel wave in the DGP.



(a) Covariate x on the axis

(b) Predicted expected utility of second alternative on the axis

Figure 2.24: Additional heatmap score plots of data with an auto regressive error process in the DGP with temporal position of the observations as the independent plot variable.

CHAPTER

Using Lagrange Multiplier Type Tests to Detect Structural Intra-Person Heterogeneity in Composite Marginal Likelihood Estimation in Panel Data Sets

3

Abstract

Gradient-based Lagrange multiplier-type tests represent a valuable tool for discriminating between nested models, obviating the necessity to estimate the unrestricted model. This is particularly advantageous when testing for pooling in panel data sets, as it permits the testing of multiple groupings without the necessity of re-estimating the model for each grouping.

In this paper, we demonstrate that the use of pairwise composite marginal likelihood (CML) estimation enables the comparison of gradients between different CML contributions of pairs of observations for individuals. This allows for the testing of

pooling over time, as well as the identification of neglected temporal correlation. The CML approach thus offers a degree of flexibility that is not present in the classical likelihood setting.

Theoretical derivations of the asymptotic distribution of the test statistics under the null hypothesis are provided for the special case of multinomial probit models, thereby forming the basis for the statistical interpretation of the test statistic.

Moreover, a comprehensive simulation study was conducted to assess the finite-sample performance of the test statistics. In particular, the distribution of the test statistic under the null hypothesis and the rejection rates of the tests under various types and degrees of violations of the null hypothesis were evaluated using synthetic panel data sets of varying sizes. This empirical evaluation provides insights into the effectiveness and reliability of the proposed tests in detecting intra-personal heterogeneity and into causes of misspecifications in the deterministic utility structure.

Keywords: probit modelling, composite marginal likelihood, Lagrange multiplier test, score test, heterogeneity

3.1 Introduction

Multinomial probit (MNP) models are used for modelling discrete choices made by decision-makers, with the understanding that these choices depend on a set of covariates which describe the characteristics of the decision-maker in question, as well as the specific choice alternatives. In accordance with the random utility interpretation of these models, it is assumed that decision-maker n forms an utility $U_{n,t,j}$ for each alternative $j \in \{1, \dots, J\}$ in their t-th choice situation, where $J \in \mathbb{N}$ denotes the number of possible alternatives. This utility is composed of a systematic part $V_{n,t,j}$ and a random (unobservable) part $\varepsilon_{n,t,j}$. The decision-makers then selects the alternative that provides the highest utility. MNP models assume additive separability and normal distribution for the error term, such that the utility can be expressed as

$$U_{n,t,j} = V_{n,t,j} + \varepsilon_{n,t,j}, \quad \varepsilon_{n,t,\cdot} = (\varepsilon_{n,t,j})_{j=1,\dots,J} \in \mathbb{R}^J, \quad \varepsilon_{n,t,\cdot} \sim \mathcal{N}(0, \Sigma), \tag{3.1}$$

where $\mathcal{N}(0, \Sigma)$ denotes the multivariate normal distribution with expectation zero and variance $\Sigma \in \mathbb{R}^{J \times J}$. It is then assumed that the observed choice $y_{n,t}$ represents the alternative with the highest random utility, therefore $y_{n,t} = \arg\max_j U_{n,t,j}$. Such models have been used extensively in a range of disciplines, including transportation [Bhat and Koppelman, 2003; Büscher et al., 2019], marketing [see, for example, Al-

bert and Chib, 1993; Chintagunta, 1992], and psychology [Johnson and Bruce, 1997; Berkowitsch et al., 2014], to mention just a few sources.

In many cases it may be questioned whether preferences (as encoded in the systematic part of the utility) are homogeneous across decision-makers. In addition, the assumption that preferences remain stable over time may warrant further investigation. The random term of the utility, nevertheless, is typically assumed to be independent, identically distributed (iid) over subsequent decisions, ignoring behaviour such as variety seeking (leading to negative autocorrelation) or habit formation (related to positive autocorrelation).

In certain instances, the heterogeneity of decision-makers can be directly modelled by incorporating observed characteristics of the decision-makers into the model. One illustrative example in this regard is the incorporation of income as a factor influencing purchase decisions. Other forms of preference heterogeneity have been addressed through the utilisation of fixed effects in panel data scenarios with a large number of repeated choices for relatively few decision-makers (small N, large T panels). More often, random effects are used in large N small T panels, assuming that the preference heterogeneity is independent of the other regressors.

It is evident that, when starting with an initial homogeneous model, there are numerous potential deviations that may need to be taken into account: (I) unobserved heterogeneity of decision-makers, (II) changes in preferences over time, (III) autocorrelation in the random term. In order to detect these deviations, it is necessary to employ diagnostic instruments. In many instances, the diagnostic assessment entails selecting between a more parsimonious model and an extension within a nested framework, such that the simpler model represents a restriction of the extended model. To discriminate between an unrestricted and a linearly restricted model, in the context of maximum likelihood estimation, one can choose from the classical triad of tests [Silvey, 1959]:

- A Wald-Test [Wald, 1943], where the standardised distance from the unrestricted parameter estimate $\hat{\theta}_{\text{ur}} \in \mathbb{R}^P$ to the restriction $R\theta = r \in \mathbb{R}^k$, with $R \in \mathbb{R}^{k \times P}$ is assessed using the test statistic $W = (R\hat{\theta}_{\text{ur}} - \text{r})'(R\hat{V}_{\hat{\theta}_{\text{ur}}}R')^{-1}(R\hat{\theta}_{\text{ur}} - \text{r}) \sim \chi^2_k$ under the null hypothesis $H_0 : \ R\theta = r$, where k denotes the number of restrictions and $V_{\hat{\theta}_{\text{ur}}} \in \mathbb{R}^{P \times P}$ the estimated covariance matrix of the maximum likelihood estimator of the unrestricted model.

- The Likelihood-Ratio (LR) test [Wilks, 1938], is used to compare the ratio between the maximum likelihood values of the restricted and the unrestricted model. More specifically, the test statistic consists of the natural logarithm of the ratio with $LR = -2\log(L(\hat{\theta}_{\text{r}})/L(\hat{\theta}_{\text{ur}})) \sim \chi^2_k$ under the null hypothesis, where $L(\hat{\theta})$ denotes the likelihood function evaluated at $\hat{\theta}$ and $\hat{\theta}_{\text{r}} \in \mathbb{R}^P$ denotes

the estimator in the restricted model.

- The Lagrange multiplier (LM) test [Breusch and Pagan, 1980], also referred to as score test, assesses the gradient of the unrestricted model at the maximum likelihood parameter estimate of the restricted model

$$LM = \left(\partial_{\theta_{\text{ur}}} \log L(\hat{\theta}_{\text{r}})\right)' I(\hat{\theta}_{\text{r}})^{-1} \left(\partial_{\theta_{\text{ur}}} \log L(\hat{\theta}_{\text{r}})\right) \sim \chi^2_k, \tag{3.2}$$

with $I(\hat{\theta}_{\text{r}})$ denoting an estimator of the Fisher information matrix $-\mathbb{E}\left(\partial_{\theta_{\text{ur}}}\partial_{\theta'_{\text{ur}}} \log L(\hat{\theta}_{\text{r}})\right)$.

Among these three tests, which are asymptotically equivalent, the LM test has the advantage that it requires only the estimation of the restricted model, as the gradient and Hessian with respect to the unrestricted parameter vector θ_{ur} are evaluated at the restricted maximum likelihood estimate $\hat{\theta}_{\text{r}}$.

The aforementioned deviations (I)-(III) from the restricted model correspond to changes in the parameter vector, either over decision-makers (in case of unobserved heterogeneity), over decisions taken (for example, for structural breaks in preferences over time), or to nonzero correlations of the error terms.

In order to illustrate the main idea of this paper, consider a standard linear panel data model with standard normally distributed error terms:

$$y_{n,t} = X'_{n,t}\beta + u_{n,t}, \quad n = 1, ..., N,\ t = 1, ..., T,\ u_{n,t} \sim \mathcal{N}(0, 1). \tag{3.3}$$

The logarithm of the Gaussian likelihood and the score function for this model can be written as (using shorthand notation for the data set)

$$ll(\beta; y, X) = c - 1/2 \sum_{n=1}^{N}\sum_{t=1}^{T}(y_{n,t} - X'_{n,t}\beta)^2, \tag{3.4}$$

$$\partial_\beta ll(\beta; y, X) = \sum_{n=1}^{N}\sum_{t=1}^{T} X_{n,t}(y_{n,t} - X'_{n,t}\beta), \tag{3.5}$$

from which it follows that at the maximum likelihood estimate $\hat{\beta}$

$$0 = \partial_\beta ll(\hat{\beta}; y, X) = \sum_{n=1}^{N}\sum_{t=1}^{T} X_{n,t}(\underbrace{y_{n,t} - X'_{n,t}\hat{\beta}}_{\hat{u}_{n,t}}), \tag{3.6}$$

and thus $\hat{\beta} = (\sum_{n=1}^{N}\sum_{t=1}^{T} X_{n,t}X'_{n,t})^{-1}\sum_{n=1}^{N}\sum_{t=1}^{T} X_{n,t}y_{n,t}$.

Let us now assume that there is reason to believe that the set of individuals can be partitioned into two groups, designated G_1 and G_2, where the parameter vector β differs between the groups. In this case, the model would be extended to

$$y_{n,t} = X'_{n,t}\beta + X'_{n,t}\beta_2 z_n + u_{n,t}, \quad n = 1, ..., N,\ t = 1, ..., T,\ u_{n,t} \sim \mathcal{N}(0, 1), \tag{3.7}$$

where z_n indicates whether the individual n is in group G_2. The restricted model corresponds to $\beta_2 = 0$, while $\beta_2 \neq 0$ implies heterogeneous parameters between the groups. The score of the extended unrestricted model is comprised of two elements: the zero derivative provided above, and the derivative with respect to β_2, which is equal to

$$\sum_{n=1}^{N}\sum_{t=1}^{T} \tilde{z}_n X_{n,t}(y_{n,t} - X'_{n,t}\hat{\beta}) = \sum_{n=1}^{N}\sum_{t=1}^{T} \tilde{z}_n X_{n,t} y_{n,t} - \left(\sum_{n=1}^{N}\sum_{t=1}^{T} \tilde{z}_n X_{n,t} X'_{n,t}\right)\hat{\beta}, \tag{3.8}$$

where $\tilde{z}_n = z_n - 1/2$. Now assume that the population is matched, that is to say, for each individual n_1 in group $G_1 = \{1, 2, ..., N/2\}$, there exists an individual $n_2 = N/2 + n_1$ in group $G_2 = \{N/2 + 1, ..., N\}$, such that

$$\sum_{t=1}^{T} X_{n_1,t} X'_{n_1,t} = \sum_{t=1}^{T} X_{n_2,t} X'_{n_2,t}, \tag{3.9}$$

implying $\sum_{n=1}^{N}\sum_{t=1}^{T} \tilde{z}_n X_{n,t} X'_{n,t} = 0$, which in turn means that the score with respect to β_2 equals

$$\begin{aligned}\sum_{n=1}^{N}\sum_{t=1}^{T} \tilde{z}_n X_{n,t} y_{n,t} &= \sum_{n=1}^{N}\sum_{t=1}^{T} \tilde{z}_n X_{n,t} \hat{u}_{n,t} \\ &= 1/2 \sum_{n=1}^{N/2} \underbrace{\left(\sum_{t=1}^{T} X_{n+N/2,t}\hat{u}_{n+N/2,t} - \sum_{t=1}^{T} X_{n,t}\hat{u}_{n,t}\right)}_{\Delta\bar{g}_n}.\end{aligned} \tag{3.10}$$

Under the null hypothesis, it follows that $\mathbb{E}\Delta\bar{g}_n = 0$ and the score is asymptotically normaly distributed. Its variance can be estimated from $\Delta\bar{g}_n$, given the assumption of iid sampling over individuals. This provides the foundation for the construction of an LM test statistic, which can be used to investigate the potential for pooling the data in this straightforward scenario.

In this case, it is evident that the roles of N and T can be switched to compile an LM test statistic for testing a structural break in time using the test statistic

$$LM = 1/2 \sum_{t=1}^{T/2} \underbrace{\left(\sum_{n=1}^{N} X_{n,t+T/2} \hat{u}_{n,t+T/2} - \sum_{n=1}^{N} X_{n,t} \hat{u}_{n,t} \right)}_{\Delta \bar{g}_t}. \tag{3.11}$$

Both statistics primarily utilise the gradient $X_{n,t}\hat{u}_{n,t}$ of the log-likelihood contribution of each individual at each time point, taking differences between different time points or different groups (depending on the effect under investigation) in conjunction with the assumption of iid sampling over these dimensions.

This concept is extended in this paper for MNP models in a panel structure with random effects, with the aim of detecting unobserved heterogeneity, structural breaks in time, as well as temporal autocorrelation in the random component. In order to generalise the argument, it should be noted that the systematic part of the utility is most often assumed to be linear and analogous to the linear model above, given as $V_{n,t,j} = \beta' X_{n,t,j}$. It is evident that the extension to the unrestricted model

$$V_{n,t,j} = \beta' X_{n,t,j} + z_n \beta_2' X_{n,t,j} \tag{3.12}$$

is appropriate when considering two groups of decision-makers with different preferences. This is then typically estimated by maximising the log-likelihood function

$$\begin{aligned} ll_{\mathrm{ML}}(\theta; y, X) &= \sum_{n=1}^{N} \log \Pr(U_{n,1,y_{n,1}} \geq U_{n,1,i} \; \forall i, \ldots, U_{n,T,y_{n,T}} \geq U_{n,T,i} \; \forall i \mid y_n, X_n) \\ &= \sum_{n=1}^{N} \log L_n(\theta). \end{aligned} \tag{3.13}$$

Assuming conditional independence of the observations across time, conditional on observed covariates, this simplifies to

$$ll_{\mathrm{ML}}(\theta; y, X) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log \Pr(U_{n,t,y_{n,t}} \geq U_{n,t,i} \; \forall i \mid y_{n,t}, X_{n,t}) = \sum_{n=1}^{N} \sum_{t=1}^{T} \log L_{n,t}(\theta). \tag{3.14}$$

However, mixing, for example due to random coefficients β_n in the systematic part $\beta_n' X_{n,t,j}$ of the utility (where β_n is drawn iid from some underlying distribution independent of all observed $X_{n,t,j}$), implies that the log-likelihood contribution for

decision-maker n corresponding to the mixed MNP cannot be written as a sum of the terms for different choice occasions t, but only within the joint term $L_n(\theta)$. This limits the flexibility since a switching of the roles of T and N is no longer possible.

Using composite marginal likelihood (CML) instead of maximum likelihood estimation entails the replacement of the joint log-likelihood of all observations $\log L_n(\theta)$ of an individual n, with a weighted sum of the log-likelihoods of margins. In the following, as is typically the case, pairs of observations are considered for the margins, resulting in the CML function

$$\begin{aligned} ll_{\text{CML}}(\theta; y, X) &= \sum_{n=1}^{N} \sum_{a=1}^{T_n-1} \sum_{b=a+1}^{T_n} w_{n,a,b} \log \Pr(U_{n,a,y_{n,a}} \geq U_{n,a,i}, \\ &\qquad U_{n,b,y_{n,b}} \geq U_{n,b,i} \; \forall i \mid y_{n,a}, y_{n,b}, X_{n,a}, X_{n,b}) \\ &= \sum_{n=1}^{N} \sum_{a=1}^{T_n-1} \sum_{b=a+1}^{T_n} w_{n,a,b} \log L_{n,a,b}(\theta), \end{aligned} \tag{3.15}$$

where T_n can vary over different individuals, allowing for unbalanced panel settings. The second equation defines the term $L_{n,a,b}(\theta)$. Analogously to the linear model case previously discussed, a number of different groupings can be used to provide different LM tests. The application of CML estimation permits the extension of LM pooling tests from the individual level to the level of margins. This flexibility allows for the investigation of unaccounted temporal effects and intra-individual heterogeneity without the necessity of incorporating these effects into the model. However, for valid inference, it is essential to recognise that pairs of observations from a single individual are not independent, as the same choice occasion may be included in multiple pairs of observations.

In this paper, we present a new Lagrange multiplier-type test, capable of testing for structural intra-person heterogeneity in CML estimation. In Section 3.2, we present a gradient-based test, similar to the Lagrange multiplier test, that accounts for the fact that pairs of observations from one individual are not independent. Furthermore, we derive the asymptotic distribution of the test statistic. In Section 3.3, we employ the discrete choice modelling setting, specifically MNP models, to generate synthetic data to demonstrate the finite sample properties of the proposed test procedure. MNP models are particularly well-suited to this purpose, given that they are computationally demanding such that CML estimation and its advancements have become popular methods for estimating these models [see, for instance, Bhat, 2011; Varin, 2008; Varin et al., 2011]. In Section 3.3.1, we first confirm the theoretical distribution of the test statistic under the null hypothesis for finite samples. In Sections 3.3.2 and 3.3.3, we then proceed to implement violations of the null hypothesis

of different types and degrees of severity in the data generating process (DGP) and evaluate the power properties of the test in these cases. The results of these simulations are presented in Section 3.4, and conclusions regarding the applicability of the presented testing method are presented in Section 3.5.

The `R` and `C++` code used for the estimation processes in this paper is bundled into an R-package named *Rprobit*, which is available on `https://github.com/dbauer72/Rprobit`.

3.2 CML pair gradient contributions

The core of this paper lies in the understanding of the score of the CML function and how it is composed of the contributions of the individual CML margins of all decision-makers. This section is therefore dedicated to understanding and dissecting the score function of the CML function, analysing the distribution of its components and then combining the components to form LM type tests.

The score corresponding to the CML function (3.15) can be expressed as

$$\partial_\theta ll_{CML}(\theta; y, X) = \sum_{n=1}^{N} \sum_{a=1}^{T_n-1} \sum_{b=a+1}^{T_n} w_{n,a,b} \partial_\theta \log L_{n,a,b}(\theta). \tag{3.16}$$

When calculating the CML estimator, $ll_{CML}(\theta; y, X)$ is maximised w.r.t. θ, without restrictions, and thus the score (3.16) at the optimum $\hat{\theta}$ is equal to zero. The aforementioned score is composed of the sum of the gradient contributions of the individual CML pairs of observations

$$g_{n,a,b}(\theta) = \partial_\theta \log L_{n,a,b}(\theta). \tag{3.17}$$

In the context of an exchangeable MNP (under the null hypothesis H_0), whereby all pairs of observations are assumed to follow the same distribution, it can be shown that $\mathbb{E}(g_{n,a,b}(\theta_\circ)) = 0$, where $\theta_\circ$ denotes the true parameter vector of the DGP. However, unobserved heterogeneities, structural breaks in time, as well as temporal autocorrelation in the random terms (deviations (I)-(III)) all lead to systematic differences in the joint choice probabilities for pairs of choices that can be detected from the log-CML gradient contributions $g_{n,a,b}(\hat{\theta})$.

This paper's primary result discusses the properties of $g_{n,a,b}(\hat{\theta})$ under a set of assumptions regarding the DGP and the CML function.

Assumption 3.1 (Data generating process)**.** *The data set $(y_n, X_n), n = 1, ..., N$ is generated by the following mechanism:*

(I) A number T_n of choice occasions are drawn from a discrete random distribution, supported in the set $\{2, 3, ..., \bar{T}\}$, and are independent of covariate values. In this context, $\bar{T} \in \mathbb{N}_{>1}$ denotes the maximum number of observed choices for one individual.

(II) For each decision-maker facing T_n choice occasions, a matrix $X_n = [X_{n,1,.}, ..., X_{n,T_n,.}] \in \mathbb{R}^{PJ \times T_n}$ of regressors is chosen iid across decision-makers, such that for each pair of choice occasions (a, b), the matrix $[X_{n,a,.}, X_{n,b,.}]$ is distributed identically over pairs $(a, b), 1 \leq a < b \leq T_n$. Furthermore, $\|X_{n,a,.}\|_\infty \leq M, 1 \leq a \leq T_n$ for some scalar $M < \infty$ (uniform norm bound).

(III) For given T_n and X_n, the vector of choices $y_n = [y_{n,1}, ..., y_{n,T_n}]' \in \mathcal{J}^{T_n}, \mathcal{J} = \{1, ..., J\}$ is chosen according to the mixed MNP model with normally distributed random effects corresponding to parameter vector $\theta_0 \in \mathbb{R}^P$ for appropriate integer P.

(IV) The regressors X are chosen from a distribution in a manner that ensures that the model parameters are locally identifiable (i.e. no multicollinearity between regressors and the model is identified w.r.t. the scale and level of the utilities).

Assumption 3.2 (CML Weights)**.** *The CML weights $w_{n,a,b}$ are chosen according to one of the following five schemes:*

(I) $w_{n,a,b} = f(a, b)$ for some bounded positive function $f : \mathbb{N}^2 \to [\underline{w}, \overline{w}]$, $0 \leq \underline{w} \leq \overline{w} < \infty$.

*(II) $w_{n,a,b} = \mathring{w}_{T_n} \in [\underline{w}, \overline{w}]$, $0 < \underline{w} \leq \overline{w} < \infty$ (*groupwise *CML weights).*

(III) $w_{n,a,b} \in \{0, 1\}$ chosen randomly independent of all other variables, iid over decision-makers, such that

$$\sum_{a=1}^{T_n - 1} \sum_{b=a+1}^{T_n} w_{n,a,b} = w_{T_n} \tag{3.18}$$

(selecting w_{T_n} random pairs from within $T_n(T_n - 1)/2$ possible pairs).

(IV) Stratification weights $w_{n,a,b} = w_n$ drawn iid over decision-makers from some underlying distribution supported on $[\underline{w}, \overline{w}]$, $0 < \underline{w} \leq \overline{w} < \infty$.

(V) A combination of (I)-(IV).

Under these assumptions, the following theorem can be obtained (for the proof of the theorem, see Appendix 3.A.):

Theorem 3.1. *Let the data be generated from a mixed MNP model with utility* $U_{n,t,j} = X'_{n,t,j}\beta + X'_{n,t,j}\gamma_n + \varepsilon_{n,t,j}$ *for* $n = 1,..,N$, $t = 1,..,T_n$, *where* $\gamma_n \sim \mathcal{N}(0,\Omega)$ *independent of* $X_{n,t,j}$, *and* $\varepsilon_{n,t,j}$ *iid over decision-makers, such that Assumption 3.1 is satisfied. Let the model be estimated using CML with weights* $w_{n,a,b}$ *in accordance with Assumption 3.2, providing the estimator* $\hat{\theta}$ *maximising the CML such that the score is zero and the estimator is asymptotically normal with* $\sqrt{N}(\hat{\theta} - \theta_\circ) \xrightarrow{d} Z \sim \mathcal{N}(0, V_\theta)$.
Then

$$g_{n,a,b}(\hat{\theta}) = \partial_\theta \log L_{n,a,b}(\hat{\theta}) = g_{n,a,b}(\theta_\circ) + \partial_\theta g_{n,a,b}(\bar{\theta}_n)(\hat{\theta} - \theta_\circ), \tag{3.19}$$

where $\mathbb{E}(g_{n,a,b}(\theta_\circ)) = 0$. *Furthermore,* $\mathbb{E}\partial_\theta g_{n,a,b}(\theta_\circ) = H_g > 0$ *and* $\bar{\theta}_n$ *is an intermediate value such that* $\bar{\theta}_n \xrightarrow{p} \theta_\circ$ *as* $N \to \infty$.
Finally,

$$N^{-1}\sum_{n=1}^{N} \partial_\theta g_{n,a,b}(\bar{\theta}_n)(\hat{\theta} - \theta_\circ) = H_g(\hat{\theta} - \theta_\circ) + o_P(N^{-1/2}). \tag{3.20}$$

The theorem demonstrates that the score contribution of each pair is essentially equal to the score contribution evaluated at the true parameter vector, plus an additional term that depends on the estimation error multiplied by the Hessian, which does not depend on the individual.

The aforementioned assumptions imply that pairs can be used interchangeably, as they are based on regressors drawn from the same underlying stochastic mechanism as are the random parts of the utility. This is, of course, a strong assumption, which is mostly adequate in panel studies with repeated choices of the same decision problem.

It should be noted, however, that the MNP model is employed here merely for illustrative purposes. The theorem does not utilise this structure and is also applicable for multinomial logit models or any other model class, provided that pairs of observations employ the same underlying conditional (on the regressors) and unconditional likelihoods.

The theorem provides us with the means to use for diagnostic checking: In the context of the restricted model, all pairs of choice occasions are identically distributed. In particular, the corresponding contribution to the score, apart from a common term accounting for the estimation error, consists of the score contribution evaluated at the true parameter vector, which has mean zero. If the population, however, can be partitioned into two groups corresponding to different parameter vectors generating the data, the common estimate will be a compromise between these two different parameter vectors. The corresponding score contributions will point in

opposite directions, and this phenomenon can be observed through a number of different diagnostic tests, depending on the grouping of the score contributions:

- tests for pooling: in this situation groups are formed for different individuals.
- tests for structural breaks: in this situation choice pairs (a, b), with $\tau_a < \tau_b \leq \tau_0$ from early in the sample prior to a specified break date τ_0, exhibit discrepancies when compared to late pairs (a', b'), with $\tau_0 < \tau'_a < \tau'_b \leq \tau_T$, where τ_a denotes the time at which observation a occurred.
- tests for autocorrelation of the error term: as demonstrated by Büscher and Bauer [2024], the autocorrelation of the error term can be erroneously identified as a mixture in the alternative specific constants (ASCs). Consequently, discrepancies in the autocorrelation can be detected by comparing pairs with close by decisions (a, b), $|\tau_a - \tau_b| < \Delta\tau_c$ with pairs exhibiting a considerable temporal interval between them (a', b'), $|\tau_{a'} - \tau_{b'}| > \Delta\tau_d$, where $\Delta\tau_c$ and $\Delta\tau_d$ specify which temporal distances are considered to be close or distant, respectively.

Under the null hypothesis of the restricted model, the score contributions are drawn from the same distribution in all cases. Since tests for pooling at the individual level are also a standard and readily applicable procedure in the maximum likelihood setting, the subsequent analysis will focus on tests comparing different pairs of choice occasions within each individual.

For the considered groups G_1 and G_2 of observation pairs, the following assumptions are made.

Assumption 3.3 (Groups)**.** *The groups of pairs of observations G_1 and G_2 are constructed in such a way that:*

(I) The groups G_1 and G_2 are disjoint ($G_1 \cap G_2 = \varnothing$), such that they do not contain the same pair of observations (a, b).

(II) The groups G_1 and G_2 are chosen independent of the observations y and the covariate values X_n, such that the matrix $[X_{n,a}, X_{n,b}]$ is distributed identically over all pairs (a, b) within both groups, ensuring that the model is identified locally within the respective groups (e.g. there is no perfect colinearity between regressors within groups, for example due to constant covariate values within a group when ASCs are also estimated).

(III) The groups G_1 and G_2 are chosen such that the sum of the CML weights of the pairs in group i for individual n is nonzero, i.e. $|G_{i,n}| = \sum_{a,b:(a,b)\in G_i} w_{n,a,b} > 0, i \in \{1, 2\}$.

Remark 3.1. *It is possible for pairs with a shared observation to be contained within different groups. For instance, the pair* $(1,2)$ *could be contained in the first group of subsequent observations, while the pair* $(1,T)$ *would belong to the group of distant observations.*

Remark 3.2. *If Assumption 3.3(III) is not satisfied, it is possible to remove individuals* n *for which* $|G_{i,n}| = 0$ *for some* i *from the subsequent calculations and adjust* N *in order to represent the number of individuals for which* $|G_{i,n}| > 0, i \in \{1,2\}$.

