Structural Econometric Methods in Industrial Organization and Quantitative Marketing

Ali Hortaçsu[1]
University of Chicago and NBER

Joonhwi Joo[2]
University of Texas at Dallas

September 25, 2020

(Click Here for the Latest Version)

[1]hortacsu@uchicago.edu
[2]joonhwi.joo@utdallas.edu

Preface

This manuscript is prepared for a course on structural econometric methods used in empirical industrial organization and quantitative marketing. Targeted readers are second-year Ph.D. students, who are assumed to have basic knowledge on first-year-level microeconomics and econometrics. We tried to give a concise and coherent overview of the literature, while not losing the key insights and features of the important models and estimation methods.

The current manuscript consists of six chapters and the appendix. After a brief introduction (Chapter 1), we discuss the basics of the single-agent static and dynamic discrete choice (Chapter 2), which is now a standard baseline modeling framework in empirical industrial organization, quantitative marketing, and other related fields. We move on to study demand estimation with market data, where we introduce demand-estimation methods in the product space and characteristics space, respectively (Chapter 3). In the following two chapters, we focus on strategic interactions of firms in the static and dynamic setup (Chapters 4 and 5). We then move our focus back to consumers to study the empirical frameworks of consumer search (Chapter 6). For completeness, we also summarize basic features of the most commonly used baseline estimation frameworks in the appendix (Appendix A). Treatments on the subjects are far from complete, although we tried to include as detailed a list of references as possible. The references at the end of each chapter include all the cited works in the body, and some additional readings related to the topic. We recommend that readers interested in a particular topic consult the references listed at the end of each chapter, and the relevant handbook chapters as well.

The purpose of this manuscript is to provide a compact overview of toolkits that can be used for a structural empirical analysis in industrial organization, quantitative marketing, and other related fields. We therefore focus more on the formulations of economic and econometric models, estimation methods, and estimation algorithms. In many instances, we abstract away from the concrete empirical applications. However, we note that it is not because we think the applications are unimportant. This manuscript is still in progress, and we are planning to give a treatment on more topics in the subsequent versions.

Contents

Chapter 1

Introduction

1.1 Some Philosophy: Theory, Data, and Estimation

Economists often define themselves as scientists.[1] Attempts to define a scholarly discipline outside the natural sciences as "science" is not only for economics. One might immediately question why such attempts are so widespread.

The definition of science is still an ongoing area of debate in the philosophy of science, and we are not going to pursue a deeper discussion on the subject here. However, the consensus is that science has been the most effective set of ideologies and methodologies to reveal the causal relationships, which has been extremely successful for the last two centuries. Then, how does a scientific methodology reveal causality? Before directly answer this question, we consider the following two thought experiments.

Experiment 1. Suppose one has collected a data on the height and weight of a randomly selected group in the population. Let y_i be the weight, and let x_i be the height of each individual. The researcher runs the following regression:

$$y_i = \alpha + \beta x_i + \epsilon_i.$$

The OLS estimate $\hat{\beta}$ turns out to be positive and highly statistically significant. Does this finding imply a causal relationship between height and weight?

Experiment 2. Suppose one conducted a repeated Hooke's experiment and recorded the result. Let y_i be the length of the spring, and let x_i be the weight of the pendulum. Again, the researcher runs the following regression:

$$y_i = \gamma + \delta x_i + \epsilon_i.$$

The OLS estimate $\hat{\delta}$ is positive and highly statistically significant. Does this find imply a causal relationship between the weight of the pendulum and the length of the spring?

The answer to the first question is definitely no.[2] But the answer to the second question is possibly yes. Positive and highly statistically significant $\hat{\delta}$ may be taken as evidence of a causal relationship. The

[1] For example, "The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel."

[2] If you are not convinced, recall Procrustes, the stretcher, in the *Odyssey*. When Procrustes stretches the guest to fit him in his bed, will the guest's weight increase?

structures of the two thought experiments seem to be quite similar; a dataset is collected, a simple linear regression is ran, and the coefficients have the same sign and both are statistically significant. But the implications on the causality can be quite different. Where does this seemingly stark difference come from? A short answer is the existence of a credible scientific theory, which can be either implicit or explicit. Rather than just being satisfied with this simple answer, further explanation would be worthwhile.

The OLS coefficients, $\hat{\beta}$ and $\hat{\delta}$, are, respectively, the measures of correlation between the two corresponding variables.[3] If more than two variables have a joint distribution, correlations can always be calculated.[4] In the above two examples, we can safely conclude the two observed variables are positively correlated. However, correlations have absolutely no information on causality. Scientific theory is what makes the correlation evidence of causality, which never comes from the data itself. We have a good theory that establishes the causality between the weight of the pendulum and the length of the spring, namely, Hooke's law. By contrast, we do not have a good theory for causality between the weight and the height; in fact, another possibly exogenous factor, such as good nutrition, is likely to simultaneously affect both height and weight.

As emphasized, the scientific theory always comes *a priori*, even for the completely controlled and repeated experiments. Establishing causality is possible only through the theory. The scientific theory is essentially summarized as models, which is a simplification of the world. The power of science in revealing the causality between distinct phenomena comes from the power and the logic of rejecting the false models. The procedure of rejecting the false models is often referred to as testing the models using data.

Science stands on two legs: building a model and testing it using the data. Models are abstractions of the real world, selecting and incorporating only the relevant features of the world. Models that are not rejected by more tests are considered more reliable. More reliable models are considered to provide more reliable predictions. Scientific models are most commonly formulated using the mathematical language, and are tested using the statistical methods. In many cases, the whole purpose of a statistical inference boils down to test a model. A (statistical) test of a scientific model is expressed in terms of the null and alternative hypotheses. Very roughly speaking, the probability that the null hypothesis is not true boils down to the form of p-value. It gives the probability that a certain statistic is obtained not by coincidence, given that (i) the null and alternative hypotheses are set up correctly, and (ii) adequate estimation method and estimators are used to compute the p-value.

Economics stands on the same ground. Economists build the economic models and test them using the data with the econometric methods. An immediate question might arise: What defines a model as an economic model? We suggest two key ingredients of an economic model: (i) optimizing behaviors of (ii) the rational agent(s).[5] Economic theory begins from preferences, technology, and different equilibrium concepts. As a result of the optimizing behavior of one or multiple rational agents, observable/testable equilibrium outcomes are derived in the form of mathematical statements. Those outcomes are tested

[3] Recall that the OLS coefficient estimate is the sample covariance of x_i and y_i scaled by the sample variance of x_i.

[4] Given that each variable has nonzero, finite variance.

[5] Recent advances in several fields such as behavioral economics allow for the violations on those two key ingredients. For instance, the rationality might be bounded, or optimization failure might occur. Although we focus mostly on the conventional concept of economic theory, we do consider such advances as an important progress in the profession.

using the real-world data with econometric methods.

The goal of econometric methods is to provide tools that are used to infer the correlation between two or more variables, to test the hypotheses generated by an economic theory, and to make predictions based on the inferred correlations. However, economists disagree about how to use the econometric methods on revealing the causality. A fundamental reason for the disagreement is because conducting repeated and controlled experiments on economic agents is difficult. If all the data in economics are obtained from controlled experiments on every variable in an economic model, such a disagreement would not occur.

An ideal identification strategy for revealing the causality would be a fully controlled experiment on every aspect of an explicitly specified economic model. Most natural scientists do conduct such experiments, and some experimental economists have attempted to do so relatively recently as well. They conduct the field experiments and laboratory experiments in relation to economic theories. Although they have made great progress toward making reliable causal inferences in economic behaviors, fundamental limitations remain. The primary reason is that the subjects are human beings, and the experimenters can never gain full control over the detailed aspects of human behavior in the real world, particularly when an intervention involves a possible ethical problem. In addition, it is often reported that subjects behave differently in a laboratory than in the real world. Natural experiments, such as unexpected policy change or natural disasters, can be a great alternative to field experiments. However, usually it is very difficult to find a natural experiment that directly alters the variable of the researcher's interest.

The intuition behind the importance of experimental variation in establishing causality between two variables can be more easily illustrated in the context of the omitted-variable bias in linear regression. Suppose a causal and linear relationship exists between the vector of explanatory variables (x_i, v_i) and y_i, where v_i is unobserved to a researcher. Furthermore, assume the correlation between x_i and v_i is nonzero, which is usual. If the sign and magnitude of the causal effect of interest is about variable x_i, a researcher may be tempted to run the following OLS regression:

$$y_i = \alpha + \beta x_i + \epsilon_i,$$

and claim $\hat{\beta}$ represents the causal effect of x_i on y_i. This claim is unarguably false,[6] unless (i) the correlation between x_i and v_i is zero or (ii) the correlation between v_i and y_i is zero. The problem with virtually all the observational data is that infinitely many v_i's are possible that are not observed, and the best way to avoid this situation is to have x_i generated by some experiment so that it has zero correlation with any possible omitted variables.

Given that the experiments only provide economists with a limited solution, economists often use the nonexperimental data to reveal the causality. The common identifying restriction of causality when using the nonexperimental data is that some exogenous variations must exist in the explanatory variables or at least in the instrumental variables. The exogeneity is, in general, argued via economic theory to gain the justification. In most circumstances that an econometrician has only one set of observations, formally testing which variable is a cause for the other is not possible.

[6] See any undergraduate-level econometrics textbook.

Causality comes from economic theory. However, there are two different ends on building an econometric model from an economic model:[7] the reduced-form econometric models and the structural econometric models in a narrow sense.[8] Reduced-form econometric models abstract away from the rational agents, optimization, and equilibrium. They specify the simple relationships between the variables of interests, and use relevant estimation methods to back out the parameters. Their econometric specifications are mostly linear, which can be justified because linear functions are a first-order approximation of any smooth function. The strength of reduced-form econometric models is their simplicity and possible robustness to the model misspecifications. On the other hand, a structural econometric model begins from stating the economic model specifications such as the objective functions, optimizing variable, the equilibrium concept, the degree of information of the agents and of the econometrician, and the possible source of endogeneity. Then, the model is solved step by step. As a result, the relations between the variables are specified in terms of the moment (in)equalities, likelihoods, or quantile restrictions. Lastly, the relevant estimation methods for such specifications are used to back out the model parameters. By explicitly specifying the economic models, one can make the out-of-sample predictions, policy counterfactuals, and so on. A downside is that the predictions or policy counterfactuals can be sensitive to the model misspecification. Ideally, every ingredient in the model could be tested, but they cannot, especially when using observational data.

Because randomized experiments of a policy at the industry level are impossible in most circumstances,[9] modern empirical industrial organization and quantitative marketing rely extensively on the structural econometric modeling with observational data. The goal of this text is to give an overview on how the various literatures in empirical industrial organization and quantitative marketing use structural econometric modeling to estimate the model parameters, give the model-based predictions and the policy counterfactuals. In the following section, we illustrate the basic rationale of the structural econometric modeling in the empirical industrial organization and quantitative marketing.

1.2 Structural Econometric Modeling in Industrial Organization and Quantitative Marketing

We explained that an economic model is the primary source of the structure in structural econometric modeling. However, the economic model is not the only source of structure in econometric models; as in all other econometric models, a statistical structure should be in place to explain the observed data. It comes from an added unobservable to the economic model, which is very often referred to as the error term.

Many economic models cannot completely explain the observed data without an added stochastic-

[7] The following section makes clear what the econometric modeling is.

[8] In a wide sense, even the linear instrumental-variable method is structural, implicitly imposing a very specific structure on how the instrumental variables affect the outcome variables. This point has been thoroughly investigated by Heckman and Vytlacil (2005).

[9] In the treatment-evaluation literature in labor economics, which is based on the individual-level randomized experiments, such experiments are possible.

ity. Thus, unobservables (to the econometrician and/or the economic agents) are incorporated into an economic model to form an econometric model. We note that an economic model and an econometric model are sometimes indistinguishable because in some stochastic economic models, the unobservables (to the economic agents) are inherent in the economic model.

Conceptually, structural econometric models have three different kinds of error terms. The first kind is the researcher uncertainty, which is sometimes referred to as the structural error or unobserved heterogeneity. This kind of error term is observable to the economic agents but not to the econometrician. The structural errors affect the decision of the economic agents in the same way that the observables do. The second kind is the agent uncertainty. It is observable to neither the economic agents nor the econometrician. However, the variable may affect the economic agent's decision, mostly in terms of ex-ante expectations. The third kind is the error term that is added merely for the rationalization of the data, or for the tractability of estimation. This type of error term may include the measurement errors. Although distinguishing between these concepts during the estimation is often difficult or even impossible, being clear about these conceptual distinctions in the modeling stage is very important.

Formulating and estimating a structural econometric model typically follows four steps: (i) Formulate a well-defined economic model of the environment under consideration, (ii) add a sufficient number of stochastic unobservables to the economic model, (iii) identify and estimate the model parameters, and (iv) verify the adequacy of the resulting structural economic model as a description of the observed data. In step (ii), a researcher should decide whether to fully specify the distribution of the unobservables. Related to (ii) and (iii), estimation of structural econometric models often boils down to backing out the point-identified, finite-dimensional, and policy-invariant model parameters.[10] A few possibilities exists for step (iv), model validation. For example, the researcher can split the sample, estimate the model only using a subset of the sample, and examine the accuracy of the out-of-sample prediction. Another brilliant idea is to match the predictions of structural models with the data from a randomized experiment. We think an appropriate model validation is crucial to the credibility of the results from estimating a structural econometric model. A simple sensitivity analysis alone may not be enough to persuade the audience that the model is a credible and realistic approximation of the world.

Modern empirical industrial organization is based on the spirit of the structure-conduct-performance paradigm. In the paradigm, the market environment has a direct effect on the market structure. Then, it affects the firms' conduct. Firms' conduct in turn affects the industry performance. Recent trends of new empirical industrial organization focus more on a single industry rather than the economy as a whole. Even when evaluating the performance of a single industry, the following factors should be taken into consideration: the profit-maximizing firms, the demand function, and the different equilibrium concepts (e.g., Walrasian, Cournot/Bertrand Nash). An essential factor here is the policy invariance. To conduct the welfare counterfactuals or other policy experiments, the economic model and the estimated model parameters should be invariant to the hypothetical policy changes. The researcher should carefully persuade the audience how and why a structure can be invariant to the policy counterfactuals of interest. In the following chapters, we study various structural econometric models that can be used to evaluate

[10] The literature on the partially identified or nonparametric structural econometric models is growing. We study some examples of them in subsequent chapters.

the industry performance and to conduct the policy counterfactuals.

1.3 Further Readings

Debates between "structural" and "reduced-form" econometricians are acute. Interested readers should refer to the list of references for this chapter. A handbook chapter by Reiss and Wolak (2007) has a clear and detailed overview on the structural econometric modeling in industrial organization. Much of our discussions in the previous section were drawn from their chapter.

References

ACKERBERG, D., C. L. BENKARD, S. BERRY, AND A. PAKES (2007): *Econometric tools for analyzing market outcomes*, Elsevier, vol. 6A of *Handbook of Econometrics*, chap. 63, 4171–4276.

ANGRIST, J. D. AND J.-S. PISCHKE (2010): "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics," *Journal of Economic Perspectives*, 24, 3–30.

BLUNDELL, R. (2010): "Comments on: Michael P. Keane 'Structural vs. atheoretic approaches to econometrics'," *Journal of Econometrics*, 156, 25–26.

DEATON, A. (2010): "Instruments, randomization, and learning about development," *Journal of Economic Literature*, 48, 424–455.

HECKMAN, J. J. AND E. VYTLACIL (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73, 669–738.

IMBENS, G. W. (2010): "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48, 399–423.

KEANE, M. P. (2010): "Structural vs. atheoretic approaches to econometrics," *Journal of Econometrics*, 156, 3–20.

KEANE, M. P., P. E. TODD, AND K. I. WOLPIN (2011): *The structural estimation of behavioral models: Discrete choice dynamic programming methods and applications*, Elsevier, vol. 4a of *Handbook of Labor Economics*, chap. 4, 331–461.

LAIBSON, D. AND J. A. LIST (2015): "Principles of (Behavioral) Economics," *American Economic Review: Papers and Proceedings*, 105, 385–390.

LEAMER, E. E. (1983): "Let's take the con out of econometrics," *American Economic Review*, 73, 31–43.

NEVO, A. AND M. D. WHINSTON (2010): "Taking the dogma out of econometrics: Structural modeling and credible inference," *Journal of Economic Perspectives*, 24, 69–82.

REISS, P. C. AND F. A. WOLAK (2007): *Structural econometric modeling: Rationales and examples from industrial organization*, Elsevier, vol. 6A of *Handbook of Econometrics*, chap. 64, 4277–4415.

RUST, J. (2010): "Comments on: "Structural vs. atheoretic approaches to econometrics" by Michael Keane," *Journal of Econometrics*, 156, 21–24.

——— (2014): "The limits of inference with theory: A review of Wolpin (2013)," *Journal of Economic Literature*, 52, 820–850.

WOLPIN, K. I. (2013): *The limits of inference without theory*, Cambridge, Massachusetts: MIT Press.

Chapter 2

Static and Dynamic Discrete Choice

The discrete-choice framework, often referred to as the qualitative response models, has become a major workhorse in diverse contexts of empirical industrial organization and other fields in applied microeconomics. In this chapter, we review the basic theory on the binary choice and multinomial choice models. Then, we proceed to study the dynamic discrete-choice models pioneered by Rust (1987); Hotz and Miller (1993); Hotz, Miller, Sanders, and Smith (1994) to study how the discrete-choice framework incorporates the forward-looking behavior of an economic agent. Throughout this chapter, we assume individual choice data on a finite set of alternatives are available. We focus mostly on the fully parametric setup, of which the main goal often boils down to derive the likelihood function for the maximum-likelihood estimation.

2.1 Binary Choice

2.1.1 Motivation: Linear Probability Model

In this section, we consider the binary choice data of individuals indexed by $i \in \mathcal{I} := \{1, 2, ..., I\}$, for alternatives indexed by $j \in \mathcal{J} := \{1, 2, ..., J\}$. Throughout this section, we assume data for each consumer's choice on *each alternative* are available, where the set of *alternatives* $\mathcal{J}$ *is not exclusive*. Let $y_{i,j}$ be a discrete outcome variable with only two possibilities, 0 and 1. If consumer i chooses to buy product j, $y_{i,j} = 1$. Otherwise, $y_{i,j} = 0$. For notational convenience, our discussion focuses on a single individual i.

Let us formalize the setup. Consider a latent utility $y^*_{i,j}$ specified by

$$\begin{aligned} y^*_{i,j} &= \delta_j + \epsilon_{i,j} \qquad \epsilon_{i,j} \sim i.i.d.\ G(\cdot) \\ y_{i,j} &= \begin{cases} 1 & \text{if } y^*_{i,j} \geq 0 \\ 0 & \text{if } y^*_{i,j} < 0 \end{cases}, \end{aligned} \tag{2.1.1}$$

where δ_j is the mean utility index that represents the observable part of $y^*_{i,j}$ to the econometrician. $\epsilon_{i,j}$ represents the idiosyncratic shocks on the latent utility that is unobservable to the econometrician, which is fully known to the consumer. We assume δ_j and $\epsilon_{i,j}$ are orthogonal. Let $\delta_j = \mathbf{x}_j'\boldsymbol{\theta}$,[1] where $\mathbf{x}_j$ is a vector of covariates that shifts the choice probability, observable to the econometrician. Throughout the chapter, we focus on this single linear mean utility index unless noted otherwise.

Consider the following linear estimation equation:

$$y_{i,j} = \mathbf{x}_j'\boldsymbol{\theta} + \eta_{i,j}.$$

The OLS estimator $\hat{\boldsymbol{\theta}}_{OLS}$ is consistent for $\boldsymbol{\theta}$ and asymptotically normal. Because this model has the prediction $\hat{y}_j = \mathbf{x}_j'\hat{\boldsymbol{\theta}}_{OLS}$, the OLS estimator provides us with an easy interpretation on the marginal effects: A one-unit increase in $x_j^{(l)}$ will increase the latent utility $y_{i,j}^*$ by $\hat{\theta}_{OLS}^{(l)}$. The model is called the linear probability model because the prediction $\hat{y}_j$ is such that

$$\hat{y}_j = \hat{E}\left[y_{i,j}|\mathbf{x}_j\right] = \hat{\Pr}\left(y_{i,j} = 1|\mathbf{x}_j\right) = \mathbf{x}_j'\hat{\boldsymbol{\theta}}_{OLS}.$$

An immediate drawback of the linear probability model is that the model may predict $\hat{y}_{i,j} > 1$ or $\hat{y}_{i,j} < 0$. It motivates the choice of $G(\cdot)$ to be a legitimate probability distribution with an unrestricted support, as we study below.