In order to test the null hypothesis that there are no significant differences in the expected value of the gradient contribution between two disjoint groups of CML pairs G_1 and G_2 ($G_1 \cap G_2 = \varnothing$), so

$$H_0: \quad \mathbb{E}(g_{n,a,b}(\theta_\circ) \mid (a,b) \in G_1) = \mathbb{E}(g_{n',a',b'}(\theta_\circ) \mid (a',b') \in G_2), \tag{3.21}$$

one can calculate the average gradient contribution of each of the groups for decision-maker n, as well as the difference between these:

$$\bar{g}_{G_i,n} = |G_{i,n}|^{-1} \sum_{a,b:(a,b)\in G_i} w_{n,a,b}\; g_{n,a,b}(\hat{\theta}), \tag{3.22}$$

$$\Delta\bar{g}_{G_1,G_2,n} = \bar{g}_{G_1,n} - \bar{g}_{G_2,n}. \tag{3.23}$$

From Theorem 3.1, we have

$$\begin{aligned} \bar{g}_{G_i,n} &= |G_{i,n}|^{-1} \sum_{a,b:(a,b)\in G_i} w_{n,a,b} g_{n,a,b}(\theta_\circ) + |G_{i,n}|^{-1} \sum_{a,b:(a,b)\in G_i} w_{n,a,b} H_g(\hat{\theta} - \theta_\circ) + o_P(N^{-1/2}) \\ &= \qquad \bar{g}_{G_i,n,\circ} \qquad + \qquad H_g(\hat{\theta} - \theta_\circ) \qquad + o_P(N^{-1/2}), \end{aligned} \tag{3.24}$$

where the latter equation defines $\bar{g}_{G_i,n,\circ}$. It thus follows that the score contribution of the group of pairs is equal to a term that, under the null hypothesis, has zero expectation and is iid over decision-makers, plus a term that does not depend on n and is of order $O_p(N^{-1/2})$, plus a negligible term of order $o_P(N^{-1/2})$. Consequently, the difference between two groups, $\Delta\bar{g}_{G_1,G_2,n}$, is essentially equal to $\bar{g}_{G_1,n,\circ} - \bar{g}_{G_2,n,\circ}$. By taking differences, the joint term cancels under the null hypothesis.

Consequently, under the null hypothesis, a sample $\Delta\bar{g}_{G_1,G_2,n}$, $n = 1,...,N$ of iid observations with expectation zero is obtained. In contrast, under the alternative, the expectation will be nonzero.

Remark 3.3. *In the aforementioned evaluations, a balanced panel situation was considered, such that the same pairs of choice occasions* (a,b) *were observed for each decision-maker. The weights can be specific to individual decision-makers and pairs of choices, however, provided*

that Assumption 3.2 is fulfilled. This makes the results generally applicable to unbalanced panel data and unequally weighted observations.

Although it is assumed that the pairs are interchangeable, it is challenging to quantify the distribution of the average over different pairs of decisions involved in the formation of $\bar{g}_{G_i,n}$ through analytical means. To illustrate, under the null hypothesis, the correlation between $g_{n,a,b}(\theta_\circ)$ and $g_{n,a',b'}(\theta_\circ)$ may be nonzero when $a = a'$ and zero for $a \neq a'$. Consequently, instead of utilising an estimate of the variance of $\Delta\bar{g}_{G_1,G_2,n}$ based on analytic expressions, we propose to use standard one-sample t-tests (if a single component of the gradient vector is being evaluated) and F-tests (if multiple components are assessed) for the mean, with the variance being estimated using the sample variance of the N observations. The tests are thus applicable primarily in situations where N is large.

Therefore, we propose to use the following test statistic:

Definition 3.1. *The test statistic LM_P is defined as follows:*

$$\Delta\bar{g}_{G_1,G_2} = N^{-1}\sum_{n=1}^{N}\Delta\bar{g}_{G_1,G_2,n}, \tag{3.25}$$

$$\hat{V}_{\Delta g} = (N-1)^{-1}\sum_{n=1}^{N}(\Delta\bar{g}_{G_1,G_2,n} - \Delta\bar{g}_{G_1,G_2})(\Delta\bar{g}_{G_1,G_2,n} - \Delta\bar{g}_{G_1,G_2})', \tag{3.26}$$

$$LM_P = N\Delta\bar{g}'_{G_1,G_2}\hat{V}^{-1}_{\Delta g}\Delta\bar{g}_{G_1,G_2}, \tag{3.27}$$

where $\Delta\bar{g}_{G_1,G_2} \in \mathbb{R}^P$.
For testing the j-th parameter individually, we define the analogue of the t-test statistic:

$$t = \frac{\sqrt{N}\left(\Delta\bar{g}_{G_1,G_2,n}\right)_{(j)}}{\sqrt{(\hat{V}_{\Delta g})_{(j,j)}}}, \tag{3.28}$$

where $(\Delta\bar{g}_{G_1,G_2,n})_{(j)}$ denotes the j-th coordinate of the vector, and $(\hat{V}_{\Delta g})_{(j,j)}$ the j-th diagonal element.

With these definitions we obtain:

Theorem 3.2. *Under the assumptions of Theorem 3.1 and Assumptions 3.3, the test statistic LM_P has an asymptotic χ^2_P distribution, given that $\hat{V}_{\Delta g}$ is regular.*
The distribution of the test statistic t converges to a standard normal distribution as $N \to \infty$, provided that $(\Delta g)_{(j)}$ is not constant over pairs.

For the proof of the theorem, see Appendix 3.A.

Remark 3.4. *It should be noted that the test statistics can be readily calculated for a number of different groupings, as the input is primarily comprised of the score contributions for each decision-maker for each pair of choices. These data are typically already calculated in gradient-based numerical optimisation procedures. Therefore, the additional computational burden associated with calculating the test statistics entails forming averages over pairs, estimating the sample mean and variance for a sample of size N, and inverting a $P \times P$ matrix.*

Remark 3.5. *To enhance the numerical stability of the calculations, it is recommended to standardise $\Delta\bar{g}_{G_1,G_2,n}$ and $\Delta\bar{g}_{G_1,G_2}$ by dividing each component by the corresponding standard deviation of $\Delta\bar{g}_{G_1,G_2,n}$ across individuals n prior to calculating $\hat{V}_{\Delta\bar{g}}$ and the test statistics. This has no impact on the test statistics; however, it does render the computation of $\hat{V}_{\Delta\bar{g}}^{-1}$ more numerically stable.*

Remark 3.6. *In case that $\hat{V}_{\Delta g}$ is singular, it is not possible to calculate the test statistics. The fact that $\hat{V}_{\Delta g}$ is singular is of intrinsic value, as it indicates that the two groups are not structurally distinct but, on the contrary, may be structurally linked. This in itself calls for further investigation by the researcher, as would the rejection of the null hypothesis by the proposed test.*

A possible (extreme) example is when observations originate from a periodically repeated plan, unknown to the researcher, which results in observations from two different panel waves being identical. If groups are selected based on the panel waves, the model would be locally identified within each group, but the average score differences between the groups would be zero for all individuals, leading to $\hat{V}_{\Delta g} = 0$.

Remark 3.7. *Under the assumptions that the $\Delta\bar{g}_{G_1,G_2,n}$ are iid multivariate normally distributed, it holds for finite samples*

$$\frac{N-P}{P(N-1)} LM_P \sim F_{P,N-P}, \tag{3.29}$$

$$t \sim t_{N-1}. \tag{3.30}$$

For details, see Hotelling [1931]. Asymptotically, the use of these distributions is equivalent to those in Theorem 3.2. However, despite the fact that the assumption of normally distributed $\Delta\bar{g}_{G_1,G_2,n}$ does not necessarily hold, the statistics in Eqs. (3.29) *and* (3.30) *have been shown to be favourable in the simulation studies presented in Section 3.3.*

3.3 Simulation evaluation of test properties

In order to evaluate the distributional properties of the statistics $\frac{N-P}{P(N-1)} LM_P$ and t in a finite sample setting, we employ the discrete choice modelling setting, specifically MNP models, as described in Section 3.1. This enables us to generate synthetic data that adheres to Assumption 3.1, thus allowing us to evaluate the finite sample properties of the test statistics in a controlled environment.

The synthetic data sets have been constructed in such a way as to emulate panel data sets comprising two waves of observations, with a one-year interval between the start of each wave. Specifically, when an individual has a total of T_n observations, the first set of $T_n/2$ observations starts at $\tau_1 = 1$, while the second set of $T_n/2$ observations begins at $\tau_{T_n/2+1} = 366$. Within each wave of observations, unit time steps are employed, such that $\tau_{t+1} = \tau_t + 1$.

To evaluate the properties of the statistical test devised, balanced panel data sets ($T_n = T$) with different numbers of observations per individual $T \in \{10, 14, 20, 30\}$ and different numbers of individuals $N \in \{100, 500, 1000\}$ were simulated.

To test the null hypothesis (3.21), two different time-dependent groupings of the CML pairs were considered [1]:

$$\begin{aligned} \textit{FirstLast:} \quad (a,b) \in G_1 &\iff (|\tau_a| < 366) \wedge (|\tau_b| < 366) \\ (a,b) \in G_2 &\iff (|\tau_a| \geq 366) \wedge (|\tau_b| \geq 366) \end{aligned}$$

$$\begin{aligned} \textit{NearFar:} \quad (a,b) \in G_1 &\iff |\tau_a - \tau_b| < 100 \\ (a,b) \in G_2 &\iff |\tau_a - \tau_b| \geq 100 \end{aligned}$$

Should evidence for a difference between the gradient contributions of group G_1 and group G_2 be detected in the *FirstLast* groupings, this would indicate a structural change in the parameters of the DGP between panel waves. When considering *NearFar* groupings, a difference between the gradient contributions from the two groups could be indicative of distance-dependent correlation patterns between observations, for example induced by either structural changes between panel waves or an autoregressive error structure.

[1] Additionally, a third type of grouping, termed *EvenOddClose*, was explored. This involved classifying pairs of observations with even temporal distances from the same panel wave in one group and pairs of observations with odd temporal distances between them from the same panel wave in a second group. These, however, did not demonstrate any advantages compared to the other groupings described in this paper, in any of the scenarios considered in the finite sample simulation studies. Further details can be provided upon request.

3.3.1 Finite sample distribution of test statistic under the null hypothesis

In order to evaluate the distribution of the test statistic under the null hypothesis, the model

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{3.31}$$
$$U_{n,t,2} = \beta_0 + x_{n,t}\beta_1 + \gamma_n + \varepsilon_{n,t,2}, \tag{3.32}$$

was employed, with $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_\varepsilon^2 \end{pmatrix}\right)$ and $\gamma_n \sim \mathcal{N}(0, \sigma_\gamma^2)$. For the DGP we set $\beta_0 = 1$, $\beta_1 = 1$, $\sigma_\gamma = 1$, and $\sigma_\varepsilon = 1$. In order to ensure model identification, we kept $\beta_1 = 1$ fixed during estimation and estimated $\theta = (\beta_0, \sigma_\gamma, \sigma_\varepsilon)$. The regressors $x_{n,t}$ were drawn iid from a standard normal distribution.

3.3.2 Detecting autoregressive errors

In this scenario, the DGP is specified as

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{3.33}$$
$$U_{n,t,2} = \beta_0 + x_{n,t}\beta_1 + \varepsilon_{n,t,2}, \tag{3.34}$$
$$\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' = \rho\varepsilon_{n,t-1} + \tilde{\varepsilon}_{n,t}, \tag{3.35}$$

with $\tilde{\varepsilon}_{n,t} = (\tilde{\varepsilon}_{n,t,1}, \tilde{\varepsilon}_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 1-\rho^2 & 0 \\ 0 & 1-\rho^2 \end{pmatrix}\right)$, $\beta_0 = 1$, $\beta_1 = 1$, and varying values of ρ between -0.99 and 0.99 for different simulation setups[2]. The regressors $x_{n,t}$ are drawn iid from a standard normal distribution.

The estimation is then based on the misspecified model

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{3.36}$$
$$U_{n,t,2} = \beta_0 + x_{n,t}\beta_1 + \gamma_n + \varepsilon_{n,t,2}, \tag{3.37}$$

with $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_\varepsilon^2 \end{pmatrix}\right)$ and $\gamma_n \sim \mathcal{N}(0, \sigma_\gamma^2)$, with $\beta_1 = 1$ fixed for identification and estimated $\theta = (\beta_0, \sigma_\gamma, \sigma_\varepsilon)$.

The covariance matrix $\Sigma_{n,a,b} = \mathrm{Cov}(\varepsilon_{n,a}, \varepsilon_{n,b})$ between the random errors of two observations from one individual can now be explicitly expressed for both mod-

[2]Values used for ρ were $\{-0.990, -0.950, -0.900, -0.750, -0.625, -0.500, -0.400, -0.300, -0.200, -0.100, -0.050, -0.025, -0.010, 0.000, 0.010, 0.025, 0.050, 0.100, 0.200, 0.300, 0.400, 0.500, 0.625, 0.750, 0.900, 0.950, 0.990\}$

els. In general, for an autoregressive error process $\varepsilon_{n,t} = \Psi^{\tau_t - \tau_{t-1}} \varepsilon_{n,t-1} + \tilde{\varepsilon}_{n,t}$ with $\tilde{\varepsilon}_{n,t} \sim \mathcal{N}(0, \tilde{\Sigma})$, we have

$$\Sigma_{n,a,b} = \begin{pmatrix} \Sigma & \Sigma(\Psi^{\tau_b - \tau_a})' \\ \Psi^{\tau_b - \tau_a}\Sigma & \Sigma \end{pmatrix}, \tag{3.38}$$

where Σ solves the Lyapunov equation $\Sigma = \Psi\Sigma\Psi' + \tilde{\Sigma}$, provided that all eigenvalues of the matrix $\Psi \in \mathbb{R}^{J \times J}$ are inside the unit circle (stability condition).

For the DGP model, we have $\Psi = \rho I_J$ and $\tilde{\Sigma} = (1 - \rho^2) I_J$, where I_J represents the $J-$dimensional identity matrix. Consequently,

$$\Sigma_{n,a,b} = \begin{pmatrix} \frac{1}{1-\rho^2}\tilde{\Sigma} & \frac{\rho^{|\tau_b - \tau_a|}}{1-\rho^2}\tilde{\Sigma} \\ \frac{\rho^{|\tau_b - \tau_a|}}{1-\rho^2}\tilde{\Sigma} & \frac{1}{1-\rho^2}\tilde{\Sigma} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \rho^{|\tau_b - \tau_a|} & 0 \\ 0 & 1 & 0 & \rho^{|\tau_b - \tau_a|} \\ \rho^{|\tau_b - \tau_a|} & 0 & 1 & 0 \\ 0 & \rho^{|\tau_b - \tau_a|} & 0 & 1 \end{pmatrix}. \tag{3.39}$$

In contrast, the estimated model yields the following results:

$$\Sigma_{n,a,b} = \begin{pmatrix} 1 + \sigma_\gamma^2 & 0 & \sigma_\gamma^2 & 0 \\ 0 & \sigma_\varepsilon^2 + \sigma_\gamma^2 & 0 & \sigma_\gamma^2 \\ \sigma_\gamma^2 & 0 & 1 + \sigma_\gamma^2 & 0 \\ 0 & \sigma_\gamma^2 & 0 & \sigma_\varepsilon^2 + \sigma_\gamma^2 \end{pmatrix} \tag{3.40}$$

and

$$\mathrm{Cor}(\varepsilon_{n,a}, \varepsilon_{n,b}) = \begin{pmatrix} 1 & 0 & \sigma_\gamma^2/(1+\sigma_\gamma^2) & 0 \\ 0 & 1 & 0 & \sigma_\gamma^2/(\sigma_\varepsilon^2 + \sigma_\gamma^2) \\ \sigma_\gamma^2/(1+\sigma_\gamma^2) & 0 & 1 & 0 \\ 0 & \sigma_\gamma^2/(\sigma_\varepsilon^2 + \sigma_\gamma^2) & 0 & 1 \end{pmatrix}. \tag{3.41}$$

Under the null hypothesis of an exchangeable probit model, ρ is equal to zero, thereby rendering the two models identical for $\sigma_\gamma = 0$. When $\rho \neq 0$, a notable distinction between the two models emerges: the estimated model assumes a constant covariance matrix for all pairs of observations, irrespective of the temporal distance between them. In contrast, the DGP model's the covariance depends on the distance $|\tau_b - \tau_a|$. When observations a and b are in different panel waves, we have $|\rho^{|\tau_b - \tau_a|}| \leq 0.99^{366-30} \approx 0.0342$, whilst for $\rho = 0.99$ we have for observations within the same panel wave, $0.99^{|\tau_b - \tau_a|} \geq 0.99^{|30-1|} \approx 0.7472$.

Furthermore, the estimated model is not sufficiently flexible to estimate negative correlations between errors at different time points. However, this is the case for pairs of observations with $|\tau_b - \tau_a| \equiv 1 \pmod 2$ for negative values of ρ.

3.3.3 Detecting structural breaks in the model

In this scenario, the DGP under consideration is

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{3.42}$$
$$U_{n,t,2} = \beta_0 + x_{n,t}\beta_{1,n,t} + \gamma_n + \varepsilon_{n,t,2}, \tag{3.43}$$

with $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_\varepsilon^2 \end{pmatrix}\right)$ and $\gamma_n \sim \mathcal{N}(0, \sigma_\gamma^2)$. For the DGP, the values $\beta_0 = 1$, $\sigma_\gamma = 1$, and $\sigma_\varepsilon = 1$ were used. In order to examine the effects of different groupings of individuals with varying parameter values between the groups and changes in these parameter values over time, we implemented a series of scenarios involving $\beta_{1,n,t}$. This temporal change was chosen to represent a discrete change in the parameter values between the two simulated panel waves. Two distinct configurations were employed.

The first configuration represents an overall shift of the parameter value for the entire population between the panel waves, so

$$\beta_{1,n,t} = \begin{cases} 1 - \alpha, & \tau_t < 365, \\ 1 + \alpha & \tau_t \geq 365. \end{cases} \tag{3.44}$$

The second setup represents a split of the population. The population is partitioned into two equally sized groups P_1 and P_2. In the initial panel wave, the parameter value was identical for all individuals. In contrast, in the second panel wave, the parameter value was increased for those individuals in P_1 and decreased for those in P_2, resulting in

$$\beta_{1,n,t} = \begin{cases} 1, & \tau_t < 365, \\ 1 + \alpha, & \tau_t \geq 365 \text{ and } n \in P_1, \\ 1 - \alpha & \tau_t \geq 365 \text{ and } n \in P_2. \end{cases} \tag{3.45}$$

In order to evaluate the rejection rate of the proposed test for different magnitudes of violations of the null hypothesis of an exchangeable probit model ($H_0 : \alpha = 0$), data sets with varying values of α were simulated.[3]

[3]The values used for α were $\{0.0000, 0.0100, 0.0175, 0.0250, 0.0500, 0.0750, 0.1000, 0.1750, 0.2500, 0.3750,$

For estimation, the misspecified model

$$U_{n,t,1} = 0 + \varepsilon_{n,t,1}, \tag{3.46}$$
$$U_{n,t,2} = \beta_0 + x_{n,t}\beta_1 + \gamma_n + \varepsilon_{n,t,2}, \tag{3.47}$$

was employed, with $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \varepsilon_{n,t,2})' \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_\varepsilon^2 \end{pmatrix}\right)$ and $\gamma_n \sim \mathcal{N}(0, \sigma_\gamma^2)$. The value of $\sigma_\varepsilon = 1$ was fixed for identification purposes, while $\theta = (\beta_0, \beta_1, \sigma_\gamma)$ was estimated. Here, a common value of $\beta_{1,n,t} = \beta_1$ was assumed for all individuals and time points.

3.4 Simulation Results

This section presents a discussion of the results obtained from the simulation studies described in Section 3.3. The results are presented in tabular form in 3.5 and illustrated in the accompanying plots.

3.4.1 Distribution under the null hypothesis

In the plots in Figure 3.1, the empirical cumulative distribution functions (CDFs) of the test statistics are compared to their theoretical CDFs, as deduced in Section 3.2. Each figure represents a single test statistic. The Figure 3.1a for the joint test statistic is based on the specification in Eq. (3.29), while the three Figures 3.1b-3.1d for the componentwise tests are based on Eq. (3.28). Each plot contains a matrix of 12 plots, with one plot for each simulated combination of the number of individuals $N \in \{100, 500, 1000\}$ and the number of observations per individual $T \in \{10, 14, 20, 30\}$. The different groupings that were considered are represented in the plots by differently coloured solid lines, whilst the theoretical distribution is overlain as a black dashed line. The x-axis represents the value of the test statistic, and the y-axis represents the corresponding CDF value. The plots demonstrate that for all 12 considered combinations of N and T, the empirical distribution of the test statistic closely follows its theoretical counterpart, irrespective of the number of individuals and the number of observations per individual.

Kolmogorov–Smirnov tests [for details, see Kolmogorov, 1933; Marsaglia et al., 2003] were performed for each of the empirical distributions against their theoretical asymptotic counterparts, resulting in test statistics in the range of 0.016 to 0.061 with

0.5000, 0.7500, 0.9500, 1.0000, 1.5000, 2.0000}.

a critical value at a 5% significance level of $K_{0.05} = 1.3581/\sqrt{N} = 0.043$. To address the issue of multiple testing, a Bonferroni correction was applied [Bonferroni, 1936]. This revealed that the null hypothesis, namely that the simulated test statistics are realisations of the theoretical asymptotic test distributions, was not rejected for any of the cases.

3.4.2 Rejections rates under violation of the null hypothesis

Figures 3.2-3.4 illustrate the rejection rates of the tests at a significance level of 5% on the y-axis, with the degree of violation of the null hypothesis on the x-axis (autocorrelation coefficient ρ for the autoregressive error case and the shift/split value α for the simulations with shifts/splits between panel waves). Each of the figures represents one of the simulated violations of the null hypothesis. These are as follows: 3.2 for the autoregressive error process, 3.3 for shifts in a parameter between panel waves, and 3.4 for a split of the population between panel waves, with both halves of the population having a unique parameter value.

Figures 3.2a-3.2b, 3.3a-3.3b, and 3.4a-3.4b contain a grid of twelve graphs arranged by number of individuals N and number of observations T per individual. Each point in the plots is based on tests on 100 simulated data sets, except for the cases with $N = 500$ and $T = 20$, where 1000 simulated data sets were used for each point. Figures 3.2c-3.2d, 3.3c-3.3d, and 3.4c-3.4d are larger-sized plots for the aforementioned cases with $N = 500$ and $T = 20$, which also contain data for a greater number of values of ρ and α, respectively. The different tests (joint and componentwise) are represented in different colours within the plots.

Additionally, for each simulation scenario and grouping of CML pairs, the componentwise test with the largest test statistic was identified. Tables 3.1-3.6 provide an overview of the respective shares across the 100 (1000 for $N = 500, T = 20$) simulated cases where the largest test statistic among the componentwise tests was observed for the given component.

Rejection rate with autoregressive errors

For the cases with an autoregressive error structure, the tests with the *FirstLast* groupings do not exhibit rejection rates that are substantially larger than 5% for any values of ρ, as illustrated in Figures 3.2a and 3.2c. This is to be expected, given that there are no structural differences between the first and second waves of observations. This is also reflected in the fact that no componentwise test yielded a clear majority of shares with the largest test statistics, as can be seen in Table 3.1.

The tests with the *NearFar* groupings demonstrate that, for both the joint test and the test for the parameter σ_γ (the variance parameter for the mixed ASC effect),

the rejection rates increase with the absolute value of the autocorrelation coefficient ρ. As the number of individuals N increases, this effect becomes more pronounced. However, an increase in the number of observations, coupled with an extreme negative autocorrelation, results in a decline in the rejection rate, as illustrated in Figures 3.2b and 3.2d. This can be attributed to the fact that with an increased number of observations per decision-maker, the maximum distance between observations within a group increases, and the average correlation between observations within these pairs already diminishes. Furthermore, the models specified for estimation are not capable of accounting for negative correlation between error terms of different observations.

The tests for β_0 and σ_ε fail to detect violations of the null hypothesis for negative autocorrelation coefficients ρ. However, they do exhibit an increase in rejection rates with larger positive autocorrelations.

The results presented in Table 3.2 also demonstrate that the test for σ_γ, which represents the variance parameter for the mixed effects, predominantly yields the largest test statistic. Notably, the share of cases where σ_γ has the largest test statistic approaches one as the absolute value of the autocorrelation coefficient ρ increases. This suggests to the practitioner that when the largest componentwise test statistic is connected to the correlation between observations from the same individual, the violation of the null hypothesis is also connected to how the correlation between observations is implemented in the model.

The sole exceptions are cases with extreme negative autocorrelation ($\rho \leq -0.95$), where the share of cases where the test for β_0 had the largest test statistic does rise with rising number of observations per individual T.

Rejection rate with shift of parameter value between panel waves

For the cases where there is a shift in the β_1 parameter between panel waves (Figure 3.3), the tests with the *FirstLast* groupings demonstrate the strongest increase in rejection rates for an increasing degree of violation of the null hypothesis, reaching a rejection rate of 1 for $\alpha \geq 0.5$ across all tests and combinations of N and T. The only exception is the test for the ASC parameter β_0, which does not achieve a perfect rejection rate. As the number of individuals N or the number of observations per individual T increases, the perfect rejection rate is reached for even smaller values of α, as illustrated in Figures 3.3a and 3.3c. The share of cases in which the test for the β_1 parameter exhibits the largest test statistic increases rapidly to one as the magnitude α of the shift in the parameter β_1 between panel waves is increased. This effect is more pronounced in larger sample sizes (larger values of N or T).

The results obtained with the *NearFar* groupings are less pronounced, yet still demonstrate a clear and conclusive pattern. For all tests, the rejection rates increase

with rising values of α and with larger numbers of individuals N and observations per individual T. Across the board, the joint test shows the highest rejection rates, followed by the test for the variance parameter of the mixed ASC σ_γ (see Figures 3.3b and 3.3d). This is consistent with expectations, as pairs of observations within the same panel wave are expected to exhibit higher correlation and, consequently, higher values of σ_γ than pairs of observations in different panel waves. This is due to the fact that the β_1 parameter changes values between two panel waves for all individuals. The effect is less pronounced for tests of the parameters of β_0 and β_1, and the dependence of high rejection rates on larger numbers of individuals N and observations per individual T is stronger. For the *NearFar* groupings, the share of cases in which the test for the σ_γ parameter exhibits the largest test statistic increases when the magnitude α of the shift in the parameter β_1 between panel waves is increased, reaching a value of one for large data sets. This is in contrast to the results obtained with the *FirstLast* groupings, where the test for β_1 had the largest shares.

The collective findings assist the practitioner in comprehending the underlying cause of the violation of the null hypothesis. The evidence suggests that pairs of observations from different panel waves exhibit different β_1 values, which, in turn, would explain the observed variation in correlations between observations with varying temporal distances, attributable to the differing σ_γ parameters. Observation pairs from the same panel display greater similarity in their β_1 parameters, leading to higher levels of correlation compared to observation pairs from different panel waves, which exhibit greater variation between their β_1 parameters.

Rejection rate with split of parameter value in second panel wave

The cases with a split in the β_1 parameter in the second panel wave (Figure 3.4) demonstrate comparable outcomes for the *FirstLast* grouping to those observed for the rejection rates with the shift violations detailed in Section 3.4.2, albeit a more pronounced degree of violation is required for the tests to reject the null hypothesis. Even for the largest combination of number of individuals ($N = 1000$) and number of observations per individual ($T = 30$) tested, the rejection rates did not differ significantly from 0.05 for $\alpha \leq 0.175$. The observed rejection rates depending on the degree of violation of the null hypothesis α can be roughly compared to sigmoid functions, where larger values of N and T result in inflection points for lower values of α. The joint test and the tests for β_1 and σ_γ exhibit comparable rejection rates, whereas the rejection rate of the test for the ASC parameter β_0 requires higher values of α to reliably reject the null hypothesis (see Figures 3.4a and 3.4c). As demonstrated in Table 3.5, the test for the β_1 parameter exhibits a similar pattern to that observed in the case of a shift of the β_1 parameter as discussed in Section 3.4.2. Specifically, the test for the β_1 parameter has the largest share of cases with the highest test

statistic when the magnitude of the violation of the null hypothesis, as measured by the value of α, increases.

The tests with the *NearFar* groupings (Figures 3.4b and 3.4d) demonstrate the greatest dependency on the number of individuals and the number of observations per individual. Notably, the tests for $N = 100$ individuals exhibit no higher rejection rates than would be expected under the null hypothesis, regardless of the degree of violation α. For $N = 1000$ and $T = 30$, however, a rejection rate of 1 is reached with $\alpha \geq 1.0$ for the joint test. For this grouping and type of violation, our simulations did not indicate that a single parameter was dominating the test statistics, as evidenced by the results in Table 3.6.

3.5 Conclusion

In this paper, we have constructed a Lagrange multiplier-type test for the composite marginal likelihood estimation framework, to test for intra-personal heterogeneity, derived the asymptotic test distribution under the null hypothesis, and, using synthetic data sets, verified the test distribution in finite samples and evaluated the size and power of the test under different types and severities of violations of the null hypothesis.

The theoretical derivations demonstrate how the concept of testing for pooling using LM tests can be extended to the CML estimation setting, thereby enabling testing at the pair level rather than the individual level. The proposed procedures require minimal additional computation time and permit the testing of multiple hypotheses without the necessity of estimating or even implementing the unrestricted models, thereby making them a versatile testing tool.

The simulation results demonstrate that the proposed tests follow the asymptotic test distributions under the null hypothesis already for relatively small samples closely. Furthermore, with an appropriate sample size, the violations of the null hypothesis considered in this paper can be detected reliably, even when the degree of violation is modest. In the case of smaller sample sizes (small N and/or small T), the rejection rate of the proposed tests only approaches 1 for stronger degrees of violation of the null hypothesis.

In practice, when the type of violation that may be present in the data is typically unknown, the test can also be used to infer which violation may be responsible for a rejection of the null hypothesis.
In instances where the tests utilising the *NearFar* groupings reject the null hypothesis, yet tests employing the *FirstLast* groupings do not, this may be indicative of a change

in the correlation between pairs of observations with different temporal lags, for example as induced by an autoregressive error process.
In instances where tests for both types of groupings, *FirstLast* and *NearFar*, reject the null hypothesis, a potential explanation for this could be a change in parameter values between panel waves, as this introduces structural differences between the two panel waves, but also reduces the correlation between observations from different panel waves.

In examining which individual parameter test yielded the largest test statistic, a practitioner may gain greater insight than would be possible through a joint test. The parameter with the largest test statistic can provide information regarding the location of the largest discrepancy between the two groups of observational pairs. Depending on the chosen grouping, this can inform the practitioner of where in the model the misspecification responsible for the rejection of the null hypothesis may occur.

In accordance with standard diagnostic procedures, the subsequent steps for a practitioner would be to adjust the model according to the detected misspecification, compare the adjusted model with the original model via a model selection criterion [e.g., the composite likelihood ratio test (CLRT) or the composite likelihood Bayesian information criterion (CLBIC), see Gao and Song, 2010], and possibly accept the new model. This would then be tested in turn for misspecifications, thus starting the loop anew. Alternatively, one could employ estimation techniques that are robust against the detected type of misspecification, like the weighted CML estimator described by [Büscher and Bauer, 2024] for panel data with an autoregressive error structure.

Avenues for further research include investigating the possibility of using the proposed tests to test for seasonality in discrete choice data sets, which is an important topic in the time series literature [compare Franses and Paap, 2004; Hyndman and Athanasopoulos, 2018, for example]. Investigating a preliminary stage of model diagnostics to be used prior to the proposed tests, comparable to residual plots in the linear model setting, would be beneficial and could help practitioners to identify groups of observational pairs to be used in the tests proposed here.

CRediT author statement

Sebastian Büscher: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original draft, Writing - Review & Editing, Visualization. **Dietmar Bauer:** Methodology, Software, Investigation, Resources, Writing - Review & Editing, Project administration, Funding acquisition.

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 356500581 which is gratefully acknowledged. The authors would also like to extend their gratitude to Manuel Batram and Lennart Oelschläger, who contributed to the codebase used for the calculations, and Kaja Balzereit for proofreading and providing an outside perspective.

Declarations of interest

Declarations of interest: none

The funding agency Deutsche Forschungsgemeinschaft (DFG) had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Appendix 3.A

Theorem 3.1. *Let the data be generated from a mixed MNP model with utility* $U_{n,t,j} = X'_{n,t,j}\beta + X'_{n,t,j}\gamma_n + \varepsilon_{n,t,j}$ *for* $n = 1, .., N$, $t = 1, .., T_n$, *where* $\gamma_n \sim \mathcal{N}(0, \Omega)$ *independent of* $X_{n,t,j}$, *and* $\varepsilon_{n,t,j}$ *iid over decision-makers, such that Assumption 3.1 is satisfied. Let the model be estimated using CML with weights* $w_{n,a,b}$ *in accordance with Assumption 3.2, providing the estimator* $\hat{\theta}$ *maximising the CML such that the score is zero and the estimator is asymptotically normal with* $\sqrt{N}(\hat{\theta} - \theta_\circ) \xrightarrow{d} Z \sim \mathcal{N}(0, V_\theta)$.
Then

$$g_{n,a,b}(\hat{\theta}) = \partial_\theta \log L_{n,a,b}(\hat{\theta}) = g_{n,a,b}(\theta_\circ) + \partial_\theta g_{n,a,b}(\bar{\theta}_n)(\hat{\theta} - \theta_\circ), \tag{3.19}$$

where $\mathbb{E}(g_{n,a,b}(\theta_\circ)) = 0$. *Furthermore,* $\mathbb{E}\partial_\theta g_{n,a,b}(\theta_\circ) = H_g > 0$ *and* $\bar{\theta}_n$ *is an intermediate value such that* $\bar{\theta}_n \xrightarrow{p} \theta_\circ$ *as* $N \to \infty$.
Finally,

$$N^{-1} \sum_{n=1}^{N} \partial_\theta g_{n,a,b}(\bar{\theta}_n)(\hat{\theta} - \theta_\circ) = H_g(\hat{\theta} - \theta_\circ) + o_P(N^{-1/2}). \tag{3.20}$$

Proof of Theorem 3.1. The intermediate value theorem provides the approximation of $g_{n,a,b}(\hat{\theta})$. Correct specification of the model implies that the score at the true parameter vector has zero expectation and the Hessian is positive definite. The data generating process implies that the Hessian H_g is identical for all pairs.
The last statement then holds from standard asymptotic theory since $N^{-1}\sum_{n=1}^{N}\partial_\theta g_{n,a,b}(\bar{\theta}) \to V_g$ for each sequence $\bar{\theta} \to \theta_\circ$. Since the intermediate points all lie between $\hat{\theta}$ and $\theta_\circ$ the maximal distance of the intermediate points decreases to zero. □

Theorem 3.2. *Under the assumptions of Theorem 3.1 and Assumptions 3.3, the test statistic LM_P has an asymptotic χ^2_P distribution, given that $\hat{V}_{\Delta g}$ is regular.*
The distribution of the test statistic t converges to a standard normal distribution as $N \to \infty$, provided that $(\Delta g)_{(j)}$ is not constant over pairs.

Proof of Theorem 3.2. The theorem follows from the iid property of $\Delta\bar{g}_{G_1,G_2,n}$ in combination with bounds on the variances. Since the regressor vector has bounded support, the gradients are uniformly bounded, such that standard central limit theorems for iid sequences can be applied.
The consistency of the estimation of the variance follows from the law of large numbers applied to the empirical variance.
For the squared test statistic LM_P, we hence obtain asymptotic χ^2_P distribution from standard arguments. The t-statistic converges to a standard normal distribution. □

Appendix 3.B

This appendix section contains the tables indicating the shares of the largest test statistic among the componentwise tests, the plots showcasing the distribution of the test statistics under the null hypothesis, and the plots of the rejection rates depending on the severity of violation of the null hypothesis, summarising the empirical results of the paper.

N	T	test	-0.99	-0.95	-0.5	-0.2	-0.05	0	0.05	0.2	0.5	0.95	0.99
100	10	β_0	0.37	0.43	0.46	0.38	0.40	0.43	0.40	0.41	0.34	0.39	0.29
		σ_γ	0.34	0.23	0.29	0.28	0.38	0.29	0.38	0.34	0.31	0.26	0.38
		σ_ε	0.29	0.34	0.25	0.34	0.22	0.29	0.22	0.25	0.35	0.35	0.33
	14	β_0	0.42	0.40	0.44	0.40	0.45	0.43	0.46	0.43	0.37	0.41	0.33
		σ_γ	0.29	0.34	0.33	0.41	0.30	0.29	0.28	0.30	0.34	0.31	0.34
		σ_ε	0.29	0.26	0.23	0.19	0.25	0.29	0.26	0.27	0.29	0.28	0.33
	20	β_0	0.49	0.51	0.40	0.43	0.36	0.43	0.36	0.40	0.39	0.38	0.42
		σ_γ	0.29	0.20	0.32	0.29	0.37	0.29	0.31	0.34	0.30	0.34	0.28
		σ_ε	0.22	0.29	0.28	0.28	0.27	0.29	0.33	0.26	0.31	0.28	0.30
	30	β_0	0.52	0.42	0.53	0.47	0.37	0.43	0.33	0.41	0.41	0.38	0.31
		σ_γ	0.18	0.28	0.19	0.23	0.31	0.29	0.34	0.32	0.31	0.35	0.37
		σ_ε	0.30	0.30	0.28	0.30	0.32	0.29	0.33	0.27	0.28	0.27	0.32
500	10	β_0	0.39	0.50	0.46	0.38	0.38	0.43	0.44	0.40	0.34	0.46	0.42
		σ_γ	0.33	0.24	0.27	0.29	0.34	0.29	0.30	0.35	0.30	0.28	0.32
		σ_ε	0.28	0.26	0.27	0.33	0.28	0.29	0.26	0.25	0.36	0.26	0.26
	14	β_0	0.46	0.41	0.45	0.50	0.50	0.43	0.40	0.47	0.42	0.43	0.43
		σ_γ	0.26	0.26	0.30	0.28	0.29	0.29	0.35	0.29	0.33	0.27	0.27
		σ_ε	0.28	0.33	0.25	0.22	0.21	0.29	0.25	0.24	0.25	0.30	0.30
	20	β_0	0.46	0.46	0.43	0.44	0.44	0.43	0.42	0.42	0.41	0.40	0.43
		σ_γ	0.29	0.26	0.30	0.29	0.30	0.29	0.31	0.30	0.31	0.34	0.31
		σ_ε	0.25	0.28	0.27	0.27	0.26	0.29	0.27	0.28	0.28	0.26	0.26
	30	β_0	0.47	0.49	0.38	0.46	0.47	0.43	0.52	0.45	0.44	0.34	0.39
		σ_γ	0.26	0.26	0.33	0.32	0.27	0.29	0.24	0.32	0.34	0.32	0.30
		σ_ε	0.27	0.25	0.29	0.22	0.26	0.29	0.24	0.23	0.22	0.34	0.31
1000	10	β_0	0.47	0.41	0.38	0.32	0.39	0.43	0.39	0.38	0.37	0.46	0.44
		σ_γ	0.32	0.32	0.32	0.32	0.29	0.29	0.33	0.27	0.33	0.24	0.30
		σ_ε	0.21	0.27	0.30	0.36	0.32	0.29	0.28	0.35	0.30	0.30	0.26
	14	β_0	0.38	0.42	0.43	0.38	0.48	0.43	0.46	0.41	0.35	0.35	0.39
		σ_γ	0.28	0.25	0.37	0.38	0.25	0.29	0.27	0.30	0.33	0.32	0.35
		σ_ε	0.34	0.33	0.20	0.24	0.27	0.29	0.27	0.29	0.32	0.33	0.26
	20	β_0	0.37	0.44	0.45	0.41	0.45	0.43	0.48	0.42	0.38	0.42	0.42
		σ_γ	0.31	0.29	0.21	0.28	0.29	0.29	0.25	0.24	0.31	0.28	0.34
		σ_ε	0.32	0.27	0.34	0.31	0.26	0.29	0.27	0.34	0.31	0.30	0.24
	30	β_0	0.45	0.43	0.49	0.52	0.51	0.43	0.45	0.47	0.42	0.39	0.37
		σ_γ	0.26	0.30	0.27	0.32	0.24	0.29	0.26	0.26	0.31	0.28	0.32
		σ_ε	0.29	0.27	0.24	0.16	0.25	0.29	0.29	0.27	0.27	0.33	0.31

Table 3.1: Shares of times that a particular componentwise test had the largest test statistic for simulations with an autoregressive error process, by number of individuals N, number of observations per individual T and autocorrelation coefficient ρ for the test with the *FirstLast* groupings.

N	T	test	-0.99	-0.95	-0.5	-0.2	-0.05	0	0.05	0.2	0.5	0.95	0.99
100	10	β_0	0.09	0.06	0.07	0.18	0.30	0.37	0.39	0.30	0.02	0.00	0.00
		σ_γ	**0.84**	**0.87**	**0.78**	0.58	0.39	0.31	0.30	0.44	**0.72**	**0.97**	**0.96**
		σ_ε	0.07	0.07	0.15	0.24	0.31	0.32	0.31	0.26	0.26	0.03	0.04
	14	β_0	0.24	0.16	0.08	0.24	0.42	0.37	0.49	0.29	0.03	0.00	0.00
		σ_γ	0.56	**0.70**	**0.83**	0.53	0.34	0.31	0.26	0.43	**0.71**	**0.96**	**0.95**
		σ_ε	0.20	0.14	0.09	0.23	0.24	0.32	0.25	0.28	0.26	0.04	0.05
	20	β_0	0.34	0.20	0.04	0.19	0.28	0.37	0.30	0.35	0.04	0.00	0.00
		σ_γ	0.43	**0.68**	**0.92**	**0.69**	0.45	0.31	0.34	0.34	0.62	**0.99**	**0.98**
		σ_ε	0.23	0.12	0.04	0.12	0.27	0.32	0.36	0.31	0.34	0.01	0.02
	30	β_0	**0.81**	0.39	0.03	0.19	0.32	0.37	0.33	0.31	0.07	0.00	0.00
		σ_γ	0.14	0.39	**0.89**	0.52	0.36	0.31	0.30	0.30	0.65	**0.98**	**0.98**
		σ_ε	0.05	0.22	0.08	0.29	0.32	0.32	0.37	0.39	0.28	0.02	0.02
500	10	β_0	0.00	0.00	0.00	0.07	0.29	0.37	0.30	0.03	0.00	0.00	0.00
		σ_γ	**0.99**	**1.00**	**1.00**	**0.86**	0.38	0.31	0.40	**0.69**	**0.85**	**1.00**	**1.00**
		σ_ε	0.01	0.00	0.00	0.07	0.33	0.32	0.30	0.28	0.15	0.00	0.00
	14	β_0	0.01	0.00	0.00	0.07	0.29	0.37	0.39	0.07	0.00	0.00	0.00
		σ_γ	**0.98**	**1.00**	**1.00**	**0.88**	0.50	0.31	0.28	0.65	**0.85**	**1.00**	**1.00**
		σ_ε	0.01	0.00	0.00	0.05	0.21	0.32	0.33	0.28	0.15	0.00	0.00
	20	β_0	0.12	0.00	0.00	0.04	0.30	0.37	0.35	0.06	0.00	0.00	0.00
		σ_γ	**0.80**	**0.99**	**1.00**	**0.89**	0.40	0.31	0.35	0.61	**0.88**	**1.00**	**1.00**
		σ_ε	0.08	0.01	0.00	0.07	0.31	0.32	0.30	0.33	0.12	0.00	0.00
	30	β_0	**0.89**	0.18	0.00	0.03	0.31	0.37	0.39	0.05	0.00	0.00	0.00
		σ_γ	0.09	**0.68**	**1.00**	**0.91**	0.37	0.31	0.33	**0.67**	**0.87**	**1.00**	**1.00**
		σ_ε	0.02	0.14	0.00	0.06	0.32	0.32	0.28	0.28	0.13	0.00	0.00
1000	10	β_0	0.00	0.00	0.00	0.01	0.22	0.37	0.31	0.01	0.00	0.00	0.00
		σ_γ	**1.00**	**1.00**	**1.00**	**0.99**	0.44	0.31	0.37	**0.71**	**0.97**	**1.00**	**1.00**
		σ_ε	0.00	0.00	0.00	0.00	0.34	0.32	0.32	0.28	0.03	0.00	0.00
	14	β_0	0.00	0.00	0.00	0.02	0.25	0.37	0.27	0.02	0.00	0.00	0.00
		σ_γ	**1.00**	**1.00**	**1.00**	**0.96**	0.54	0.31	0.46	0.63	**0.97**	**1.00**	**1.00**
		σ_ε	0.00	0.00	0.00	0.02	0.21	0.32	0.27	0.35	0.03	0.00	0.00
	20	β_0	0.05	0.00	0.00	0.01	0.24	0.37	0.33	0.00	0.00	0.00	0.00
		σ_γ	**0.93**	**1.00**	**1.00**	**0.98**	0.50	0.31	0.39	**0.69**	**0.93**	**1.00**	**1.00**
		σ_ε	0.02	0.00	0.00	0.01	0.26	0.32	0.28	0.31	0.07	0.00	0.00
	30	β_0	**0.95**	0.07	0.00	0.00	0.24	0.37	0.39	0.01	0.00	0.00	0.00
		σ_γ	0.05	**0.87**	**1.00**	**0.98**	0.49	0.31	0.32	**0.72**	**0.93**	**1.00**	**1.00**
		σ_ε	0.00	0.06	0.00	0.02	0.27	0.32	0.29	0.27	0.07	0.00	0.00

Table 3.2: Shares of times that a particular componentwise test had the largest test statistic for simulations with an autoregressive error process, by number of individuals N, number of observations per individual T and autocorrelation coefficient ρ for the test with the *NearFar* groupings. Values larger or equal to 0.66 are highlighted in bold.

N	T	test	0	0.01	0.05	0.1	0.25	0.5	0.75	0.95	1	1.5	2
		β_0	0.33	0.35	0.31	0.15	0.03	0.00	0.00	0.00	0.00	0.00	0.00
	10	β_1	0.31	0.32	0.42	0.57	**0.95**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.36	0.33	0.27	0.28	0.02	0.00	0.00	0.00	0.00	0.00	0.00
		β_0	0.32	0.32	0.30	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	14	β_1	0.35	0.32	0.44	**0.72**	**0.96**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.33	0.36	0.26	0.14	0.04	0.00	0.00	0.00	0.00	0.00	0.00
100		β_0	0.36	0.36	0.28	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	20	β_1	0.31	0.36	0.39	**0.70**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.33	0.28	0.33	0.17	0.01	0.00	0.00	0.00	0.00	0.00	0.00
		β_0	0.33	0.28	0.19	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	β_1	0.31	0.39	0.53	**0.86**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.36	0.33	0.28	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		β_0	0.35	0.34	0.20	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10	β_1	0.30	0.36	0.65	**0.97**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.35	0.30	0.15	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		β_0	0.35	0.34	0.14	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	14	β_1	0.31	0.30	**0.74**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.34	0.36	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
500		β_0	0.37	0.33	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	20	β_1	0.32	0.34	**0.78**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.32	0.33	0.13	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		β_0	0.33	0.23	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	β_1	0.31	0.46	**0.91**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.36	0.31	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		β_0	0.34	0.37	0.10	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	10	β_1	0.31	0.32	**0.80**	**0.98**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.35	0.31	0.10	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		β_0	0.36	0.32	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	14	β_1	0.33	0.39	**0.84**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.31	0.29	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000		β_0	0.32	0.29	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	20	β_1	0.41	0.35	**0.87**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.27	0.36	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		β_0	0.40	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	30	β_1	0.31	0.45	**0.97**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.29	0.25	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 3.3: Shares of times that a particular componentwise test had the largest test statistic for simulations with a shift in the $\beta_{1,n,t}$ parameter between panel waves, by number of individuals N, number of observations per individual T and size of the shift α for the test with the *FirstLast* groupings. Values larger or equal to 0.66 are highlighted in bold.

N	T	test	0	0.01	0.05	0.1	0.25	0.5	0.75	0.95	1	1.5	2
100	10	β_0	0.30	0.33	0.36	0.34	0.31	0.35	0.33	0.31	0.30	0.30	0.33
		β_1	0.36	0.31	0.28	0.34	0.31	0.23	0.21	0.13	0.16	0.17	0.29
		σ_γ	0.34	0.36	0.36	0.32	0.38	0.42	0.46	0.56	0.54	0.53	0.38
	14	β_0	0.41	0.36	0.37	0.35	0.35	0.34	0.26	0.26	0.24	0.24	0.31
		β_1	0.27	0.29	0.33	0.32	0.31	0.26	0.17	0.16	0.15	0.16	0.18
		σ_γ	0.32	0.35	0.30	0.33	0.34	0.40	0.57	0.58	0.61	0.60	0.51
	20	β_0	0.35	0.38	0.35	0.31	0.33	0.33	0.30	0.26	0.25	0.23	0.32
		β_1	0.37	0.35	0.40	0.40	0.37	0.24	0.12	0.12	0.12	0.12	0.18
		σ_γ	0.28	0.27	0.25	0.29	0.30	0.43	0.58	0.62	0.63	0.65	0.50
	30	β_0	0.33	0.34	0.33	0.29	0.32	0.29	0.18	0.15	0.14	0.17	0.25
		β_1	0.26	0.25	0.27	0.24	0.21	0.08	0.04	0.04	0.03	0.13	0.19
		σ_γ	0.41	0.41	0.40	0.47	0.47	0.63	**0.78**	**0.81**	**0.83**	**0.70**	0.56
500	10	β_0	0.45	0.47	0.37	0.42	0.37	0.39	0.32	0.20	0.22	0.13	0.25
		β_1	0.33	0.31	0.34	0.30	0.31	0.13	0.05	0.06	0.02	0.13	0.19
		σ_γ	0.22	0.22	0.29	0.28	0.32	0.48	0.63	**0.74**	**0.76**	**0.74**	0.56
	14	β_0	0.28	0.32	0.28	0.28	0.34	0.20	0.18	0.14	0.12	0.20	0.23
		β_1	0.37	0.36	0.34	0.33	0.26	0.18	0.05	0.04	0.04	0.12	0.15
		σ_γ	0.35	0.32	0.38	0.39	0.40	0.62	**0.77**	**0.82**	**0.84**	**0.68**	0.62
	20	β_0	0.33	0.32	0.34	0.34	0.33	0.22	0.11	0.05	0.04	0.06	0.18
		β_1	0.35	0.37	0.34	0.34	0.28	0.05	0.01	0.01	0.01	0.05	0.14
		σ_γ	0.32	0.31	0.31	0.32	0.40	**0.73**	**0.88**	**0.94**	**0.94**	**0.89**	0.67
	30	β_0	0.33	0.32	0.35	0.34	0.32	0.13	0.04	0.00	0.00	0.01	0.08
		β_1	0.36	0.38	0.26	0.30	0.17	0.02	0.00	0.00	0.00	0.00	0.13
		σ_γ	0.31	0.30	0.39	0.36	0.51	**0.85**	**0.96**	**1.00**	**1.00**	**0.99**	**0.79**
1000	10	β_0	0.38	0.42	0.41	0.43	0.35	0.26	0.10	0.05	0.07	0.11	0.28
		β_1	0.29	0.30	0.33	0.30	0.24	0.07	0.01	0.02	0.01	0.07	0.11
		σ_γ	0.33	0.28	0.26	0.27	0.41	**0.67**	**0.89**	**0.93**	**0.92**	**0.82**	0.61
	14	β_0	0.31	0.32	0.26	0.29	0.36	0.20	0.08	0.03	0.04	0.07	0.19
		β_1	0.38	0.35	0.35	0.32	0.28	0.08	0.03	0.01	0.01	0.05	0.21
		σ_γ	0.31	0.33	0.39	0.39	0.36	**0.72**	**0.89**	**0.96**	**0.95**	**0.88**	0.60
	20	β_0	0.31	0.28	0.24	0.25	0.37	0.22	0.03	0.00	0.00	0.01	0.12
		β_1	0.31	0.34	0.35	0.32	0.19	0.00	0.00	0.00	0.00	0.03	0.11
		σ_γ	0.38	0.38	0.41	0.43	0.44	**0.78**	**0.97**	**1.00**	**1.00**	**0.96**	**0.77**
	30	β_0	0.36	0.31	0.34	0.30	0.25	0.04	0.00	0.00	0.00	0.00	0.06
		β_1	0.33	0.38	0.33	0.31	0.14	0.00	0.00	0.00	0.00	0.00	0.05
		σ_γ	0.31	0.31	0.33	0.39	0.61	**0.96**	**1.00**	**1.00**	**1.00**	**1.00**	**0.89**

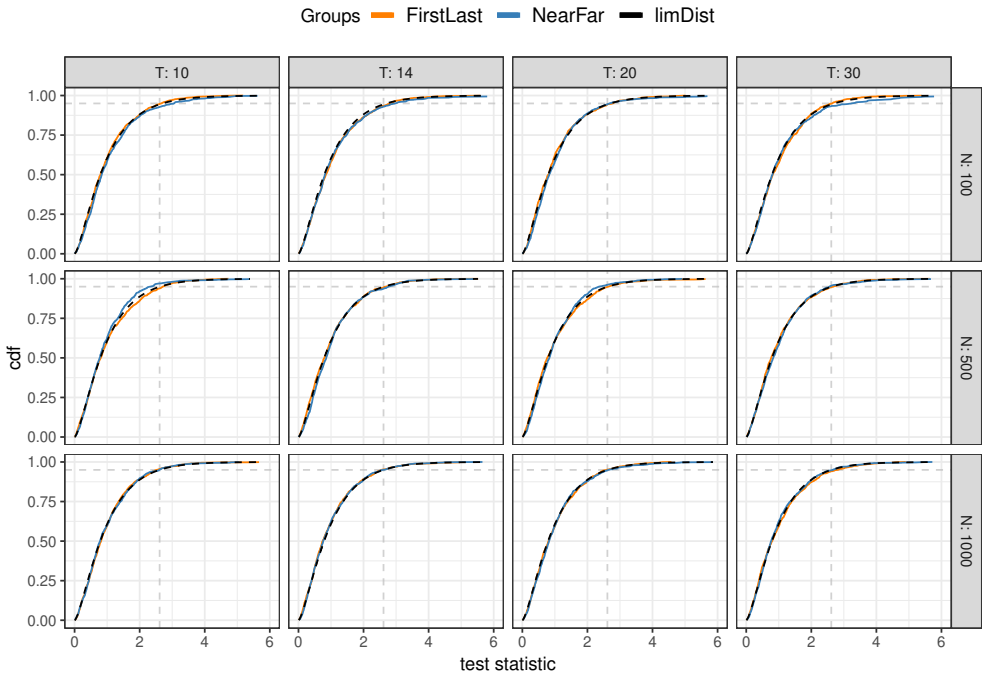
Table 3.4: Shares of times that a particular componentwise test had the largest test statistic for simulations with a shift in the $\beta_{1,n,t}$ parameter between panel waves, by number of individuals N, number of observations per individual T and size of the shift α for the test with the *NearFar* groupings. Values larger or equal to 0.66 are highlighted in bold.