2.1.2 Logit and Probit

The two most often used parametric specifications of $G(\cdot)$ are standard Gaussian and logistic $(0,1)$. Assume $g(.)$, the pdf of $G(.)$, is symmetric around 0. It is indeed the case for the standard Gaussian and the logistic distribution with location parameter 0. Consider the conditional probability:

$$\begin{aligned}
\Pr\left(y_{i,j} = 1|\mathbf{x}_j\right) &= \Pr\left(\mathbf{x}_j'\boldsymbol{\theta} + \epsilon_{i,j} > 0\right) \\
&= \Pr\left(\epsilon_{i,j} > -\mathbf{x}_j'\boldsymbol{\theta}\right) \\
&= 1 - \Pr\left(\epsilon_{i,j} \leq -\mathbf{x}_j'\boldsymbol{\theta}\right) \\
&= 1 - G\left(-\mathbf{x}_j'\boldsymbol{\theta}\right) \\
&= G\left(\mathbf{x}_j'\boldsymbol{\theta}\right).
\end{aligned}$$

The last equality follows from $\Pr\left(-\epsilon_{i,j} \leq \mathbf{x}_j'\boldsymbol{\theta}\right) = \Pr\left(\epsilon_{i,j} \leq \mathbf{x}_j'\boldsymbol{\theta}\right)$ by the symmetry of $g(.)$. The Probit model assumes $\epsilon_{i,j} \sim i.i.d.\ \mathcal{N}(0,1)$, the logit model assumes $\epsilon_{i,j} \sim i.i.d.$ Logistic $(0,1)$[2]. In both models,

[1] Throughout the section, we index the vector of characteristics $\mathbf{x}$ only by j, which implies $\mathbf{x}$ can vary only with the alternatives, not with individuals. However, this restriction is purely for notational convenience. In principle, $\mathbf{x}$ can be indexed by both i and j.

[2] That is, $\Pr\left(y_{i,j} = 1|\mathbf{x}_j\right) = G\left(\mathbf{x}_j'\boldsymbol{\theta}\right) = \frac{\exp\left(x_j'\theta\right)}{1+\exp\left(x_j'\theta\right)}$.

either the scale of $\boldsymbol{\theta}$ or $G(\cdot)$'s scale parameter σ cannot be identified. To see why,

$$\begin{aligned}\Pr\left(y_{i,j}=1|\mathbf{x}_j\right) &= \Pr\left(\frac{\epsilon_{i,j}}{\sigma} > -\frac{\mathbf{x}_j'\boldsymbol{\theta}}{\sigma}\right) \\ &= \Pr\left(\epsilon_{i,j} \leq \mathbf{x}_j'\boldsymbol{\theta}\right).\end{aligned}$$

As long as $\sigma > 0$, any change in σ does not affect the choice probability. The convention is to set $\sigma = 1$ rather than adjusting the scale of $\boldsymbol{\theta}$.

Fix i. Suppose we have the observations $\left\{y_{i,j}, \mathbf{x}_j\right\}_{j=1}^J$. The likelihood function of the binary choice models for individual i is

$$\begin{aligned}L_i\left(\left\{y_{i,j}, \mathbf{x}_j\right\}_{j=1}^J|\boldsymbol{\theta}\right) &= \prod_{j=1}^J \left[\Pr\left(y_{i,j}=1|\mathbf{x}_j\right)\right]^{\mathbf{1}\left(y_{i,j}=1\right)} \left[\Pr\left(y_{i,j}=0|\mathbf{x}_j\right)\right]^{\mathbf{1}\left(y_{i,j}=0\right)} \\ &= \prod_{j=1}^J \left[G\left(\mathbf{x}_j'\boldsymbol{\beta}\right)\right]^{y_{i,j}} \left[1-G\left(\mathbf{x}_j'\boldsymbol{\beta}\right)\right]^{\left(1-y_{i,j}\right)}. \qquad (2.1.2)\end{aligned}$$

The log-likelihood function follows by taking logarithm:

$$l_i\left(\left\{y_{i,j}, \mathbf{x}_j\right\}_{j=1}^J|\boldsymbol{\theta}\right) = \sum_{j=1}^J \left\{y_{i,j}\ln G\left(\mathbf{x}_j'\boldsymbol{\theta}\right) + \left(1-y_{i,j}\right)\ln\left[1-G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]\right\}. \qquad (2.1.3)$$

Taking derivatives with respect to the parameter vector $\boldsymbol{\theta}$ on (2.1.3) yields the sum of the scores:

$$\begin{aligned}\mathbf{s}_i\left(\left\{y_{i,j}, \mathbf{x}_j\right\}_{j=1}^J|\boldsymbol{\theta}\right) &= \sum_{j=1}^J \left\{\frac{y_{i,j}}{G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)} g\left(\mathbf{x}_j'\boldsymbol{\theta}\right) - \frac{1-y_{i,j}}{1-G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)} g\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right\}\mathbf{x}_j \\ &= \sum_{j=1}^J \left\{\frac{g\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\left[1-G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]}\left[y_{i,j}-G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]\right\}\mathbf{x}_j. \qquad (2.1.4)\end{aligned}$$

Setting (2.1.4) as $\mathbf{0}$ yields the first-order condition for the maximum-likelihood estimation for the binary choice models. $\frac{g\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\left[1-G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]}$ part can be interpreted as a weighting function. $\left[y_{i,j}-G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]$ is the prediction error, the expectation of which is zero.

In the logit model, the first-order condition $\mathbf{s}_i\left(\left\{y_{i,j}, \mathbf{x}_j\right\}_{j=1}^J|\boldsymbol{\theta}\right) = \mathbf{0}$ simplifies further, using the fact that $\forall z \in \mathbb{R}$, $\frac{g(z)}{G(z)[1-G(z)]} = 1$.[3] Thus, (2.1.4) with the first-order condition simplifies as

$$\begin{aligned} \mathbf{s}_i\left(\left\{y_{i,j}, \mathbf{x}_j\right\}_{j=1}^{J} | \boldsymbol{\theta}\right) &= \sum_{j=1}^{J}\left[y_{i,j} - G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\right]\mathbf{x}_j \\ &= \mathbf{0}. \end{aligned}$$

If $\mathbf{x}_j$ contains 1 in its row, the first-order condition also contains $\bar{y} = \overline{G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}$.

2.1.3 Marginal Effects

The marginal effect of the binary choice model, $\frac{\partial \Pr\left(y_{i,j}=1|\mathbf{x}_j\right)}{\partial \mathbf{x}_j^{(l)}}$, is

$$\begin{aligned} \frac{\partial \Pr\left(y_{i,j} = 1|\mathbf{x}_j\right)}{\partial x_j^{(l)}} &= \frac{\partial G\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{\partial x_j^{(l)}} \\ &= g\left(\mathbf{x}_j'\boldsymbol{\theta}\right)\theta^{(l)}. \end{aligned} \tag{2.1.5}$$

Unlike the linear probability model, the marginal effect varies across observations. Heterogeneity in responses exists in this model because of the nonlinearity of $G(\cdot)$. One may report (2.1.5) for each observation j in principle. Alternatively, one can consider either (i) average marginal effect $\frac{1}{J}\sum_{j=1}^{J} g\left(\mathbf{x}_j'\hat{\boldsymbol{\theta}}\right)\hat{\boldsymbol{\theta}}^{(l)}$ or (ii) marginal effect on average (or median) observation $g\left(\bar{\mathbf{x}}'\hat{\boldsymbol{\theta}}\right)\hat{\theta}^{(l)}$. What the researcher is reporting must be clear.

2.2 Multinomial Choice: Multinomial Logit Model

To model a discrete choice over multiple alternatives, the usual choice in the literature has been the multinomial logit model developed in a series of works by McFadden (1974, 1981, 1989); McFadden and Train (2000) among others. The multinomial logit model is the major workhorse in a diverse context

[3] For a logistic pdf $g(\cdot)$ with location parameter 0 and scale parameter 1, the following holds:

$$\begin{aligned} g(z) &= \frac{\exp(z)}{1+\exp(z)} - \frac{\exp(2z)}{[1+\exp(z)]^2} \\ &= \frac{[1+\exp(z)]\exp(z) - \exp(2z)}{[1+\exp(z)]^2} \\ &= \frac{\exp(z) + \exp(2z) - \exp(2z)}{[1+\exp(z)]^2} \\ &= \frac{\exp(z)}{[1+\exp(z)]^2} \end{aligned}$$

and

$$G(z)[1-G(z)] = \frac{\exp(z)}{1+\exp(z)}\left[\frac{1}{1+\exp(z)}\right].$$

Taking the ratio yields $\frac{g(z)}{G(z)[1-G(z)]} = 1$.

of applied microeconomics when multiple and exclusive alternatives exist. A common way to derive the multinomial logistic choice probabilities is to begin from the additive i.i.d. type I extreme value distributed idiosyncratic utility shocks. We present some preliminary results on type I extreme value distribution in section 2.2.1, and then present our main results in the following subsections.

2.2.1 Preliminary Results: Type I Extreme Value Distribution and Its Properties

Definition. (Type I Extreme Value Distribution) $\epsilon_i \sim T1EV(\alpha)$ if ϵ_i follows the cdf

$$\begin{aligned}\Pr(\epsilon_i \leq \epsilon) &= F_\alpha(\epsilon) \\ &= \exp\left[-\exp\left[-(\epsilon-\alpha)\right]\right].\end{aligned}$$

Note this distribution is also referred to as a Gumbel distribution or double exponential distribution.[4] When $\alpha = 0$, the expectation of a type I extreme value random variable is the Euler-Mascheroni constant $\gamma \approx 0.5772$.

Lemma 2.2.1. *(Density Function of Type I Extreme Value Distribution) Let* $F_\alpha(\epsilon)$ be the cdf of the $T1EV(\alpha)$. Then, the pdf $f_\alpha(\epsilon) = \exp(\alpha-\epsilon) F_\alpha(\epsilon)$.

Lemma 2.2.2. *(Distribution of Maximum over i.i.d. T1EV Random Variables) Let $\epsilon_{i,j} \sim T1EV(\alpha_j)$, where $\epsilon_{i,j}$ are independent over j. Let $\alpha = \ln\left[\sum_{j=1}^{J}\exp(\alpha_j)\right]$. Then,*

$$\max_j \{\epsilon_{i,j}\} \sim T1EV(\alpha).$$

Proof. Let $\epsilon \in \mathbb{R}$. We have

$$\begin{aligned}\Pr\left(\max_{j\in\mathcal{J}}\{\epsilon_{i,j}\} \leq \epsilon\right) &= \prod_{j=1}^{J}\Pr(\epsilon_{i,j} \leq \epsilon) \\ &= \prod_{j=1}^{J}\exp\left[-\exp\left[-(\epsilon-\alpha_j)\right]\right] \\ &= \exp\left[-\sum_{j=1}^{J}\exp\left[-(\epsilon-\alpha_j)\right]\right] \\ &= \exp\left[-\exp(-\epsilon)\sum_{j=1}^{J}\exp(\alpha_j)\right] \\ &= \exp\left[-\exp(-\epsilon)\exp(\alpha)\right] \\ &= F_\alpha(\epsilon).\end{aligned}$$

□

[4] In principle, type I extreme value distribution is a two-parameter distribution, location and scale. We normalize the scale parameter as 1 because it cannot be identified in general.

Corollary. *(Expectation of Maximum over T1EV Random Variables) Let* $j \in \mathcal{J}$. *Let* $\epsilon_{i,j} \sim T1EV(0)$. *Let* $u_{i,j} = \delta_j + \epsilon_{i,j}$ *where* δ_j *is the additive deterministic component of choice j. Then,*

$$E\left[\max_{j\in\mathcal{J}}\left\{u_{i,j}\right\}\right] = \ln\left[\sum_{j\in\mathcal{J}}\exp\left(\delta_j\right)\right] + \gamma,$$

where the Euler-Mascheroni constant $\gamma \approx 0.5772$.

Lemma 2.2.3. *(Subtraction of Two Independent T1EV Random Variables) Suppose* $u_{i,j} \sim T1EV\left(\delta_j\right)$ *and* $u_{i,k} \sim T1EV\left(\delta_k\right)$ *where* $u_{i,j} \perp\!\!\!\perp u_{i,k}$. *Then,* $u_{i,j} - u_{i,k} \sim Logistic\left(\delta_j - \delta_k\right)$.

Proof. Let $F\left(u_j\right)$ be the cdf of $T1EV\left(\delta_j\right)$, and let $f_{\delta_j}\left(u_j\right)$ be the corresponding pdf. By Lemma 2.2.1,

$$\begin{aligned} f_{\delta_k}\left(u_k\right) &= \exp\left[-\left(u_k-\delta_k\right)\right]F_{\delta_k}\left(u_k\right) \\ &= \exp\left[-\left(u_k-\delta_k\right)\right]\exp\left[-\exp\left[-\left(u_k-\delta_k\right)\right]\right]. \end{aligned}$$

Thus,

$$\begin{aligned} & \Pr\left(u_j-u_k<u\right) \\ =\ & \Pr\left(u_j<u_k+u\right) \\ =\ & \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{u_k+u} f_{\delta_j}\left(u_j\right)du_j\right\} f_{\delta_k}\left(u_k\right)du_k \\ =\ & \int_{-\infty}^{\infty} f_{\delta_k}\left(u_k\right)F_{\delta_j}\left(u_k+u\right)du_k \\ =\ & \int_{-\infty}^{\infty}\exp\left[-\left(u_k-\delta_k\right)\right]F_{\delta_k}\left(u_k\right)F_{\delta_j}\left(u_k+u\right)du_k \\ =\ & \int_{-\infty}^{\infty}\exp\left[-\left(u_k-\delta_k\right)\right]\exp\left[-\exp\left[-\left(u_k-\delta_k\right)\right]\right]\exp\left[-\exp\left[-\left(u_k+u-\delta_j\right)\right]\right]du_k \\ =\ & \int_{-\infty}^{\infty}\exp\left[-\left(u_k-\delta_k\right)\right]\exp\left[-\exp\left[-\left(u_k-\delta_k\right)\right]-\exp\left[-\left(u_k+u-\delta_j\right)\right]\right]du_k \\ =\ & \int_{-\infty}^{\infty}\exp\left[-\left(u_k-\delta_k\right)\right]\exp\left[-\exp\left[-\left(u_k-\delta_k\right)\right]-\exp\left[-\left(u_k-\delta_k+u+\delta_k-\delta_j\right)\right]\right]du_k \\ =\ & \int_{-\infty}^{\infty}\exp\left[-\left(u_k-\delta_k\right)\right]\exp\left[-\exp\left[-\left(u_k-\delta_k\right)\right]\left\{1+\exp\left[-\left(u+\delta_k-\delta_j\right)\right]\right\}\right]du_k. \end{aligned}$$

Denote $a := \left\{1+\exp\left[-\left(u+\delta_k-\delta_j\right)\right]\right\}$ for notational simplicity.

$$\begin{aligned} & \int_{-\infty}^{\infty}\exp\left[-\left(u_k-\delta_k\right)\right]\exp\left[-a\exp\left[-\left(u_k-\delta_k\right)\right]\right]du_k \\ =\ & \frac{1}{a}\int_{-\infty}^{\infty}a\exp\left[-\left(u_k-\delta_k\right)\right]\exp\left[-a\exp\left[-\left(u_k-\delta_k\right)\right]\right]du_k \\ =\ & \frac{1}{1+\exp\left[-\left(u+\delta_k-\delta_j\right)\right]} \qquad (2.2.1) \\ =\ & \frac{\exp\left(u-\left(\delta_j-\delta_k\right)\right)}{\exp\left(u-\left(\delta_j-\delta_k\right)\right)+1}, \end{aligned}$$

which is a logistic cdf with mean $\left(\delta_j - \delta_k\right)$. Note that (2.2.1) follows because

$$a \exp\left[-\left(u_k - \delta_k\right)\right] \exp\left[-a \exp\left[-\left(u_k - \delta_k\right)\right]\right]$$

is the pdf of a type I extreme value distribution with scale parameter a^{-1}, which integrates to 1. □

2.2.2 Multinomial Simple Logit Model

Let i denote individual, and let j denote alternative where $j \in \mathcal{J} := \{1, 2, ..., J\}$. Consumer i is assumed to choose up to one product in the set of alternatives $\mathcal{J}$. That is, now the *consumer's choice is exclusive over the set of alternatives*.[5] $\mathcal{J}$ may contain product 0, which is most commonly interpreted as choosing an outside option.

The latent utility is modeled as

$$u_{i,j} = \delta_j + \epsilon_{i,j}, \tag{2.2.2}$$

where δ_j is the utility from the observed product characteristics of product j,[6] $\epsilon_{i,j}$ represents the unobserved idiosyncratic utility shocks. The most common functional form used is the linear utility specification $\delta_j = \mathbf{x}_j'\boldsymbol{\theta}$. When $\mathcal{J}$ contains the outside option, normalizing $\mathbf{x}_0 = \mathbf{0}$ so the mean utility of an outside option δ_0 is zero is common.

Analogous with the binary choice model, the probability of individual i choosing product j is

$$\begin{aligned}
\Pr\left(i \text{ Chooses } j\right) &= \Pr\left(u_{i,j} > u_{i,k}, \forall k \neq j\right) \\
&= \Pr\left(\delta_j + \epsilon_{i,j} > \delta_k + \epsilon_{i,k}, \forall k \neq j\right) \\
&= \Pr\left(\delta_j + \epsilon_{i,j} > \max_{k \neq j}\left\{\delta_k + \epsilon_{i,k}\right\}\right) \\
&= \Pr\left(\mathbf{x}_j'\boldsymbol{\theta} + \epsilon_{i,j} > \max_{k \neq j}\left\{\mathbf{x}_k'\boldsymbol{\theta} + \epsilon_{i,k}\right\}\right), \qquad (2.2.3)
\end{aligned}$$

where the last equality used the functional form $\delta_j = \mathbf{x}_j'\boldsymbol{\theta}$.

For the estimation, we assume the following to derive the closed-form probability of individual i choosing alternative j. Note MLM(1) and MLM(2) are often jointly abbreviated as $\epsilon_{i,j} \sim i.i.d.\ T1EV(0)$.

MLM(1) $\forall i, \forall k \neq j, \epsilon_{i,j} \perp\!\!\!\perp \epsilon_{i,k}$.

MLM(2) $\epsilon_{i,j} \sim T1EV(0)$.

[5] The setup and the assumptions on the data availability are different from the binary choice model we have discussed. In the binary choice model, the choice over the set of alternatives was not exclusive. We assumed the choice data on each alternative are available in the form of $\{0, 1\}$. However, if we consider $\{1\}$ as the set of alternatives where 0 denotes the choice of choosing nothing, the discrete choice model can be reformulated in terms of the multinomial logit model that we study in this subsection.

[6] δ_j may include the unobserved (to the econometrician) characteristics. We discuss including the unobserved characteristics in Chapter 3. For now, we take δ_j only composed with the observed characteristics.

Theorem 2.2.1. *(Multinomial Logit Likelihood) Suppose MLM(1) and MLM(2) holds. Then,*

$$\Pr\left(i \text{ Chooses } j\right) = \frac{\exp\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{\sum_{k\in\mathcal{J}}\exp\left(\mathbf{x}_k'\boldsymbol{\theta}\right)}. \tag{2.2.4}$$

Proof. Let $u_{i,j} = \delta_j + \epsilon_{i,j}$, $u_{i,-j} = \max_{k\neq j}\left\{\delta_k + \epsilon_{i,k}\right\}$. Let $\delta_{-j} = \ln\left[\sum_{k\neq j}\exp\left(\delta_k\right)\right]$. By Lemma 2.2.2, we know $u_{i,-j} \sim T1EV\left(\delta_{-j}\right)$. Under MLM(1) and MLM(2), we have

$$\begin{aligned}
\Pr\left(\delta_j + \epsilon_{i,j} \geq \max_k\left\{\delta_k + \epsilon_{i,k}\right\}\right) &= \Pr\left(u_{i,j} \geq u_{i,-j}\right) \\
&= \Pr\left(u_{i,-j} - u_{i,j} \leq 0\right) \\
&= \frac{\exp\left(\delta_j\right)}{\exp\left(\delta_{-j}\right) + \exp\left(\delta_j\right)} \qquad (2.2.5) \\
&= \frac{\exp\left(\delta_j\right)}{\sum_{k\in\mathcal{J}}\exp\left(\delta_k\right)}.
\end{aligned}$$

(2.2.5) is by applying (2.2.1) with $u = 0$. Substituting $\delta_k = \mathbf{x}_k'\boldsymbol{\theta}\ \forall k$, we get the desired result. □

It is straightforward from 2.2.3 that when $\mathcal{J} = \{0, 1\}$ and $\mathbf{x}_0 = \mathbf{0}$, (2.2.3) boils down to the logit likelihood.

Corollary. *(Multinomial Logit Likelihood with Nonzero Mean Parameter) Suppose MLM(1) holds. Suppose MLM(2) is replaced with.* $\epsilon_{i,j} \sim T1EV\left(\alpha_j\right)$. *Then,*

$$\Pr\left(i \text{ Chooses } j\right) = \frac{\exp\left(\mathbf{x}_j'\boldsymbol{\theta} + \alpha_j\right)}{\sum_{k\in\mathcal{J}}\exp\left(\mathbf{x}_k'\boldsymbol{\theta} + \alpha_k\right)}.$$

The individual choice probability equation (2.2.4) derived from the i.i.d. additive type I extreme value shocks on the preferences plays a central role in many different contexts. The choice probability itself can be used as a likelihood for the maximum-likelihood estimation, or can be equated with data that approximate the individual choice probabilities. We study the models of the latter type in depth in section 3.2.