N	T	test	0	0.01	0.05	0.1	0.25	0.5	0.75	0.95	1	1.5	2
		β_0	0.33	0.34	0.35	0.35	0.38	0.35	0.18	0.07	0.08	0.01	0.03
	10	β_1	0.31	0.30	0.30	0.32	0.30	0.37	0.61	**0.77**	**0.74**	**0.90**	**0.96**
		σ_γ	0.36	0.36	0.35	0.33	0.32	0.28	0.21	0.16	0.18	0.09	0.01
		β_0	0.32	0.30	0.34	0.36	0.37	0.22	0.07	0.03	0.03	0.01	0.02
	14	β_1	0.35	0.36	0.37	0.33	0.32	0.54	**0.74**	**0.82**	**0.83**	**0.94**	**0.95**
		σ_γ	0.33	0.34	0.29	0.31	0.31	0.24	0.19	0.15	0.14	0.05	0.03
100		β_0	0.36	0.37	0.37	0.40	0.34	0.31	0.07	0.01	0.00	0.00	0.04
	20	β_1	0.31	0.32	0.32	0.27	0.28	0.48	**0.71**	**0.82**	**0.85**	**0.91**	**0.91**
		σ_γ	0.33	0.31	0.31	0.33	0.38	0.21	0.22	0.17	0.15	0.09	0.05
		β_0	0.33	0.33	0.35	0.38	0.39	0.23	0.04	0.01	0.01	0.00	0.07
	30	β_1	0.31	0.34	0.27	0.26	0.31	0.47	**0.80**	**0.80**	**0.81**	**0.86**	**0.90**
		σ_γ	0.36	0.33	0.38	0.36	0.30	0.30	0.16	0.19	0.18	0.14	0.03
		β_0	0.35	0.32	0.32	0.28	0.30	0.07	0.00	0.00	0.00	0.00	0.00
	10	β_1	0.30	0.33	0.32	0.36	0.41	0.72	**0.93**	**0.98**	**0.98**	**0.99**	**1.00**
		σ_γ	0.35	0.35	0.36	0.36	0.29	0.21	0.07	0.02	0.02	0.01	0.00
		β_0	0.35	0.37	0.36	0.37	0.27	0.05	0.00	0.00	0.00	0.00	0.00
	14	β_1	0.31	0.31	0.33	0.31	0.38	0.71	**0.94**	**0.99**	**1.00**	**1.00**	**1.00**
		σ_γ	0.34	0.32	0.31	0.32	0.35	0.24	0.06	0.01	0.00	0.00	0.00
500		β_0	0.37	0.37	0.38	0.38	0.31	0.02	0.00	0.00	0.00	0.00	0.00
	20	β_1	0.32	0.32	0.32	0.32	0.41	**0.82**	**0.94**	**0.98**	**0.98**	**1.00**	**1.00**
		σ_γ	0.32	0.31	0.30	0.30	0.28	0.16	0.06	0.02	0.02	0.00	0.00
		β_0	0.33	0.33	0.32	0.33	0.18	0.01	0.00	0.00	0.00	0.00	0.00
	30	β_1	0.31	0.31	0.30	0.28	0.56	**0.87**	**0.93**	**0.94**	**0.95**	**0.99**	**1.00**
		σ_γ	0.36	0.36	0.38	0.39	0.26	0.12	0.07	0.06	0.05	0.01	0.00
		β_0	0.34	0.33	0.30	0.35	0.26	0.01	0.00	0.00	0.00	0.00	0.00
	10	β_1	0.31	0.34	0.34	0.34	0.41	**0.81**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.35	0.33	0.36	0.31	0.33	0.18	0.01	0.00	0.00	0.00	0.00
		β_0	0.36	0.36	0.36	0.35	0.22	0.00	0.00	0.00	0.00	0.00	0.00
	14	β_1	0.33	0.33	0.31	0.32	0.48	**0.89**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.31	0.31	0.33	0.33	0.30	0.11	0.00	0.00	0.00	0.00	0.00
1000		β_0	0.32	0.34	0.38	0.38	0.27	0.00	0.00	0.00	0.00	0.00	0.00
	20	β_1	0.41	0.39	0.34	0.32	0.54	**0.93**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.27	0.27	0.28	0.30	0.19	0.07	0.01	0.00	0.00	0.00	0.00
		β_0	0.40	0.41	0.40	0.40	0.16	0.00	0.00	0.00	0.00	0.00	0.00
	30	β_1	0.31	0.31	0.29	0.35	0.58	**0.96**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**
		σ_γ	0.29	0.28	0.31	0.25	0.26	0.04	0.01	0.00	0.00	0.00	0.00

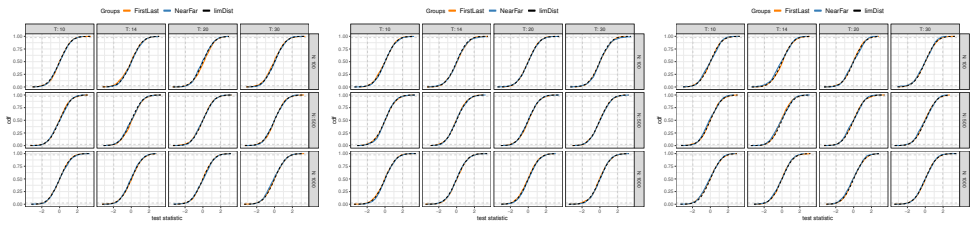
Table 3.5: Shares of times that a particular componentwise test had the largest test statistic for simulations with a split in the $\beta_{1,n,t}$ parameter in second panel wave, by number of individuals N, number of observations per individual T and size of the shift α for the test with the *FirstLast* groupings. Values larger or equal to 0.66 are highlighted in bold.

N_total	Tp	test	0	0.01	0.05	0.1	0.25	0.5	0.75	0.95	1	1.5	2
		β_0	0.30	0.31	0.29	0.29	0.31	0.27	0.30	0.35	0.32	0.30	0.29
	10	β_1	0.36	0.36	0.36	0.37	0.35	0.35	0.38	0.33	0.34	0.40	0.44
		σ_γ	0.34	0.33	0.35	0.34	0.34	0.38	0.32	0.32	0.34	0.30	0.27
		β_0	0.41	0.38	0.35	0.40	0.41	0.42	0.34	0.32	0.32	0.30	0.32
	14	β_1	0.27	0.26	0.27	0.23	0.24	0.28	0.27	0.30	0.29	0.32	0.36
		σ_γ	0.32	0.36	0.38	0.37	0.35	0.30	0.39	0.38	0.39	0.38	0.32
100		β_0	0.35	0.38	0.34	0.37	0.33	0.31	0.23	0.26	0.26	0.28	0.17
	20	β_1	0.37	0.35	0.37	0.34	0.38	0.34	0.40	0.34	0.33	0.39	0.34
		σ_γ	0.28	0.27	0.29	0.29	0.29	0.35	0.37	0.40	0.41	0.33	0.49
		β_0	0.33	0.34	0.37	0.30	0.31	0.30	0.35	0.32	0.31	0.26	0.21
	30	β_1	0.26	0.22	0.20	0.23	0.30	0.33	0.25	0.26	0.27	0.39	0.40
		σ_γ	0.41	0.44	0.43	0.47	0.39	0.37	0.40	0.42	0.42	0.35	0.39
		β_0	0.45	0.48	0.39	0.40	0.34	0.39	0.37	0.32	0.30	0.18	0.26
	10	β_1	0.33	0.30	0.34	0.34	0.46	0.26	0.19	0.28	0.30	0.46	0.35
		σ_γ	0.22	0.22	0.27	0.26	0.20	0.35	0.44	0.40	0.40	0.36	0.39
		β_0	0.28	0.26	0.30	0.31	0.30	0.27	0.30	0.25	0.23	0.15	0.21
	14	β_1	0.37	0.38	0.35	0.35	0.32	0.24	0.26	0.32	0.30	0.58	0.43
		σ_γ	0.35	0.36	0.35	0.34	0.38	0.49	0.44	0.43	0.47	0.27	0.36
500		β_0	0.33	0.33	0.35	0.32	0.33	0.35	0.31	0.26	0.24	0.10	0.16
	20	β_1	0.35	0.36	0.35	0.37	0.35	0.31	0.26	0.30	0.32	0.57	0.37
		σ_γ	0.32	0.31	0.30	0.32	0.32	0.35	0.43	0.44	0.44	0.33	0.47
		β_0	0.33	0.32	0.34	0.31	0.29	0.29	0.22	0.16	0.16	0.01	0.15
	30	β_1	0.36	0.37	0.34	0.36	0.32	0.28	0.25	0.31	0.33	0.70	0.30
		σ_γ	0.31	0.31	0.32	0.33	0.39	0.43	0.53	0.53	0.51	0.29	0.55
		β_0	0.38	0.44	0.44	0.39	0.35	0.34	0.29	0.25	0.26	0.15	0.18
	10	β_1	0.29	0.28	0.24	0.33	0.28	0.30	0.27	0.33	0.35	0.46	0.39
		σ_γ	0.33	0.28	0.32	0.28	0.37	0.36	0.44	0.42	0.39	0.39	0.43
		β_0	0.31	0.30	0.30	0.27	0.35	0.34	0.25	0.26	0.26	0.12	0.14
	14	β_1	0.38	0.34	0.34	0.37	0.35	0.26	0.31	0.34	0.37	0.58	0.37
		σ_γ	0.31	0.36	0.36	0.36	0.30	0.40	0.44	0.40	0.37	0.30	0.49
1000		β_0	0.31	0.30	0.27	0.27	0.32	0.43	0.29	0.27	0.21	0.05	0.12
	20	β_1	0.31	0.30	0.36	0.31	0.28	0.28	0.21	0.27	0.30	**0.67**	0.42
		σ_γ	0.38	0.40	0.37	0.42	0.40	0.29	0.50	0.46	0.49	0.28	0.46
		β_0	0.36	0.38	0.36	0.35	0.35	0.33	0.23	0.15	0.13	0.02	0.09
	30	β_1	0.33	0.32	0.29	0.29	0.22	0.15	0.13	0.19	0.27	**0.68**	0.26
		σ_γ	0.31	0.30	0.35	0.36	0.43	0.52	0.64	**0.66**	0.60	0.30	0.65

Table 3.6: Shares of times that a particular componentwise test had the largest test statistic for simulations with a split in the $\beta_{1,n,t}$ parameter in second panel wave, by number of individuals N, number of observations per individual T and size of the shift α for the test with the *NearFar* groupings. Values larger or equal to 0.66 are highlighted in bold.

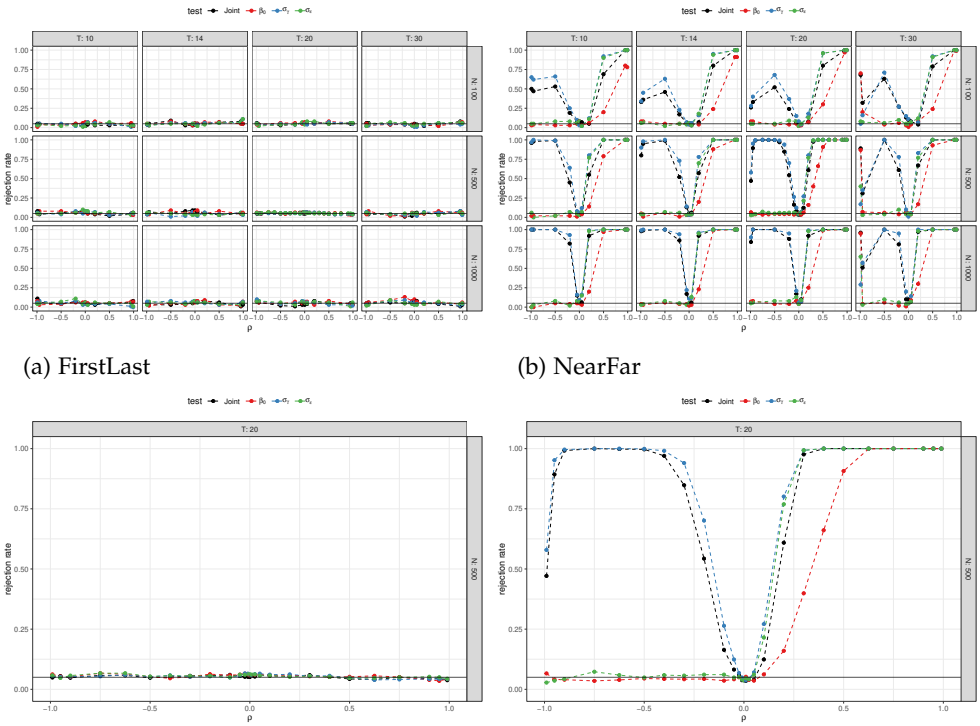


(a) Joint test



(b) $\theta_1 = \beta_0$ (c) $\theta_2 = \sigma_\gamma$ (d) $\theta_3 = \sigma_\varepsilon$

Figure 3.1: CDFs of the test statistics for the joint test and the tests for the individual parameter components of $\theta = (\beta_0, \sigma_\gamma, \sigma_\varepsilon)'$ for the three different groupings. These are based on 1000 simulated data sets each, with the CDF of the limiting distribution of the test statistic shown in dashed black. The dashed vertical lines represent the respective 95% quantile for the joint test and the 2.5% and 97.5% quantiles for the componentwise tests of the theoretical and empirical distributions. The dashed horizontal lines represent the value of the respective empirical CDFs evaluated at the 95% quantile for the joint test and the 2.5% and 97.5% quantiles for the componentwise tests of the theoretical distribution.



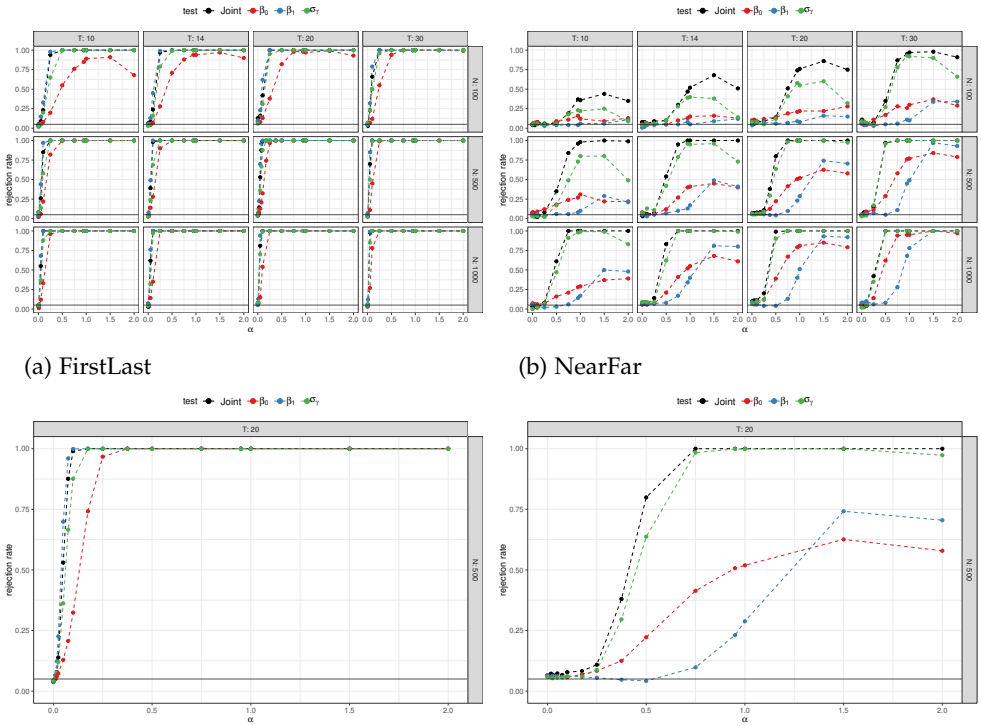
(a) FirstLast

(b) NearFar

(c) FirstLast with $T = 20$, $N = 500$

(d) NearFar with $T = 20$, $N = 500$

Figure 3.2: Rejection rates of the proposed tests for different CML-pair groupings, with the different tests separated by colour and the autocorrelation coefficient of the autoregressive errors on the x-axis. The rejection rates presented herein are based on 100 simulated data sets, with the exception of those with $T = 20$ and $N = 500$, for which 1000 simulations were employed.



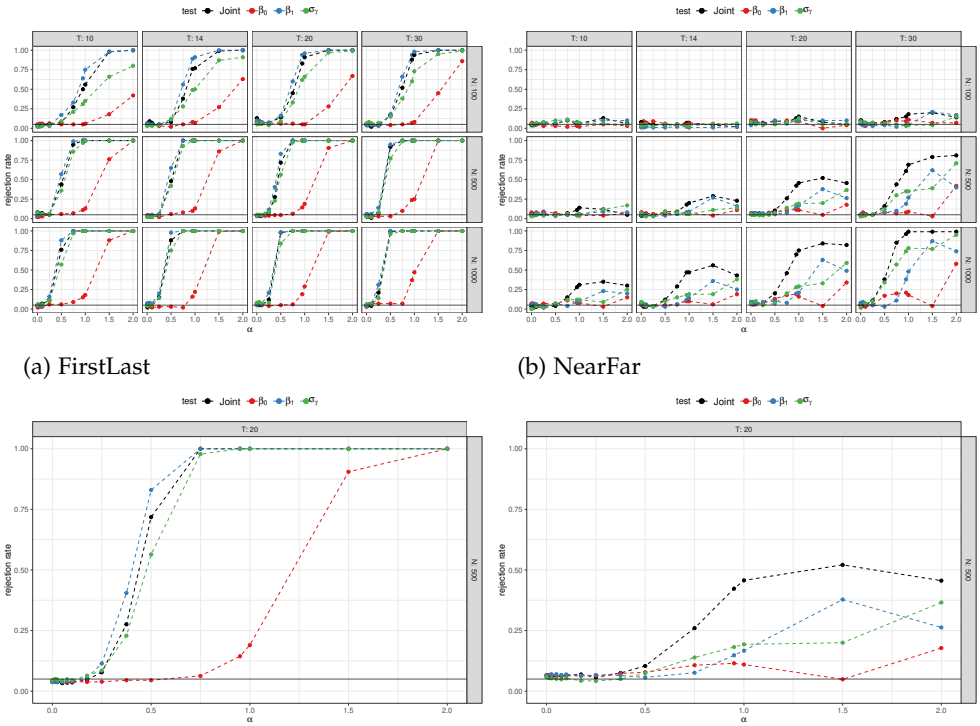
(a) FirstLast

(b) NearFar

(c) FirstLast with $T = 20$, $N = 500$

(d) NearFar with $T = 20$, $N = 500$

Figure 3.3: Rejection rates of the proposed tests for different CML pair groupings, with the different tests separated by colour and the severity of the violation of the null hypothesis (shift of $\beta_{1,n,t}$ value between panel waves) on the x-axis. The rejection rates presented herein are based on 100 simulated data sets, with the exception of those with $T = 20$ and $N = 500$, for which 1000 simulations were employed.



(a) FirstLast

(b) NearFar

(c) FirstLast with $T = 20$, $N = 500$

(d) NearFar with $T = 20$, $N = 500$

Figure 3.4: Rejection rates of the proposed tests for different CML autocorrelation-pair groupings, with the different tests separated by colour and the severity of the violation of the null hypothesis (split of $\beta_{1,n,t}$ value between groups of individuals in second panel wave) on the x-axis. The rejection rates presented herein are based on 100 simulated data sets, with the exception of those with $T = 20$ and $N = 500$, for which 1000 simulations were employed.

CHAPTER

Weighting strategies for pairwise composite marginal likelihood estimation in case of unbalanced panels and unaccounted autoregressive structure of the errors

4

Abstract

Composite marginal likelihood (CML) estimation and its advancements are popular ways to reduce the computational burden involved in the estimation of multinomial probit (MNP) models. CMLs use the product of marginal likelihoods of decision makers instead of the complete joint likelihood, reducing the numerical load. This allows for the estimation of models for larger and more complex data sets. The definition of the CML involves power weights on the marginal likelihoods that in-

fluence the statistical properties of the estimator. In this paper, we discuss how to effectively use the power weights in the cases of (1) unbalanced panel settings, where the weights help to reduce the variance of the estimator, and (2) unaccounted autoregressive structure of the errors, where the weights help to reduce the asymptotic bias of the estimator due to misspecification.

Keywords: probit modelling, composite marginal likelihood, weighting, unbalanced panel data, unaccounted autoregressive error process, efficiency

4.1 Introduction

The mobility behaviour of persons depends heavily on a great number of choices ranging from the choice of the origin and destination as well as the chosen mode for a particular trip to the acquisition of vehicles. In many cases, these choices involve selecting one option from a finite number of alternatives. These decisions depend on the characteristics of the various alternatives as well as the preferences and characteristics of the deciders. In order to learn about these preferences one typically collects data on repeated choice situations.

When modelling (potentially repeated) discrete choices, the two dominant model families are the multinomial logit (MNL) and the multinomial probit models [MNP; see, for example, Train, 2009]. Both can be formulated as random utility models (RUMs) as

$$y^*_{n,t,j} = X'_{n,t,j}\beta_{n,j} + \varepsilon_{n,t,j}, \tag{4.1}$$

where $y^*_{n,t,j} \in \mathbb{R}$ denotes the unobserved utility the individual $n \in \{1,\dots,N\}$ attributes at time $t \in \{1,\dots,T\}$ to choice alternative $j \in \{1,\dots,J\}$, $X_{n,t,j} \in \mathbb{R}^{R\times 1}$ denotes the set of regressors, and $\beta_{n,j} \in \mathbb{R}^{R\times 1}$ denotes a parameter vector, which can potentially be individual specific. Consequently, $R \in \mathbb{N}$ denotes the number of regressors and, hence, also the number of parameters in $\beta_{n,j}$. Assuming that $\beta_{n,j}$ is drawn from a parametric family of distributions $f_j(\cdot;\theta_j)$ (for parameter vector θ_j) independent of the regressors $X_{n,t,k}$ and the noise terms $\varepsilon_{n,t,k}$, the model allows the representation of unobserved taste heterogeneity [compare Train, 2009, Chapter 6, for the logit case].

The alternative with the highest random utility is then assumed to be chosen; therefore, for observation $y_{n,t} = k_{n,t}$, we assume $y^*_{n,t,k_{n,t}} \geq y^*_{n,t,j}, j = 1,...,J$. By assuming different distributions for the random error terms $\varepsilon_{n,t,j}$, we obtain the different choice model families, both of which are typically estimated using likelihood maximisation. Between these two model families, the MNP offers more modelling

flexibility whilst suffering from increased computational costs due to the need to evaluate multivariate normal cumulative distribution functions (MVNCDFs) [see, for example, Train, 2009] to evaluate the choice probabilities.

Moderately sized MNP models were estimated using maximum simulated likelihood (MSL) methods, which show computational difficulties with a growing number of choice alternatives and choice occasions, as demonstrated in Bhat [2014]. Whilst developments in quasi-Monte Carlo (QMC) methods [see Bhat, 2003; Hess et al., 2006; Dick et al., 2016], sparse grid quadrature (SGQ) methods [see Heiss and Winschel, 2008], and, most recently, designed quadrature (DQ) methods [see Ryu and Boyd, 2015; Keshavarzzadeh et al., 2018; Bansal et al., 2022] have been shown to reduce computation time significantly in MSL estimation, they continue to be subject to the curse of dimensionality.

To mitigate these difficulties an alternative approach is to use a composite marginal likelihood (CML) [suggested in Varin, 2008]. In this approach, the product of the probabilities of subsets of the choices of one individual is used instead of the joint probability. These subsets are called margins and the related probabilities marginal probabilities. In most cases a pairwise CML is used, where the margins consist of pairs of choices. In this approach, the logarithm of the joint probability of the T_n observations $y_{n,1} = k_{n,1},\ \ldots,\ y_{n,T_n} = k_{n,T_n}$ from one individual n is replaced in the criterion function by the logarithm of the product of marginal likelihoods of pairs of observations. The difference becomes apparent when comparing the probit log-likelihood function to the resulting CML quasi-log-likelihood function

$$ll(y, X; \theta) = \sum_{n=1}^{N} \log\left(\int \mathbb{P}(y_{n,1} = k_{n,1}, \ldots, y_{n,T_n} = k_{n,T_n}; \tilde{\theta}) f_\theta(\tilde{\theta}) d\tilde{\theta}\right), \tag{4.2}$$

$$ll_{\text{CML}}(y, X; \theta) = \sum_{n=1}^{N} \sum_{a=1}^{T_n - 1} \sum_{b=a+1}^{T_n} \log \underbrace{\left(\int \mathbb{P}(y_{n,a} = k_{n,a},\ y_{n,b} = k_{n,b}; \tilde{\theta}) f_\theta(\tilde{\theta}) d\tilde{\theta}\right)}_{\mathbb{P}(y_{n,a}=k_{n,a},\ y_{n,b}=k_{n,b};\theta)}, \tag{4.3}$$

where θ denotes the vector of all estimable parameters of the model, considering a mixed effects model with f_θ denoting the probability density function (PDF) of the mixing distribution. In the following, probabilities indexed using θ refer to the mixed choice probabilities (neglecting the dependence on $X_{n,t,j}$ in the notation). Using this quasi-likelihood $ll_{\text{CML}}(y, X; \theta)$ reduces the dimension of the MVNCDFs required for calculating the choice probabilities and can thus significantly reduce the computational burden. Usage of the CMLs and advancements of these, such as maximum approximate composite marginal likelihood (MACML) [see Bhat, 2011],

have thus become popular [as evidenced by a 'Web of Science' search returning a total of 3717 papers at the time of writing, with an almost steady increase from 34 papers in 1999 to 301 in 2022, with a peak of 354 in 2021, see Clarivate, 2022].

The CML formulation allows for the assignment of different power weights to each bivariate margin. This results in the weighted CML quasi-log-likelihood function

$$ll_{\mathrm{CML}}(y, X; \theta, W) = \sum_{n=1}^{N} \sum_{a=1}^{T_n-1} \sum_{b=a+1}^{T_n} w_{n,a,b} \log \mathbb{P}(y_{n,a} = k_{n,a}, y_{n,b} = k_{n,b}; \theta) \tag{4.4}$$

where W denotes the collection of weights $w_{n,a,b}$, $n = 1, \ldots, N$, $a = 1, \ldots, T_n - 1$, and $b = a + 1, \ldots, T_n$.

The chosen weights affect the statistical properties of the CML estimator [compare Lindsay et al., 2011] and, hence, have the potential to improve the statistical properties of the estimator compared to the unweighted CML estimator with $w_{n,a,b} \equiv 1$. A discussion of the impact of the weights can be found in Section 4.2 of the survey Varin et al. [2011]. More general CML formulations combine marginal and bivariate margins in the estimation, which can be traced back to early papers in psychology such as Muthén [1984]. A more recent survey can be found in Cox and Reid [2004]. In the transportation literature the pairwise CML formulation as in equation (4.4) is dominant. Beside the unweighted CML estimator also adjacent pairwise formulations are popular, wherein bivariate margins for pairs of adjacent observations are included and all other margins excluded. This reduces the computational cost a lot at the price of reducing the information (and thus increasing the estimation variance) in the CML.

It is not hard to construct synthetic examples (see 4.4) where the variance can be approximately halved (compared to using the unweighted CML) by including weights.[1] Furthermore, by setting a portion of the weights to zero, the number of bivariate probabilities to be computed can significantly be reduced, which further reduces the computational burden.

Thus, by adopting various weighting strategies, the impact of different pairs of observations on the estimation can be tuned. In this paper, we examine how to optimise some of the most commonly used weighting strategies.

This tuning of the weights first will be investigated under the assumption of correctly specified models. However, in panel data situations, oftentimes some choice occasions arise in spurts, leading to small time differences for some occasions, whereas others arise well separated in time. This is the case, for example,

[1] These examples not necessarily are typical for real world applications where the gain can be more modest.

in panel waves of mobility household surveys, such as the German mobility panel [Zumkeller et al., 1999; Zumkeller and Chlond, 2009], wherein the daily mobility of the participants is surveyed for one week every year.

In these cases, it appears plausible that the error terms $\varepsilon_{n,t,j}$ are sampled from an underlying continuous-time stationary process with non-zero autocorrelation being present for short time lags, but correlation diminishing over time. Such dependencies typically are not modelled explicitly. But by ignoring these correlations, the model is misspecified. Misspecifications typically lead to an asymptotic bias, resulting in an inconsistent estimator. In such cases, we investigate whether weighting schemes can be used to reduce the expected asymptotic bias despite the misspecification, providing a diagnostic tool to check for temporal correlation of error terms.

Whereas there has been a fair amount of research into weighting schemes in the context of CML estimation in the past (see, for example, Bhat [2014] for an overview of this topic, Pedeli and Varin [2020] for the use of weighted pairwise CML in latent autoregressive models, Crastes dit Sourd et al. [2020] for the use on autoregressive ordered probit models), this paper aims to contribute to the existing literature by examining the effects of weighting strategies in two specific cases:

First, in Section 4.2, we introduce a two-step optimal group-weight CML estimator to optimise (in a specific sense) the estimator variance in the case of unbalanced panel data. This is important because, in contrast to the maximum likelihood (ML) probit estimator, the CML estimator is not efficient in general [see Varin, 2008].