Now suppose we have the individual choice data $\left\{y_{i,j}, \mathbf{x}_j\right\}_{i\in\mathcal{I}, j\in\mathcal{J}}$. The likelihood of observing the data is

$$L\left(\left\{y_{i,j}, \mathbf{x}_j\right\}_{i\in\mathcal{I}, j\in\mathcal{J}} | \boldsymbol{\theta}\right) = \prod_{i\in\mathcal{I}}\left\{\prod_{j\in\mathcal{J}}\Pr\left(i \text{ Chooses } j\right)^{\mathbf{1}(i \text{ Chooses } j)}\right\},$$

where $\Pr\left(i \text{ Chooses } j\right)$ is as in (2.2.4). The log-likelihood and score function, which is required for the maximum-likelihood estimation, can be derived as usual. We emphasize that the model parameter $\boldsymbol{\theta}$ can be estimated using maximum likelihood only when δ_j in (2.2.2) does not contain any unknown or unobservables. A more sophisticated method is required when an unobservable is included. We discuss some of them in section 2.3.4 and in Chapter 3.

2.2.3 Independence of Irrelevant Alternatives and Nested Logit

Consider the ratio of multinomial logit choice probabilities :

$$\frac{\Pr(i \text{ Chooses } j)}{\Pr(i \text{ Chooses } k)} = \frac{\exp\left(\mathbf{x}_j'\boldsymbol{\theta}\right)}{\exp\left(\mathbf{x}_k'\boldsymbol{\theta}\right)}. \tag{2.2.6}$$

The ratio (2.2.6), often referred to as the odds ratio of choices j and k, is constant regardless of the utility from other choices. The property is called the independence of irrelevant alternatives (IIA henceforth) property, which is pioneered by Luce (1959). IIA restricts the substitution pattern over the alternatives substantially. We study how and why IIA may not be desirable in section 3.2.3. What we want to emphasize at this point is that given that the mean utility $\delta_j = \mathbf{x}_j'\boldsymbol{\theta}$, the individual choice probability equation (2.2.4) is the only legitimate choice probability equation that satisfies (2.2.6) for each alternative in $\mathcal{J}$ and sums up to 1. In that sense, the individual choice probability equation (2.2.4) can be derived from IIA, not from the additive idiosyncratic type I extreme value distributed shocks.

One way to avoid the IIA when individual choice-level data are available is nesting the choice set and modeling the individual's choice in multiple stages. Suppose the choice set $\mathcal{J}$ can be divided into S modules. Each module is denoted by $\mathcal{B}_s$, where $s \in \{1, 2, ..., S\}$. If the joint distribution of $\left\{\epsilon_{i,j}\right\}_{j\in\mathcal{J}}$ is of the form

$$F\left(\left\{\epsilon_j\right\}_{j\in\mathcal{J}}\right) = \exp\left\{-\sum_{s=1}^{S}\alpha_s\left[\sum_{j\in\mathcal{B}_s}\exp\left(-\rho_s^{-1}\epsilon_j\right)\right]^{\rho_s}\right\},$$

it can be shown that the individual choice probabilities have the following closed-form formula:

$$\begin{aligned}
\Pr(i \text{ Chooses } \mathcal{B}_s) &= \frac{\alpha_s\left(\sum_{k\in\mathcal{B}_s}\exp\left(\rho_s^{-1}\delta_k\right)\right)^{\rho_s}}{\sum_{\tau=1}^{S}\alpha_\tau\left(\sum_{k\in\mathcal{B}_\tau}\exp\left(\rho_\tau^{-1}\delta_k\right)\right)^{\rho_\tau}} && (2.2.7)\\
\Pr(i \text{ Chooses } j | i \text{ Chooses } \mathcal{B}_s) &= \frac{\exp\left(\rho_s^{-1}\delta_j\right)}{\sum_{k\in\mathcal{B}_s}\exp\left(\rho_s^{-1}\delta_k\right)} && (2.2.8)\\
\Pr(i \text{ Chooses } j) &= \Pr(i \text{ Chooses } j | i \text{ Chooses } \mathcal{B}_s)\Pr(i \text{ Chooses } \mathcal{B}_s). && (2.2.9)
\end{aligned}$$

$\{\alpha_s, \rho_s\}_{s=1}^{S}$ are the parameters to be estimated. The nesting structures can also be extended to more than two stages in an analogous way.

Given the utility specification $\delta_j = \mathbf{x}_j'\boldsymbol{\theta} + \epsilon_{i,j}$, the nested logit model (2.2.7)-(2.2.9) can be estimated either by the full maximum likelihood or by the two-stage method. Although both methods yield the consistent estimates, the asymptotic variance formulas are different. For further discussions on implementing nested logit with the two-stage methods, see McFadden (1981).

Nested logit allows for a more complex choice structure than the simple logit, but two major weaknesses remain in the model (2.2.7)-(2.2.9). First, it does not exhibit IIA across modules, but it still exhibits IIA within a module. Next, to implement the nested logit model, the econometrician has to impose a prior knowledge on the choice structures, and thus on the composition of modules.

If one wants to just avoid the IIA, one could consider using the multinomial probit. Another benefit of using the multinomial probit is that the Gaussian error term naturally allows for the correlation of errors across alternatives. However, it leads to greater computational burden than multinomial logit, because the likelihood has no closed-form solution, and it usually involves evaluating the integral numerically. That is one reason why the somewhat restrictive simple multinomial logit model is still the major workhorse in practice.

2.2.4 Discussion

Historically, the development of the multinomial simple logit and nested logit model has taken the opposite direction as our presentation. The type I extreme value distribution is carefully reverse engineered to yield the odds ratio of the form (2.2.4).[7] Generalizations of the simple multinomial logit model to the generalized extreme-value class models[8] are made by McFadden (1974, 1981, 1989); Cardell (1997). As a last remark, the generalized extreme-value class choice models are often referred to as the random utility maximization (RUM) model, to emphasize the connection between the resulting choice probabilities and the axioms of stochastic choice, which is a set of axioms to rationalize a stochastic choice.

2.3 Dynamic Discrete Choice

Agents are forward-looking when state variables affect the current and future payoffs, and they change with respect to accumulation flow. The dynamic discrete-choice models are a combination of discrete choice with forward-looking agents. In this section, we study the extension of the static discrete-choice framework to incorporate the dynamics. We focus on the setups in which the modeled dynamics are stationary, so the value function boils down to a two-period Bellman's equation.

We first study the method of Rust (1987)'s bus-engine-replacement problem, which has become the standard baseline dynamic discrete-choice framework. Then, we discuss the key idea of Hotz and Miller (1993); Hotz, Miller, Sanders, and Smith (1994), which is based on inverting the conditional choice probabilities, and then we study the nested pseudo-likelihood method by Aguirregabiria and Mira (2002), which is based on finding the dual problem of the value function fixed-point equations in the conditional choice probability space. Lastly, we overview Arcidiacono and Miller (2011)'s method to accommodate the unobserved state variables into Rust (1987)'s framework. The single-agent dynamic discrete-choice framework studied in this section will be extended and applied to richer settings in the later chapters.

2.3.1 Full-Solution Method with Fixed-Point Iteration

2.3.1.1 Formulation

In this subsection, we study the full maximum-likelihood method developed by Rust (1987). Consider an infinite-horizon stochastic stationary dynamic utility-maximization problem, which has a two-period

[7] Consider the odds ratios of the choice set $\mathcal{J} = \{0, 1\}$ with $\delta_0 = \epsilon_{i,0}$, $\delta_1 = \mathbf{x}_1'\boldsymbol{\theta} + \epsilon_{i,1}$, where $\epsilon_{i,j} \sim$ *i.i.d.* $T1EV(0)$. Lemma 2.2.3 yields that $\epsilon_{i,1} - \epsilon_{i,0}$ follows standard logistic distribution.

[8] Generalized extreme-value class includes nested logit.

Bellman equation representation. The agent maximizes the discounted sum of the utilities. Let $\mathcal{J} = \{1, ..., J\}$ be the set of alternatives, where the agent's action $a_t \in \mathcal{J}$.[9] Let $\mathbf{s}_t \, (\in \mathcal{S})$ be a state vector that has both an observed component $x_t \, (\in \mathcal{X})$ and an unobserved component $\boldsymbol{\epsilon}_t \equiv \left(\epsilon_{j,t}\right)_{j \in \mathcal{J}}$ to the econometrician but not to the agent. That is, the agent fully observes the realization of $\boldsymbol{\epsilon}_t$ before she takes the action in each period t. We assume $\mathcal{X}$, the support of x_t, is finite, that is, $\mathcal{X} = \{1, 2, ..., X\}$.[10] For each j, t pair, a corresponding observed component and the unobserved utility shock exist, and thus the state vector is $\mathbf{s}_t = (x_t, \epsilon_{1,t}, ..., \epsilon_{J,t})$. The agent solves

$$\max_{\{a_t\}_{t=1}^{\infty}} E\left[\sum_{t=1}^{\infty} \beta^{t-1} u\left(x_t, a_t, \boldsymbol{\epsilon}_t; \boldsymbol{\theta}\right)\right],$$

where $\beta \in (0, 1)$ is the discount factor, which has to be a prior knowledge of the econometrician.[11] Two sets of parameter vectors, $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, exist. $\boldsymbol{\theta}$ is the parameter vector that governs the utility function of the agent, and $\boldsymbol{\varphi}$ is the parameter vector that governs the state-transition probability $\Pr\left(x_{t+1}, \boldsymbol{\epsilon}_{t+1} | x_t, \boldsymbol{\epsilon}_t, a_t; \boldsymbol{\varphi}\right)$. Both parameter vectors $(\boldsymbol{\theta}, \boldsymbol{\varphi})$ are known to the agent, consistent estimation of which is our goal. The following assumptions on the unobserved state vector are imposed henceforth.

DDC(1) (Additive Separability of Preferences) Let $u\left(\mathbf{s}_t, a_t = j; \boldsymbol{\theta}\right)$ be the utility of choosing action j at state $\mathbf{s}_t$. Then,

$$u\left(\mathbf{s}_t, a_t = j; \boldsymbol{\theta}\right) = \bar{u}_j\left(x_t; \boldsymbol{\theta}\right) + \epsilon_{j,t}.$$

DDC(2) ($\boldsymbol{\epsilon}_t$ is i.i.d. Over t) $\forall \boldsymbol{\epsilon}_t \left(\in \mathbb{R}^J\right)$,

$$\Pr\left(\boldsymbol{\epsilon}_{t+1} | \boldsymbol{\epsilon}_t\right) = \Pr\left(\boldsymbol{\epsilon}_{t+1}\right).$$

DDC(3) (Conditional Independence of State-Transition Probabilities) Given (x_t, a_t) observed, $x_{t+1} \perp\!\!\!\perp (\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t+1})$, which implies

$$\begin{aligned}
\Pr\left(x_{t+1}, \boldsymbol{\epsilon}_{t+1} | x_t, \boldsymbol{\epsilon}_t, a_t; \boldsymbol{\varphi}\right) &\equiv \Pr\left(\boldsymbol{\epsilon}_{t+1} | x_{t+1}, x_t, \boldsymbol{\epsilon}_t, a_t; \boldsymbol{\varphi}\right) \Pr\left(x_{t+1} | x_t, \boldsymbol{\epsilon}_t, a_t; \boldsymbol{\varphi}\right) \\
&= \Pr\left(\boldsymbol{\epsilon}_{t+1}\right) \Pr\left(x_{t+1} | x_t, a_t; \boldsymbol{\varphi}\right).
\end{aligned}$$

Roughly stated, the assumptions are imposed to restrict the influence of $\boldsymbol{\epsilon}_t$ on $(x_{t+1}, \boldsymbol{\epsilon}_{t+1})$ such that $\boldsymbol{\epsilon}_t$ affects x_{t+1} only through a_t, and it does not affect $\boldsymbol{\epsilon}_{t+1}$ at all. It allows us to derive the conditional choice probability of the form (2.3.1).

As in the static discrete-choice problem, our goal is to derive the closed-form likelihoods using the properties of the i.i.d. type I extreme value distributed utility shocks. To that end, we expect the condi-

[9] Rust (1987) considered only the binary choice set, but we allow for finite alternatives. The extension from binary to finite alternatives is straightforward as we studied before.

[10] The discrete state space assumption is useful not only because it simplifies the illustration, but also because it simplifies the estimation algorithm. We note the assumption can be relaxed to the compact support with the continuous value function. However, an approximation has to be taken during the estimation when the state space is continuous, which may complicate the estimation algorithm and incur an additional source of error in the estimates.

[11] Usually, β is not separately identified from the data. We discuss this point later.

tional choice probability that action j is chosen in period t tis

$$\begin{aligned}\Pr\left(a_t = j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right) &= \frac{\exp\left(\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)}{\sum_{k\in\mathcal{J}}\exp\left(\bar{v}_k\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)} \qquad (2.3.1)\\ &= \frac{\exp\left(\bar{u}_j\left(x_t;\boldsymbol{\theta}\right)+\beta E\left[v\left(x_{t+1},\boldsymbol{\epsilon}_{t+1};\boldsymbol{\theta},\boldsymbol{\varphi}\right)|a_t = j, x_t,\boldsymbol{\epsilon}_t;\boldsymbol{\varphi}\right]\right)}{\sum_{k\in\mathcal{J}}\exp\left(\bar{u}_k\left(x_t;\boldsymbol{\theta}\right)+\beta E\left[v\left(x_{t+1},\boldsymbol{\epsilon}_{t+1};\boldsymbol{\theta},\boldsymbol{\varphi}\right)|a_t = k, x_t,\boldsymbol{\epsilon}_t;\boldsymbol{\varphi}\right]\right)}, \qquad (2.3.2)\end{aligned}$$

where $v_j(\cdot)$ is the choice-specific value function, $\bar{v}_j(\cdot)$ is the choice-specific expected value function, and $\bar{u}_j(\cdot)$ is the choice-specific expected utility function. The precise definition of $v_j(\cdot)$ and $\bar{v}_j(\cdot)$ will be given later. The likelihood of observing the single time-series data $\{(x_t, a_t)\}_{t\in\mathcal{T}}$ where $\mathcal{T} = \{1, 2, ..., T\}$ is

$$\prod_{t=2}^{T}\left\{\prod_{j\in\mathcal{J}}\Pr\left(a_t = j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)^{\mathbf{1}(a_t=j)}\right\}\Pr\left(x_t|x_{t-1},a_{t-1};\boldsymbol{\varphi}\right). \qquad (2.3.3)$$

$\Pr\left(x_{t+1}|x_t,a_t;\boldsymbol{\varphi}\right)$ is the stationary transition probability of the state variables. We assume $\Pr\left(x_{t+1}|x_t,a_t;\boldsymbol{\varphi}\right)$ is known to the econometrician up to parameter $\boldsymbol{\varphi}$.

2.3.1.2 Information and Timeline

We summarize the informational assumption and the timeline of the decision here. The agent and the econometrician both know the shapes of $u\left(x_t, a_t, \boldsymbol{\epsilon}_t;\boldsymbol{\theta}\right)$, $\Pr\left(\boldsymbol{\epsilon}_{t+1}\right)$ and $\Pr\left(x_{t+1}|x_t,a_t;\boldsymbol{\varphi}\right)$, respectively. β is also known to both the agent and the econometrician. The value of $(\boldsymbol{\theta},\boldsymbol{\varphi})$ and the realization of $\boldsymbol{\epsilon}_t$ is not known to the econometrician but is known to the agent. At the beginning of each period t, the agent observes x_t and the full realization of $\boldsymbol{\epsilon}_t$ and then makes the decision a_t.

2.3.1.3 The Bellman-Equation Representation

The above three assumptions DDC(1) through DDC(3) yield the stationary, Markov state-transition probability $\Pr\left(x_{t+1}|x_t,a_t;\boldsymbol{\varphi}\right)$. Under the regularity conditions including the infinite time horizons and $\beta < 1$, Blackwell's theorem asserts that stationary and Markov decision rule exists that depends only on $(x_t, \boldsymbol{\epsilon}_t)$, and such a decision rule is unique.

In addition to DDC(2), assume $\epsilon_{j,t}$ follows the mean-zero i.i.d. type I extreme value distribution.[12] Assuming the regularity conditions for the stochastic dynamic programming are satisfied,[13] the problem

[12]Not the location parameter zero. Recall that the location parameter zero Type I extreme value distribution has mean$\approx$ 0.5772.

[13]These include the stationarity of the transition probability among others. For details, see, e.g., Stokey and Lucas (1989).

has a Bellman-equation representation:

$$
\begin{aligned}
& v\left(x_t, \boldsymbol{\epsilon}_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \\
= & \max_{a_t \in \mathcal{J}} \left\{u\left(x_t, a_t, \boldsymbol{\epsilon}_t; \boldsymbol{\theta}\right) + \beta E\left[v\left(x_{t+1}, \boldsymbol{\epsilon}_{t+1}; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) | a_t, x_t, \boldsymbol{\epsilon}_t; \boldsymbol{\varphi}\right]\right\} && (2.3.4) \\
= & \max_{a_t \in \mathcal{J}} \left\{u\left(x_t, a_t, \boldsymbol{\epsilon}_t; \boldsymbol{\theta}\right) + \beta \int \int v\left(x_{t+1}, \boldsymbol{\epsilon}_{t+1}; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \Pr\left(d\boldsymbol{\epsilon}_{t+1}\right) \Pr\left(dx_{t+1} | x_t, a_t; \boldsymbol{\varphi}\right)\right\} && (2.3.5) \\
= & \max_{j \in \mathcal{J}} \left\{\bar{u}_j\left(x_t; \boldsymbol{\theta}\right) + \epsilon_{j,t} + \beta \int \int v\left(x_{t+1}, \boldsymbol{\epsilon}_{t+1}; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \Pr\left(d\boldsymbol{\epsilon}_{t+1}\right) \Pr\left(dx_{t+1} | x_t, a_t = j; \boldsymbol{\varphi}\right)\right\} && (2.3.6) \\
\equiv & \max_{j \in \mathcal{J}} \left\{\bar{v}_j\left(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) + \epsilon_{j,t}\right\}. && (2.3.7)
\end{aligned}
$$

Note we invoked DDC(3) and DDC(2) in (2.3.5), DDC(1) in (2.3.6). (2.3.7) follows from defining $\bar{v}_j\left(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right)$ as the deterministic part of the choice-specific value function.[14] Then, the mean-zero i.i.d. Type-I extreme value assumption on $\epsilon_{j,t}$ yields

$$
\begin{aligned}
& \bar{v}_j\left(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \\
= & \bar{u}_j\left(x_t; \boldsymbol{\theta}\right) + \beta \int \int v\left(x_{t+1}, \boldsymbol{\epsilon}_{t+1}; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \Pr\left(d\boldsymbol{\epsilon}_{t+1}\right) \Pr\left(dx_{t+1} | x_t, a_t = j, \boldsymbol{\varphi}\right) \\
= & \bar{u}_j\left(x_t; \boldsymbol{\theta}\right) + \beta \int \int \max_{a_{t+1} \in \mathcal{J}} \left\{v_j\left(x_{t+1}, \epsilon_{j,t+1}; \boldsymbol{\theta}, \boldsymbol{\varphi}\right)\right\} \Pr\left(d\boldsymbol{\epsilon}_{t+1}\right) \Pr\left(dx_{t+1} | x_t, a_t = j, \boldsymbol{\varphi}\right) \\
= & \bar{u}_j\left(x_t; \boldsymbol{\theta}\right) + \beta \int \ln\left(\sum_{k \in \mathcal{J}} \exp \bar{v}_k\left(x_{t+1}; \boldsymbol{\theta}, \boldsymbol{\varphi}\right)\right) \Pr\left(dx_{t+1} | x_t, a_t = j, \boldsymbol{\varphi}\right). && (2.3.8)
\end{aligned}
$$

(2.3.8) is a fixed-point equation, and solving it yields the value of $\bar{v}_j\left(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right)$ for each possible x_t given $(\boldsymbol{\theta}, \boldsymbol{\varphi})$. To see the fact more clearly, consider the stacked matrix form of (2.3.8) over the alternatives $\mathcal{J}$ and the possible states $\mathcal{X}$:

$$
\begin{aligned}
& \operatorname{vec}\begin{pmatrix} \bar{v}_1\left(1; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) & \cdots & \bar{v}_J\left(1; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \\ \vdots & \ddots & \vdots \\ \bar{v}_1\left(X; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) & \cdots & \bar{v}_J\left(X; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \end{pmatrix} \\
= & \operatorname{vec}\begin{pmatrix} \bar{u}_1\left(1; \boldsymbol{\theta}\right) & \cdots & \bar{u}_J\left(1; \boldsymbol{\theta}\right) \\ \vdots & \ddots & \vdots \\ \bar{u}_1\left(X; \boldsymbol{\theta}\right) & \cdots & \bar{u}_J\left(X; \boldsymbol{\theta}\right) \end{pmatrix} + \beta \mathbf{P}\left(x_{t+1} | x_t, a_t; \boldsymbol{\varphi}\right) \operatorname{vec}\begin{pmatrix} \bar{w}\left(1; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) & \cdots & \bar{w}\left(1; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \\ \vdots & \ddots & \vdots \\ \bar{w}\left(X; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) & \cdots & \bar{w}\left(X; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \end{pmatrix} && (2.3.9)
\end{aligned}
$$

[14]Existing literature often denote the choice specific value function as $\bar{v}_j\left(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) + \epsilon_{j,t} - \bar{u}_j\left(x_t; \boldsymbol{\theta}\right)$ in our notation. If one wants to follow the conventional notation, the conditional choice probabilities and the fixed point equation below should be adjusted accordingly. In particular, the fixed point equation should be altered to numerically solve for the functions $\bar{w}\left(\cdot; \boldsymbol{\theta}, \boldsymbol{\varphi}\right)$ in our notations. This was the original algorithm of Rust (1987), followed by many others.

where $\mathbf{P}(x_{t+1}|x_t,a_t;\boldsymbol{\varphi})$ returns the state-transition probability matrix given (x_t,a_t),[15] and $\bar{w}(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}) = \ln\left(\sum_{k\in\mathcal{J}}\exp\bar{v}_k(x_t;\boldsymbol{\theta},\boldsymbol{\varphi})\right)$. Because $\beta<1$, the fixed-point equation (2.3.9) has a solution, which is unique as well. Solving (2.3.9) for $\bar{v}_j(x_t;\boldsymbol{\theta},\boldsymbol{\varphi})$ and plugging back into (2.3.3) for each $(\boldsymbol{\theta},\boldsymbol{\varphi})$ fully specifies the likelihood.

Note that the additive i.i.d. Type I extreme value assumption simplifies the derivation of the log-likelihood (2.3.10) substantially in two distinct aspects: (i) the conditional choice probability equation (2.3.1) has such a simple closed-form solution, and (ii) the expected value function (2.3.8) also has the simple closed form.

2.3.1.4 Estimation Algorithms: Nested Fixed Point (NFP) and Mathematical Programming with Equilibrium Constraints (MPEC)

The goal of maximum-likelihood estimation is to find the parameter values $(\boldsymbol{\theta},\boldsymbol{\varphi})$ such that

$$\begin{aligned}
&(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\varphi}})\\
=\;& \arg\max_{(\boldsymbol{\theta},\boldsymbol{\varphi})}\prod_{t=2}^{T}\left\{\prod_{j\in\mathcal{J}}\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)^{\mathbf{1}(a_t=j)}\right\}\Pr\left(x_t|x_{t-1},a_{t-1};\boldsymbol{\varphi}\right)\\
=\;& \arg\max_{(\boldsymbol{\theta},\boldsymbol{\varphi})}\sum_{t=2}^{T}\left\{\sum_{j\in\mathcal{J}}\mathbf{1}\left(a_t=j\right)\ln\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right\}+\sum_{t=2}^{T}\ln\Pr\left(x_t|x_{t-1},a_{t-1};\boldsymbol{\varphi}\right)\\
=\;& \arg\max_{(\boldsymbol{\theta},\boldsymbol{\varphi})}\sum_{t=2}^{T}\left\{\sum_{j\in\mathcal{J}}\mathbf{1}\left(a_t=j\right)\ln\left(\frac{\exp\left(\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)}{\sum_{k\in\mathcal{J}}\exp\left(\bar{v}_k\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)}\right)\right\}+\sum_{t=2}^{T}\ln\Pr\left(x_t|x_{t-1},a_{t-1};\boldsymbol{\varphi}\right)
\end{aligned}\tag{2.3.10}$$

holds. We emphasize again that $\Pr(x_t|x_{t-1},a_{t-1};\boldsymbol{\varphi})$ is known up to $\boldsymbol{\varphi}$.

The original nested fixed-point (NFP henceforth) algorithm suggested by Rust (1987) is based on computing the likelihood value $\Pr(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi})$ at each guess of $(\boldsymbol{\theta},\boldsymbol{\varphi})$ by solving the set of fixed-point equations (2.3.9). The inner loop of NFP consists of iterating over the values of $\bar{v}_j(x_t;\boldsymbol{\theta},\boldsymbol{\varphi})$ for a given guess of $(\boldsymbol{\theta},\boldsymbol{\varphi})$. The outer loop consists of guessing a new value of $(\boldsymbol{\theta},\boldsymbol{\varphi})$ to maximize the objective function. As one may easily imagine, the process is computationally very intensive, especially when β is close to 1 and the size of the state space is large.[16]

Recent development of the modern derivative-based numerical optimizers provided another solution that is possibly more accurate, and computationally much less intensive. The key idea is to convert the unconstrained likelihood maximization problem (2.3.10) to the constrained maximization problem by adding more parameters $\left\{\bar{v}_j(x)\right\}_{\forall j\in\mathcal{J},x\in\mathcal{X}}$ that the optimizer may freely adjust. To that end, the prob-

[15]This should be a sort of block-diagonal matrix such that

$$\begin{pmatrix}
\mathbf{P}\left(x_{t+1}|x_t,a_t=1;\boldsymbol{\varphi}\right) & \mathbf{O} & \cdots & \mathbf{O}\\
\mathbf{O} & \mathbf{P}\left(x_{t+1}|x_t,a_t=2;\boldsymbol{\varphi}\right) & \cdots & \mathbf{O}\\
\vdots & \vdots & \ddots & \vdots\\
\mathbf{O} & \mathbf{O} & \cdots & \mathbf{P}\left(x_{t+1}|x_t,a_t=J;\boldsymbol{\varphi}\right)
\end{pmatrix},$$

where each $\mathbf{P}(x_{t+1}|x_t,a_t=j;\boldsymbol{\varphi})$ is a $J\times X$ matrix.

[16]Recall the fixed-point equation is the contraction mapping of modulus β.

lem may be converted to

$$
\begin{aligned}
&\arg \max_{\left(\boldsymbol{\theta},\boldsymbol{\varphi},\left\{\bar{v}_j(x)\right\}_{\forall j\in\mathcal{J},x\in\mathcal{X}}\right)} \sum_{t=2}^{T}\left\{\sum_{j\in\mathcal{J}}\mathbf{1}\left(a_t=j\right)\ln\left(\frac{\exp\left(\bar{v}_j\left(x_t=x\right)\right)}{\sum_{k\in\mathcal{J}}\exp\left(\bar{v}_k\left(x_t=x\right)\right)}\right)\right\}+\sum_{t=2}^{T}\ln\Pr\left(x_t|x_{t-1},a_{t-1};\boldsymbol{\varphi}\right)\\
s.t.\quad & \bar{v}_j\left(x_t=x\right)=\bar{u}_j\left(x_t=x;\boldsymbol{\theta}\right)+\beta\int\ln\left(\sum_{k\in\mathcal{J}}\exp\bar{v}_k\left(x_{t+1}=x'\right)\right)\Pr\left(dx_{t+1}=x'|x_t=x,a_t=j;\boldsymbol{\varphi}\right)\\
&\forall j\in\mathcal{J},x,x'\in\mathcal{X},
\end{aligned}
$$

which yields the same solution with (2.3.10). This approach is referred to as mathematical programming with equilibrium constraints (MPEC henceforth), developed by Su and Judd (2012). MPEC conducts the maximum-likelihood optimization at one step, which does not require iterating over inner and outer loops. The accuracy gain of MPEC comes from removing the numerical errors of the inner loop that propagates to the outer loop. Notice the number of control variables has drastically increased because $\bar{v}_j\left(x_t=x\right)$ are treated as free parameters, which the optimizer can adjust. But it does not increase the computation time, or even reduces it drastically. The speed advantage of MPEC comes from not iterating for the fixed-point equation at every guess of $(\boldsymbol{\theta},\boldsymbol{\varphi})$. Unlike in the nested fixed-point algorithm, the fixed-point equation has to be satisfied only at the optimum, and modern numerical optimizers allow for the constraint violations during the trajectory. Su and Judd (2012) report that the speed advantage can be an order of 10^3. Contrarily, Iskhakov, Lee, Rust, Schjerning, and Seo (2016) report that the nested fixed-point algorithms combined with the Newton-Kantorovich iterations can be as fast as MPEC. Finding a faster and more reliable algorithm for this type of problem is an ongoing area of research. We revisit the use of this constrained optimization approach in section 3.2.2.3 in the context of demand estimation.

2.3.2 Estimation with Conditional Choice Probability Inversion

2.3.2.1 Choice Probability Inversion

The choice probability inversion theorem and the estimators for finite-horizon models were developed in Hotz and Miller (1993). The major advantage of their method is that evaluating the fixed-point equation is unnecessary, even when depending on the estimation scheme. We review a simplified version for illustration.

Recall the conditional choice probability equation (2.3.1) under the type I extreme value assumption:

$$
\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)=\frac{\exp\left(\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)}{\sum_{k\in\mathcal{J}}\exp\left(\bar{v}_k\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)}. \tag{2.3.11}
$$

The major challenge associated with this conditional choice probability equation is the difficulty and computational intensity to estimate the continuation values $E\left[v\left(x_{t+1},\boldsymbol{\epsilon}_{t+1};\boldsymbol{\theta},\boldsymbol{\varphi}\right)|a_t=k,x_t,\boldsymbol{\epsilon}_t;\boldsymbol{\varphi}\right]$ for each possible choice k.

The breakthrough of Hotz and Miller (1993) comes from the observation that a sample analogue of $\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ can be observed from the data under the stationarity assumption. Given that a long enough time series of choices are observed, those conditional choice probabilities can be approximated precisely from data. Denote $\hat{\Pr}(a_t = j|x_t)$ and $\hat{v}_j(x_t)$ by the sample analogue of $\Pr(a_t = j|x_t; \boldsymbol{\theta}, \boldsymbol{\varphi})$ and $\bar{v}_j(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi})$, respectively.[17][18] Then, we can formulate a sample analogue of (2.3.11) as

$$\hat{\Pr}(a_t = j|x_t) = \frac{\exp(\hat{v}_j(x_t))}{\sum_{k\in\mathcal{J}} \exp(\hat{v}_k(x_t))}.$$

Denote J as the reference choice, or the outside option, of which the choice-specific expected value $\bar{v}_J(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi})$ is normalized to 0. We have

$$\ln\left(\frac{\hat{\Pr}(a_t = j|x_t)}{\hat{\Pr}(a_t = J|x_t)}\right) = \hat{v}_j(x_t) - 0 \tag{2.3.12}$$

for each possible state x_t. This reveals the values of $\hat{v}_j(x_t)$ relative to $\hat{v}_J(x_t)$ for each j, which will be taken as data during the estimation. Note the level of $\hat{v}_j(x_t)$'s cannot be identified, because the conditional choice probability is invariant to the uniform location shift of $\hat{v}_j(x_t)$. We return to the identification problem of these models in section 2.3.5.

2.3.2.2 Maximum-Likelihood and Minimum Distance Estimators

To form $\bar{v}_j(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}) - \bar{v}_J(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi})$, we iterate the value function once by plugging $\hat{v}_j(x_t)$ back into the expectations. It yields

$$\begin{aligned} & \bar{v}_j(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}) - \bar{v}_J(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}) \\ = \; & \bar{u}_j(x_t; \boldsymbol{\theta}) - \bar{u}_J(x_t; \boldsymbol{\theta}) \\ & + \beta \int \ln\left(\sum_{k\in\mathcal{J}} \exp \hat{v}_k(x_{t+1})\right) \left\{\Pr(dx_{t+1}|x_t, a_t = j, \boldsymbol{\varphi}) - \Pr(dx_{t+1}|x_t, a_t = J, \boldsymbol{\varphi})\right\}. \end{aligned} \tag{2.3.13}$$

Then, one can proceed with either maximum likelihood, or with the moment matching and minimum distance estimation. The maximum-likelihood estimator is

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\varphi}}) = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\varphi}} \prod_t \prod_{j\in\mathcal{J}} \left\{\frac{\exp(\bar{v}_j(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}))}{\sum_{k\in\mathcal{J}} \exp(\bar{v}_k(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}))}\right\}^{\mathbf{1}(a_t = j|x_t)}. \tag{2.3.14}$$

The moment matching estimator is

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\varphi}}) = \arg\min_{\boldsymbol{\theta}, \boldsymbol{\varphi}} \sum_t \left\{\ln\left(\frac{\hat{\Pr}(a_t = j|x_t)}{\hat{\Pr}(a_t = J|x_t)}\right) - \bar{v}_j(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi})\right\}^2, \tag{2.3.15}$$

[17] Note that $\hat{\Pr}(a_t = j|x_t)$ and $\hat{v}_j(x_t)$ are no longer functions of $\boldsymbol{\theta}, \boldsymbol{\varphi}$, because we treat them to be directly observable from data.

[18] Several possibilities exist for forming these, including a simple frequency estimator and Nadaraya-Watson-type kernel-smoothing estimator.

or alternatively,

$$\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\varphi}}\right) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\varphi}} \sum_t \left\{ \frac{\exp\left(\bar{v}_j\left(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right)\right)}{\sum_{k\in\mathcal{J}} \exp\left(\bar{v}_k\left(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right)\right)} - \mathbf{1}\left(a_t = j | x_t\right) \right\}^2. \tag{2.3.16}$$

The estimators defined in (2.3.14)-(2.3.16) will be $\sqrt{T}$-consistent and asymptotically normal estimators, given that the mean-zero i.i.d. type I error assumption is valid and the errors from the first-stage estimators $\hat{v}_j\left(x_t\right)$ vanish at a fast enough rate.

2.3.2.3 Discussion

Hotz, Miller, Sanders, and Smith (1994) suggest forward simulation to find the right-hand side of (2.3.13). The discreteness of the action set is essential in both Hotz and Miller (1993); Hotz, Miller, Sanders, and Smith (1994), because they compute the discounted sum of utilities using the tree structure. Although we assumed $\epsilon_{j,t}$ follows the i.i.d. type I extreme value distribution, the inversion result holds for more general distributions supported on the entire real line. The idea to invert the choice probabilities was exploited in diverse contexts, including demand estimation, static/dynamic game model estimation, and so on.

2.3.3 Nested Pseudo-Likelihood Estimation

2.3.3.1 Duality of Dynamic Discrete-Choice Models: Inverting the Value-Function Fixed-Point Equation in Probability Terms

Aguirregabiria and Mira (2002) further elaborate on the idea to invert the choice probabilities. They consider inverting the value-function fixed-point equation of Rust (1987), in terms of the conditional choice probabilities. The setup and the assumption is the same with Rust (1987), whereas Aguirregabiria and Mira's arguments apply for a generic error-term distribution other than i.i.d. type I extreme value.

We work with a slightly different notion of value function here. Consider the ex-ante expected value function:

$$\begin{aligned}
& \bar{\bar{v}}\left(x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \\
:= & \int \max_{j\in\mathcal{J}} \left\{ \bar{u}_j\left(x_t; \boldsymbol{\theta}\right) + \epsilon_{j,t} + \beta \int\int v\left(x_{t+1}, \boldsymbol{\epsilon}_{t+1}; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \Pr\left(d\boldsymbol{\epsilon}_{t+1}\right) \Pr\left(dx_{t+1} | x_t, a_t = j; \boldsymbol{\varphi}\right) \right\} \Pr\left(d\boldsymbol{\epsilon}_t\right) \\
= & \int \max_{j\in\mathcal{J}} \left\{ \bar{u}_j\left(x_t; \boldsymbol{\theta}\right) + \epsilon_{j,t} + \beta \int \bar{\bar{v}}\left(x_{t+1}; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \Pr\left(dx_{t+1} | x_t, a_t = j; \boldsymbol{\varphi}\right) \right\} \Pr\left(d\boldsymbol{\epsilon}_t\right) \\
= & \sum_{j\in\mathcal{J}} \Pr\left(a_t = j | x_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \Big\{ \bar{u}_j\left(x_t; \boldsymbol{\theta}\right) + E\left[\epsilon_{j,t} | a_t = j, x_t\right] \\
& \qquad + \beta \sum_{x'\in\mathcal{X}} \bar{\bar{v}}\left(x_{t+1} = x'; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \Pr\left(x_{t+1} = x' | x_t, a_t = j; \boldsymbol{\varphi}\right) \Big\}.
\end{aligned} \tag{2.3.17}$$

[19] It can be shown that, for a given $(\boldsymbol{\theta},\boldsymbol{\varphi})$, the value function $\bar{\bar{v}}\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ satisfying (2.3.17) is a unique fixed point. Next, note that

$$E\left[\epsilon_{j,t}|a_t=j,x_t\right]=-\ln\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right) \tag{2.3.18}$$

holds, where $\epsilon_{j,t}$ is j-th element of $\boldsymbol{\epsilon}_t$.[20] Substituting back, (2.3.17) becomes

$$\begin{aligned}\bar{\bar{v}}\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right) \quad = \quad & \sum_{j\in\mathcal{J}}\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\Big\{\bar{u}_j\left(x_t;\boldsymbol{\theta}\right)-\ln\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right) \\ & \left.+\beta\sum_{x'\in\mathcal{X}}\bar{\bar{v}}\left(x_{t+1}=x';\boldsymbol{\theta},\boldsymbol{\varphi}\right)\Pr\left(x_{t+1}=x'|x_t,a_t=j;\boldsymbol{\varphi}\right)\right\}.\end{aligned} \tag{2.3.21}$$

We stack the fixed-point equations (2.3.21) in order to express it in terms of the conditional choice probabilities $\Pr\left(a_t|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ and the state-transition probabilities $\Pr\left(x_{t+1}|x_t,a_t;\boldsymbol{\varphi}\right)$. For an illustration,

[19] Note the difference between $\bar{\bar{v}}\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ and the choice-specific expected value function $\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ that we worked with before.