Second, in Section 4.3, we discuss weighting strategies to reduce the asymptotic bias due to misspecification of the model in case of unaccounted autocorrelation of the error terms. In correctly specified models, the use of close-by observations in CML pairs is favoured [see, for example, Crastes dit Sourd et al., 2020; Joe and Lee, 2009; Varin and Vidoni, 2005], whereas the case of misspecification due to unaccounted autocorrelation is, to the best of our knowledge, still to be investigated and benefits from an opposing approach.

In both cases, we first present the theory of the proposed weighting approaches, including asymptotic properties (Sections 4.2.1 and 4.3.2), before demonstrating the properties of the proposed weighting schemes in finite-sample simulation studies (Sections 4.2.2 and 4.3.3). Proofs of the theorems introduced in the paper can be found in 4.4.

4.2 Weighting in case of unbalanced panel data

When using a *full-pairwise* weighting strategy $w_{n,a,b} \equiv 1$, each of the T_n observations of individual n is part of $(T_n - 1)$ CML margins, and there is a total of $(T_n - 1)T_n/2$ pairs of choice occasions for this individual. Consequently, in an unbalanced panel setting, the number of margins, and thus also the total sum of weights, for one individual scales quadratically with the number of observations. Hence, individuals with more observations have a disproportionately larger influence on the CML function. A priori, it is not clear if this weighting scheme is optimal.

To correct for a possible misrepresentation due to an unequal number of choice occasions, we may group individuals n into clusters according to their number of observations $T_n = s$ and use cluster weights $w_{n,a,b} = w_s$ for $T_n = s$.

Asymptotically, the choice of the weights w_s typically does not affect the consistency of the estimators but will affect the accuracy of the estimator. Cessie and Houwelingen [1994], Joe and Lee [2009] and Kuk and Nott [2000] investigated the effects on the asymptotic variance of the choice of w_s in the context of binary choice occasions and proposed different weighting strategies to account for the imbalance in weights.

Cessie and Houwelingen [1994] used $w_s = (s-1)^{-1}$ for clustered binary data, with $s \in \{2, \ldots, \overline{S}\}$, arguing that this weighting strategy results in each cluster's contribution to the likelihood being relative to its size and that for independent observations within the clusters, the CML is equal to the full likelihood.

Kuk and Nott [2000] argue for using the same weights as Cessie and Houwelingen [1994] to estimate the parameters of the utility function but to use the unweighted CML function to estimate the correlation between observations. They argue that, for the estimation of the correlation between observations within one cluster, the number of pairs is relevant instead of the number of observations.

Joe and Lee [2009] derived optimal power weights minimising the variance of the parameter estimators for an exemplary binary choice model $y_n^* = (y_{n,1,1}^*, \ldots, y_{n,s,1}^*) \sim \mathcal{N}(\mu, \Sigma_s(\sigma^2, \rho))$, with μ denoting the mean utility of the nonreference alternative, σ^2 the variance of the errors and ρ the correlation between the errors at different choice occasions. For the estimation of μ with known correlation ρ between observations within clusters, they find the optimal weights to be $w_s = (s-1)^{-1}[1 + (s-1)\rho]^{-1}$. For uncorrelated observations, this coincides with the weights used by Cessie and Houwelingen [1994] and Kuk and Nott [2000]. For both uncorrelated ($\rho = 0$) and perfectly correlated ($\rho = 1$) observations within clusters, these weights result in the CML being equivalent to the full likelihood. They study these weights with different choices for ρ in $w_s = (s-1)^{-1}[1 + (s-1)\rho]^{-1}$ in the context of multivariate clustered exchangeable probit models (which lead to a similar correlation structure

as the mixed MNP models in the panel case) by performing asymptotic relative efficiency analysis in a simulation study with different correlation coefficients ρ and mixtures of numbers of observations s and find that for unknown correlation ρ using a midway option with $w_s = (s-1)^{-1}[1+1/2(s-1)]^{-1}$, whilst not being the best choice when the correlation is known, performs well over a range of moderate to strong correlation. They, hence, suggest to use $w_s = (s-1)^{-1}[1+1/2(s-1)]^{-1}$ in general for clustered data.

4.2.1 Two-Step Optimal Group-Weight CML Estimator

In this section we will introduce a new two-step optimal group-weight CML estimator, derived from asymptotic properties of the variance of the weighted CML estimator. In order to do so, we will first introduce some assumptions on the initial weights (used in the first step) for the CML estimator (Assumption 4.1) and on the data generating process (Assumption 4.2), both of which are reasonable and can be assumed to hold for a wide range of practical model applications. Under these assumptions we introduce and proof two theorems, Theorem 4.1 on properties of the asymptotic variance of the CML estimator with unbalanced panel data, and Theorem 4.2 on theoretical optimal group-specific weights for the CML estimator. Based on these theorems, we propose two versions of a new two-step optimal group-weight CML estimator in Algorithm 4.1 and Algorithm 4.2.

Following Joe and Lee [2009], also in our setting, the asymptotic variance of the estimator can be calculated as a mixture of variances for deciders with an identical number of choice occasions. We use this to derive variance optimal (in a certain sense) weights depending on the number of choice occasions of an individual.

To calculate the optimal weights, we use a two-step approach commencing from an initial weighting scheme $\mathring{w}_{n,a,b}$ that is subsequently adjusted, using group weights in order to optimise the asymptotic variance.

For the initial weights, several options exist, including those listed in the following assumption:

Assumption 4.1 (Initial Weights)**.** *The initial weights for N deciders, where the n-th decider faces T_n choice occasions, are chosen according to one of the following five schemes:*

(I) $\mathring{w}_{n,a,b} = f(a,b)$ *for some bounded positive function* $f : \mathbb{N}^2 \to [\underline{w}, \overline{w}]$, $0 \leq \underline{w} \leq \overline{w} < \infty$.

(II) $\mathring{w}_{n,a,b} = \mathring{w}_{T_n} \in [\underline{w}, \overline{w}], 0 < \underline{w} \leq \overline{w} < \infty$ (groupwise *CML weights).*

(III) $\mathring{w}_{n,a,b} \in \{0,1\}$ *chosen randomly independent of all other variables, independent, identically distributed (iid) over deciders, such that*

$$\sum_{a=1}^{T_n-1} \sum_{b=a+1}^{T_n} \mathring{w}_{n,a,b} = \mathring{w}_{T_n} \tag{4.5}$$

(selecting $\mathring{w}_{T_n}$ *random pairs from within* $T_n(T_n-1)/2$ *possible pairs).*

(IV) Stratification weights $\mathring{w}_{n,a,b} = \mathring{w}_n$ *drawn iid over deciders from some underlying distribution supported on* $[\underline{w}, \overline{w}], 0 < \underline{w} \leq \overline{w} < \infty$.

(V) A combination of (I)-(IV).

These weights include all commonly used weighting schemes: The full pairwise case $\mathring{w}_{n,a,b} \equiv 1$ can be seen as a special case of any of the schemes above and *adjacent pairwise weighting* is a special case of (I), where $f(a,b) = \mathbb{I}(b = a+1)$ (the indicator function that b equals $a+1$). The proposals of Joe and Lee [2009] correspond to groupwise CML weights (II). (IV) allows for stratification weights, which are typically uniformly bounded from above and below. The various bounds are needed to ensure that (a) each decider has an impact on the criterion function and (b) no decider dominates all others.

In the following, we will use the weights $w_{n,a,b} = w_{T_n}\mathring{w}_{n,a,b}$, where $\mathring{w}_{n,a,b}$ denotes an arbitrary initial weighting scheme from the list above. Subsequently, we use the groupwise weights $w_s, s = 2, ..., \overline{S}$ to optimise certain aspects of the asymptotic variance.

In the subsequent derivation of these groupwise weights, the total initial weight per decider $C_{n,T_n}(\mathring{W}) = \sum_{a=1}^{T_n} \sum_{b=a+1}^{T_n} \mathring{w}_{n,a,b}$ will come into play. In many examples, this is a function of T_n only. However, for stratification weights, it varies over different deciders facing $T_n = s$ choice occasions. Thus, let $C_{N,s}(\mathring{W}) := {N_s}^{-1} \sum_{n:T_n=s} C_{n,T_n}$ denote the average total weight of all deciders with s choices. In all cases fulfilling Assumption 4.1, we have $C_{N,s}(\mathring{W}) \to C_s(\mathring{W})$ almost surely.

Standard theory then implies that choosing positive weights under appropriate assumptions (independent of the sample size and not depending on θ; these assumptions are satisfied for all choices discussed above) implies consistent estimators $\hat{\theta}(W) \to \theta_0$ [see Lindsay et al., 2011] and asymptotic normality derived from mean value theorems [see Cox and Reid, 2004; Joe and Lee, 2009], specifically

$$\sqrt{N}(\hat{\theta}(W) - \theta_0) = -(\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W)/N)^{-1}(\partial_\theta ll_{CML}(y, X; \theta_0, W)/\sqrt{N}), \tag{4.6}$$

where $\bar{\theta}$ denotes an intermediate value between $\hat{\theta}(W)$ and θ_0.

Assumption 4.2 (Data generating process)**.** *The data set $(y_n, X_n), n = 1, ..., N$ is generated by the following mechanism:*

1. *A number T_n of choice occasions are drawn from a discrete random distribution supported in $\{2, 3, ..., \overline{S}\}$.*
2. *For each decider facing T_n choice occasions, a matrix $X_n = [X_{n,1}, ..., X_{n,T_n}] \in \mathbb{R}^{JR \times T_n}$ of regressors is chosen iid over deciders such that for each pair of choice occasions (a, b), the distribution of the matrix $[X_{n,a}, X_{n,b}]$ is identical. Furthermore, $\|X_{n,a}\| \leq M$ (uniform norm bound).*
3. *For given T_n and X_n, the vector of choices $y_n = [y_{n,1}, ..., y_{n,T_n}]' \in \mathcal{J}^{T_n}, \mathcal{J} = \{1, ..., J\}$ is chosen according to the mixed MNP model corresponding to parameter vector $\theta_0 \in \mathbb{R}^d$ for appropriate integer d.*

Note that the assumptions on the regressor variables imply iid sampling over deciders but allow for dependence for the choice occasions faced by one decider. The ordering of the choice occasions must be of no relevance, however, as the distribution of $[X_{n,a}, X_{n,b}]$ for all pairs of choices needs to be identical. This allows for decider-specific regressors that do not vary over choice occasions. Temporal dependence, as would be achieved from systematically posing choices depending on previous answers, is excluded by the assumptions.

A second remark relates to the mechanism for drawing the number of choice occasions: We assume that this choice is performed independently of the regressors or the choice process. This excludes study designs that decide on the number of choice tasks based on either regressors or choices taken so far.

One implication of this data generation is that the number N_s of deciders facing s choice occasions is random, but the fraction N_s/N converges to $f_s \geq 0$, the corresponding probability. The result also holds if $f_s = 0$ for some s. In this case, $\hat{V}_s$ (defined below) is not estimated consistently, whilst $\hat{H}_0$ (see below) still is consistent.

Under this data generating process (DGP), the key to understanding the asymptotic properties then lies in the following representation of the CML, the corresponding score, and the Hessian matrix when using the weighting $w_{n,a,b} = w_{T_n} \mathring{w}_{n,a,b}$:

$$\begin{aligned} ll_{CML}(y, X; \theta_0, W) &= \sum_{s=2}^{\overline{S}} \sum_{n:T_n=s} ll_{CML}(y_n, X_n; \theta_0, W) \\ &= \sum_{s=2}^{\overline{S}} w_s \sum_{n:T_n=s} \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) \right), \end{aligned} \tag{4.7}$$

$$\partial_\theta ll_{CML}(y, X; \theta_0, W) = \sum_{s=2}^{\overline{S}} w_s \sum_{n:T_n=s} \underbrace{\left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) \right)}_{:=g_n(\mathring{W})}$$

$$= \sum_{s=2}^{\overline{S}} w_s \sum_{n:T_n=s} g_n(\mathring{W}), \tag{4.8}$$

$$\partial_\theta^2 ll_{CML}(y, X; \theta_0, W) = \sum_{s=2}^{\overline{S}} w_s \sum_{n:T_n=s} \partial_\theta g_n(\mathring{W}), \tag{4.9}$$

where the next to last equation defines $g_n(\mathring{W})$. Here, we assume that $2 \leq T_n \leq \overline{S}$, $n = 1, ..., N$, such that every decider faces at least two choice occasions and at most $\overline{S}$. Note that

$$\sum_{n:T_n=s} ll_{CML}(y_n, X_n; \theta_0, W) = w_s \sum_{n:T_n=s} ll_{CML}(y_n, X_n; \theta_0, \mathring{W}) \tag{4.10}$$

defines a CML for the balanced subset $\mathcal{S}_s := \{n : T_n = s\}$ whose optimising argument depends on the initial weights $\mathring{W}$ rather than the groupwise weights w_s. Therefore, under standard assumptions (see below), the usual asymptotic properties hold such that $g_n(\mathring{W}), n \in \mathcal{S}_s$ constitutes an iid sequence with $\mathbb{E}g_n(\mathring{W}) = 0$ and variance $V_s(\mathring{W})$. Further independent sampling implies independence for different values of s.

With this notation, we obtain the following result:

Theorem 4.1 (Asymptotic Variance)**.** *Let the data be generated according to Assumption 4.2 with parameter vector θ_0 and let $\hat{\theta}$ be the CML estimator maximising the weighted CML function* (4.4) *using the weights $w_{n,a,b} = w_{T_n} \mathring{w}_{n,a,b}$, where the initial weights $\mathring{w}_{n,a,b}$ adhere to Assumption 4.1, where $C_s(\mathring{W}) > 0, s = 2, ..., \overline{S}$.*

Further, let $w_s \geq 0$ denote group-specific weights according to the number of observations $T_n = s$, $s \in \{2, \ldots, \overline{S}\}$ of decider n, such that $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$. Then the following hold:

(I) $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_\theta(W_s))$*, where the asymptotic variance-covariance matrix $V_\theta(W_s)$ for a given vector $W_s = (w_2, \ldots, w_{\overline{S}})'$ of group-specific weights has the form*

$$V_\theta(W_s) = \sum_{s=2}^{\overline{S}} f_s w_s^2 H_0^{-1} V_s H_0^{-1}, \tag{4.11}$$

with

$$V_s = \mathbb{E}\, g_n(\mathring{W}) g_n(\mathring{W})', \quad H_0 = \mathbb{E}\, \partial_\theta^2 \log \mathbb{P}(y_{n,1}, y_{n,2}, X_n; \theta_0). \tag{4.12}$$

(II) H_0 *can be estimated consistently as*

$$\hat{H}_0 = \frac{1}{\sum_{n,a,b} \mathring{w}_{n,a,b}} \sum_{n,a,b} \mathring{w}_{n,a,b} \partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}). \tag{4.13}$$

(III) If $f_s > 0$ *and* N_s *denotes the number of deciders facing* s *choice occasions, then* V_s *can be estimated consistently using*

$$\hat{V}_s = {N_s}^{-1} \sum_{n:T_n=s} \hat{g}_n(\mathring{W}) \hat{g}_n(\mathring{W})', \tag{4.14}$$

where $\hat{g}_n(\mathring{W}) := \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}) \right)$.

For the proof, see Appendix 4.B.

Since $V_\theta(W)$ is a matrix, it is not obvious that a weighting scheme exists that minimises the variance in the sense of positive definite matrices. Instead, we will investigate optimal choices with respect to linear functions of the form $l : \mathbb{R}^{d \times d} \to \mathbb{R}$; $V \mapsto \mathrm{tr}(VA)$ (for a positive semidefinite matrix $A \in \mathbb{R}^{d \times d}$), mapping positive definite matrices V to the real line. That is, we want to find weights W that minimise $\mathrm{tr}(V_\theta(W)A)$ for given matrix A.

The form of the linear functional $l(V) = \mathrm{tr}(VA)$ covers, with an appropriate choice of the matrix A, a wide range of possible options. With $A = e_j e_j'$, $e_j \in \mathbb{R}^d$ denoting the j-th standard basis vector, we get $l(V_\theta(W_s)) = V_\theta(W_s)_{j,j} = \mathrm{Var}(\theta_j)$, enabling the minimisation of the variance of individual parameters. Kessels et al. [2006] outline four criteria to evaluate choice design efficiency[2], which are related to A-, D-, G-, and V-optimality. Kessels et al. [2006] uses a Bayesian setting and hence integrate the optimality criteria over the prior distribution. As we work in a frequentist setting, we will neglect this.

All these concepts take estimation accuracy as measured by the variance into account, but differ in the form of dependence. The criterion related to A-optimality, which aims to minimise the average variance of the parameters, is $l(V_\theta(W_s)) = \mathrm{tr}\left(V_\theta(W_s)\right)$, which is obtained by choosing $A = I_d$. The criterion termed V-optimality is defined as $l(V_\theta(W_s)) = \int_X \sum_{y=1}^{J} c'(y, X) V_\theta(W_s) c(y, X) p_y(X) dF(X) = \mathrm{tr}\left(V_\theta(W_s) \int_X \sum_{y=1}^{J} c(y, X) c'(y, X) p_y(X) dF(X) \right)$, which is covered by our methodo-

[2]This reference has been pointed out to us by a referee for which we are grateful.

logy with $A = \int_X \sum_{y=1}^{J} c(y,X)c'(y,X)p_y(X)dF(X)$. Here, $c(y,X) = \partial_\theta \mathbb{P}(y,X;\hat{\theta})$ for $X \in \mathbb{R}^{R\times 1}$ and $y \in \{1,\dots,J\}$, $p_y(X) = \mathbb{P}(y,X;\hat{\theta})$, and $F(X)$ denotes the cdf of X. Using the Delta rule $c'(y,X)V_\theta(W_s)c(y,X)$ equals the variance of the predicted choice probability for choice y and regressor matrix X. V-optimality hence assesses the average variance of the estimation of choice probabilities.

The criterion for D-optimality is $l(V_\theta(W_s)) = \det(V_\theta(W_s))$. As the determinant is a non-linear functional, our approach does not cover this particular criterion. G-optimality is defined using $l(V_\theta(W_s)) = \max_{y,X} c'(y,X)V_\theta(W_s)c(y,X)$, which again our methodology does not cover since the maximum function is not linear.

Theorem 4.2 (Optimal Group-Specific Weights)**.** *Let $l : \mathbb{R}^{d\times d} \to \mathbb{R}$ be a linear mapping of the form $l(V) = \mathrm{tr}(VA)$, with $A \in \mathbb{R}^{d\times d}$, $A \neq 0$ symmetric and positive semidefinite, $\hat{\theta}$ be the CML estimator maximising the weighted CML function* (4.4)*, and let the data be generated subject to Assumption 4.2. Let $\mathring{w}_{n,a,b}$ be the initial weights fulfilling Assumption 4.1, where $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$.*

Then $l(V_\theta(W_s))$ is minimised over W_s by

$$w_s^* = \left(\sum_{s=2}^{\overline{S}} f_s C_s(\mathring{W})^2 / v_s(\mathring{W}) \right)^{-1} C_s(\mathring{W})/v_s(\mathring{W}) \propto C_s(\mathring{W})/v_s(\mathring{W}), \tag{4.15}$$

with $v_s(\mathring{W}) = l\left(H_0^{-1} V_s(\mathring{W}) H_0^{-1}\right)$.

For the proof, see Appendix 4.B.

These weights W_s^* are, in general, different from $w_s = 1$ or $w_s = C_s(\mathring{W})$. Note that the proportionality constant is not relevant for the estimation but influences the formulas for the asymptotic variance (as the restriction $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$ is required in Theorem 4.1).

Note that the optimal weight may differ for each diagonal entry of V_θ. Since CML is not efficient, we potentially gain from using different functions for each parameter.

The next natural question is how to determine the optimal weights w_s^* in a practical setting, as H_0 and $V_s(\mathring{W})$ need to be estimated. The formula shows that it depends on $C_s(\mathring{W})$, which does not depend on the data, but the CML method we use. Equally weighted full pairwise weighting, for example, implies $C_s(\mathring{W}) = s(s-1)/2$; adjacent pairwise weights lead to $C_s(\mathring{W}) = s-1$. The second factor is the variance entry $v_s(\mathring{W})$: deciders with choice occasions where we do not learn much from (large $v_s(\mathring{W})$) get a smaller weight, whilst others get a larger weight. This, of course, is reminiscent of generalised least square (GLS) estimation.

From the data, we can estimate V_s as the empirical variance of the derivative of the likelihood contribution of the deciders with a particular number of observations

s (compare Theorem 4.1 (III), as well as H_0).

With these estimates, the optimisation above can easily be applied in order to obtain a more efficient second-step estimator. The detailed procedure to calculate such an estimator can be seen in Algorithm 4.1 in 4.4.

The estimation accuracy of $\hat{V}_s$ depends on the number of individuals facing s choice occasions. In many data sets, this number will differ heavily between values of $s = 2, ..., \overline{S}$. It is well known that the estimation of variances is generally noisy. Good estimators, thus, need to rely on sufficient data support. Therefore, we suggest to enhance the estimation accuracy by imposing smooth variation of W_s as a function of s. This is achieved by using a parametric model $w_s = g(s)$ (for details in the simulations, see Section 4.2.2) and leads to a variation of the two-step optimal group-weight CML estimator, as described in Algorithm 4.2 in 4.4.The choice of parametric model $g(s)$ is dependent on the available data (it is not useful, for example, if only two or three different values for s occur in a dataset) and is ultimately at the discretion of the practitioner, similar to Feasible Generalised Least Square [FGLS; see Wooldridge, 2015]. Weights and estimates calculated using this algorithm will henceforth be denoted with BB_param.

Even though – in contrast to the probit ML estimator – this two-step optimal group-weight CML estimator is not efficient, it is computationally feasible and potentially has lower variance than the standard unweighted CML estimator.

Note also that in the following finite sample simulation example (Section 4.2.2), we use $l(\cdot) = \text{tr}(\cdot)$, so the weights are calculated to minimise the trace of the variance-covariance matrix of the estimator. Another possible approach could be to optimise differently weighted CMLs for each parameter dimension to obtain the best possible estimator. To do so, one could alternate between the estimation of the parameters of β, Ω, and Σ, with different weighting strategies for each, whilst keeping the other parameters fixed. Joe and Lee [2009] proposed a similar procedure in their paper.

As a further note, observe that calculating the optimal weights w_s^* according to equation (4.15) requires the calculation of the second-order derivative $\partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta})$, which is computationally expensive. An alternative is to use the second Bartlett identity

$$\mathbb{E}\partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}) = -\mathbb{E}\left(\partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta})\right) \left(\partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta})\right)'. \tag{4.16}$$

This requires only the calculation of first-order derivatives and leads to a slightly altered algorithm, which has led to similar results as compared to using the exact second derivative.[3]

[3]Results can be obtained from the authors on request.

4.2.2 Finite-sample simulation study

To investigate the effects of the individual weights on the CML estimator in the MACML setting, we simulate unbalanced panel data sets with an underlying MNP model with mixed effects and estimate the model using MACML with different weighting strategies for individuals with different numbers of observations.

Simulation setup

For $J = 6$ choice alternatives, we simulate 100 panel data sets with 300 individuals, each with the underlying DGP

$$y_{n,t} = \underset{j \in \{1,\dots,6\}}{\arg\max}\, y^*_{n,t,j} \tag{4.17}$$

$$y^*_{n,t,j} = \beta_{0,j} + \beta_{1,n} x_{1,n,t,j} + \beta_{2,j} x_{2,n,t} + \varepsilon_{n,t,j}, \tag{4.18}$$

with the parameters

$$\beta_{0,j} = (1 - (j-1)/6)^2 - 1, \quad j = 1, \dots, 6, \tag{4.19}$$

$$\beta_{1,n} \sim \mathcal{N}(\mu_1 = 1,\ \omega_1 = 0.25) \quad \text{iid drawn for each individual } n, \tag{4.20}$$

$$\beta_{2,j} = \begin{cases} \sin(30j) & 2 \le j \le 5 \\ 0 & \text{else} \end{cases}. \tag{4.21}$$

The regressors x were iid drawn for each observation such that

$$x_{1,n,t,j} \sim \begin{cases} \mathcal{N}(0,1) & 1 \le j \le 3 \\ \mathcal{N}(0,0.5) & \text{else} \end{cases} \tag{4.22}$$

$$x_{2,n,t,j} \in \{0,1\}, \quad \text{with} \quad \mathbb{P}(x_{2,n,t,j} = 1) = \mathbb{P}(x_{2,n,t,j} = 0) = 0.5. \tag{4.23}$$

The error terms $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \dots, \varepsilon_{n,t,6})'$ are iid distributed over time and individuals such that $\varepsilon_{n,t} \sim \mathcal{N}(0, \Sigma)$ with Σ as a diagonal matrix with entries $\Sigma_{jj} = |\tilde{\Sigma}_{jj}/\tilde{\Sigma}_{11}|$, $\tilde{\Sigma}_{jj} \sim \mathcal{N}(0,1)$. The entries of Σ are drawn once for each of the 100 data sets and then kept fixed within each data set. To simulate unbalanced data sets, each of the 300 individuals n within each data set gets assigned a random number of choice occasions T_n with $T_n \sim \mathcal{P}_{2 \le k \le 20}(\lambda = 10)$, where $\mathcal{P}_{2 \le k \le 20}$ denotes a truncated Poisson distribution.

Estimation

We used a version of the MACML estimation procedure with a Solow-Joe (SJ) approximation [Joe, 1995; Solow, 1990] of the MVNCDFs [as proposed by Bhat, 2011] to evaluate the pairwise marginal likelihoods and specified the estimated model according to the DGP. For identification, we fixed $\hat{\beta}_{0,1} = 0$, $\hat{\beta}_{2,1} = 0$, and $\hat{\Sigma}_{1,1} = 1$ at the true values, leaving $\theta = (\beta_{0,2}, \ldots, \beta_{0,6}, \mu_1, \omega_1, \beta_{2,2}, \ldots, \beta_{2,6}, \Sigma_{22}, \ldots, \Sigma_{66})'$ with $3(6-1)+2 = 17$ free parameters to estimate.

Details on the code used for the estimation can be found in 4.4.

The proposed two-step optimal group-weight CML estimator was estimated according to Algorithm 4.1 with $\mathring{W} = 1$. The optimal weights and corresponding estimates will be denoted with the abbreviation BB.

To calculate the weights as proposed in Algorithm 4.2, we used the parametric model

$$1/g(s) = \gamma_0 + \gamma_1 s + \gamma_2 s^2 + \nu_s, \tag{4.24}$$

with ν_s denoting the random error of the model, and estimated the restricted least squares estimator with the restrictions

$$\gamma_0 + \gamma_1 s + \gamma_2 s^2 \geq \min(\hat{v}_s / C_s(\mathring{W})), \quad \gamma_1 + 2\gamma_2 s \geq 0, \tag{4.25}$$

ensuring positive and in s monotonically decreasing weights with

$$\breve{w}_s = (\hat{\gamma}_0 + \hat{\gamma}_1 s + \hat{\gamma}_2 s^2)^{-1}, \tag{4.26}$$

which has the same form as the optimal weights calculated by Joe and Lee [2009]. The estimation was done in R using the COBYLA algorithm as implemented in the *nloptr* package [see Johnson, 2022; Powell, 1994]. The optimal weights and the corresponding estimates calculated according to Algorithm 4.2, with the parametric model for the weights as described above in equations (4.24)-(4.26), will be abbreviated with BB_param.

For comparison, another weighted model was estimated using the heuristic weights suggested by Joe and Lee [2009] with $\bar{w}_s = (s-1)^{-1}[1 + 0.5(s-1)]^{-1}$, which will be abbreviated with JL_0.5.

These simulations of the data and estimations of the models were repeated 100 times, of which 77 times all five models were successfully estimated. Successfully estimated means, in this context, that the `nlm()` function from the R package `stats` minimising the CML had as exit code either 1 ("relative gradient is close to zero, current iterate is probably solution.") or 2 ("successive iterates within tolerance, current iterate is probably solution."). In the remaining cases, optimisation ended with

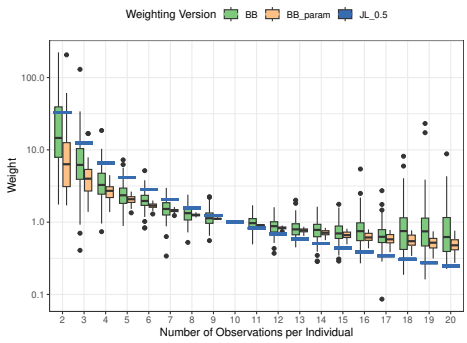


Figure 4.1: Boxplots of the distribution of calculated optimal weights for 100 simulated data sets compared to heuristic weights according to Joe and Lee [2009], represented by horizontal lines. Weights are scaled such that individuals with 10 observations get a unit weight. The y-axis has a logarithmic scale.