[20] We invoked mean-zero i.i.d. type I extreme value assumption in (2.3.18). By the corollary to Lemma 2.2.2, we know that

$$\begin{aligned}&\ln\left(\sum_{j\in\mathcal{J}}\exp\left(\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)\right)\\ &=\int\max_{j\in\mathcal{J}}\left\{\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)+\epsilon_{j,t}\right\}\Pr\left(d\boldsymbol{\epsilon}_t\right)\\ &=\sum_{j\in\mathcal{J}}\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\left\{\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)+E\left[\epsilon_{j,t}|a_t=j,x_t\right]\right\}.\end{aligned} \tag{2.3.19}$$

On the other hand,

$$\begin{aligned}&\sum_{j\in\mathcal{J}}\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\left\{\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)-\ln\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right\} \\ &=\sum_{j\in\mathcal{J}}\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\left\{\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)-\ln\frac{\exp\left(\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)}{\sum_{k\in\mathcal{J}}\exp\left(\bar{v}_k\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)}\right\}\\ &=\left\{\ln\left(\sum_{k\in\mathcal{J}}\exp\left(\bar{v}_k\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)\right)\right\}\sum_{j\in\mathcal{J}}\Pr\left(a_t=j|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\\ &=\ln\left(\sum_{j\in\mathcal{J}}\exp\left(\bar{v}_j\left(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right)\right).\end{aligned} \tag{2.3.20}$$

Combining (2.3.19) and (2.3.20) yields the desired conclusion.

consider the case $x_t = 1$:

$$
\begin{aligned}
\bar{\bar{v}}(1;\boldsymbol{\theta},\boldsymbol{\varphi}) &= \mathbf{p}(a_t|1;\boldsymbol{\theta},\boldsymbol{\varphi})' \begin{pmatrix} \bar{u}_1(1;\boldsymbol{\theta}) - \ln \Pr(a_t = 1|1;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \bar{u}_2(1;\boldsymbol{\theta}) - \ln \Pr(a_t = 2|1;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \vdots \\ \bar{u}_J(1;\boldsymbol{\theta}) - \ln \Pr(a_t = J|1;\boldsymbol{\theta},\boldsymbol{\varphi}) \end{pmatrix} \\
&\quad + \beta \mathbf{p}(a_t|1;\boldsymbol{\theta},\boldsymbol{\varphi})' \mathbf{P}(x_{t+1}|1,a_t;\boldsymbol{\varphi}) \begin{pmatrix} \bar{\bar{v}}(1;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \bar{\bar{v}}(2;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \vdots \\ \bar{\bar{v}}(X;\boldsymbol{\theta},\boldsymbol{\varphi}) \end{pmatrix},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{p}(a_t|1;\boldsymbol{\theta},\boldsymbol{\varphi}) &= \begin{pmatrix} \Pr(1|1;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \Pr(2|1;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \vdots \\ \Pr(J|1;\boldsymbol{\theta},\boldsymbol{\varphi}) \end{pmatrix} \\
\mathbf{P}(x_{t+1}|1,a_t;\boldsymbol{\varphi}) &= \begin{pmatrix} \Pr(1|1,1;\boldsymbol{\varphi}) & \Pr(2|1,1;\boldsymbol{\varphi}) & \cdots & \Pr(X|1,1;\boldsymbol{\varphi}) \\ \Pr(1|1,2;\boldsymbol{\varphi}) & \Pr(2|1,2;\boldsymbol{\varphi}) & \cdots & \Pr(X|1,2;\boldsymbol{\varphi}) \\ \vdots & \vdots & \ddots & \vdots \\ \Pr(1|1,J;\boldsymbol{\varphi}) & \Pr(2|1,J;\boldsymbol{\varphi}) & \cdots & \Pr(X|1,J;\boldsymbol{\varphi}) \end{pmatrix}.
\end{aligned}
$$

[21] Notice that $\mathbf{p}(a_t|1;\boldsymbol{\theta},\boldsymbol{\varphi})' \mathbf{P}(x_{t+1}|1,a_t;\boldsymbol{\varphi})$ yields a $1 \times X$ vector. Analogous observation leads us to form an $X \times X$ unconditional state-transition probability matrix $\mathbf{P}(x_{t+1}|x_t;\boldsymbol{\theta},\boldsymbol{\varphi})$. Define

$$
\bar{\mathbf{u}}(x_t = x;\boldsymbol{\theta}) := \begin{pmatrix} \bar{u}_1(x;\boldsymbol{\theta}) \\ \bar{u}_2(x;\boldsymbol{\theta}) \\ \vdots \\ \bar{u}_J(x;\boldsymbol{\theta}) \end{pmatrix}.
$$

By stacking $\bar{\bar{v}}(x_t;\boldsymbol{\theta},\boldsymbol{\varphi})$, we have

$$
\begin{pmatrix} \bar{\bar{v}}(1;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \bar{\bar{v}}(2;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \vdots \\ \bar{\bar{v}}(X;\boldsymbol{\theta},\boldsymbol{\varphi}) \end{pmatrix} = \begin{pmatrix} \mathbf{p}(a_t|1;\boldsymbol{\theta},\boldsymbol{\varphi})' (\bar{\mathbf{u}}(x_t = 1;\boldsymbol{\theta}) - \ln \mathbf{p}(a_t|1;\boldsymbol{\theta},\boldsymbol{\varphi})) \\ \mathbf{p}(a_t|2;\boldsymbol{\theta},\boldsymbol{\varphi})' (\bar{\mathbf{u}}(x_t = 2;\boldsymbol{\theta}) - \ln \mathbf{p}(a_t|2;\boldsymbol{\theta},\boldsymbol{\varphi})) \\ \vdots \\ \mathbf{p}(a_t|X;\boldsymbol{\theta},\boldsymbol{\varphi})' (\bar{\mathbf{u}}(x_t = X;\boldsymbol{\theta}) - \ln \mathbf{p}(a_t|X;\boldsymbol{\theta},\boldsymbol{\varphi})) \end{pmatrix} + \beta \mathbf{P}(x_{t+1}|x_t;\boldsymbol{\theta},\boldsymbol{\varphi}) \begin{pmatrix} \bar{\bar{v}}(1;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \bar{\bar{v}}(2;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \vdots \\ \bar{\bar{v}}(X;\boldsymbol{\theta},\boldsymbol{\varphi}) \end{pmatrix}. \tag{2.3.22}
$$

[21] In practice, one should let free only $J-1$ of the conditional choice probabilities.

We invert (2.3.22), which yields:

$$\begin{pmatrix} \bar{\bar{v}}(1;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \bar{\bar{v}}(2;\boldsymbol{\theta},\boldsymbol{\varphi}) \\ \vdots \\ \bar{\bar{v}}(X;\boldsymbol{\theta},\boldsymbol{\varphi}) \end{pmatrix} = \left[\mathbf{I} - \beta \mathbf{P}(x_{t+1}|x_t;\boldsymbol{\theta},\boldsymbol{\varphi})\right]^{-1} \begin{pmatrix} \mathbf{p}(a_t|1;\boldsymbol{\theta},\boldsymbol{\varphi})' \left(\bar{\mathbf{u}}(x_t=1;\boldsymbol{\theta}) - \ln \mathbf{p}(a_t|1;\boldsymbol{\theta},\boldsymbol{\varphi})\right) \\ \mathbf{p}(a_t|2;\boldsymbol{\theta},\boldsymbol{\varphi})' \left(\bar{\mathbf{u}}(x_t=2;\boldsymbol{\theta}) - \ln \mathbf{p}(a_t|2;\boldsymbol{\theta},\boldsymbol{\varphi})\right) \\ \vdots \\ \mathbf{p}(a_t|X;\boldsymbol{\theta},\boldsymbol{\varphi})' \left(\bar{\mathbf{u}}(x_t=X;\boldsymbol{\theta}) - \ln \mathbf{p}(a_t|X;\boldsymbol{\theta},\boldsymbol{\varphi})\right) \end{pmatrix}. \tag{2.3.23}$$

The right-hand side of the equation (2.3.23) involves only the model primitives $\{\beta, \bar{u}(x_t;\boldsymbol{\theta})\}$, the state-transition probabilities $\Pr(x_{t+1}|x_t, a_t;\boldsymbol{\varphi})$, and the conditional choice probabilities $\Pr(a_t|x_t;\boldsymbol{\theta},\boldsymbol{\varphi})$. We write (2.3.23) compactly in a matrix notation as

$$\bar{\bar{\mathbf{v}}}(x_t;\boldsymbol{\theta},\boldsymbol{\varphi}) = \left[\mathbf{I} - \beta \mathbf{P}(x_{t+1}|x_t;\boldsymbol{\theta},\boldsymbol{\varphi})\right]^{-1} \mathbf{w}(a_t, x_t;\boldsymbol{\theta},\boldsymbol{\varphi}), \tag{2.3.24}$$

where

$$\mathbf{w}(a_t, x_t;\boldsymbol{\theta},\boldsymbol{\varphi}) := \begin{pmatrix} \mathbf{p}(a_t|1;\boldsymbol{\theta},\boldsymbol{\varphi})' \left(\bar{\mathbf{u}}(x_t=1;\boldsymbol{\theta}) - \ln \mathbf{p}(a_t|1;\boldsymbol{\theta},\boldsymbol{\varphi})\right) \\ \mathbf{p}(a_t|2;\boldsymbol{\theta},\boldsymbol{\varphi})' \left(\bar{\mathbf{u}}(x_t=2;\boldsymbol{\theta}) - \ln \mathbf{p}(a_t|2;\boldsymbol{\theta},\boldsymbol{\varphi})\right) \\ \vdots \\ \mathbf{p}(a_t|X;\boldsymbol{\theta},\boldsymbol{\varphi})' \left(\bar{\mathbf{u}}(x_t=X;\boldsymbol{\theta}) - \ln \mathbf{p}(a_t|X;\boldsymbol{\theta},\boldsymbol{\varphi})\right) \end{pmatrix}.$$

[22]

2.3.3.2 *K*-stage Policy-Iteration Estimator

Now we get to the pseudo-likelihood estimation. First, it is assumed that $\hat{\boldsymbol{\varphi}}$, a consistent estimator for the parameters of state-transition probabilities, can be obtained separately. Plugging $\hat{\boldsymbol{\varphi}}$ back into the conditional choice probabilities yields

$$\Pr(a_t = j|x_t;\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}) = \frac{\exp\left(\bar{v}_j(x_t;\boldsymbol{\theta},\hat{\boldsymbol{\varphi}})\right)}{\sum_{k\in\mathcal{J}} \exp\left(\bar{v}_k(x_t;\boldsymbol{\theta},\hat{\boldsymbol{\varphi}})\right)}, \tag{2.3.25}$$

and

$$\bar{v}_j(x_t;\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}) = \bar{u}_j(x_t;\boldsymbol{\theta}) + \beta \sum_{x'\in\mathcal{X}} \bar{\bar{v}}\left(x_{t+1}=x';\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right) \Pr\left(x_{t+1}=x'|x_t, a_t=j;\hat{\boldsymbol{\varphi}}\right). \tag{2.3.26}$$

Substituting (2.3.23) back into (2.3.26) for all possible pairs of (a_t, x_t, x_{t+1}), and then again substituting back into (2.3.25), we have

$$\mathbf{P}(a_t|x_t;\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}) = \Psi\left(\mathbf{P}(a_t|x_t;\boldsymbol{\theta},\hat{\boldsymbol{\varphi}})\right).$$

$\Psi(\cdot)$ is what we define as the policy-iteration operator.

The estimation algorithm proceeds as follows. Begin from an initial guess $\mathbf{P}^0(a_t|x_t;\boldsymbol{\theta},\hat{\boldsymbol{\varphi}})$. The initial

[22]Chiong, Galichon, and Shum (2016) also derive (2.3.24) through simpler steps using the concept of McFadden (1974, 1981)'s social surplus function and its differentials.

guess can be, for example, obtained by Hotz and Miller (1993) inversion. The first stage is to obtain

$$\hat{\boldsymbol{\theta}}^K = \arg\max_{\boldsymbol{\theta}} \prod_i \mathcal{L}_i \left(\Psi \left(\mathbf{P}^{K-1} \left(a_t | x_t; \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}} \right) \right) \right),$$

where $\mathcal{L}_i(\cdot)$ denotes the likelihood induced by the argument, and i denotes the data.[23] The second stage is to update $\mathbf{P}(a_t|x_t;\boldsymbol{\theta},\hat{\boldsymbol{\varphi}})$ using the right-hand side of (2.3.23), then plugging back into (2.3.26) and (2.3.25) sequentially. To that end, we have

$$\mathbf{P}^K \left(a_t | x_t; \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}} \right) = \Psi \left(\mathbf{P}^{K-1} \left(a_t | x_t; \hat{\boldsymbol{\theta}}^K, \hat{\boldsymbol{\varphi}} \right) \right).$$

It can be shown that the Hotz-Miller estimator can be nested as a special case of this estimator when $K = 1$. When $K \to \infty$, the estimator is equivalent to Rust's full solution provided that $\hat{\boldsymbol{\theta}}^K$ converges. This NPL estimator can achieve the potential accuracy gain over Hotz-Miller when the first-stage estimator is imprecise. Furthermore, it can achieve some speed advantage over Rust's full-solution method, because the researcher can decide the stopping criterion. As in MPEC, the speed advantage comes from the fact that iterating over the fixed point at each iteration is unnecessary. The authors show that if the sequence $\hat{\boldsymbol{\theta}}^K$ converges, it is $\sqrt{T}$-consistent, asymptotically normal, and asymptotically equivalent to the full-solution MLE. They also report that solving Rust's original problem does not seem to lead to much accuracy gain for $K > 3$.

2.3.4 Extension to Incorporate Unobserved State Variables

The dynamic discrete-choice models we studied thus far do not allow for an unobserved state variable. That is, they are de facto assuming the agents and the econometrician have the same degree of ex-ante information,[24] which is often unrealistic.

Agents may observe persistent state variables that are not observable to the econometrician. Such state variables may or may not be invariant over time. In this subsection, we study the estimation method developed by Arcidiacono and Miller (2011), which may incorporate unobserved heterogeneity in the dynamic discrete-choice models. Arcidiacono and Miller allow for unobserved heterogeneity that has the following two properties: (i) finitely many possible unobserved states exist, and (ii) either the unobserved states are time invariant or they follow the first-order Markov chain. The estimation combines the Hotz and Miller (1993) inversion with the expectation-maximization algorithm. The setup we study below simplifies (ii), where the unobserved state variable h_t is i.i.d. over time.

[23]This will take the usual form, i.e.,

$$\prod_{x,j,t} \Pr \left(a_{i,t} = j | x_{i,t} = x; \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}} \right)^{\mathbf{1}(a_{i,t}=j|x_{i,t}=x)}.$$

[24]By ex-ante information, we refer to the information right before the realization of $\boldsymbol{\epsilon}_t$'s.

2.3.4.1 Formulation

Denote $\mathbf{z}_t := (x_t, h_t)$ by the state vector at time t, where x_t is observable to the econometrician and h_t is unobservable. All other notations are the same as before. The state space is assumed to be a finite set. We denote $x_t \in \mathcal{X} := \{1, 2, ..., X\}$, $h_t \in \mathcal{H} := \{1, 2, ..., H\}$, and $a_t \in \mathcal{J} := \{1, 2, ..., J\}$. The flow utility $\bar{u}_j(\mathbf{z}_t; \boldsymbol{\theta})$ is now a function of the unobserved state h_t, as well as of the observed state x_t, and it is assumed that the functional form is known up to the parameter vector $\boldsymbol{\theta}$. The econometrician observes the evolution of x_t, but the evolution of h_t is hidden to the econometrician.

The idea is to implement the Hotz and Miller (1993) inversion (2.3.13) to form the likelihood, that is,

$$\begin{aligned}
& \bar{v}_j(\mathbf{z}_t; \boldsymbol{\theta}, \boldsymbol{\varphi}) - \bar{v}_J(\mathbf{z}_t; \boldsymbol{\theta}, \boldsymbol{\varphi}) \\
= \; & \bar{u}_j(\mathbf{z}_t; \boldsymbol{\theta}) - \bar{u}_J(\mathbf{z}_t; \boldsymbol{\theta}) \\
& + \beta \int \ln\left(\sum_{k \in \mathcal{J}} \exp \hat{v}_k(\mathbf{z}_{t+1})\right) \left\{\Pr(d\mathbf{z}_{t+1} | \mathbf{z}_t, a_t = j, \boldsymbol{\varphi}) - \Pr(d\mathbf{z}_{t+1} | \mathbf{z}_t, a_t = J, \boldsymbol{\varphi})\right\}, \qquad (2.3.27)
\end{aligned}$$

with

$$\ln\left(\frac{\hat{\Pr}(a_t = j | \mathbf{z}_t)}{\hat{\Pr}(a_t = J | \mathbf{z}_t)}\right) = \hat{v}_j(\mathbf{z}_t) - \hat{v}_J(\mathbf{z}_t). \qquad (2.3.28)$$

The problem here is that the sample analogue of $\Pr(\mathbf{z}_{t+1} | \mathbf{z}_t, a_t = j, \boldsymbol{\varphi})$ and $\hat{\Pr}(a_t = j | \mathbf{z}_t)$ are not observable to the econometrician. Only $\Pr(x_{t+1} | x_t, a_t = j, \boldsymbol{\varphi})$ and $\Pr(a_t = j | x_t)$ are observable.

2.3.4.2 Expectation-Maximization (EM) Algorithm

The likelihood contribution when the observed data $\{a_t = j, x_t = x\}$ is

$$L(\{a_t = j, x_t = x\} | \boldsymbol{\theta}, \boldsymbol{\varphi}) = E_{h_t}\left[L(\{a_t = j, x_t = x, h_t\} | \boldsymbol{\theta}, \boldsymbol{\varphi})\right]. \qquad (2.3.29)$$

However, the product of (2.3.29) over t cannot be directly maximized, because h_t is hidden. In turn, the sample analogue of the conditional choice probability $\Pr(a_t = j | \mathbf{z}_t)$, the state-transition probability $\Pr(\mathbf{z}_{t+1} | \mathbf{z}_t, a_t, \boldsymbol{\varphi})$, and the mean utility $\bar{u}_j(\mathbf{z}_t; \boldsymbol{\theta})$ are also unknown even if we know the parameter $\boldsymbol{\theta}$. Therefore, we proceed with the expectation-maximization algorithm, taking h_t as the hidden state. For notational simplicity, we assume $\boldsymbol{\varphi}$ is the H-dimensional vector such that $\varphi_{(h)} = \Pr(h_t = h)$, and no other variable is included in $\boldsymbol{\varphi}$.

Suppose the iteration is at the m-th stage, so we have $\hat{\boldsymbol{\theta}}^{(m)}$ and $\hat{\boldsymbol{\varphi}}^{(m)}$. We begin the m-th iteration by assuming $\hat{\Pr}^{(m)}(a_t = j | x_t, h_t)$ is known, which will be updated in Step 3.

Step 0. Formulate the Likelihood It is immediate that we can formulate the likelihood contribution by combining (2.3.27), (2.3.28), and (2.3.29):

$$L\left(\{a_t = j, x_t = x, h_t = h\} | \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right) = \frac{\exp\left(\bar{v}_j\left(x_t = x, h_t = h; \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right)\right)}{\sum_{k \in \mathcal{J}} \exp\left(\bar{v}_k\left(x_t = x, h_t = h; \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right)\right)}. \qquad (2.3.30)$$

Step 1. Update $\hat{\Pr}^{(m)}\left(h_t = h|a_t, x_t; \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right)$ Next, we find $\hat{\Pr}^{(m+1)}\left(h_t = h|a_t, x_t; \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right)$, by using the Bayes' rule:

$$\hat{\Pr}^{(m+1)}\left(h_t = h|a_t = j, x_t = x; \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right) = \frac{\hat{\varphi}_h^{(m)} L\left(\{a_t = j, x_t = x, h_t = h\} | \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right)}{\sum_{h' \in \mathcal{H}} \hat{\varphi}_{h'}^{(m)} L\left(\{a_t = j, x_t = x, h_t = h'\} | \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right)}. \quad (2.3.31)$$

Step 2. Update $\hat{\boldsymbol{\varphi}}^{(m)}$ We update $\hat{\boldsymbol{\varphi}}^{(m)}$ to $\hat{\boldsymbol{\varphi}}^{(m+1)}$. By definition,

$$\hat{\varphi}_h^{(m+1)} = \frac{1}{N_{j,x}} \sum_{j \in \mathcal{J}} \sum_{x \in \mathcal{X}} \hat{\Pr}^{(m+1)}\left(h_t = h|a_t = j, x_t = x; \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right), \quad (2.3.32)$$

where $N_{j,x}$ is the observed sample size for action j and state x.

Step 3. Update $\hat{\Pr}^{(m)}(a_t = j|x_t, h_t)$ Similarly, we also update $\hat{\Pr}^{(m)}(a_t = j|x_t, h_t)$ as

$$\hat{\Pr}^{(m+1)}(a_t = j|x_t, h_t = h) = \frac{\sum_{x \in \mathcal{X}} \mathbf{1}(a_t = j, x_t = x) \hat{\Pr}^{(m+1)}\left(h_t = h|a_t = j, x_t = x; \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right)}{\sum_{x \in \mathcal{X}} \mathbf{1}(x_t = x) \hat{\Pr}^{(m+1)}\left(h_t = h|a_t = j, x_t = x; \hat{\boldsymbol{\theta}}^{(m)}, \hat{\boldsymbol{\varphi}}^{(m)}\right)}.$$

Step 4. Update $\hat{\boldsymbol{\theta}}^{(m)}$ Using $\hat{\Pr}^{(m+1)}(a_t = j|x_t, h_t)$, form the updated likelihood by substituting $\hat{\boldsymbol{\varphi}}^{(m+1)}$ back in (2.3.30) and solving the likelihood maximization problem to update $\hat{\boldsymbol{\theta}}^{(m)}$

$$\hat{\boldsymbol{\theta}}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} \sum_t \sum_{h \in \mathcal{H}} \mathbf{1}(a_t = j|x_t = x) \ln\left\{E_{h_t}^{(m+1)}\left[L\left(\{a_t = j, x_t = x, h_t\} | \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}^{(m+1)}\right)\right]\right\}, \quad (2.3.33)$$

where

$$E_{h_t}^{(m+1)}\left[L\left(\{a_t = j, x_t = x, h_t\} | \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}^{(m+1)}\right)\right] = \sum_{h \in \mathcal{H}} \hat{\varphi}_h^{(m+1)} \frac{\exp\left(\bar{v}_j\left(x_t = x, h_t = h; \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}^{(m+1)}\right)\right)}{\sum_{k \in \mathcal{J}} \exp\left(\bar{v}_k\left(x_t = x, h_t = h; \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}^{(m+1)}\right)\right)}.$$

Iterating over Steps 0-4 will give the consistent and asymptotically normal estimator for $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$.

2.3.4.3 Discussion and Further Readings

The expectation-maximization algorithm illustrated above implicitly assumes the hidden state h_t is stationary so the model parameters can be consistently estimated with a single time series. It can be easily extended to a setup where the hidden state follows the first-order Markov chain or is persistent over all periods,[25] which will in general require a panel data for identification and estimation.

[25]This way, the problem boils down to estimating the Markov transition matrix along with the model parameters. The estimation algorithm should be modified to update the transition matrix Ξ and initialization probability $\boldsymbol{\Pi}$. Ξ is a matrix of which the (h, τ)'th entry is the transition probability from state h to state τ; that is,

$$\Xi_{h,\tau} := \Pr(h_{t+1} = \tau | h_t = h).$$

Incorporating the unobservable state variables in the dynamic econometric models is an ongoing area of research. Connault (2016); Hu and Shum (2012); Hu, Shum, Tan, and Xiao (Forthcoming); Kasahara and Shimotsu (2009) provide identification results under different setups, respectively. Notably, Kasahara and Shimotsu provide the identification results when the unobserved heterogeneity is time invariant, which Arcidiacono and Miller (2011) build upon. Norets (2009) develops an estimation and inference procedure based on the Bayesian perspective.

2.3.5 (Non)-Identification of the Discount Factor

Magnac and Thesmar (2002) study the identification problem of the dynamic discrete-choice framework. The parameter vector $\boldsymbol{\theta}$ of interest is identified if the observed joint distribution of data $\Pr(a_t|x_t)$[26] and the structural restrictions of the economic model uniquely pin down $\boldsymbol{\theta}$. To put the conclusions first, Magnac and Thesmar show the model is generically underidentified.[27] To identify the model parameters, one should have prior knowledge on (i) the discount factor β, (ii) the state-transition probability $\Pr(x_{t+1}|x_t, a_t)$, and (iii) functional forms of the current utility and the value function for at least one reference alternative. (ii) and (iii) do not pose a concern when the functional forms of the current utility and the transition probability distribution is specified. Yet, the nonidentifiability of β is a difficult problem to resolve even in the fully parametric dynamic discrete-choice models.

A heuristic argument of the nonidentifiability of β is as below. Assume $\boldsymbol{\epsilon}_t \sim i.i.d.\ T1EV$ and

$$\bar{u}(x_t, a_t; \theta) = \begin{cases} 0 & \text{if } a_t = 1 \\ x_t\theta & \text{if } a_t = 2 \end{cases}.$$

We have shown that the conditional choice probability is of the following form:

$$\begin{aligned} \Pr(a_t = 1|x_t) &= \frac{\exp(\beta\bar{\varpi}_1(x_t))}{\exp(\beta\bar{\varpi}_1(x_t)) + \exp(x_t\theta + \beta\bar{\varpi}_2(x_t))} \\ &= \frac{1}{1 + \exp(x_t\theta + \beta(\bar{\varpi}_2(x_t) - \bar{\varpi}_1(x_t)))}. \end{aligned}$$

Furthermore, suppose we know the exact value of $\bar{\varpi}_2(x_t) - \bar{\varpi}_1(x_t)$ for all possible values of x_t. Now the problem boils down to finding (θ, β) by "inverting" the conditional cumulative distribution function $F_{a_t|x_t}(a = 1|x)$, which turns out to be impossible. To see why, fix $x_t = x$, and assume we have infinitely many data, so the value $\Pr(a_t = 1|x_t = x)$ can be pinned down exactly from data. We can always find

$\boldsymbol{\Pi}$ is a matrix of which the (x, h)'th entry is the conditional probability of h_1 given x_1; that is,

$$\Pi_{x,h} := \Pr(h_1 = h|x_1 = x).$$

Now $\boldsymbol{\varphi}$ contains Ξ, $\boldsymbol{\Pi}$, and $\Pr(h_t = h)$, but otherwise the estimation routine is similar to what is described before.

[26] It is essentially assuming the situation where an econometrician has an infinite amount of data.

[27] In the context of dynamic discrete-choice models, the functional form of $\bar{u}(x_t, a_t; \boldsymbol{\theta})$ and the fact that the agent is engaging in the dynamic optimization problem is the main structural restriction of consideration.

some $\alpha_1 \neq 1$ and $\alpha_2 \neq 1$ such that the following holds:

$$\begin{aligned}\Pr\left(a_t = 1 | x_t = x\right) =& \frac{1}{1 + \exp\left(x\theta + \beta\left(\bar{w}_2\left(x\right) - \bar{w}_1\left(x\right)\right)\right)} \\ =& \frac{1}{1 + \exp\left(\alpha_1 x\theta + \alpha_2\beta\left(\bar{w}_2\left(x\right) - \bar{w}_1\left(x\right)\right)\right)},\end{aligned}$$

which means uniquely pinning down (θ, β) by "inverting" $F_{a_t|x_t}(1|x)$ is impossible. (θ, β) and $(\alpha_1\theta, \alpha_2\beta)$ in this example are *observationally equivalent*, in that they both result in the same probability distribution of observable data. (Point) identification of the model parameters requires that the observationally equivalent set is a singleton.[28] For this reason, we specified the value of β ex ante, often with some a priori knowledge.[29]

A common solution to identify β is to impose the exclusion restriction. Magnac and Thesmar suggest finding a set of variables $\mathbf{z}_j$, which is an argument of the future value, but not of the current utility. Abbring and Daljord (2017) suggest imposing the exclusion restriction on the current utility. Finding such a variable that only affects the current flow utility but not the future payoffs (or vice versa) is generally difficult.

[28]The parameter is said to be partially identified when the observationally equivalent set is nonempty and not a singleton.

[29]For example, the most usual choice of β is 0.95 when the data are yearly.

References

ABBRING, J. H. AND Ø. DALJORD (2017): "Identifying the discount factor in dynamic discrete choice models," *Working Paper*.

AGUIRREGABIRIA, V. AND P. MIRA (2002): "Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models," *Econometrica*, 70, 1519–1543.

——— (2010): "Dynamic discrtete choice structural models: A survey," *Journal of Econometrics*, 156, 38–67.

ARCIDIACONO, P. AND R. A. MILLER (2011): "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity," *Econometrica*, 79, 1823–1867.

CARDELL, N. S. (1997): "Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity," *Econometric Theory*, 13, 185–213.

CHIONG, K. X., A. GALICHON, AND M. SHUM (2016): "Duality in dynamic discrete-choice models," *Quantitative Economics*, 7, 83–115.

CONNAULT, B. (2016): "Hidden rust models," *Working Paper*, 2014.

HOTZ, V. J. AND R. A. MILLER (1993): "Conditional choice probabilities and the estimation of dynamic models," *Review of Economic Studies*, 60, 497–529.

HOTZ, V. J., R. A. MILLER, S. SANDERS, AND J. SMITH (1994): "A simulation estimator for dynamic models of discrete choice," *Review of Economic Studies*, 61, 265–289.

HU, Y. AND M. SHUM (2012): "Nonparametric identification of dynamic models with unobserved state variables," *Journal of Econometrics*, 171, 32–44.

HU, Y., M. SHUM, W. TAN, AND R. XIAO (Forthcoming): "A simple estimator for dynamic models with serially correlated unobservables," *Journal of Eonometric Methods*.

ISKHAKOV, F., J. LEE, J. RUST, B. SCHJERNING, AND K. SEO (2016): "Constrained optimization approaches to estimation of structural models: Comment," *Econometrica*, 84, 365–370.

KASAHARA, H. AND K. SHIMOTSU (2009): "Nonparametric identification of finite mixture models of dynamic discrete choices," *Econometrica*, 77, 135–175.

LUCE, R. D. (1959): *Individual Choice Behavior: A Theoretical Analysis*, New York: Wiley.

MAGNAC, T. AND D. THESMAR (2002): "Identifying dynamic discrete decision processes," *Econometrica*, 70, 801–816.

MCFADDEN, D. (1974): *Conditional logit analysis of qualitative choice behavior*, Academic Press, chap. 4, 105–142.

——— (1981): *Econometric models of probabilistic choice*, The MIT Press, chap. 5, 198–272, Structural analysis of discrete data with econometric applications.

——— (2001): "Economic choices," *American Economic Review*, 91, 351–378.

MCFADDEN, D. AND K. TRAIN (2000): "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15, 447–470.

NORETS, A. (2009): "Inference in dynamic discrete choice models with serially correlated unobserved state variables," *Econometrica*, 77, 1665–1682.

RUST, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica*, 55, 999–1033.

SMALL, K. A. AND H. S. ROSEN (1981): "Applied welfare economics with discrete choice models," *Econometrica*, 49, 105–130.

STOKEY, N. L. AND R. E. LUCAS (1989): *Recursive Methods in Economic Dynamics*, Cambridge, Massachusetts: Harvard University Press.

SU, C.-L. AND K. L. JUDD (2012): "Constrained optimization approaches to estimation of structural models," *Econometrica*, 80, 2213–2230.

Chapter 3

Demand Estimation with Market Data

In this chapter, we study estimation methods of demand systems derived under different assumptions on the consumer preferences and optimization structures. According to conventional consumer theory, the Marshallian and Hicksian demand systems are derived as a result of consumers' budget-constrained optimization problem. A "modern" approach to the empirical demand systems begins from a discrete-choice problem of consumers, and aggregate the individual choice probabilities to form a market demand. An empirical demand system has certain restrictions to be imposed, which are inherent from the consumers' respective optimization problem. Empirically, those restrictions are tested or imposed as a priori assumptions. We study how different setups on the consumer preferences and optimization structures lead to different empirical demand systems and restrictions imposed on them, and how they can be estimated or tested.

Demand estimation has been a central problem in the new empirical industrial organization literature because it is fundamental in structural analyses in various contexts. The primary aim of the demand estimation is to estimate the own- and cross-price elasticities. Combined with cost-parameter estimates of an industry, they can be used for applied analyses, including computing the price-cost margins, merger analysis, welfare changes due to the introduction of a new product, and so on.

Throughout the chapter, we assume that only market-level data on the prices, quantities, and other product characteristics of an alternative in the choice set are available to an econometrician. Although not reviewed in this chapter, remarkably, recent advances in the literature make extensive use of individual-level purchase data, taking individual-level heterogeneity more seriously.

3.1 Product-Space Approach

The product-space approach, from which our discussion on demand systems begins, attempts to directly estimate the own- and cross-price elasticities of a demand system. In a sense, the product-space approach is more flexible than the characteristics-space approach, at a cost of a rapidly increasing number of parameters to be estimated. The approach is often referred to as estimating the reduced-form demand systems.

3.1.1 Linear and Log-Linear Demand Model

We start from the simplest linear demand model. Let j, k denote alternatives. In the linear demand model, the market demand for product j is specified as

$$q_j = a_j + \sum_{k=1}^{J} \beta_{j,k} p_{j,k} + \epsilon_{j,k}. \tag{3.1.1}$$

(3.1.1) can be interpreted as a first-order approximation to any demand system. In that regard, $\epsilon_{j,k}$ can be an approximation error or measurement error. Each product j has its own intercept a_j. Summing over k for each fixed j also allows the nonzero cross-price elasticities. However, this model can predict even a negative market demand q_j, which is problematic.

A straightforward modification of (3.1.1) is the log-linear specification. Consider

$$\ln q_j = a_j + \eta_j \ln w + \sum_{k=1}^{J} \beta_{j,k} \ln p_{j,k} + \epsilon_{j,k}, \tag{3.1.2}$$

where w is the expenditure within the specified demand system. Because $\varepsilon_{j,k} := \frac{d \ln q_j}{d \ln p_k} = \beta_{j,k}$ is the cross-price elasticity between products j and k, it is immediate that the own- and cross-price elasticities are constant. This functional form imposes a testable restriction $\eta_j = 1$ for all j. If this restriction is met, the system (3.1.2) can be regarded as a first-order approximation to a Marshallian demand system.

3.1.2 The Almost Ideal Demand System

Deaton and Muellbauer (1980) elaborate on the problem in the context of the aggregation theory. The motivation for the development of the Almost Ideal Demand System (AIDS)[1] was to rationalize the aggregate consumer demand as if it were the outcome of a single representative utility-maximizing consumer.

Consider the following form of the individual expenditure function:

$$c_i(\mathbf{p}, u_i) = \kappa_i \left(a(\mathbf{p})^{\alpha} (1 - u_i) + b(\mathbf{p})^{\alpha} u_i \right)^{\frac{1}{\alpha}},$$

where the real-valued functions $a(\mathbf{p})$ and $b(\mathbf{p})$ are concave and homogeneous of degree 1 in $\mathbf{p}$. This form of expenditure function is called the price-independent generalized linearity (PIGL) form, which allows for a nonlinear aggregation. It can be shown that there exists a representative expenditure function

$$c(\mathbf{p}, u) = \left(a(\mathbf{p})^{\alpha} (1 - u) + b(\mathbf{p})^{\alpha} u \right)^{\frac{1}{\alpha}},$$

which aggregates c_i's in a certain nonlinear way. Sending α to zero and taking logarithms yields the

[1] This AIDS (1980) was developed before the other AIDS was first discovered (1981).

price-independent generalized log-linear (PIGLOG) expenditure function, which is

$$\ln c(\mathbf{p}, u) = (1-u) \ln a(\mathbf{p}) + u \ln b(\mathbf{p}). \tag{3.1.3}$$

The AIDS specification begins from (3.1.3). The following functional form specification allows for a second-order approximation of an arbitrary expenditure function:

$$\begin{aligned} \ln a(\mathbf{p}) &= a_0 + \sum_k a_k \ln p_k + \frac{1}{2} \sum_k \sum_j \gamma^*_{k,j} \ln p_k \ln p_j \\ \ln b(\mathbf{p}) &= \ln a(\mathbf{p}) + \beta_0 \prod_k p_k^{\beta_k}. \end{aligned}$$

Hence, the AIDS expenditure function is defined by

$$\ln c(\mathbf{p}, u) = a_0 + \sum_k a_k \ln p_k + \frac{1}{2} \sum_k \sum_j \gamma^*_{k,j} \ln p_k \ln p_j + \beta_0 \prod_k p_k^{\beta_k}. \tag{3.1.4}$$

Differentiating the left-hand side of (3.1.4) with respect to $\ln p_j$ yields

$$\frac{\partial \ln c(\mathbf{p}, u)}{\partial \ln p_j} = \frac{p_j q_j}{c(\mathbf{p}, u)} = w_j,$$

which is the expenditure share of product j. Differentiating the right-hand side and equating, we have

$$\begin{aligned} w_j &= \alpha_j + \sum_k \gamma_{j,k} \ln p_k + \beta_j u \beta_0 \prod_k p_k^{\beta_k} & (3.1.5) \\ &= \alpha_j + \sum_k \gamma_{j,k} \ln p_k + \beta_j \ln \left(\frac{w}{P}\right), & (3.1.6) \end{aligned}$$

where $\gamma_{j,k} = \frac{1}{2}\left(\gamma^*_{j,k} + \gamma^*_{k,j}\right)$, w is the total expenditure, and P is the price index such that

$$\ln P := \alpha_0 + \sum_k \alpha_k \ln p_k + \frac{1}{2} \sum_k \sum_j \gamma^*_{k,j} \ln p_k \ln p_j. \tag{3.1.7}$$

[2] This price index can be estimated either in its exact form or in a first-order approximation.

(3.1.6) is the AIDS demand function in the expenditure-share form, given that the following testable parameter restrictions are met: (i) $\sum_j \alpha_j = 1$ and $\sum_j \gamma^*_{k,j} = \sum_k \gamma^*_{k,j} = \sum_j \beta_j = 0$ in order for (3.1.4) to be a legitimate expenditure function, and (ii) $\gamma_{k,j} = \gamma_{j,k}$ for Slutsky symmetry. Given these restrictions are met, (3.1.6) represents a system of Marshallian demand functions that add up to the total expenditure.

[2]Equation (3.1.6), in the price index form, is obtained by inverting (3.1.4) and plugging back to (3.1.5).

The Marshallian and Hicksian price elasticities and the income elasticity have the following expressions:

$$
\begin{aligned}
\varepsilon_{j,k}^{M} &= -\mathbf{1}(j=k) + \left(\frac{\gamma_{j,k}}{w_j}\right) - \left(\frac{\beta_j}{w_j}\right) w_k \\
\varepsilon_{j,k}^{H} &= \varepsilon_{j,k}^{M} + \eta_j w_k \\
\eta_j &= 1 + \left(\frac{\beta_j}{w_j}\right).
\end{aligned}
$$

As an arbitrary second-order approximation to any PIGLOG class of demand systems, AIDS has several theoretically desirable properties. Furthermore, estimating the system looks simple, and thus it has been popular in the literature. Exemplar applications include Hausman, Leonard, and Zona (1994) and Chaudhuri, Goldberg, and Jia (2006). Major challenges for the estimation include the number of parameters and the number of instruments for a consistent estimation. Specifically, the number of endogenous variables to be estimated increases in the order of J^2.

3.1.3 Discussion on the Product-Space Approach

The product-space approach has two major limitations. First, a curse-of-dimensionality problem exists, originating from the flexibility that all the cross-price elasticities should be estimated. Second, this approach does not allow counterfactuals. For example, it cannot answer the questions about the response of the market when a new product is introduced. These two major shortcomings were the motivations for the development of the characteristics-space approach, which we discuss in the following section.

We also mention two potential problems that still persist in the characteristics-space approach here. First is the endogeneity of the prices. The observed market data are from the supply-demand equilibrium. Although none of the works mentioned above pay explicit attention to what the error term includes, the price and the error term are likely to be positively correlated. An upward-sloping demand curve is often estimated when this type of endogeneity is not appropriately taken account of. A common solution is to find a cost shifter as an instrument. Second, in the log-log/multiplicative models, the zero demand is inherently ruled out. Rationalizing zeroes in the multiplicative models[3] is an ongoing area of research (see, e.g., Dubé, Hortaçsu, and Joo (2020); Gandhi, Lu, and Shi (2019)).

3.2 Characteristics-Space Approach

In the characteristics-space approach, a product is defined by a bundle of observed and unobserved product characteristics. It reduces the parameter dimensions substantially, and allows for the counterfactual analyses such as consumer welfare changes by new product entry. It has become the de facto standard method for demand estimation during the last two decades since Berry (1994); Berry, Levinsohn, and Pakes (1995).

[3]Not only for the demand models, but also for the gravity models in the international trade literature.

3.2.1 Microfoundation: Discrete-Choice Random Utility Maximization

In the characteristics-space approach, a product is decomposed into the bundle of characteristics $(\mathbf{x}_j, p_j, \xi_j)$, where $\mathbf{x}_j, p_j$ is observed but ξ_j is unobserved. Demand-estimation literature based on this approach has been developed upon the discretechoice random utility maximization (RUM) framework that we studied in section 2.2. We briefly review the discrete-choice random utility approach in the context of demand estimation.

Consumer i's utility from consuming product j is given by

$$u_{i,j} = \alpha_i (w_i - p_j) + \mathbf{x}_j' \boldsymbol{\beta}_i + \xi_j + \epsilon_{i,j}, \tag{3.2.1}$$

where w_i is the consumer's income, p_j is the price of product j, $\mathbf{x}_j$ is the observed characteristics of product j, ξ_j represents utility from the unobserved (only to the econometrician, consumers observe ξ_j fully) characteristics of product j, and $\epsilon_{i,j}$ is the idiosyncratic utility shocks.[4] The explicit distinction between the two error terms ξ_j and $\epsilon_{i,j}$ was one of the key breakthroughs made by Berry (1994); Berry, Levinsohn, and Pakes (1995). Note also that we allowed the coefficients $(\alpha_i, \boldsymbol{\beta}_i)$ to vary over individual i. The detailed implications of these features are discussed in section 3.2.2.

Individual i chooses product j if the product maximizes the utility over the finite choice set $\mathcal{J}_t$, that is,

$$\begin{aligned} j &= \arg\max_{k \in \mathcal{J}_t} \{u_{i,k}\} && (3.2.2) \\ &= \arg\max_{k \in \mathcal{J}_t} \{u_{i,k} - \alpha_i w_i\}. && (3.2.3) \end{aligned}$$

The solution for the unconstrained utility-maximization problem (3.2.2) is invariant to a location adjustment, which we made in (3.2.3) as a normalization. We also note that $\mathcal{J}_t$ must include the numeraire denoted by product 0 in order for the demand system derived from the choice problem (3.2.3) to be sensible. Otherwise, the consumer has no option to choose not to buy anything, and thus this linear utility specification produces no income effect.[5] In that sense, product 0 is often interpreted as the *composite outside goods*. p_0, $\mathbf{x}_0$, and ξ_0 are normalized to zero, so that $E[u_{i,0}] = 0$.

As studied in section 2.2, we obtain the following expression of the individual choice probability $\Pr(i \text{ Chooses } j)$ by assuming $\epsilon_{i,j}$ follows the mean-zero i.i.d. Type-I extreme value distribution:

$$\Pr(i \text{ Chooses } j) = \frac{\exp\left(-\alpha_i p_j + \mathbf{x}_j' \boldsymbol{\beta}_i + \xi_j\right)}{\sum_{k \in \mathcal{J}_t} \exp\left(-\alpha_i p_k + \mathbf{x}_k' \boldsymbol{\beta}_i + \xi_k\right)}. \tag{3.2.4}$$

[4] Early literature tends to interpret $u_{i,j}$ in (3.2.1) as an indirect utility, to emphasize the similarity between discrete-choice demand systems and the Marshallian demand systems. It is partly because price p_j is a direct argument of $u_{i,j}$, and partly because a relationship similar to the Roy's identity holds when $u_{i,j}$ is linear in income and prices. However, we prefer to interpret $u_{i,j}$ as direct utility, because recent applications of the discrete-choice rRUM framework are not confined to the setups that require the resulting choice probability system as a Marshallian demand system. One can simply interpret that there is a negative "direct utility" from price .

[5] We discuss the arbitrariness of the income effect later.

This individual choice probability equation, which shall be equalized with the predicted individual quantity shares and combined with a few approximation assumptions, completely specifies the logit demand system.