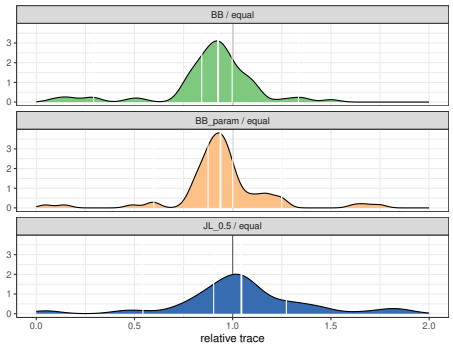


Figure 4.2: Distribution of relative trace of variance-covariance matrix $\hat{V}_\theta$ of the weighted models compared to the unweighted model from 77 successful simulations. Vertical breaks in the filling indicate 5%, 25%, 50%, 75%, and 95% percentiles. The x-axis is scaled to $[0, 2]$ for better visibility. 5 of the JL_0.5 weighted and 2 of the BB weighted, as well as the 95% percentile of the JL_0.5 version, are out of scale.

code 5 ("maximum step size `stepmax` exceeded five consecutive times. Either the function is unbounded below, becomes asymptotic to a finite value from above in some direction or `stepmax` is too small."), implying that the numerical optimisation algorithm did not successfully converge to a minimum [for details, see R Core Team, 2021]. Interestingly, the `nlm()` function was in this study more frequently successful in locating a maximum when using the proposed weighted BB and BB_param procedures, compared to the equally weighted or JL_0.5 versions. Table 4.2 in the appendix provides an overview of the reported `nlm()` exit codes by weighting scheme. To ensure a fair assessment of the weighting methods, the 23 simulations in which at least one method did not lead to an exit code of 1 or 2 were excluded from the subsequent analysis.

An overview of the resulting weights calculated for the successfully estimated models can be seen in Figure 4.1.

Results

The following results are calculated for the 77 simulations, in which all five estimators were successfully estimated. Using the optimal BB weighted estimator, as described in Algorithm 4.1, we can reduce the trace of the covariance matrix V_θ on average by 4.38% over the 77 simulations compared to the unweighted model. In the median over the 77 different simulations, the reduction is, however, by 7.42%, and the BB weighted model has in 58 of the 77 cases a lower trace of the covariance matrix $\hat{V}_\theta$ than the unweighted estimator.

Utilising a parametric function to estimate the weights BB_param, as described in Algorithm 4.2, leads to more stable weights compared to the BB version, as showcased by the distribution of the weights in Figure 4.1. Using this estimator leads to a reduction of the trace of the covariance matrix $\hat{V}_\theta$ on average by 5.25%, in the median by 6.2% and leads to a smaller trace in 72.7% of the cases (56 out of 77), all compared to the trace of the covariance matrix of the equally weighted estimator.

In comparison, the estimator with the JL_0.5 weights has, on average, in the 77 simulations, a 23.02% larger trace of the covariance matrix $\hat{V}_\theta$ compared to the unweighted estimator, in the median a 4.4% larger trace and has in just 41.56% of the cases (32 out of 77) a lower trace than the covariance matrix of the unweighted estimator.

In Figure 4.2, the distribution of the quotient between the traces of the covariance matrix of the weighted estimator and that of the unweighted estimator is shown for the different simulated models. An overview of the distributions over the different simulated data sets of the quotients between the traces of differently weighted estimators can be seen in Table 4.3 in the appendix.

Although the optimal weights, as described in Algorithm 4.1, are asymptotically independent of the group weights for the initial weights $\mathring{W}$ (for constant relative weight of the pairs within a group, that is), this may not be the case in finite samples. In addition to the optimal group weights according to Algorithm 4.1 with a uniformly weighted initial model with $\mathring{W} = 1$, a second set of optimal group weights with $\mathring{W}$ set to the deterministic group weights JL_0.5 according to Joe and Lee [2009] was computed, which will be denoted BB_fJL. The average ratios between the resulting weights and the ratios between the traces of the variance-covariance matrices were calculated and an overview of the distributions can be seen in Tables 4.4 (comparing weights) and 4.5 (comparing traces of the variance matrix) in the appendix. Whilst this example demonstrates that the optimal group weights do depend on initial group weights, the effects are in most cases small.

Beside comparing the trace of the variances one can also measure the precision of the estimated choice probabilities. For this we use the mean squared l_2 distance between the predicted choice probabilities given the model and the choice probabil-

ities according to the true DGP

$$\bar{l}_2^2(p_0, \hat{p}(W)) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\sum_{t=1}^{T_n}\sum_{j=1}^{J}\left(\mathbb{P}\left(y_{n,t}=j;\theta_0\right)-\mathbb{P}(y_{n,t}=j;\hat{\theta}(W))\right)^2, \quad (4.27)$$

depending on the weighting schemes W. We calculate the percentual change compared to the measure for the equally weighted model with $W_{\text{equal}} = 1$ as

$$\Delta_{\%}\bar{l}_2^2(W) = \frac{\bar{l}_2^2(p_0, \hat{p}(W_{\text{equal}})) - \bar{l}_2^2(p_0, \hat{p}(W))}{\bar{l}_2^2(p_0, \hat{p}(W_{\text{equal}}))}. \quad (4.28)$$

An overview of the distribution of the results can be see in Table 4.6 in the appendix. These indicate that using weights designed to reduce the variance of the estimator also results in an improvement in recovering choice probabilities.

Note, however, that this measure does not account for the panel nature of the data.

4.3 Weighting in case of unaccounted autoregressive error structure

In the previous section, we focused on the relative weighting between observations from deciders with a different number of choice occasions. The weights, however, are also relevant for the relative importance of the different pairs for one decider.

In some situations, it is plausible that the error process follows a continuous time stationary process with nonzero autocorrelation function across time. Such behaviour can be interpreted as taste perseverance if it generates positive correlations.

Typically, for stationary processes, the autocorrelation will decrease with increasing temporal distance between the two observations [compare, for example, Lütkepohl, 2005]. An autoregressive process of order 1 (AR(1)), for example, is given as the stationary solution to the difference equation $\varepsilon_t = \rho\varepsilon_{t-1} + \tilde{\varepsilon}_t$ for iid process $\tilde{\varepsilon}_t, t = 1, 2, ...$, where $|\rho| < 1$ is assumed (the stability assumption). The correlation between observations at time 1 and at time t is then given as ρ^{t-1} and, hence, decreases with increasing distance t.

The MNP models with random effects, as discussed in the previous section, however, introduce correlation between observations by including a random constant in the error term of the utility: Random effects in the alternative specific constant (ASC), for example, introduce autocorrelation for the various error terms of one individual

[see, for example, Train, 2009]. These error terms then also form a stationary sequence with random mean vector and constant correlation for nonzero temporal differences, which consequently does not decrease to zero for $t \to \infty$.

This difference in behaviour has consequences in situations where the underlying DGP is not known and a misspecified mixed MNP model is used, neglecting the temporal correlation. Misspecification of the model typically leads to inconsistent estimators. The degree of misspecification depends on the strength of the correlation, which differs between different pairs of observations.

Below, we investigate how the weights for different pairs can be used to reduce the influence of the aforementioned type of misspecification on the estimation of the model. This will first be done in a simplified example where asymptotic properties of the estimators are illustrated, showing the magnitude of asymptotic biases that can occur in conjunction with proposals to limit the corresponding inconsistency. Subsequently, a simulation exercise will show that similar effects can also be observed in finite samples.

4.3.1 Variance-covariance structure of autocorrelated errors

In this section, we deal with stationary vector processes of the autoregressive type for a given decider n. We will always assume independence and identical distribution across different deciders, and hence consider the properties of the error process only over time for a representative decider n. As we assume a balanced panel data set with independent and identically distributed regressors $x_{n,t}$ and errors $\varepsilon_{n,t}$ over different deciders n, we omit the index n in the remainder of this subsection, as well as in Section 4.3.2.

Here, to fix notation, the error terms $\varepsilon_t = (\varepsilon_{t,1}, \ldots, \varepsilon_{t,J})'$, $t = 1, ..., T$, constitute a stationary vector autoregressive process of order 1 (VAR(1)) if they are a stationary solution to the difference equation

$$\varepsilon_t = \Psi \varepsilon_{t-1} + \tilde{\varepsilon}_t \tag{4.29}$$

where $\tilde{\varepsilon}_t = (\tilde{\varepsilon}_{t,1}, \ldots, \tilde{\varepsilon}_{t,J})' \sim \mathcal{N}_J(0, \tilde{\Sigma})$ is iid over time $t = 1, \ldots, T$ [see, for example, Lütkepohl, 2005].[4]

In this situation, the stationary distribution is a normal distribution with expectation zero and variance Σ solving the Lyapunov equation $\Sigma = \Psi \Sigma \Psi' + \tilde{\Sigma}$ if and only if all eigenvalues of the matrix $\Psi \in \mathbb{R}^{J \times J}$ are inside the unit circle (stability condition).

For a pair of choice observations $p = (a, b)'$ at time points $t_a < t_b$, we can now

[4] The distribution of the error is not important for the stationarity properties, but due to the consideration of probit models.

express the covariance matrix Σ_p of the joint process $(\varepsilon'_{t_a}, \varepsilon'_{t_b})'$ as

$$\Sigma_p = \begin{pmatrix} \Sigma & \Sigma(\Psi^{t_b-t_a})' \\ \Psi^{t_b-t_a}\Sigma & \Sigma \end{pmatrix}. \tag{4.30}$$

In case of no autocorrelation, where $\Psi = 0$, we have $\Sigma_p = \begin{pmatrix} \tilde{\Sigma} & 0 \\ 0 & \tilde{\Sigma} \end{pmatrix}$. In case of a simple autocorrelation structure, where $\Psi = \rho I_J$, we have the simplified form

$$\Sigma_p = \frac{1}{1-\rho^2} \begin{pmatrix} \tilde{\Sigma} & \rho^{t_b-t_a}\tilde{\Sigma} \\ \rho^{t_b-t_a}\tilde{\Sigma} & \tilde{\Sigma} \end{pmatrix}. \tag{4.31}$$

Under the stability assumption $\Psi^k \to 0$ for $k \to \infty$, the covariance $\Psi^{t_b-t_a}\Sigma$ of the errors decreases with increasing temporal distance. This implies that, for distant pairs of observations, the misspecification of the autocorrelation of the error terms (for example by assuming zero correlation) is of less importance.

A DGP with autocorrelated errors and no random effects thus possesses a covariance between two observations equal to Σ_p.

Similarly, a model with randomly mixed ASCs with variance-covariance matrix Ω and iid error terms has the variance matrix

$$\Sigma_p = \begin{pmatrix} \tilde{\Sigma} + \Omega & \Omega \\ \Omega & \tilde{\Sigma} + \Omega \end{pmatrix} \tag{4.32}$$

for the joint vector of the error terms.

These two matrices coincide when $\Psi^{t_b-t_a}\Sigma = \Omega$ and then $\tilde{\Sigma} = \Sigma - \Omega$. For a given DGP, this is true, for example, for the following special cases:

1. Ψ is such that $\Psi^{t_b-t_a}\Sigma = \Omega$ for some $t_b - t_a$. In this case, symmetry must hold, such that $\Psi^{t_b-t_a}\Sigma = \Sigma(\Psi^{t_b-t_a})'$. This holds, for example, for $\Psi = \rho I_J$. Clearly, stability implies that this may hold for at most one value of $t_b - t_a$ for nonzero Ψ.

2. $\Psi^{t_b-t_a} = 0$ and, hence, $\Omega = 0$.

It follows that the two models lead to different forms of correlation over time. The expressions can only be identical for one value of $t_b - t_a$. Since $\Psi^k \to 0$ for $k \to \infty$, the second case approximately holds for large temporal distances.

In all other cases, misspecifying the model by mistakenly using randomly mixed ASCs when the data generating process uses a stationary but correlated error process will lead to inconsistent estimators.

In the next subsection (Section 4.3.2), we investigate the corresponding asymptotic bias in a special case, whereas, in the following subsection (Section 4.3.3), we examine the finite sample properties in a simulation study.

4.3.2 Deriving the asymptotic bias in misspecified cases

As a starting point, consider the following simple setup: The decision problem involves a number of consecutive binary decisions. This allows us to consider only the difference between the utilities of the two alternatives. Thus, the utility of the first alternative is zero, whilst the utility for the second alternative is assumed to equal $y^*_{t,2} = \beta_1 + x_t + v_t$, with $v_t = \gamma + u_t, u_t = \rho u_{t-1} + \varepsilon_t$, and $\varepsilon_t \sim \mathcal{N}(0, \tilde{\sigma}^2)$. This implies that there exists one alternative specific constant for alternative 2 (the one for alternative 1 is set to zero to fix the level), which contains a random individual specific part γ (with expectation equal to zero and variance ω^2) included in the error term. Additionally, we include a second regressor $x_{n,t}$ drawn iid standard normally distributed over both individuals and choice situations with a corresponding coefficient normalised to 1 (to fix the scale). The error term u_t follows an AR(1) process with autocorrelation coefficient $|\rho| < 1$.

The specification encompasses both the data generating process as well as the model:

- For the DGP, the data are considered to be generated with $\omega = 0$, such that there are no random effects, but $\rho \in (-1, 1)$, such that the variance of u_t equals $\sigma^2 = 1$, and the correlation between u_t and u_{t-k} equals $\rho^{|k|}$.

- For the model, we assume that the temporal dependence is neglected, and thus $\rho = 0$ is used. Instead, a random effect in the ASC is postulated. Consequently, three parameters are estimated: β_1, ω^2, and σ^2.

In this setting, we investigate the asymptotic bias of the misspecified estimator, assuming random effects but ignoring the temporal correlation structure.

The calculation of the limiting estimator is achieved using the following setting: We calculate the binary choices observed at days $t \in \mathcal{T} := \{1, 2, 3, 366, 367, 368\}$ corresponding to two waves (in adjacent years) of observations for three days each. The regressors x_t are drawn from an underlying finite set of vectors (we use $M = 100$ in our calculations). For a given vector $x_\cdot = [x_t]_{t\in\mathcal{T}}$, we then calculate the choice probabilities for all possible combinations of choices $[y_t]_{t\in\mathcal{T}}$ according to the DGP. These are obtained using the combined random utility vector ($\iota = [1, \ldots, 1]'$)

$$U_{:} = \beta_{1,\circ}\iota + x_{:} + v_{:}, \quad \text{with } \mathrm{E}(v_{:}) = 0, \tag{4.33}$$

$$\mathrm{Var}(v_{:}) = \sigma_{\circ}^2 \left(\begin{array}{ccc|ccc} 1 & \rho_\circ & \rho_\circ^2 & \rho_\circ^{365} & \rho_\circ^{366} & \rho_\circ^{367} \\ \rho_\circ & 1 & \rho_\circ & \rho_\circ^{364} & \rho_\circ^{365} & \rho_\circ^{366} \\ \rho_\circ^2 & \rho_\circ & 1 & \rho_\circ^{363} & \rho_\circ^{364} & \rho_\circ^{365} \\ \hline \rho_\circ^{365} & \rho_\circ^{364} & \rho_\circ^{363} & 1 & \rho_\circ & \rho_\circ^2 \\ \rho_\circ^{366} & \rho_\circ^{365} & \rho_\circ^{364} & \rho_\circ & 1 & \rho_\circ \\ \rho_\circ^{367} & \rho_\circ^{366} & \rho_\circ^{365} & \rho_\circ^2 & \rho_\circ & 1 \end{array}\right). \tag{4.34}$$

Note here that, for the off-diagonal blocks, we have $|\rho_\circ^k| \leq |\rho_\circ^{363}| \approx 0$ for practically all values of $|\rho_\circ| < 1$. Using these variances, we calculate $\mathbb{P}(y_{:}; x_{:}, \beta_{1,\circ}, \sigma_\circ^2, \rho_\circ)$, the choice probabilities for all possible combinations according to the data generating process.

In the estimation, we assume for the model random effects as well as $\rho = 0$, which results in $U_{:} = \beta_1\iota + x_{:} + v_{:}$ with $\mathrm{E}(v_{:}) = 0$, $\mathrm{Var}(v_{:}) = \omega^2\iota\iota' + \sigma^2 I_6$, $\iota = [1, \ldots, 1]'$.

Using this variance, we obtain for every pair of choices (y_a, y_b) the model choice probabilities $\mathbb{P}(y_a, y_b; x_{:}, \beta_1, \omega^2, \sigma^2)$. Clearly, this is different from the variance due to the DGP, and thus the model is misspecified. As the criterion function, we use the pairwise CML with weights $w_{a,b}$ for the pair (y_a, y_b). For a discrete uniform distribution over $x_{:}$ in the set $\{X_i, i = 1, ..., 100\}$ this converges to [5]

$$\begin{aligned} &Q_\circ(\beta_1, \omega^2, \sigma^2) \\ &= \frac{1}{100}\sum_{i=1}^{100}\sum_{y_{:}} \mathbb{P}(y_{:}; X_i, \beta_{1,\circ}, \sigma_\circ^2, \rho_\circ) \left(\sum_{a=1}^{5}\sum_{b=a+1}^{6} w_{a,b} \log \mathbb{P}(y_a, y_b; X_i, \beta_1, \omega^2, \sigma^2)\right). \end{aligned} \tag{4.35}$$

This limiting asymptotic function is then maximised with respect to the parameters β_1, ω^2, and σ^2 in order to calculate the asymptotic value of the estimators.

[5] The choice of 100 different regressor vectors is arbitrary and only done in order to reduce the impact of the regressors. One way to view this is the approximation of the expectation over continuously uniformly distributed regressors vectors.

In the comparison, we include four different weighting schemes:

FP	The full pairwise CML uses $w_{a,b} \equiv 1$.
growth	The distant pairs, or step growth, CML uses $w_{a,b} = \mathbb{I}(\lvert t_a - t_b \rvert > 7)$. Only pairs measured with more than a week time difference are included.
adj.	The adjacent pairwise CML uses $w_{a,b} = \mathbb{I}(\lvert a - b \rvert = 1)$. Only consecutive observations, irrespective of the distance between them, are used.
decay	The step decay CML uses $w_{a,b} = \mathbb{I}(\lvert t_a - t_b \rvert \leq 7)$. Only pairs of observations less than a week apart are used.

Since the correlation between two pairs is – according to the DGP – given by $\rho^{|t_a - t_b|}$, we see that distant pairs are practically uncorrelated, whilst, for adjacent observations, we obtain a correlation of ρ when $|t_a - t_b| = 1$ and a correlation of almost zero for $|t_a - t_b| = 363$. According to the model, however, independent of the temporal distance, we obtain a variance of $\omega^2 + \sigma^2$ and a covariance of all pairs of ω^2, leading to a correlation of $\omega^2/(\omega^2 + \sigma^2)$.

For the calculations, we use $\beta_{1,\circ} = 1, \sigma_\circ^2 = 1$ and vary $\rho_\circ \in (-1, 1)$. It follows that, in order to mimic the correlation and the variance according to the DGP, for adjacent pairs we must have $\omega^2 + \sigma^2 = 1$ and $\omega^2 = \rho_\circ$, if $\rho_\circ \geq 0$. For negative $\rho_\circ$, the closest ω^2 equals $\omega^2 = 0$.

For distant pairs with one year between observations, the true correlation amounts to almost zero, whilst the model still implies a correlation of $\omega^2/(\omega^2 + \sigma^2)$. Thus, the fit is perfect if $\omega^2 = 0$ and $\sigma^2 = 1$.

Therefore, we expect that the adjacent weighting results in an asymptotic bias for positive values of $\rho_\circ$, where the bias gets larger with larger values of $\rho_\circ$. For the distant weighting approach, we expect that $\omega^2 = 0$ and $\sigma^2 = 1$ throughout. Whilst the temporal correlation is not estimated correctly, it does not lead to an inconsistent estimation for ω if time points are included such that the temporal correlation has vanished. If only close pairs are considered, however, the estimator for ω^2 compensates to match the correlation.

This behaviour can be seen in the three plots in Figure 4.3.

4.3.3 Finite-sample simulation

In this subsection we demonstrate that the asymptotic effects derived in Section 4.3.2 in a simple case can also be observed in finite-samples in more complex cases. We use a simulation study with autoregressive VAR(1) errors and no mixed effects in the DGP, but with mixed ASCs and no auto-correlated errors in the estimated model.

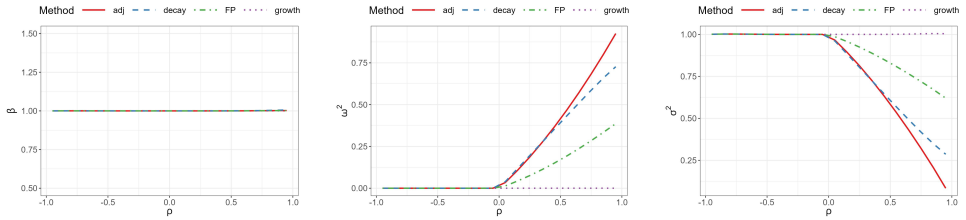


Figure 4.3: Visualisation of the asymptotic bias in β, ω^2, and σ^2, respectively, for the four different CML pair structures.

Simulation setup

For $J = 6$ choice alternatives, we simulated 100 panel data sets with 300 individuals and 10 observations for each individual at time points $t = 1, 2, 3, 4, 5, 366, 367, 368, 369, 370$, resulting in two waves of five observations each in adjacent years, mimicking the study design of the German Mobility panel; see, for example, Zumkeller and Chlond [2009]. For all data sets and individuals, we used the same DGP as described in Section 4.2.2 in equations (4.17)-(4.23), with the modification that $\omega_1 = 0$ instead of $\omega_1 = 1$, such that we have fixed $\beta_{1,n} = 1$.

The error terms $\varepsilon_{n,t} = (\varepsilon_{n,t,1}, \ldots, \varepsilon_{n,t,6})'$ are, however, randomly drawn from a VAR(1) process, as described in equation (4.29), and scaled such that $\text{Var}(\varepsilon_{n,t,1}) = 1$. The errors $\varepsilon_{n,1}$ for the first observation of an individual were drawn from the stationary distribution $\mathcal{N}(0, \Sigma)$. The following errors are calculated as

$$\varepsilon_{n,t+1} = \Psi \varepsilon_{n,t} + \tilde{\varepsilon}_{n,t+1}, \tag{4.36}$$

with $\tilde{\varepsilon}_{n,t+1} \sim \mathcal{N}(0, \tilde{\Sigma})$, $\tilde{\Sigma}$ as a diagonal matrix with entries $\tilde{\Sigma}_{jj} = (1 - \rho^2)|\tilde{\sigma}_{jj}|$, $\tilde{\sigma}_{jj} \sim \mathcal{N}(0, 1)$ for $j > 1$ and $\tilde{\Sigma}_{11} = (1 - \rho^2)$. The entries of $\tilde{\Sigma}$ are drawn once for each of the 100 data sets and are then kept fixed for all individuals within one data set.

Five types of errors were simulated for each of the 100 data sets of regressors, resulting in different observations for the choice variable y, leading to a total of 500 distinct data sets. To simulate the errors, five different parameter matrices Ψ were used, with $\Psi = \rho \mathbb{I}_6$, $\rho \in \{-0.95, -0.2, 0, 0.2, 0.95\}$.

Estimation

We used a version of the MACML estimation procedure with an SJ approximation of the MVNCDFs to evaluate the pairwise marginal likelihoods and used a misspecified

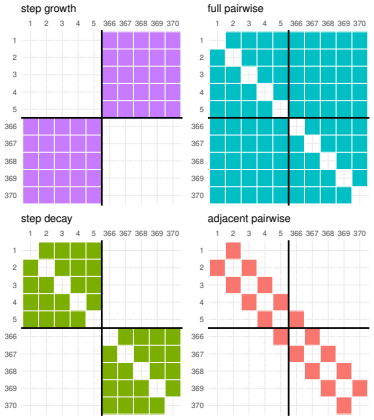


Figure 4.4: Visualisation of the four different CML pair structures applied for the estimation process.

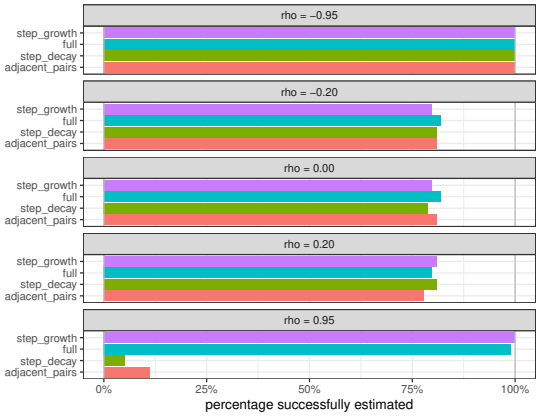


Figure 4.5: Percentage of successfully estimated models by autocorrelation coefficient ρ and CML pair-type.

model, where instead of using autocorrelated errors, we assumed iid errors over individuals and observations. To introduce correlation between the observations, the estimated model assumes mixed effects for the ASCs; thus, $\beta_{0,j} \sim \mathcal{N}(\mu_{0,j}, \omega^2_{0,j})$. For identification, we fixed $\hat{\mu}_{0,1} = 0$, $\hat{\beta}_{2,1} = 0$, and $\hat{\Sigma}_{1,1} = 1$, leaving $\theta = (\mu_{0,2}, \ldots, \mu_{0,6}, \omega_{0,1}, \ldots, \omega_{0,6}, \beta_1, \beta_{2,2}, \ldots, \beta_{2,6}, \Sigma_{22}, \ldots, \Sigma_{66})'$ with $3(6-1)+6+1 = 22$ free parameters to estimate.

The model was then estimated using four different CML pair-types, as described in Section 4.3.2: *step growth*, *full pairwise*, *step decay*, and *adjacent pairwise*. A visualisation of the different pair-types is shown in Figure 4.4.

For the MACML estimations the same setup was used as above (see 4.4).

Results

Possibly owing to the misspecification of the estimated model, our procedures did not converge to a minimum in some cases. This means, in this case, that the `nlm()` function minimising the CML had as exit code neither 1 (“relative gradient is close to zero, current iterate is probably solution."), nor 2 (“successive iterates within tolerance, current iterate is probably solution."). Figure 4.5 visualises the relative success rates of the four different models with different pair-types for the five different autocorrelation structures of the errors.

These results indicate that, in case of large positive autocorrelation coefficients in

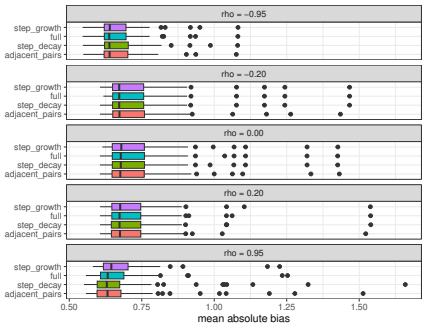


Figure 4.6: Distribution of the mean absolute deviation of the estimated β parameters from the true parameters over all estimated models by autocorrelation coefficient ρ and CML pair-type.

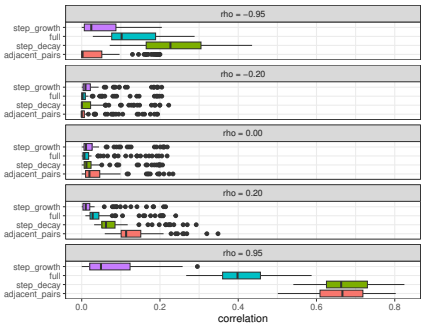


Figure 4.7: Distribution of mean estimated correlation $\tilde{\rho}$ over all estimated models by autocorrelation coefficient ρ of the true autoregressive error process and CML pair-type.

the unaccounted autoregressive process of the error, the inclusion of distant pairs in the CML increases the rate of successfully estimated models substantially.

To evaluate the accuracy of the estimated model further, we calculated the mean absolute deviation of the estimated parameters from the true parameters for each model. To avoid bias in the comparison due to problems with the normalisation implied by fixing the scale via $\Sigma_{1,1} = 1$, we decided to re-scale the estimated models to minimise the sum of squared deviations of the estimated parameters, including the deviation introduced in Σ_{11}, and combine the parameters of Σ and Ω to account for their joint effect in the model. Therefore, we effectively examined the distance within the equivalence class of systems differing only in the scale of the utility.

When looking at the mean absolute deviation of the estimated β parameters of the successfully estimated models, we do not see a significant difference in the deviations of the parameters across the different CML pair-types used for the estimation. This aligns with the asymptotic results presented in Section 4.3.2. An overview of the results is shown in Figure 4.6.

4.3.4 Discussion

When looking at the estimated correlation between the different observations of one individual – in the estimated model introduced by random effects in the ASCs, in the DGP introduced by autocorrelation of the errors – we observe in the finite-sample simulations results similar to those shown in the analysis of the asymptotic behaviour in Section 4.3.2. This can be observed in Figure 4.7, which summarises the

distribution of the estimated average correlation $\tilde{\rho} = (6-1)^{-1} \sum_{j=2}^{6} \tilde{\Omega}_{jj} / (\tilde{\Omega}_{jj} + \tilde{\Sigma}_{jj})$.

The estimated correlation between the observations of one individual is due to the mixed ASCs and would, in the standard interpretation, be attributed to individual taste heterogeneity. In the estimated models, this correlation is the same for any pair of observations of one individual, irrespective of the temporal distance between the observations. In the DGP, however, the correlation was introduced due to an autoregressive process of the error terms; hence, the correlation between different observations of one individual decreases as a function of their temporal distance.

In the case of positive autocorrelation of the errors, the models using `full`, `step_decay`, or `adjacent_pairs` CML pair-types in the estimation process estimate significantly larger average correlations between the observations compared to the models using `step_growth` CML pair-types.

In the case of no or small negative autocorrelation, there was no significant difference between the different pair-type structures used in the estimation.

In the case of large negative autocorrelation, both the `full` and `step_decay` models estimate a relatively large positive autocorrelation between the observations. This could be attributed to two factors. First, the models are – due to their structure – incapable of estimating any negative correlation between observations of one individual since $\tilde{\Omega}_{jj} / (\tilde{\Omega}_{jj} + \tilde{\Sigma}_{jj}) \geq 0$. Second, the observations with an even number $2k$ of steps between them have a positive correlation $\rho^{2k} \geq 0$, even for negative autocorrelation of the errors. This is then captured by models that include such pairs in their CML pair structure. The `adjacent_pairs` model has only pairs with temporal distances equal to one or equal to 361 in the CML pair structure, where $\rho < 0$ for $\rho^{361} < 0$; hence, there are no pairs with positive correlation in the estimation process. For this reason, a model with `adjacent_pairs` estimates only a very small correlation between observations in the case of negative autocorrelation of the errors, similar to the `step_growth` models.

Overall the simulation study has shown that when the errors follow an autoregressive process, which is not accounted for in the model, the inclusion of distant observations leads to a reduction of the deviation of the estimated correlation coefficients and makes the estimation process more reliable in terms of estimation success rate. In case of a large positive autocorrelation coefficient in the autoregressive error process, the estimators only relying on distant pairs performed by far the best. Since the direction of the autocorrelation is, however, usually not a priori known, it is, in any case, worth considering using distant observations in the CML when an autoregressive error process in the DGP cannot be ruled out but is also not explicitly modelled.

Note that these findings are due to the misspecification of the model not including temporal autocorrelation in the model, which hence is picked up partially by

the random error term. The situation would change if the correct model, including temporal autocorrelation of the error term, were to be used. In that case, Varin and Vidoni [2006] show that pairs with large temporal distances contain less information on the autocorrelation parameter. In the correctly specified case, this is a disadvantage. In the misspecified case, however, it is an advantage.

4.4 Conclusion

In this paper, we have shown how different weighting strategies for pairwise CML estimation can be used to effectively improve the properties of the estimator in case of (1) unbalanced panel data, and (2) unaccounted autocorrelation of the errors.

In the case of unbalanced panel data, we introduced a new two-step optimal group-weight CML estimator, which exploits the panel structure by grouping individuals with the same number of observations together to assign group weights to effectively reduce the variance of the estimator. The estimator is based on asymptotic theory and has shown to be effective in a finite sample simulation study in which the two-step optimal group-weight CML estimator improved upon the unweighted pairwise CML estimator in 75.3% of the cases when measured in the trace of the variance-covariance matrix of the estimator. Utilising a parametric model to estimate the weights, as described in Algorithm 4.2, leads in the median to less of a reduction in the trace of V_θ than using the optimal weights directly. However, in the rare cases when the estimated optimal weights lead to an increase of the trace, this effect is less pronounced when using a parametric model for the weights. As a consequence the estimator described in Algorithm 4.2 has, on average, the lowest trace of the variance-covariance matrix compared to the unweighted estimator out of all tested weighted estimators. Using a parametric model to estimate the optimal weights, hence, seems to lead to more stable results and reduces the dependence on initial group weights $\mathring{W}$.

Although the reduction in estimator variance accomplished by utilizing the proposed optimal group weights is less pronounced in this simulation study than the 50% reduction demonstrated in 4.4, the gains attained are still noteworthy, and the extra computational cost is minor in comparison to the possible benefits.

In comparison, using the weights by Joe and Lee [2009] resulted in most of the cases in larger traces of the variance matrix than the usage of other weights, including the initial weights $w_s = 1$. It resulted, however, in one case in the largest overall improvement compared to the unweighted version. A detailed overview of the results can be seen in Table 4.3.

In the case of an unaccounted autoregressive process of the error terms, we were able to reduce the effect of the misspecification of the model by choosing distant observations in the CML pair structure. This was shown in an asymptotic calculation, as well as in a finite sample simulation study. Using distant pairs reduces the effects of autocorrelation of the errors, which decreases over time and which potentially introduces an asymptotic bias in the estimation of the mixed effects, resulting in an inconsistent estimator. This effect is showcased in examples where the data represents a panel data set with two panel waves, in which the differences in covariances between observation pairs is especially apparent. It can be argued that, to distinguish between the covariance induced by individual specific effects and that induced by an autoregressive error process, the inclusion of distant pairs is most effective in case of data with distinct panel waves, such that the known data structure can be used in the estimation process.

Furthermore, the simulation study indicated that the inclusion of distant pairs in the CML results in a substantially higher rate of successfully estimated models when dealing with an unaccounted autoregressive error process with large positive autocorrelation coefficient.

Both results can also be combined using group-specific weights to account for unbalanced panels and stressing distant pairs of observations for a given number of choice occasions to robustify the estimation of the random effects. The latter weighting strategy can be used in order to detect temporal autocorrelation in the error terms, whilst the group-specific weights lead to improved accuracy at almost no numerical costs for the estimation.

CRediT author statement

Sebastian Büscher: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing. **Dietmar Bauer:** Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing - review & editing

Acknowledgement

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 356500581 which is gratefully acknowledged. The authors also thank Manuel Batram and Lennart Oelschläger who contributed to the codebase used for the calculations.

Declarations of interest

Declarations of interest: none
The funding agency Deutsche Forschungsgemeinschaft (DFG) had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Appendix 4.A

In this appendix, we construct an example in which weighting reduces the variance of the estimator by up to 50%. The main idea here is to choose a setting in which pairs of choice decisions are almost perfectly correlated. If this holds for the scores of the pairwise choice probabilities, then for a decider facing s choice occasions, the gradient of the full pairwise CML will be equal to the gradient for one pair times the number $C_s = s(s-1)/2$ of pairs. The Hessian will also be equal to C_s times the Hessian for one pair. Therefore the variance of the score will be equal to $C_s^2 V_1$ (V_1 denoting the variance of the gradient for one pair) and the Hessian equal to $C_s H_1$ (H_1 denoting the Hessian for one pair).

When combining an equal number of deciders with two choice occasions (hence only one pair of choices) and deciders with s choices (where the contribution of these deciders is weighted with a scalar w), we obtain the variance formula

$$V = \frac{1 + w^2 C_s^2}{(1 + w C_s)^2} H_1^{-1} V_1 H_1^{-1}. \tag{4.37}$$

Without weighting, corresponding to $w = 1$, we obtain a fraction of $(1 + C_s^2)/(1 + C_s)^2$, which tends to 1 for $s \to \infty$ and hence $C_s \to \infty$. Introducing

the weight $w = 1/C_s$, however, leads to a fraction $(1+1^2)/(1+1)^2 = 2/4 = 1/2$. This demonstrates that, for maximal positive correlation, we achieve a variance reduction of 50%.

A situation with a maximal positive correlation is achieved, for example, for a binary decision with no regressors but an ASC for the second choice. The ASC is modelled as $\beta + \gamma, \gamma \sim \mathcal{N}(0, \omega^2)$, where $\beta = 5$ is fixed, and $\omega = 5$ is estimated. Furthermore, we assume that $\varepsilon_{n,t,1} \sim \mathcal{N}(0, 0.01)$. In this model ω is the only parameter estimated.

The noise is negligible compared to the random effect. The choices of a decider almost exclusively are decided via the random ASC. This leads to a model where the random utilities for each decider are almost perfectly positively correlated across choice occasions. It is easy to see that this implies the same for the corresponding gradients.

We computed the ratio of the asymptotic variance of ω with the optimal weighting to the one for the unweighted case for $s = 3, ..., 8$. Table 4.1 shows that the relative variance decreases by almost 50% by using the optimal weights.

s	3	4	5	6	7	8
variance ratio	0.80	0.66	0.60	0.57	0.55	0.54

Table 4.1: Ratio of the asymptotic variance for optimal weights versus unweighted case.

Appendix 4.B

In this appendix, we proof the theorems from Section 4.2. For ease of reading, the theorems are repeated before the respective proofs.

Theorem 4.1 (Asymptotic Variance)**.** *Let the data be generated according to Assumption 4.2 with parameter vector θ_0 and let $\hat{\theta}$ be the CML estimator maximising the weighted CML function* (4.4) *using the weights $w_{n,a,b} = w_{T_n}\mathring{w}_{n,a,b}$, where the initial weights $\mathring{w}_{n,a,b}$ adhere to Assumption 4.1, where $C_s(\mathring{W}) > 0, s = 2, ..., \overline{S}$.*

Further, let $w_s \geq 0$ denote group-specific weights according to the number of observations $T_n = s$, $s \in \{2, \ldots, \overline{S}\}$ of decider n, such that $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$. Then the following hold:

(I) $\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_\theta(W_s))$, *where the asymptotic variance-covariance matrix*

$V_\theta(W_s)$ for a given vector $W_s = (w_2, \ldots, w_{\overline{S}})'$ of group-specific weights has the form

$$V_\theta(W_s) = \sum_{s=2}^{\overline{S}} f_s w_s^2 H_0^{-1} V_s H_0^{-1}, \tag{4.11}$$

with

$$V_s = \mathbb{E}\; g_n(\mathring{W}) g_n(\mathring{W})', \quad H_0 = \mathbb{E}\; \partial_\theta^2 \log \mathbb{P}(y_{n,1}, y_{n,2}, X_n; \theta_0). \tag{4.12}$$

(II) H_0 can be estimated consistently as

$$\hat{H}_0 = \frac{1}{\sum_{n,a,b} \mathring{w}_{n,a,b}} \sum_{n,a,b} \mathring{w}_{n,a,b} \partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}). \tag{4.13}$$

(III) If $f_s > 0$ and N_s denotes the number of deciders facing s choice occasions, then V_s can be estimated consistently using

$$\hat{V}_s = {N_s}^{-1} \sum_{n:T_n=s} \hat{g}_n(\mathring{W}) \hat{g}_n(\mathring{W})', \tag{4.14}$$

where $\hat{g}_n(\mathring{W}) := \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}) \right)$.

Proof of Theorem 4.1.

(I) The average Hessian $\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W_s))/N$ takes the form

$$\begin{aligned} &\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W_s))/N \\ &\qquad = N^{-1} \sum_{n=1}^{N} \left(\sum_{b=a+1}^{T_n} \sum_{a=1}^{T_n - 1} w_{n,a,b} \partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \bar{\theta}) \right). \end{aligned} \tag{4.38}$$

Hereby, $\mathbb{E}\left(\partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0)\right) = H_0$ is independent of a and b if the regressors are drawn from an identical marginal distribution. It follows that

$$\begin{aligned} &\mathbb{E}\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W_s)/N \\ &\qquad = \sum_{s=2}^{\overline{S}} N^{-1} \left(\sum_{n:T_n=s} \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} w_{n,a,b} \right) \right) H_0 \to H(W, f_s), \end{aligned} \tag{4.39}$$

where $f_s = \lim N_s/N, s = 2, ..., \overline{S}$ denotes the relative frequency of deciders

with s choice situations (with a maximum of $\overline{S}$ and a minimum of 2), and $N_s, s = 2, ..., \overline{S}$ the number of individuals with s choice situations.

Use $w_{n,a,b} = w_s \mathring{w}_{n,a,b}$ as combined weights, and $C_{N,s}(W) = {N_s}^{-1} \sum_{n:T_n=s} \sum_{a=1}^{s-1} \sum_{b=a+1}^{s} w_{n,a,b}$ as the average sum of weights for a pair of observations from an individual with s choice occasions. Then $C_{N,s}(W) \to w_s C_s(\mathring{W})$. This limit for the weights $\mathring{w}_{n,a,b}$ obviously does not depend on n. Otherwise, independent sampling of the stratification weights subject to lower and upper bounds shows that $C_{N,s}(\mathring{W}) \to w_s C_s(\mathring{W})$.

Consequently we obtain

$$H(W, f_s) = \sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) H_0 = H_0, \tag{4.40}$$

by the assumption $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$. Furthermore,

$$\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W)/N - \mathbb{E}\partial_\theta^2 ll_{CML}(y, X; \bar{\theta}, W)/N \xrightarrow{p} 0 \tag{4.41}$$

follows from independence over deciders and boundedness of the variance implied by the bound on the weights in combination with bounds on the regressors. The choice probabilities depend differentiably to any degree on the underlying parameters and regressors, implying a uniform bound on all moments involved.

With respect to the score $\partial_\theta ll_{CML}(y, X; \theta_0, W)/\sqrt{N}$, note that it is the sum of independent (assuming independent draws over deciders) scores conditional on the number of choice occasions:

$$g_n(W) := \left(\sum_{a=1}^{T_n - 1} \sum_{b=a+1}^{T_n} w_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) \right), \tag{4.42}$$

$$\begin{aligned} &\sqrt{N} \partial_\theta ll_{CML}(y, X; \theta_0, W) \\ &\qquad = \frac{1}{\sqrt{N}} \sum_{s=2}^{\overline{S}} \sum_{n:T_n=s} \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} w_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) \right) \\ &\qquad = \sum_{s=2}^{\overline{S}} \sqrt{f_s} \left(\frac{1}{\sqrt{f_s N}} \sum_{n:T_n=s} g_n(W) \right). \end{aligned} \tag{4.43}$$

The inner sum is for each value $s = 2, ..., \overline{S}$ the sum over independent terms with expectation zero and a variance $V_s(W)$ depending on the number of choice occasions (for iid draws of the regressors over deciders).

Thus the limiting normal distribution of $\sqrt{N}\partial_\theta ll_{CML}(y, X; \theta_0, W)$ has variance (due to the independence of contributions of different deciders)

$$V(W, f_s) = \sum_{s=2}^{\overline{S}} f_s V_s(W). \tag{4.44}$$

The variance of the estimator then follows the sandwich form

$$V_\theta = H(W, f_s)^{-1} V(W, f_s) H(W, f_s)^{-1} = H_0^{-1} V(W, f_s) H_0^{-1}. \tag{4.45}$$

Moreover, for n such that $T_n = s$, we have

$$\begin{aligned} g_n(W) &= \sum_{a=1}^{s-1} \sum_{b=a+1}^{s} w_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) \\ &= w_s \sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \theta_0) = w_s g_n(\mathring{W}) \end{aligned} \tag{4.46}$$

and, therefore,

$$\begin{aligned} V_s(W) &= \mathbb{E} g_n(W) g_n(W)' = \mathbb{E} w_s g_n(\mathring{W}) w_s g_n(\mathring{W})' \\ &= w_s^2 \mathbb{E} g_n(\mathring{W}) g_n(\mathring{W})' = w_s^2 V_s(\mathring{W}). \end{aligned} \tag{4.47}$$

with $\mathring{W}$ denoting the set of initial weights $\mathring{w}_{n,a,b}$. Then the variance formula simplifies to

$$V_\theta = \sum_{s=2}^{\overline{S}} f_s w_s^2 H_0^{-1} V_s(\mathring{W}) H_0^{-1}. \tag{4.48}$$

(II) Consistency for the estimator $\hat{H}_0$ follows from the consistency of $\hat{\theta} \to \theta_0$, iid sampling over individuals and the differentiability of the criterion function as a function of the parameter vector. The arguments are standard and, hence, omitted.

(III) The same applies for the estimate $\hat{V}_s$. Here consistency requires $N_s \to \infty$.

□

Theorem 4.2 (Optimal Group-Specific Weights)**.** *Let $l : \mathbb{R}^{d\times d} \to \mathbb{R}$ be a linear mapping of the form $l(V) = \operatorname{tr}(VA)$, with $A \in \mathbb{R}^{d\times d}$, $A \neq 0$ symmetric and positive semidefinite, $\hat{\theta}$ be the CML estimator maximising the weighted CML function* (4.4)*, and let the data be generated subject to Assumption 4.2. Let $\mathring{w}_{n,a,b}$ be the initial weights fulfilling Assumption 4.1, where $\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) = 1$.*

Then $l(V_\theta(W_s))$ is minimised over W_s by

$$w_s^* = \left(\sum_{s=2}^{\overline{S}} f_s C_s(\mathring{W})^2 / v_s(\mathring{W}) \right)^{-1} C_s(\mathring{W})/v_s(\mathring{W}) \propto C_s(\mathring{W})/v_s(\mathring{W}), \tag{4.15}$$

with $v_s(\mathring{W}) = l\left(H_0^{-1} V_s(\mathring{W}) H_0^{-1}\right)$.

Proof of Theorem 4.2. Due to the linearity of the function $l(\cdot)$, we obtain

$$l\left(V_\theta(W)\right) = l\left(\sum_{s=2}^{\overline{S}} f_s w_s^2 H_0^{-1} V_s H_0^{-1} \right) = \sum_{s=2}^{\overline{S}} f_s w_s^2 l\left(H_0^{-1} V_s H_0^{-1}\right) = \sum_{s=2}^{\overline{S}} f_s w_s^2 v_s(\mathring{W}). \tag{4.49}$$

It is also ensured that $v_s(\mathring{W}) = l\left(H_0^{-1} V_s(\mathring{W}) H_0^{-1}\right) = \operatorname{tr}\left(H_0^{-1} V_s H_0^{-1} A\right)$ is positive since $H_0^{-1} V_s H_0^{-1}$ is positive definite and A is, by requirement, positive semidefinite.

The Lagrange function for optimising with respect to w_s subject to the restrictions then equals

$$\mathcal{L}(W, \lambda) = \sum_{s=2}^{\overline{S}} f_s w_s^2 v_s(\mathring{W}) - \lambda \left(\sum_{s=2}^{\overline{S}} f_s w_s C_s(\mathring{W}) - 1 \right). \tag{4.50}$$

The first order condition for w_s evaluated at the optimum then reads:

$$2 f_s w_s^* v_s(\mathring{W}) - \lambda f_s C_s(\mathring{W}) = 0 \quad \Longrightarrow \quad w_s^* = \lambda \frac{f_s C_s(\mathring{W})}{2 f_s v_s(\mathring{W})} = \frac{\lambda}{2} \frac{C_s(\mathring{W})}{v_s(\mathring{W})}. \tag{4.51}$$

This implies

$$w_s^* = c C_s(\mathring{W})/v_s(\mathring{W}), \tag{4.52}$$

with the constant $c = \left(\sum_{s=2}^{\overline{S}} f_s C_s(\mathring{W})^2 / v_s(\mathring{W}) \right)^{-1}$. □

Appendix 4.C

In this appendix, algorithms to calculate the different versions of the optimal group weights are shown.

Algorithm 4.1 Algorithm to calculate two-step optimal group-weight CML estimator

Require: initial weights $\mathring{W} \geq 0$, linear functional $l(\cdot) = \mathrm{tr}(\cdot A)$, $A \in \mathbb{R}^{d \times d}$ symmetric and positive semidefinite

1: Set $w_s \leftarrow 1,\ s = 2, \ldots, \overline{S}$
2: Estimate $\hat{\theta}$, minimising $ll_{\mathrm{CML}}(y, X; \theta, \mathring{W})$
3: Estimate $\hat{H}_0 = \frac{1}{\sum_{n,a,b} \mathring{w}_{n,a,b}} \sum_{n,a,b} \mathring{w}_{n,a,b} \partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta})$
4: Calculate $C_{N_s,s}(\mathring{W}) = {N_s}^{-1} \sum_{n:T_n=s} \sum_{b=a+1}^{s} \sum_{a=1}^{s-1} \mathring{w}_{n,a,b},\ s = 2, \ldots, \overline{S}$
5: Estimate $\hat{g}_n(\mathring{W}) = \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}) \right),\ n = 1, \ldots, N$
6: Estimate $\hat{V}_s = {N_s}^{-1} \sum_{n:T_n=s} \hat{g}_n(\mathring{W}) \hat{g}_n(\mathring{W})',\ s = 2, \ldots, \overline{S}$
7: Estimate 'optimal' group-weights $\hat{w}_s^* = C_{N_s,s}(\mathring{W}) / l(\hat{H}_0^{-1} \hat{V}_s \hat{H}_0^{-1}),\ s = 2, \ldots, \overline{S}$
8: Calculate new weights $\hat{W}^*$ with $\hat{w}_{n,a,b}^* = \hat{w}_s^* \mathring{w}_{n,a,b}$ for $T_n = s$
9: Estimate $\tilde{\theta}$, minimising $ll_{\mathrm{CML}}(y, X; \theta, \hat{W}^*)$

Algorithm 4.2 Algorithm to calculate two-step optimal group-weight CML estimator with parametric variation

Require: initial weights $\mathring{W} \geq 0$, linear functional $l(\cdot) = \mathrm{tr}(\cdot A)$, $A \in \mathbb{R}^{d \times d}$ symmetric and positive semidefinite

1: Set $w_s \leftarrow 1,\ s = 2, \ldots, \overline{S}$
2: Estimate $\hat{\theta}$, minimising $ll_{\mathrm{CML}}(y, X; \theta, \mathring{W})$
3: Estimate $\hat{H}_0 = \frac{1}{\sum_{n,a,b} \mathring{w}_{n,a,b}} \sum_{n,a,b} \mathring{w}_{n,a,b} \partial_\theta^2 \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta})$
4: Calculate $C_{N_s,s}(\mathring{W}) = {N_s}^{-1} \sum_{n:T_n=s} \sum_{b=a+1}^{s} \sum_{a=1}^{s-1} \mathring{w}_{n,a,b},\ s = 2, \ldots, \overline{S}$
5: Estimate $\hat{g}_n(\mathring{W}) = \left(\sum_{a=1}^{s-1} \sum_{b=a+1}^{s} \mathring{w}_{n,a,b} \partial_\theta \log \mathbb{P}(y_{n,a}, y_{n,b}, X_n; \hat{\theta}) \right),\ n = 1, \ldots, N$
6: Estimate $\hat{V}_s = {N_s}^{-1} \sum_{n:T_n=s} \hat{g}_n(\mathring{W}) \hat{g}_n(\mathring{W})',\ s = 2, \ldots, \overline{S}$
7: Calculate $\hat{v}_s = l(\hat{H}_0^{-1} \hat{V}_s \hat{H}_0^{-1})$
8: Estimate parametric function $g : \{2, \ldots, \overline{S}\} \to \mathbb{R}_+$ to model $C_{N_s,s}(\mathring{W}) / \hat{v}_s = g(s)$ and calculate $\breve{w}_s^* = g(s)$
9: Calculate new weights $\breve{W}^*$ with $\breve{w}_{n,a,b}^* = \breve{w}_s^* \mathring{w}_{n,a,b}$ for $T_n = s$
10: Estimate $\tilde{\theta}$, minimising $ll_{\mathrm{CML}}(y, X; \theta, \breve{W}^*)$

Appendix 4.D

For the MACML estimations, the computations were done in the statistical computing language R (Version 4.1.2) [R Core Team, 2021] with the CML function calculations written in C++11, integrated in R via the *Rcpp* package by Eddelbuettel and Francois [2011]. The negative CML function is then minimised using the R function `nlm()` with analytic gradients. The variance-covariance matrix $\hat{V}_\theta$ is calculated using the analytic Hessian matrix of the CML function.

The R and C++ code used for the estimation process is bundled into an R-package named *Rprobit* and is available on `https://github.com/dbauer72/Rprobit`.

Appendix 4.E

Weighting Type	code 1	code 2	code 3	code 4	code 5	code ≤ 2
equal	0.77	0.00	0	0	0.23	0.77
BB	0.91	0.08	0	0	0.01	0.99
BB_param	0.85	0.15	0	0	0.00	1.00
JL_0.5	0.80	0.00	0	0	0.20	0.80

Table 4.2: Share of `nlm()` exit codes by employed weighting scheme for 100 unbalanced panel data set simulations.

trace quotient	mean	min	Q_05	Q_25	median	Q_75	Q_95	max	% < 1
BB / equal	0.956	0.090	**0.292**	**0.842**	**0.926**	**0.997**	1.335	4.470	**75.325**
BB / JL_0.5	1.090	0.032	0.348	0.778	0.882	0.975	1.351	13.988	81.818
BB_param / equal	**0.948**	0.046	0.599	0.874	0.938	1.000	**1.250**	**1.749**	72.727
BB_param / JL_0.5	0.909	0.243	0.457	0.806	0.900	0.952	1.135	3.682	83.117
BB_param / BB	1.299	0.060	0.768	0.971	1.020	1.060	1.535	12.250	38.961
JL_0.5 / equal	1.230	**0.039**	0.543	0.903	1.044	1.273	2.217	6.909	41.558

Table 4.3: Overview of distribution of quotients between the traces between differently weighted models. The x% percentile is denoted as Q_x. In each column, the value of the method showing the largest reduction compared to the equally weighted model is highlighted.

weight quotient	min	Q_05	Q_25	median	Q_75	Q_95	max
BB_fJL / BB	0.614	0.938	0.996	1.016	1.045	1.740	4.471
BB_fJL_param / BB_param	0.959	0.976	0.997	1.001	1.006	1.094	4.420

Table 4.4: Distribution of average ratios between the optimal group weights depending on different initial weights $\mathring{W} = 1$ and $\mathring{W} =$ JL_0.5 for 100 unbalanced panel simulations.

trace quotient	min	Q_05	Q_25	median	Q_75	Q_95	max
BB_fJL / BB	0	0.445	0.988	1.004	1.018	1.327	2.895
BB_fJL_param / BB_param	0	0.836	0.996	1.001	1.013	1.268	239.023

Table 4.5: Distribution of ratios between the traces of the variance-covariance matrix of the estimator depending on different initial weights $\mathring{W} = 1$ and $\mathring{W} =$ JL_0.5 for 100 unbalanced panel simulations.

Weighting Type	mean	min	Q_05	Q_25	median	Q_75	Q_95	max	% improved
BB / equal	0.981	0.672	0.720	0.841	**0.946**	1.054	1.393	2.193	66.2
BB_param / equal	**0.940**	0.581	0.732	0.865	0.957	**1.014**	**1.161**	**1.329**	**70.1**
JL_0.5 / equal	0.971	**0.523**	**0.605**	**0.840**	0.949	1.093	1.366	1.891	57.1

Table 4.6: Overview of distribution of relative mean squared l_2 distance of predicted choice probabilities to true choice probabilities, as quotient with mean squared l_2 distance of equally weighted model. The x% percentile is denoted as Q_x. In each column, the value of the method showing the largest reduction compared to the equally weighted model is highlighted.