3.2.2 Logit Demand Models with Aggregate Market Data

We assume an econometrician has data on market shares, prices, characteristics, and the availability of each product in each market. We also assume the existence of suitable instruments to correct for the endogeneity of prices. Throughout the section, we denote s_j by the observed market share, and π_j by the predicted market share of alternative j, respectively.

We begin our discussion by distinguishing between the two different error terms, ξ_j and $\epsilon_{i,j}$. These two error terms are included in the model for different reasons. ξ_j represents the product characteristics that are not observable to the econometrician. Because ξ_j directly enters into the consumer's utility specification, it might be regarded as consumers' mean valuations on the product characteristics that are unobservable to the econometrician. In the suppliers' side, ξ_j captures the quality of a product, which is highly likely to be correlated with the firms' pricing decision. For this reason, we need to correct for the endogeneity of ξ_j using the instruments. On the other hand, $\epsilon_{i,j}$, which is often referred to as the idiosyncratic utility shocks, is included to derive the tractable expression for the choice probability equation. It allows us to derive the individual choice probability equation in terms of the other model parameters, so it is washed out in (3.2.4). Inclusion of $\epsilon_{i,j}$ is the reason why the framework formulated by (3.2.1)-(3.2.3) is called the discrete-choice *random* utility.

3.2.2.1 Homogeneous Logit Model of Demand

We study Berry (1994)'s homogeneous logit model of demand here. Suppose $\boldsymbol{\beta}_i = \boldsymbol{\beta}$, $\alpha_i = \alpha$ for all i. That is, we are removing all the structural heterogeneity over individuals in the utility specification (3.2.1), and thus from (3.2.4) as well. We relax this homogeneity assumption later. Under this homogeneity assumption, the individual choice probability equation (3.2.4) reduces to

$$\Pr\left(i \text{ Chooses } j\right) = \frac{\exp\left(-\alpha p_j + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j\right)}{\sum_{k\in\mathcal{J}_t}\exp\left(-\alpha p_k + \mathbf{x}_k'\boldsymbol{\beta} + \xi_k\right)}.$$

If infinitely many consumers are in a market, we may equate the individual choice probability $\Pr\left(i \text{ Chooses } j\right)$ with the predicted market share π_j, and then approximate it with the observed market share s_j. Therefore, we have

$$\begin{aligned} s_j &\simeq \pi_j && (3.2.5) \\ &= \frac{\exp\left(-\alpha p_j + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j\right)}{\sum_{k\in\mathcal{J}_t}\exp\left(-\alpha p_k + \mathbf{x}_k'\boldsymbol{\beta} + \xi_k\right)} && (3.2.6) \end{aligned}$$

for each product j, which is often referred to as the market share equation.[6] By taking the ratios s_j/s_0, logarithms, and then invoking the normalization assumption on $(p_0, \mathbf{x}_0, \xi_0)$, we obtain

$$\ln\left(\frac{s_j}{s_0}\right) = -\alpha p_j + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j. \tag{3.2.7}$$

(3.2.7) is the estimation equation for the homogeneous logit model of demand. With a suitable set of instruments $\mathbf{z}_j$[7] to correct for the endogeneity of ξ_j, we can estimate the model parameters $(\alpha, \boldsymbol{\beta})$ consistently by using GMM with the moment condition $E\left[\xi_j|\mathbf{z}_j\right] = 0$. In section 3.2.3, we discuss the choice of suitable instruments.

3.2.2.2 Random Coefficients Logit Model of Demand

In the random coefficients logit model developed by Berry, Levinsohn, and Pakes (1995) (BLP henceforth), the coefficients in the individual choice probability equation may vary over i. For tractability, distributions of $(\alpha_i, \boldsymbol{\beta}_i)$ are assumed to be known up to the finite-dimensional parameters $\boldsymbol{\Pi}$ and $\boldsymbol{\Sigma}$. To that end, BLP assume individual i's utility of choosing product j is specified as

$$\begin{aligned} u_{i,j} &= \alpha_i\left(w_i - p_j\right) + \mathbf{x}_j'\boldsymbol{\beta}_i + \xi_j + \epsilon_{i,j} \qquad \epsilon_{i,j} \sim i.i.d.\ T1EV \\ \begin{pmatrix} \alpha_i \\ \boldsymbol{\beta}_i \end{pmatrix} &= \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \boldsymbol{\Pi}\mathbf{q}_i + \boldsymbol{\Sigma}\mathbf{v}_i \\ &= \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Pi}_\alpha \\ \boldsymbol{\Pi}_\beta \end{pmatrix}\mathbf{q}_i + \begin{pmatrix} \boldsymbol{\Sigma}_\alpha \\ \boldsymbol{\Sigma}_\beta \end{pmatrix}\begin{pmatrix} v_{\alpha,i} \\ \mathbf{v}_{\beta,i} \end{pmatrix}, \end{aligned} \tag{3.2.8}$$

where $\mathbf{q}_i$ is the demographic variables and $\mathbf{v}_i$ is an idiosyncratic shock. $\boldsymbol{\Pi}$ represents the correlation between demographics and the coefficients, and $\boldsymbol{\Sigma}$ is the correlation between the idiosyncratic shocks $\mathbf{v}_i$ and the coefficients. We assume the distribution of $\mathbf{q}_i$ and $\mathbf{v}_i$ are fully known. For the idiosyncratic shocks $\mathbf{v}_i$, the assumption that $\mathbf{v}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is common. The demographics $\mathbf{q}_i$ can be drawn from the population distribution such as census. See Nevo (2001); Petrin (2002); Berry, Levinsohn, and Pakes (2004a) for discussions and examples on which variables can be included in $\mathbf{q}_i$.

When we allow coefficients to vary across individuals, the right-hand side of the market share equa-

[6] Note the distinction between the market share equation and the individual choice probability equation. This market share equation is a result of aggregation, approximation, and equating with observed data in the left-hand side.

[7] $\mathbf{z}_j$ includes all the exogenous components of $\mathbf{x}_j$.

tion should involve the expectation over i. Specifically,

$$\begin{aligned} s_j &\simeq \pi_j && (3.2.9)\\ &= E\left[\Pr\left(i \text{ Chooses } j\right)\right] && (3.2.10)\\ &= \int_{\mathbf{v}_i}\int_{\mathbf{q}_i} \Pr\left(i \text{ Chooses } j\right) dF\left(\mathbf{q}_i\right) dF\left(\mathbf{v}_i\right) && (3.2.11)\\ &= \int_{\mathbf{v}_i}\int_{\mathbf{q}_i} \frac{\exp\left(-\alpha_i p_j + \mathbf{x}_j'\boldsymbol{\beta}_i + \xi_j\right)}{\sum_{k\in\mathcal{J}_t}\exp\left(-\alpha_i p_k + \mathbf{x}_k'\boldsymbol{\beta}_i + \xi_k\right)} dF\left(\mathbf{q}_i\right) dF\left(\mathbf{v}_i\right) && (3.2.12)\\ &\simeq \frac{1}{n_s}\sum_{i=1}^{n_s} \frac{\exp\left(-\alpha_i p_j + \mathbf{x}_j'\boldsymbol{\beta}_i + \xi_j\right)}{\sum_{k\in\mathcal{J}_t}\exp\left(-\alpha_i p_k + \mathbf{x}_k'\boldsymbol{\beta}_i + \xi_k\right)}. && (3.2.13) \end{aligned}$$

[8] Now the inversion does not have a simple closed-form formula as in (3.2.7). In fact, the invertibility itself of (3.2.12) to recover $(\alpha, \boldsymbol{\beta})$ is not trivial, proof which is one of the major contributions of Berry (1994); Berry, Levinsohn, and Pakes (1995).[9] In the random coefficients logit model, the market share inversion should be done numerically in the estimation, which we discuss below.

(3.2.13) is added to approximate the values of the integrals in (3.2.12) because the closed-form formula for the integral is not available in general. BLP's sollution to this problem was to approximate the exact integral with the means of the simulation draws. Because we assume the distribution $F(\mathbf{q}_i)$ and $F(\mathbf{v}_i)$ are fully known, the distribution of $\mathbf{q}_i$ and $\mathbf{v}_i$ can be simulated. Combined with (3.2.8), we can simulate the distribution of $\Pr(i \text{ Chooses } j)$ for any given set of parameter values $(\alpha, \boldsymbol{\beta}, \boldsymbol{\Pi}, \boldsymbol{\Sigma})$, which leads us to approximate the integral $E[\Pr(i \text{ Chooses } j)]$ with the sample mean (3.2.13) with n_s numbers of simulation draws. We take both *approximations* (3.2.11) and (3.2.13) as *exact* during the estimation.

Two dimensions of approximations, (3.2.11) and (3.2.13), are involved in the market share equation of BLP. In principle, the errors from both approximations should be considered in the inference but are ignored in most of the applications. At least in theory, the approximation error from taking (3.2.13) as equality causes less of a problem because n_s can be set as an arbitrarily large number provided that the computing power and time allows. Different asymptotics are applied in different circumstances for the simulation error to wash out. Berry, Linton, and Pakes (2004b) show n_s should be of an order of $\sqrt{|\mathcal{J}_t|} + \eta$ for some $\eta > 0$ when only one market exists and $|\mathcal{J}_t|$ is growing to infinity. Recently, Freyberger (2015) shows that when the number of markets T grows to infinity with a finite $|\mathcal{J}_t|$, n_s should be of an order of $T^2 + \eta$. We note the upper bound of n_s seems to be in the order of thousands in most applications.

Taking (3.2.11) as equality may cause a more serious conceptual problem. Freyberger (2015) shows that when the number of markets T grows to infinity, the number of consumers $|\mathcal{I}|$ should grow with an order of $T^2 + \eta$, in order for the approximation error from taking (3.2.11) as equality to disappear at least asymptotically. It can be more problematic when only a finite number of consumers exist. Gandhi, Lu, and Shi (2019) argue that any finite approximation that takes (3.2.11) as equality is uninformative for the

[8] Note we implicitly assume $\mathbf{q}_i \perp\!\!\!\perp \mathbf{v}_i$ in (3.2.11).

[9] The most general version of the demand-system invertibility theorem to date can be found in Berry, Gandhi, and Haile (2013). Berry and Haile (2014) study identification of the random coefficients $(\alpha_i, \boldsymbol{\beta}_i)$ based on Berry, Gandhi, and Haile's invertibility results.

probability distribution of the parameters.

3.2.2.3 Estimation of the Random Coefficients Logit Model of Demand

The objective of the BLP estimation is essentially the same with the homogeneous logit model: GMM estimation with the moment condition $E\left[\xi_j|\mathbf{z}_j\right]=0$. All the extra complications arise from the nontrivial invertibility of the market share equation (3.2.12).

Let t denote the market that ranges from 1 to T. What we want to do is find the parameter values of $(\alpha,\boldsymbol{\beta},\boldsymbol{\Pi},\boldsymbol{\Sigma})$ to minimize the following GMM objective function:

$$\sum_{j,t}\left(\xi_{j,t}\mathbf{z}_{j,t}\right)'\boldsymbol{\Omega}\left(\xi_{j,t}\mathbf{z}_{j,t}\right),$$

where $\boldsymbol{\Omega}$ is the GMM weighting matrix. Consider the unobservable $\xi_{j,t}$. It would be some function of the parameters $(\alpha,\boldsymbol{\beta},\boldsymbol{\Pi},\boldsymbol{\Sigma})$ and the data $\left(p_{j,t},\mathbf{x}_{j,t},\mathbf{q}_i,\mathbf{v}_i\right)$. Denote the function by $\xi_{j,t}\left(\alpha,\boldsymbol{\beta},\boldsymbol{\Pi},\boldsymbol{\Sigma}|p_{j,t},\mathbf{x}_{j,t},\mathbf{q}_i,\mathbf{v}_i\right)$. In the homogeneous logit model,

$$\begin{aligned}\xi_{j,t} &\equiv \xi_{j,t}\left(\alpha,\boldsymbol{\beta},\boldsymbol{\Pi},\boldsymbol{\Sigma}|p_{j,t},\mathbf{x}_{j,t},\mathbf{q}_i,\mathbf{v}_i\right)\\ &= \xi_{j,t}\left(\alpha,\boldsymbol{\beta}|p_{j,t},\mathbf{x}_{j,t}\right)\\ &= \ln\left(\frac{s_j}{s_0}\right)-\left(-\alpha p_{j,t}+\mathbf{x}'_{j,t}\boldsymbol{\beta}\right),\end{aligned}$$

which is linear. Regarding the random coefficients logit model, closed-form expression for inversion no longer exists. What can be only shown is that such a function $\xi_{j,t}\left(\alpha,\boldsymbol{\beta},\boldsymbol{\Pi},\boldsymbol{\Sigma}|p_{j,t},\mathbf{x}_{j,t},\mathbf{q}_i,\mathbf{v}_i\right)$ exists and the value at each parameter value can be calculated in a certain nonlinear way. To that end, the optimization problem that BLP solves is

$$\min_{\alpha,\boldsymbol{\beta},\boldsymbol{\Pi},\boldsymbol{\Sigma}}\sum_{j,t}\left(\xi_{j,t}\left(\alpha,\boldsymbol{\beta},\boldsymbol{\Pi},\boldsymbol{\Sigma}|p_{j,t},\mathbf{x}_{j,t},\mathbf{q}_i,\mathbf{v}_i\right)\mathbf{z}_{j,t}\right)'\boldsymbol{\Omega}\left(\xi_{j,t}\left(\alpha,\boldsymbol{\beta},\boldsymbol{\Pi},\boldsymbol{\Sigma}|p_{j,t},\mathbf{x}_{j,t},\mathbf{q}_i,\mathbf{v}_i\right)\mathbf{z}_{j,t}\right). \tag{3.2.14}$$

The original BLP paper suggested the nested fixed point (NFP) algorithm to calculate $\delta_{j,t}$ at each possible parameter value. The NFP algorithm solves the fixed-point equation to find $\xi_{j,t}$ for each candidate parameter value $(\alpha,\boldsymbol{\beta},\boldsymbol{\Pi},\boldsymbol{\Sigma})$ at each iteration. The fixed-point algorithm places substantial computational burden.

Recent developments on the numerical optimization techniques led to the mathematical programming under equality constraints (MPEC) approach, which is much faster and more reliable. The optimization problem (3.2.14) can be rewritten as

$$\min_{\{\xi_{j,t},\zeta_{j,t}\}_{j\in\mathcal{J}_t,t\in\mathcal{T}},\alpha,\boldsymbol{\beta},\boldsymbol{\Pi},\boldsymbol{\Sigma}} \sum_{j,t} \boldsymbol{\zeta}'_{j,t}\boldsymbol{\Omega}\boldsymbol{\zeta}_{j,t}$$

$$s.t. \quad s_{j,t} = \frac{1}{n_s}\sum_{i=1}^{n_s} \frac{\exp\left(-\alpha_i p_{j,t} + \mathbf{x}'_{j,t}\boldsymbol{\beta}_i + \xi_{j,t}\right)}{\sum_{k\in\mathcal{J}_t}\exp\left(-\alpha_i p_{k,t} + \mathbf{x}'_{k,t}\boldsymbol{\beta}_i + \xi_{k,t}\right)}. \tag{3.2.15}$$

$$\boldsymbol{\zeta}_{j,t} = \xi_{j,t}\mathbf{z}_{j,t} \qquad \forall j,t \tag{3.2.16}$$

Dube, Fox, and Su (2012) show (3.2.15) is equivalent to (3.2.14); that is, they both yield the same solution. Notice that (3.2.15) is an implicit function of $\xi_{j,t}$'s. In a sense, MPEC leaves the nonlinear fixed-point inversion problem of finding $\delta_{j,t}$ to the numerical optimizers. The major speed advantage of MPEC comes from the fact that the the optimizer does not need to solve for fixed point at every candidate parameter value. MPEC requires the fixed0point equation to be satisfied only at the optimum. For further details on the estimation, we encourage the readers to refer to the online Appendices of Dube, Fox, and Su (2012).

3.2.3 Discussion on the Static Logit Demand Models

3.2.3.1 Own- and Cross-Price Elasticities: Independence of Irrelevant Alternatives

Let $\pi_{i,j,t} := \Pr(i \text{ Chooses } j|t)$ be the individual predicted quantity shares. The own- and cross-price elasticities of the demand system specified by (3.2.12) are

$$\begin{aligned}\varepsilon_{jk,t} &= \frac{\partial \pi_{j,t}}{\partial p_{k,t}}\frac{p_{k,t}}{\pi_{j,t}} \\ &= \begin{cases} -\frac{p_{j,t}}{\pi_{j,t}}\int_{\mathbf{v}_i}\int_{\mathbf{z}_i}\alpha_i\pi_{i,j,t}\left(1-\pi_{i,j,t}\right)dF\left(\mathbf{q}_i\right)dF\left(\mathbf{v}_i\right) & \text{if } j=k \\ \frac{p_{k,t}}{\pi_{j,t}}\int_{\mathbf{v}_i}\int_{\mathbf{z}_i}\alpha_i\pi_{i,j,t}\pi_{i,k,t}dF\left(\mathbf{q}_i\right)dF\left(\mathbf{v}_i\right) & \text{otherwise.}\end{cases}\end{aligned} \tag{3.2.17}$$

Consider the homogeneous consumers case, that is, $\alpha_i = \alpha$, $\boldsymbol{\beta}_i = \boldsymbol{\beta}$. (3.2.17) simplifies to

$$\varepsilon_{jk,t} = \begin{cases} -\alpha p_{j,t}\left(1-\pi_{j,t}\right) & \text{if } j=k \\ \alpha p_{k,t}\pi_{k,t} & \text{otherwise.}\end{cases} \tag{3.2.18}$$

[10][11] The elasticity expression (3.2.18) can be problematic, because the own- and cross-price elasticities depend only on the prices and the market shares. In particular, the form of the cross-price elasticities restricts the substitution pattern in a very strict way, which we illustrate by the following example. Suppose only three car brands exist, KIA, Benz, and BMW, with shares of 33% each. If the price of Benz

[10] The price elasticities here could be calculated either by using the estimated model parameters or by substituting the observed market shares $s_{j,t}$'s back into $\pi_{j,t}$'s. The former method is especially relevant when a researcher is conducting policy counterfactuals.

[11] Note these elasticity expressions change with respect to the functional form of the utility $u_{i,j}$.

increases by 1%, the same number of consumers will substitute toward KIA and BMW, respectively. This pattern, often referred to as the independence of irrelevant alternatives (IIA) properties, is unrealistic because BMW is more likely to substitute Benz than KIA.[12]

The prime motivation for the extension to the random coefficients logit model from the homogeneous logit model was to overcome this IIA property. IIA is resolved in (3.2.17) at least at the market level. Yet, such a solution comes at its own cost: the increased complication in the estimation, computational burden, and possible numerical instability of the estimators. Furthermore, although the IIA is overcome at the market level in the random coefficients logit model, the problem still persists within an individual. Another way to avoid IIA is to assume a correlated error structure such as nested logit. Goldberg (1995); Goldberg and Verboven (2001) study the car market using the lested logit demand model.

3.2.3.2 Inclusion of the Numeraire in the Alternative Set

Consider the individual choice probability equation (3.2.4). Both in the homogeneous and random coefficients logit model of demand, the set of alternatives $\mathcal{J}_t$ must include the numeraire. Otherwise, a price increase of product j will only lead to switching to other products in the choice set $\mathcal{J}_t$; that is, no one will exit from the market. This problem is inherent because the utility-maximization problem of the discrete-choice random utility framework is not budget constrained. To resolve this problem, the literature has chosen to include the numeraire in the consumer's choice set $\mathcal{J}_t$ directly. The market share of the numeraire s_0, which is often referred to as the *outside shares*, is imputed by the researcher. One imminent challenge is that the researcher has to either know or be able to estimate the *potential* market size. The other tentative problem is that combined with the IIA properties, the magnitude of the income effect is in a sense predetermined by the researcher.[13]

3.2.3.3 Choice of the Instruments for Prices

Recall the utility specification and the estimation equation of the homogeneous logit model of demand:

$$\begin{aligned} u_{i,j} &= \alpha\left(w - p_j\right) + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j + \epsilon_{i,j} \\ \ln\left(\frac{\pi_j}{\pi_0}\right) &= -\alpha p_j + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j. \end{aligned} \tag{3.2.19}$$

ξ_j represents the quality of product j, which is not captured by the utility from observable product characteristics $\mathbf{x}_j'\boldsymbol{\beta}$. The assumption that different characteristics of a product are independent of each other in many circumstances is plausible. The observable product characteristics $\mathbf{x}_j$ can be taken as exogenous when we consider the consumer's choice. Thus, we can assume $\mathbf{x}_j \perp\!\!\!\perp \xi_j$, or at least $Cov\left(\mathbf{x}_j, \xi_j\right) = \mathbf{0}$.[14] However, price p_j cannot be taken as exogenous, because the firm's pricing decision should be consid-

[12] We note the source of IIA is not by the T1EV assumption, but by the i.i.d. assumption of $\epsilon_{i,j}$'s. The IIA property will persist when we assume, for example, $\epsilon_{i,j} \sim i.i.d.\,\mathcal{N}(0,1)$.