CHAPTER

Conclusion and Outlook 5

5.1 Conclusion

The work presented in this thesis demonstrates that the special structure of pairwise composite marginal likelihood (CML) functions can be utilised to develop methods and tools aimed at improving CML estimation-based discrete choice modelling at several key points along the modelling workflow, demonstrating its value beyond that of increased computational speed, for which they were originally developed.

The first paper, presented in Chapter 2, develops score plots as a diagnostic tool, analogous in concept, interpretation and application to residual plots used in linear regression models. By utilizing the pairwise CML structure, score information can be obtained at the level of observational pairs, rather than at the level of individuals as in maximum likelihood (ML) estimation, allowing a more detailed investigation of the dependence of score contributions on covariate values and the temporal position of observations. This in turn enables the generation of heatmap score plots, as introduced in Section 2.3, which facilitate a qualitative analysis of the dependence of score contributions on covariate values by colour coding deviations in average score contributions. The sliced score plots enable a quantitative analysis by combining scatter plots with the colour coding, providing further insight into the form of dependence between score contributions and the independent plot variable. These tools enable practitioners to identify model misspecification, such as non-linear covariate dependencies, missing variables, temporal dependencies and dynamic effects,

as demonstrated with synthetic and real-world data in Sections 2.5 and 2.6.

The second paper, presented in Chapter 3, extends the utilisation of score information to develop Lagrange multiplier (LM) type tests for pooling of observational pairs. In contrast to Wald or likelihood-ratio tests, these tests do not require the estimation of an unrestricted model. Instead, they rely on score information with respect to the unrestricted model at the restricted maximum. In the case of a test for pooling, this eliminates the necessity to implement the unrestricted model. In combination with the score information of observational pairs, this provides a versatile methodology for the detection of structural changes and dynamic processes in the data generating process (DGP). The theoretical asymptotic distributions of the test statistics under the null hypothesis were derived and subsequently confirmed in the finite-sample setting through extensive simulation studies. The sensitivity of the tests to detect various types and severities of model misspecification was analysed via simulation studies, and the tests were subsequently applied in a real-world context, complementing the diagnostic plots from Chapter 2 in the case study presented in Section 2.6.

The third paper, presented in Chapter 4, examines weighting strategies for CML estimation. Section 4.2 introduces a two-step optimal group-weight CML estimator for unbalanced panel data with the objective to minimise the estimator variance by constructing group-specific power weights via a first estimation, grouping individuals by their number of observations, and utilising these weights for an optimal weighted second estimation. The theoretical properties of the estimator have been confirmed in simulation studies, in which it outperformed on average the unweighted models and a reference model that applied deterministic weights as proposed by Joe and Lee [2009]. Section 4.3 examines the use of weighting strategies to emphasise observational pairs with a large temporal distance, thereby reducing estimation bias in the presence of an autoregressive error process in the DGP that was not accounted for in the model. These strategies were informed by sound asymptotic considerations and validated in the finite sample case through simulations, which demonstrated a significant reduction of the bias compared to equally weighted models.

The papers collectively present several novel methodologies that have the potential to enhance the modelling workflow for discrete choice models. As a first step, the score plots introduced in Chapter 2 inform the model specification process by identifying missing variables and non-linear dependencies in the utility function. The visual nature, interpretability, and low computational cost of the score plots make them a suitable tool for guiding the utility specification process. They can assist in reducing its subjectiveness, identifying hidden structures, and considerably speed up the process. As a second step, the same score plots serve as diagnostic

tools for assessing temporal stability of the model, regarding estimated parameter and dynamic error processes, complemented by LM tests from Chapter 3, which provide rigorous tests of the visual findings. In case of a dynamic error process the weighting strategies introduced in Section 4.3 can be employed as a third step in the modelling workflow to mitigate the effects of a potential misspecification of the model, thereby reducing the bias introduced by this misspecification. As a final step, in the case of an unbiased panel data set, the two-step optimal group-weight CML estimator, introduced in Section 4.2, can be employed to reduce the variance of the estimator.

In conclusion, the methods introduced in this thesis provide the discrete choice modeller with a number of new tools that can be employed at various stages of the modelling workflow, potentially improving the overall process and, most importantly, the final model, the derived inference, and the resulting predictions.

5.2 Outlook

The adage "The more one knows, the more one realizes how much there is still to learn" has been ascribed to various individuals throughout the ages. This sentiment is equally applicable to the work presented in this thesis, as each result opens up new avenues for research and inquiry.

Regarding the score plots introduced in Chapter 2, a prospective research avenue would be the investigation of score plots based on single observations, in contrast to observational pairs, potentially by employing single observation margins in combination with pairwise CML. Secondly, extending the application of score plots to a wider range of models, including mixed multinomial logit models, ordered probit models, or statistical models outside the field of discrete choice modelling, could prove advantageous, given that the concept of scores is not exclusive to this field. This approach has the potential to enhance the repertoire of available diagnostic tools for a wide range of statisticians. Another apparent path would be the application of score plots to a broader range of real-world data sets. This would not only serve to validate their practical applicability but also potentially enhance previously estimated models and provide new insights into the data and the underlying DGPs.

The Lagrange multiplier type tests introduced in Chapter 3 offer similar potential research directions in terms of their practical application. They could be particularly useful in cases where score plots revealed patterns of score dependence on time or temporal distance to test for dynamic effects in the error process or for structural changes in the model over time. Further development of the LM type tests could

involve adapting them for pooling tests based on groups depending on covariate values, further complementing the model selection guidance offered by the score plots.

Concerning the two-step optimal group-weight CML estimator outlined in Chapter 4, future research could explore the definition of groups based on criteria that extend beyond the number of observations per individual, such as sociodemographic variables. This approach has the potential to lead to the development of hybrid group-weighting schemes that integrate different types of group weights. Analogies to feasible generalised least square (FGLS) estimation are pertinent in this context, where observations are weighted by the estimated variance of the error term. A similar methodology could be devised for discrete choice models, wherein weights are estimated for each individual and expressed through a parametric function dependent on the individual's sociodemographic characteristics.

Beyond the direct advancements of the methods presented in this thesis, future work could focus on the inclusion and estimation of dynamic effects in discrete choice models. While this thesis demonstrated how such effects can be detected and avoided, the next step would be to develop methods for the estimation of time-dependent choice processes. Enhancing the `Rprobit` package to include these capabilities could prove to be of significant benefit to practitioners, as an understanding of the dynamic effects themselves might provide valuable insights.

Ultimately, the continued exploration and refinement of these methodologies hold great promise for the advancement of discrete choice modelling and, in turn, for empowering practitioners by providing them with new toolsets, facilitating a deeper understanding of their data, thereby enhancing the comprehension of complex decision-making processes.

> "I didn't think it would end this way."
> "End? No, the journey doesn't end here.
> [This] is just another path, one that we all
> must take."
>
> *J.R.R. Tolkien (1954)*

Bibliography

Adamowicz, W., Boxall, P., Williams, M., and Louviere, J. (1998). Stated Preference Approaches for Measuring Passive Use Values: Choice Experiments and Contingent Valuation. *American Journal of Agricultural Economics*, 80(1):64–75.

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.

Bansal, P., Keshavarzzadeh, V., Guevara, A., Li, S., and Daziano, R. A. (2022). Designed quadrature to approximate integrals in maximum simulated likelihood estimation. *The Econometrics Journal*, 25(2):301–321.

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182.

Batram, M. and Bauer, D. (2016). New results on the asymptotic and finite sample properties of the MaCML approach to multinomial probit model estimation. *ArXiv e-prints*.

Batram, M. and Bauer, D. (2017). Model selection and model averaging in MACML-estimated MNP models. *ArXiv e-prints*.

Batram, M. and Bauer, D. (2019). On consistency of the MACML approach to discrete choice modelling. *Journal of Choice Modelling*, 30:1–16.

Bauer, D., Büscher, S., and Batram, M. (2022). Non-parametric estimation of mixed discrete choice models. *ArXiv e-prints*.

Ben-Akiva, M., Bolduc, D., and Bradley, M. (1993). Estimation of Travel Choice Models with Randomly Distributed Values of Time. *Transportation Research Record*, 1413:88–97.

Ben-Akiva, M. and Lerman, S. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press series in transportation studies. MIT Press.

Ben-Akiva, M. E. and Morikawa, T. (1990). Estimation of travel demand models from multiple data sources. In *Transportation and traffic theory; proceedings of the Eleventh International Symposium, held July 18-20, 1990, in Yokohama, Japan*, Yokohama.

Berkowitsch, N. A. J., Scheibehenne, B., and Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, 143(3):1331–1348.

Bhat, C. R. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B: Methodological*, 37(9):837–855.

Bhat, C. R. (2011). The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B: Methodological*, 45(7):923–939.

Bhat, C. R. (2014). The composite marginal likelihood (CML) inference approach with applications to discrete and mixed dependent variable models. *Foundations and Trends in Econometrics*, 7(1):1–117.

Bhat, C. R. (2018). New matrix-based methods for the analytic evaluation of the multivariate cumulative normal distribution function. *Transportation Research Part B: Methodological*, 109:238–256.

Bhat, C. R. and Koppelman, F. S. (2003). Activity-Based Modeling of Travel Demand. In Hall, R., editor, *Handbook of Transportation Science*, volume 56, chapter 3, pages 39–65. Kluwer Academic Publishers, Boston.

Bonferroni, C. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze, 8 edition.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., Benoit, R. G., Berkers, R. M. W. J., Bhanji, J. P., Biswal, B. B., Bobadilla-Suarez, S., Bortolini, T., Bottenhorn, K. L., Bowring, A., Braem, S., Brooks, H. R., Brudner, E. G., Calderon, C. B., Camilleri, J. A., Castrellon, J. J., Cecchetti, L., Cieslik, E. C., Cole, Z. J., Collignon, O., Cox, R. W., Cunningham, W. A., Czoschke, S., Dadi, K., Davis, C. P., Luca, A. D., Delgado, M. R., Demetriou, L., Dennison, J. B., Di, X., Dickie, E. W., Dobryakova, E., Donnat, C. L., Dukart, J., Duncan, N. W., Durnez, J., Eed, A., Eickhoff, S. B., Erhart, A., Fontanesi, L., Fricke, G. M., Fu, S., Galván, A., Gau, R., Genon, S., Glatard, T., Glerean, E., Goeman, J. J., Golowin, S. A. E., González-García, C., Gorgolewski, K. J., Grady, C. L., Green, M. A., Guassi Moreira, J. F., Guest, O., Hakimi, S., Hamilton, J. P., Hancock, R., Handjaras, G., Harry, B. B., Hawco, C., Herholz, P., Herman, G., Heunis, S., Hoffstaedter, F., Hogeveen, J., Holmes, S., Hu, C.-P., Huettel, S. A., Hughes, M. E., Iacovella, V., Iordan, A. D., Isager, P. M., Isik, A. I., Jahn, A., Johnson, M. R., Johnstone, T., Joseph, M. J. E., Juliano, A. C., Kable, J. W., Kassinopoulos, M., Koba, C., Kong, X.-Z., Koscik, T. R., Kucukboyaci, N. E., Kuhl, B. A., Kupek, S., Laird, A. R., Lamm, C., Langner, R., Lauharatanahirun, N., Lee, H., Lee, S., Leemans, A., Leo, A., Lesage, E., Li, F., Li,

M. Y. C., Lim, P. C., Lintz, E. N., Liphardt, S. W., Losecaat Vermeer, A. B., Love, B. C., Mack, M. L., Malpica, N., Marins, T., Maumet, C., McDonald, K., McGuire, J. T., Melero, H., Méndez Leal, A. S., Meyer, B., Meyer, K. N., Mihai, G., Mitsis, G. D., Moll, J., Nielson, D. M., Nilsonne, G., Notter, M. P., Olivetti, E., Onicas, A. I., Papale, P., Patil, K. R., Peelle, J. E., Pérez, A., Pischedda, D., Poline, J.-B., Prystauka, Y., Ray, S., Reuter-Lorenz, P. A., Reynolds, R. C., Ricciardi, E., Rieck, J. R., Rodriguez-Thompson, A. M., Romyn, A., Salo, T., Samanez-Larkin, G. R., Sanz-Morales, E., Schlichting, M. L., Schultz, D. H., Shen, Q., Sheridan, M. A., Silvers, J. A., Skagerlund, K., Smith, A., Smith, D. V., Sokol-Hessner, P., Steinkamp, S. R., Tashjian, S. M., Thirion, B., Thorp, J. N., Tinghög, G., Tisdall, L., Tompson, S. H., Toro-Serey, C., Torre Tresols, J. J., Tozzi, L., Truong, V., Turella, L., van 't Veer, A. E., Verguts, T., Vettel, J. M., Vijayarajah, S., Vo, K., Wall, M. B., Weeda, W. D., Weis, S., White, D. J., Wisniewski, D., Xifra-Porxas, A., Yearling, E. A., Yoon, S., Yuan, R., Yuen, K. S. L., Zhang, L., Zhang, X., Zosky, J. E., Nichols, T. E., Poldrack, R. A., and Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88.

Bradley, M. A. and Daly, A. J. (1991). Estimation of Logit Choice Models Using Mixed Stated Preference and Revealed Preference Information. In *6th International Conference on Travel Behaviour*, pages 209–231.

Breusch, T. S. and Pagan, A. R. (1980). The lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, 47:239.

Büscher, S. (2024). Visual Guidance for Model Specification: Introducing Score Plots for Discrete Choice Models. *Available at SSRN*.

Büscher, S., Batram, M., and Bauer, D. (2019). Using Motifs for Population Synthesis in Multi-agent Mobility Simulation Models. In *Springer Proceedings in Mathematics and Statistics*, volume 294, pages 335–349.

Büscher, S. and Bauer, D. (2024). Using Lagrange Multiplier Type Tests to Detect Structural Intra-Person Heterogeneity in Composite Marginal Likelihood Estimation in Panel Data Sets. *Available at SSRN*.

Büscher, S. and Bauer, D. (2024). Weighting strategies for pairwise composite marginal likelihood estimation in case of unbalanced panels and unaccounted autoregressive structure of the errors. *Transportation Research Part B: Methodological*, 181:102890.

Cessie, S. L. and Houwelingen, J. C. V. (1994). Logistic Regression for Correlated Binary Data. *Applied Statistics*, 43(1):95.

Chintagunta, P. K. (1992). Estimating a Multinomial Probit Model of Brand Choice Using the Method of Simulated Moments. *Marketing Science*, 11(4):386–407.

Çinlar, E. (2011). *Probability and Stochastics*, volume 261 of *Graduate Texts in Mathematics*. Springer New York, New York, NY.

Clarivate (2022). Number of publications on the topic of "composite likelihood" by year. `https://www.webofscience.com/wos/woscc/analyze-results/6de73426-9e3f-4ae3-90dd-48c827eee91d-6456e209`. Last checked on 08 December 2022.

Cox, D. R. and Reid, N. (2004). Miscellanea A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737.

Crastes dit Sourd, R., Daly, A., Palma, D., and Holz-rau, C. (2020). Weighting strategies for modelling life course history events via pairwise composite marginal likelihood. working paper.

Croissant, Y. (2020). Estimation of Random Utility Models in R : The mlogit Package. *Journal of Statistical Software*, 95(11):1–41.

Delporte, M., Verbeke, G., Fieuws, S., and Molenberghs, G. (2025). Accelerating computation: A pairwise fitting technique for multivariate probit models. *Computational Statistics & Data Analysis*, 203(June 2024):108082.

Dick, J., Gantner, R. N., Gia, Q. T. L., and Schwab, C. (2016). Multilevel higher order Quasi-Monte Carlo Bayesian Estimation. *Mathematical Models and Methods in Applied Sciences*, 27(5):953–995.

Eddelbuettel, D. and Francois, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40.

Engler, D. A., Mohapatra, G., Louis, D. N., and Betensky, R. A. (2005). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics*, 7(3):399–421.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Franses, P. H. and Paap, R. (2004). *Periodic Time Series Models*. Oxford University PressOxford.

Gao, X. and Song, P. X. (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540.

Gelman, Andrew and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348:3.

Haghani, M., Bliemer, M. C., Rose, J. M., Oppewal, H., and Lancsar, E. (2021). 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 Modelling*, 41(December 2020):100309.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY.

Healy, K. (2018). *Data visualization: a practical introduction*. Princeton University Press.

Heiss, F. and Winschel, V. (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, 144(1):62–80.

Hess, S. and Daly, A. J. (2014). *Handbook of choice modelling*. Edward Elgar.

Hess, S., Train, K. E., and 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(2):147–163.

Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2:360–378.

Householder, A. S. (1941). A theory of steady-state activity in nerve-fiber networks: I. Definitions and preliminary lemmas. *The Bulletin of Mathematical Biophysics*, 3(2):63–69.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts, 2nd edition.

Joe, H. (1995). Approximations to Multivariate Normal Rectangle Probabilities Based on Conditional Expectations. *Journal of the American Statistical Association*, 90(431):957–964.

Joe, H. and Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis*, 100(4):670–685.

Johnson, J. E. V. and Bruce, A. C. (1997). A Probit Model for Estimating the Effect of Complexity on Risk Taking. *Psychological Reports*, 80(3):763–772.

Johnson, S. G. (2022). *The NLopt nonlinear-optimization package*.

Keshavarzzadeh, V., Kirby, R. M., and Narayan, A. (2018). Numerical Integration in Multiple Dimensions with Designed Quadrature. *SIAM Journal on Scientific Computing*, 40(4):A2033–A2061.

Kessels, R., Goos, P., and Vandebroek, M. (2006). A comparison of criteria to design efficient choice experiments. *Journal of Marketing Research*, 43(3):409–419.

Kim, J., Lee, H., and Lee, J. (2020). Smartphone preferences and brand loyalty: A discrete choice model reflecting the reference point and peer effect. *Journal of Retailing and Consumer Services*, 52(May 2019):101907.

Kolmogorov, A. L. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari*, 4:83–91.

Kuk, A. Y. and Nott, D. J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics & Probability Letters*, 47(4):329–335.

Lerman, S. R. and Manski, C. F. (1981). On the Use of Simulated Frequencies to Approximate Choice Probabilities. In Manski, C. F. and McFadden, D., editors, *Structural Analysis of Discrete Data with Econometric Applications*. The MIT Press.

Lindsay, B. G., Yi, G. Y., and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21(1):71–105.

Louviere, J. J., Hensher, D. A., Swait, J. D., and Adamowicz, W. (2000). *Stated Choice Methods*, volume 89. Cambridge University Press.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mariel, P., Hoyos, D., 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*. SpringerBriefs in Economics. Springer International Publishing, Cham.

Marsaglia, G., Tsang, W. W., and Wang, J. (2003). Evaluating kolmogorov's distribution. *Journal of Statistical Software*, 8.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. *Frontiers in Econometrics*.

McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica*, 57(5):995.

Meijer, E. and Rouwendal, J. (2006). Measuring welfare effects in models with random coefficients. *Journal of Applied Econometrics*, 21(2):227–244.

Mendell, N. R. and Elston, R. C. (1974). Multifactorial Qualitative Traits: Genetic Analysis and Prediction of Recurrence Risks. *Biometrics*, 30(1):41–57.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1):115–132.

Nobel Prize Outreach (2000). The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2000. `https://www.nobelprize.org/prizes/economic-sciences/2000/summary/`. Last checked on 15 December 2024.

Nova, G., van Cranenburgh, S., and Hess, S. (2024). Understanding the decision-making process of choice modellers. *ArXiv e-prints*, pages 1–35.

Oelschläger, L. and Bauer, D. (2023). Bayesian probit models for preference classification. In *Proceedings of the 37th International Workshop on Statistical Modelling*.

Oelschläger, L. and Bauer, D. (2024). *RprobitB: Bayesian Probit Choice Modeling*.

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(July 2020):100285.

Páez, A. and Boisjoly, G. (2022). *Discrete Choice Analysis with R*. Use R! Springer International Publishing, Cham.

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(November 2020):100257.

Paz, A., Arteaga, C., and Cobos, C. (2019). Specification of mixed logit models assisted by an optimization framework. *Journal of Choice Modelling*, 30(April 2018):50–60.

Pedeli, X. and Varin, C. (2020). Pairwise likelihood estimation of latent autoregressive count models. *Statistical Methods in Medical Research*, 29(11):3278–3293.

Porter, S. R., Whitcomb, M. E., and Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, 2004(121):63–73.

Powell, M. J. D. (1994). A Direct Search Optimization Method That Models the Objective and Constraint Functions by Linear Interpolation. In *Advances in Optimization and Numerical Analysis*, number 1, pages 51–67. Springer Netherlands, Dordrecht.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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.

Ryu, E. K. and Boyd, S. P. (2015). Extensions of Gauss Quadrature Via Linear Programming. *Foundations of Computational Mathematics*, 15(4):953–971.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A. (2019). Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *The Annals of Applied Statistics*, 13(3):1564–1589.

Short, J. and Kopp, A. (2005). Transport infrastructure: Investment and planning. Policy and research aspects. *Transport Policy*, 12(4):360–367.

Silvey, S. D. (1959). The lagrangian multiplier test. *The Annals of Mathematical Statistics*, 30:389–407.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11):1359–1366.

Solow, A. R. (1990). A method for approximating multivariate normal orthant probabilities. *Journal of Statistical Computation and Simulation*, 37(3-4):225–229.

The Royal Swedish Academy of Sciences (2000). The Scientific Contributions of James Heckman and Daniel McFadden.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press, 2nd edition.

Trinh, G. and Genz, A. (2015). Bivariate conditioning approximations for multivariate normal probabilities. *Statistics and Computing*, 25(5):989–996.

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., and Walker, J. (2022). Choice modelling in the age of machine learning - Discussion paper. *Journal of Choice Modelling*, 42(January 2021):100340.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1):1–28.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Varin, C. and Vidoni, P. (2005). A Note on Composite Likelihood Inference and Model Selection. *Biometrika*, 92(3):519–528.

Varin, C. and Vidoni, P. (2006). Pairwise likelihood inference for ordinal categorical time series. *Computational Statistics & Data Analysis*, 51(4):2365–2373.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54:426–482.

Wickham, H., Çetinkaya-Rundel, M., and Grolemund, G. (2023). *R for Data Science*. O'Reilly Media, Inc., 2 edition.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9:60–62.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Nelson Education, 5th edition.

Zumkeller, D. and Chlond, B. (2009). Dynamics of change: Fifteen-year german mobility panel. In *Transportation Research Board 88th Annual Meeting Compendium of Papers*. Transportation Research Board.

Zumkeller, D., Chlond, B., and Lipps, O. (1999). Das mobilitäts-panel (mop) - konzept und realisierung einer bundesweiten längsschnittbeobachtung. In Hautzinger, H., editor, *Schriftenreihe der Deutschen Verkehrswissenschaftlichen Gesellschaft / B*, volume 217, pages 33–72. DVWG.

Curriculum Vitae

Personal Data

Name:	Sebastian Büscher
Email:	sebastian.buescher@uni-bielefeld.de

Scientific Education

Since 03/2018	**Ph.D., Econometrics** at Bielefeld University, Germany
09/2015- 09/2017	**M.Sc. Mathematical Modelling in Engineering (MathMods)** at University of L'Aquila, Italy, University of Hamburg, Germany, and Autonomous University of Barcelona, Spain Degree: Joint Degree of Master of Science in Mathematical Modelling in Engineering: Theory, Numerics, Applications
10/2011- 07/2015	**B.Sc. Mathematics and Physics** at Bielefeld University, Germany Degree: Bachelor of Science in Mathematics and Physics
09/2013- 05/2014	**B.Sc. Mathematics and Physics** at Edinburgh University, United Kingdom ERASMUS program

Academic Employment

Since 03/2018	**Research Associate** Chair of Econometrics Bielefeld University, Bielefeld, Germany
10/2014- 09/2015	**Student Assistant** Faculty of Mathematics Bielefeld University, Bielefeld, Germany

Teaching Experience

04/2019- 09/2025	Statistical and Econometric Models Tutorial
10/2020- 03/2025	Regression Analysis Lecture
10/2019- 03/2025	Einführung in die Ökonometrie Tutorial
04/2024- 09/2024	Reading Course "Statistische Wissenschaften" Seminar
04/2023- 09/2023	Daten in den Medien - Information versus Manipulation Seminar
03/2023	Statistische Software: R Seminar
10/2018- 09/2019	Präsenzübung zur Mathematik I und II Tutorial
10/2014- 09/2015	Übungen zu Analysis I Tutorial

List of Publications

This thesis is comprised of the following three publications:

- Büscher, S. and Bauer, D. (2024). Weighting strategies for pairwise composite marginal likelihood estimation in case of unbalanced panels and unaccounted autoregressive structure of the errors. *Transportation Research Part B: Methodological*, 181:102890

- Büscher, S. and Bauer, D. (2024). Using Lagrange Multiplier Type Tests to Detect Structural Intra-Person Heterogeneity in Composite Marginal Likelihood Estimation in Panel Data Sets. *Available at SSRN*

- Büscher, S. (2024). Visual Guidance for Model Specification: Introducing Score Plots for Discrete Choice Models. *Available at SSRN*

During my term as a research associate at Bielefeld University I also contributed to the following papers:

- Büscher, S., Batram, M., and Bauer, D. (2019). Using Motifs for Population Synthesis in Multi-agent Mobility Simulation Models. In *Springer Proceedings in Mathematics and Statistics*, volume 294, pages 335–349

- Bauer, D., Büscher, S., and Batram, M. (2022). Non-parametric estimation of mixed discrete choice models. *ArXiv e-prints*

List of Talks

- *Visual Guidance for Model Specification: Introducing Score Plots for Composite Marginal Likelihood Estimated Discrete Choice Models*, Talk at the 14th Young Researchers Workshop of the Centre for Statistics, Bielefeld, Bielefeld, 2024.

- *Lagrange multiplier type test to detect structural intra-personal heterogeneity in Composite Marginal Likelihood estimation of discrete choice models*, Talk at the 8th International Choice Modelling Conference, Puerto Varas, 2024.

- *New Lagrange Multiplier type test for CML models*, Talk at the 13th Young Researchers Workshop of the Centre for Statistics, Bielefeld, 2024.

- *How to test for intra-personal heterogeneity and temporal effects in discrete choice models in an economic manner*, Talk at the 12th Young Researchers Workshop of the Centre for Statistics, Bielefeld, 2023.
- *Robust estimation of misspecified discrete choice models*, Talk at the 11th Young Researchers Workshop of the Centre for Statistics, Bielefeld, 2022.
- *Weighting strategies for pairwise composite marginal likelihood estimation*, Talk at the 10th Young Researchers Workshop of the Centre for Statistics, Bielefeld, 2022.
- *Weighting strategies for pairwise composite marginal likelihood estimation in case of unbalanced panels and unaccounted autocorrelation of the errors*, Talk at the 7th International Choice Modelling Conference, Reykjavik, Iceland, 2022.
- *Econometrics vs. Machine Learning: Who is better at predicting human mobility patterns?* Talk at the 9th Young Researchers Workshop of the Centre for Statistics, Bielefeld, 2021.
- *Modeling Motif Choice: Using Out of Sample Prediction Performance to Compare Different Discrete Choice Models for Model Selection to Analyze Temporal Stability*, Talk at the 8th Young Researchers Workshop of the Centre for Statistics, Bielefeld, 2019.
- *Using Motifs in person-centred mobility simulation models*, Talk at the 7th Young Researchers Workshop of the Centre for Statistics, Bielefeld, 2019.
- *Using Motifs for Population Synthesis in Multi-Agent Mobility Simulation Models*, Talk at the 14th Workshop on Stochastic Models, Statistics and their Application, Dresden, 2019.
- *Modeling Motif Choice: Comparing the prediction performance of different discrete choice models to analyze temporal stability*, Talk at the 6th International Choice Modelling Conference, Kobe, Japan, 2019.
- *An analysis of two decades of German mobility data with an emphasis on temporal stability using Multinomial Logit Models*, Talk at the 6th Young Researchers Workshop of the Centre for Statistics, Bielefeld, 2018.
- *The MACML approach to probit model estimation with application on motif choice*, Talk at the 5th Young Researchers Workshop of the Centre for Statistics, Bielefeld, 2018.