[13] Recall, for example, the price elasticities of the homogeneous logit model (3.2.18). The income effect in this context where the utility is linear in prices net of income is cross-price elasticity with the good 0, which is dictated by $\pi_{0,t}$.

[14] Incorporating endogenous product characteristics is an ongoing area of research. See, for example, Fan (2013).

ered. The firm will charge a higher price for the products that have a higher (marginal) cost of production, and the costs tend to increase with the unobserved quality term ξ_j.

In that regards, we can reasonably consider the firms' pricing function, arguments of which include the observed and unobserved product characteristics. Because price p_j is some function of $\mathbf{x}_j$ and ξ_j, an immediate consequence is that a strong endogeneity is involved for the term p_j in the estimation equation (3.2.19). Without an instrument, it is highly likely that

$$E\left[\xi_j | s_j, \mathbf{x}_j, p_j\right] > 0$$

holds, which makes estimating (3.2.19) using OLS implausible. Finding an instrument to correct for the endogeneity in prices has been a major challenge in the literature. In estimating the consumer demand, the valid instrument should be a cost shifter that is uncorrelated with the demand shocks. Because the cost shifter is not directly observable in most circumstances, researchers have tried to find some proxies. The suitable instruments differ by context, individual case, and the data availability. No global general solution exists for the problem of finding a suitable instrument. We introduce two classic examples of the proxies below.

Hausman, Leonard, and Zona (1994) use the means of other cities' prices of the same product as a proxy for the cost shifter. Their identifying assumption is the following. Consider a set of geographically segregated markets. The valuations of each product may vary over markets (demand side), but the marginal cost of a product should be common (supply side). Because the marginal costs are reflected in the prices of other markets concurrently, the means of other markets' prices can be used as a proxy for the cost shifters. Berry, Levinsohn, and Pakes (1995) use the sums of the values of the same characteristics of other products by the same firm, and the sums of the values of the same characteristics of products by other firms. The identifying assumption is similar: Some common cost-shifter component should exist in the summand.

3.2.3.4 Welfare Analysis: Compensating Variation and Equivalent Variation

The compensating variation (CV) and the equivalent variation (EV) have been the primary welfare parameters of interest in the literature, because they easily quantify the utility change with respect to the changes of the prices and/or the choice set. McFadden (1981); Small and Rosen (1981) provide formulas to calculate the compensating variations in the discrete-choice demand models. In the utility-maximization-problem context, compensating variations and equivalent variations are defined such that

$$\begin{aligned} v\left(\mathbf{p}^0, w^0\right) &= v\left(\mathbf{p}^1, w^0 + CV\right) \\ v\left(\mathbf{p}^0, w^0 - EV\right) &= v\left(\mathbf{p}^1, w^0\right) \end{aligned}$$

hold, where $v(\cdot, \cdot)$ is the indirect utility function. The superscripts 0 and 1 denote the values before and after the price/choice set change, respectively. CV and EV can be expressed in terms of the expenditure function as

$$\begin{aligned} e\left(\mathbf{p}^1, u^0\right) + CV &= e\left(\mathbf{p}^1, u^1\right) \\ e\left(\mathbf{p}^0, u^0\right) + EV &= e\left(\mathbf{p}^0, u^1\right), \end{aligned}$$

where $e(\cdot,\cdot)$ is the expenditure function.

For the purpose of illustration, we consider the simplest utility specification thatallows us to derive the closed-form expression for $E[CV]$. Consider the utility specification of the homogeneous logit model:

$$u_{i,j} = \alpha\left(w_i - p_j\right) + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j + \epsilon_{i,j}.$$

Now consider the following equality from the definition of compensating variation:

$$\max_{j\in\mathcal{J}^0}\left\{\alpha\left(w_i - p_j^0\right) + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j + \epsilon_{i,j}\right\} = \max_{j\in\mathcal{J}^1}\left\{\alpha\left(w_i + CV - p_j^1\right) + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j + \epsilon_{i,j}\right\}.$$

Rearranging and taking the expectation yields

$$E[CV] = \frac{1}{\alpha}\left\{\ln\left(\sum_{j\in\mathcal{J}^0}\exp\left(\alpha\left(w_i - p_j^0\right) + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j\right)\right) - \ln\left(\sum_{j\in\mathcal{J}^1}\exp\left(\alpha\left(w_i - p_j^1\right) + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j\right)\right)\right\}. \tag{3.2.20}$$

[15] It is also immediate that $E[CV] = E[EV]$ in this linear case. We note that if the utility from the wealth is nonlinear, the above closed-form formula no longer works.[16] However, EV and CV can be still calculated using, for example, simulation.

In the context of the distinction between Marshallian and Hicksian demand, the "exact" compensating variation and equivalent variation should be calculated according to the Hicksian demand system (see, e.g., Hausman and Newey (1995, 2016)), because the real income change has to be compensated before calculating the consumer surplus. Marshallian and Hicksian demand systems differ by the income effect. Because no such distinction exists between the Marshallian and Hicksian demand systems in the discrete-choice demand systems when the utility is linear in income net of prices,[17] the "exact" consumer surplus coincides with (3.2.20).

---

[15] We used the fact that when $\epsilon_{i,j} \sim i.i.d.\ T1EV$,

$$E\left[\max_j \delta_j + \epsilon_{i,j}\right] = \ln\left[\sum_{j=1}^{J}\exp\left(\delta_j\right)\right] + C,$$

where the Euler-Mascheroni constant $C \approx 0.5772$. McFadden (1981) calls this function the social surplus function. Note that when δ_j is linear in p_j, taking partial derivatives with respect to p_j gives the choice probability of j. This fact is often called the Williams-Daly-Zachary theorem in the literature.

[16] See Herriges and Kling (1999); Dagsvik and Karlström (2005) for the case in which utility is nonlinear in income net of prices.

[17] Mainly because the budget is not exhaustive on the products.

3.2.3.5 Aggregation and Integrability of Discrete-Choice Demand Systems

Many authors have attempted to place the aggregated discrete-choice demand systems in the context of Marshallian and Hicksian demand systems, a result of the budget-constrained utility-maximization problem. The classical integrability theorem by Hurwicz and Uzawa (1971) does not apply, because one of the critical conditions for their theorem to hold is that the demand is zero whenever income is zero.[18] Only recently, Nocke and Schutz (2017) suggested a general quasi-linear integrability theorem, which can rationalize the discrete-choice demand systems as resulting from a budget-constrained quasi-linear utility-maximization problem. We note, however, that the Walrasian (direct) utility function that can rationalize the discrete-choice demand systems does not have all the desired properties for a direct utility function in the following aspects: (i) It takes money (numeraire) as a direct argument, (ii) the utility function is not defined on a subset of the commodity space,[19] and (iii) a large subset of the commodity space has to be composed of satiation points.

3.2.4 Accommodating Zero Market Shares

Zero market shares are often observed in demand data. The assumptions of the aggregate discrete-choice demand systems that the choice probabilities are strictly positive and there are infinitely many underlying consumers are inconsistent with the observed zero market shares. When the observed market shares are zero for some alternatives, the implied utility of those alternatives should be negative infinity, which is problematic. Two papers in the literature provide solutions accommodate zero observed market shares in the literature.

The first approach, Gandhi, Lu, and Shi (2019), is to treat zero observed market shares as a result of measurement errors, caused by small number of underlying consumer population. Gandhi et al. first pushes the observed market shares s_j in to $(0, 1)$ using the Laplace correction factor, and use the set identification and bound estimation methods developed by, e.g., Manski and Tamer (2002); Imbens and Manski (2004); Chernozhukov, Hong, and Tamer (2007); Ciliberto and Tamer (2009).

The second approach, Dubé, Hortaçsu, and Joo (2020), considers the zero observed market shares as as result of consumers' selection on the choice set, possibly caused by consumers' consideration. Let d_j be the indicator function such that the value is 1 when the $s_j > 0$ and 0 when $s_j = 0$. When there is selection on unobservables ξ_j, the original GMM moment condition imposed by BLP after dropping the alternatives with zero observed market shares, $E\left[\xi_j|\mathbf{z}_j, d_j = 1\right]$, is not zero anymore even when $E\left[\xi_j|\mathbf{z}_j\right] = 0$. Ignoring the selection may bias the demand parameter estimates.

Dubé et al. considers an alternative GMM moment condition, based on the pairwise-difference estimation literature (e.g., Ahn and Powell, 1993; Blundell and Powell, 2004; Aradillas-Lopez et al., 2007; Aradillas-Lopez, 2012). Let μ_j denote the continuously-distributed propensity score for alternative j not

[18] For the discrete-choice demand systems, the sum of the demand is fixed at some positive level even when the income is zero.

[19] Or, equivalently, some commodity bundles should yield the utility of negative infinity.

being selected. First, re-write the demand shocks as follows:

$$\xi_j = \iota\left(\mu_j\right) + \zeta_j,$$

with $\mathbb{E}\left[\zeta_j | \mu_j, \mathbf{z}_j\right] = 0$ and $\iota\left(\cdot\right)$ being a smooth, monotone increasing function. Then, consider the following moment condition based on differences between the demand shocks of two goods i and j:

$$E\left[\xi_i - \xi_j | \mathbf{z}_i, \mathbf{z}_j, \mu_i = \mu_j, d_i = d_j = 1\right] = 0.$$

The moment condition above is based on differencing out the selection propensities. To illustrate, cnsider a pair of products i and j with the same $\mu_i = \mu_j$, and hence $\iota\left(\mu_i\right) = \iota\left(\mu_j\right)$ in the homogeneous Logit model:

$$\begin{aligned} \ln\left(\frac{s_i}{s_0}\right) &= -\alpha p_i + \mathbf{x}_i' \boldsymbol{\beta} + \iota\left(\mu_i\right) + \zeta_i \\ \ln\left(\frac{s_j}{s_0}\right) &= -\alpha p_j + \mathbf{x}_j' \boldsymbol{\beta} + \iota\left(\mu_j\right) + \zeta_j. \end{aligned}$$

Taking differences yields:

$$\ln\left(\frac{s_i}{s_0}\right) - \ln\left(\frac{s_j}{s_0}\right) = -\alpha\left(p_i - p_j\right) + \left(\mathbf{x}_i - \mathbf{x}_j\right)' \boldsymbol{\beta} + \left(\zeta_i - \zeta_j\right).$$

Then, the following conditional moment condition can be derived:

$$\begin{aligned} 0 &= E\left[\zeta_i - \zeta_j | \mathbf{z}_i, \mathbf{z}_j, \mu_i = \mu_j, d_i = d_j = 1\right] \\ &= E\left[\xi_i - \xi_j | \mathbf{z}_i, \mathbf{z}_j, \mu_i = \mu_j, d_i = d_j = 1\right]. \end{aligned} \tag{3.2.21}$$

The same moment condition applies to the random-coefficients Logit demand model.

(3.2.21) leads to the GMM objective function and constraints based on pairwise-difference:

$$\frac{2}{n\left(n-1\right)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \hat{\omega}_{i,j} \left\{\Delta \mathbf{z}_{i,j} \Delta \xi_{i,j}\right\}' \boldsymbol{\Omega} \left\{\Delta \mathbf{z}_{i,j} \Delta \xi_{i,j}\right\}$$

where $\hat{\omega}_{i,j}$ is the estimated nonparametric kernel-weights that places more weights to the pair of observations where μ_i and μ_j are close. The kernel weights are introduced because the event $\mu_i = \mu_j$ occurs with probability 0 in theory. The same set of market share equations should be imposed either as constraints (MPEC) or as inner-loops (NFXP), as in the Berry et al. (1995) estimation.

3.2.5 Characteristics-Space Approach without Random Utility Shocks

Bajari and Benkard (2005); Berry and Reiss (2007) develop demand-estimation frameworks in the characteristics space that do not rely on the random utility shocks. In line with Bajari and Benkard; Berry and Reiss, Dubé et al. (2020) derive the same market share equation of the homogeneous and random-choice logit demand models from the budget-constrained constant elasticity of substitution (CES henceforth) utility-maximization problem, implying the CES Marshallian demand system is just as rich as the logit demand system to incorporate the product characteristics and demographics. Furthermore, the identification results and estimation methods developed for the logit demand systems can be directly applied to estimate the CES Marshallian demand systems. Additionally, Dubé et al.'s framework can accommodate the zero individual choice probabilities by separating the intensive and extensive margins, and recognizing the zero market shares as a result of the endogenous selection. We review Dubé et al.'s result here.

3.2.5.1 Equivalence of the Predicted Market Shares Derived from CES Marshallian Demand System and Logit Demand System

Consider a differentiated-products market denoted by subscript t, composed of homogeneous consumers with a CES preference. The representative consumer solves

$$\max_{\{q_{j,t}\}_{j\in\mathcal{J}_t}} \left(\sum_{j\in\mathcal{J}_t} \left\{ \chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right) \right\}^{\frac{1}{\sigma}} q_{j,t}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}} \quad s.t. \quad \sum_{j\in\mathcal{J}_t} p_{j,t} q_{j,t} = y_t. \tag{3.2.22}$$

$\mathcal{J}_t$ is a set of alternatives in the category, which might include the numeraire that represents the outside option. $q_{j,t}$ is the quantity of product j consumed in market t. $\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right)$, defined by the quality kernel, is a non-negative function of observed and unobserved product characteristics. $\mathbf{x}_{j,t}$ and $\mathbf{w}_{j,t}$ are vectors of product j's characteristics in market t, which are observable to the econometrician. $\xi_{j,t}$ and $\eta_{j,t}$ are scalars that represent utility from product j's characteristics that are unobservable to the econometrician. $\mathbf{w}_{j,t}$ and $\eta_{j,t}$ are extensive-margin shifters that a consumer considers whether to buy the product. $\mathbf{x}_{j,t}$ and $\xi_{j,t}$ are intensive margin shifters that determine the level of utility when a consumer buys a product. $\mathbf{w}_{j,t}$ and $\mathbf{x}_{j,t}$ might have common components, but we can require an exclusion restriction on $\mathbf{w}_{j,t}$ for semiparametric identification when the extensive margin matters. In such a case, $\mathbf{w}_{j,t}$ must contain at least one component that is not in $\mathbf{x}_{j,t}$. Finally, we assume the existence of the instrument for the prices $\mathbf{z}_{j,t}$.

It is immediate that the Marshallian demand system is

$$q_{j,t} = y_t \left\{ \frac{\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right) p_{j,t}^{-\sigma}}{\sum_{k\in\mathcal{J}_t} \chi\left(\mathbf{x}_{k,t}, \xi_{k,t}, \mathbf{w}_{k,t}, \eta_{k,t}\right) p_{k,t}^{1-\sigma}} \right\} \qquad \forall j \in \mathcal{J}_t, \tag{3.2.23}$$

which leads to the predicted quantity market shares $\pi_{j,t}$ as

$$\begin{aligned} \pi_{j,t} &\equiv \frac{q_{j,t}}{\sum_{k\in\mathcal{J}_t} q_{k,t}} \\ &= \frac{\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right) p_{j,t}^{-\sigma}}{\sum_{k\in\mathcal{J}_t} \chi\left(\mathbf{x}_{k,t}, \xi_{k,t}, \mathbf{w}_{k,t}, \eta_{k,t}\right) p_{k,t}^{-\sigma}}. \end{aligned} \tag{3.2.24}$$

Now consider a functional form for $\chi(\cdot)$:

$$\chi\left(\mathbf{x}_{j,t}, \xi_{j,t}, \mathbf{w}_{j,t}, \eta_{j,t}\right) = \exp\left(\alpha + \mathbf{x}_{j,t}'\boldsymbol{\beta} + \xi_{j,t}\right). \tag{3.2.25}$$

(3.2.24) becomes

$$\pi_{i,j} = \frac{\exp\left(-\sigma \ln p_j + \mathbf{x}_j'\boldsymbol{\beta} + \xi_j\right)}{\sum_{k\in\mathcal{J}_t} \exp\left(-\sigma \ln p_k + \mathbf{x}_k'\boldsymbol{\beta} + \xi_k\right)}, \tag{3.2.26}$$

which is the predicted market share equation of the logit demand where the price is replaced by the logarithm of the prices. Taking the logarithm on the prices can be taken as a scale adjustment on the utility of each alternative.[20] If we allow the model parameters $(\sigma, \boldsymbol{\beta})$ to vary across the individuals i, we obtain the Berry, Levinsohn, and Pakes (1995) predicted individual share equation:

$$\pi_{i,j} = \frac{\exp\left(-\sigma_i \ln p_j + \mathbf{x}_j'\boldsymbol{\beta}_i + \xi_j\right)}{\sum_{k\in\mathcal{J}_t} \exp\left(-\sigma_i \ln p_k + \mathbf{x}_k'\boldsymbol{\beta}_i + \xi_k\right)}. \tag{3.2.27}$$

Aggregating (3.2.27) over individuals and equating with observed market shares yields the estimation equation of Berry, Levinsohn, and Pakes (1995).

3.2.5.2 Marshallian and Hicksian Price Elasticities and Income Elasticity

Usually, the first-order goal of demand estimation is to find the price and income elasticities. Even though the identification results and estimation methods developed by many authors including Berry (1994); Berry, Levinsohn, and Pakes (1995); Berry and Haile (2014) can be directly implemented for the model parameters $(\sigma_i, \boldsymbol{\beta}_i)$, the elasticity expressions of the CES Marshallian demand system are different from the logit demand system.

Let $b_{j,t}$ be the budget share of product j in market t. Consider the homogeneous coefficients case. Denote $\varepsilon^M_{jc,t}$ and $\varepsilon^H_{jc,t}$ by the Marshallian and Hicksian cross-price elasticities between alternatives j and c, respectively. We have the following simple closed-form formulas for the Marshallian and the Hicksian own- and cross-price elasticities:

$$\begin{aligned}
\varepsilon^M_{jj,t} &= -\sigma + (\sigma - 1)\, b_{j,t} \\
\varepsilon^M_{jc,t} &= (\sigma - 1)\, b_{c,t} \\
\varepsilon^H_{jj,t} &= -\sigma\left(1 - b_{j,t}\right) \\
\varepsilon^H_{jc,t} &= \sigma b_{c,t},
\end{aligned} \tag{3.2.28}$$

and the income elasticity is 1. When budget shares are observed in the data or calculated from the data,

[20] Note the "indirect utility" interpretation of the discrete-choice utility no longer holds if we use the logarithm of the prices instead of the raw prices. However, considering that the indirect utility interpretation in the discrete-choice problem holds only in a restricted setup (e.g., the household makes only one discrete choice among its choice problem), we argue it is reasonable to interpret the utility from each choice alternative as the direct utility.

these elasticities can be calculated easily. From these elasticity expressions, it can be immediately noticed that a version of the independence of irrelevant alternatives (IIA) property holds; the substitution pattern depends solely on the budget shares of corresponding products.[21] Welfare counterfactuals should be done accordingly following the formulas from the CES Marshallian demand system.

[21] If $\mathbf{w}_{j,t}$ includes the prices or a function of the prices so that the extensive margin is affected by the price changes, the simple closed-form expressions for the own- and cross-price elasticities cannot be derived. In practice, the corresponding price elasticities can be calculated using simulations.

References

ACKERBERG, D., C. L. BENKARD, S. BERRY, AND A. PAKES (2007): *Econometric tools for analyzing market outcomes*, Elsevier, vol. 6A of *Handbook of Econometrics*, chap. 63, 4171–4276.

AHN, H. AND J. L. POWELL (1993): "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, 58, 3–29.

ARADILLAS-LOPEZ, A. (2012): "Pairwise-difference estimation of incomplete information games," *Journal of Econometrics*, 168, 1.

ARADILLAS-LOPEZ, A., B. E. HONORÉ, AND J. L. POWELL (2007): "Pairwise difference estimation with nonparametric control variables," *International Economic Review*, 48, 1119–1158.

BAJARI, P. AND C. L. BENKARD (2005): "Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach," *Journal of Political Economy*, 113, 1239–1276.

BERRY, S. (1994): "Estimating discrete-choice models of product differentiation," *RAND Journal of Economics*, 25, 242–262.

BERRY, S., A. GANDHI, AND P. HAILE (2013): "Connected substitutes and invertibility of demand," *Econometrica*, 81, 2087–2111.

BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): "Automobile prices in market equilibrium," *Econometrica*, 63, 841–890.

——— (2004a): "Differentiated products demand systems from a combination of micro and macro data: The new car market," *Journal of Political Economy*, 112, 68–105.

BERRY, S., O. B. LINTON, AND A. PAKES (2004b): "Limit theorems for estimating the parameters of differentiated product demand systems," *Review of Economic Studies*, 71, 613–654.

BERRY, S. AND A. PAKES (2007): "The pure characteristics demand model," *International Economic Review*, 48, 1193–1225.

BERRY, S. T. AND P. A. HAILE (2014): "Identification in differentiated products markets using market level data," *Econometrica*, 82, 1749–1797.

BLUNDELL, R. W. AND J. L. POWELL (2004): "Endogeneity in semiparametric binary response models," *Review of Economic Studies*, 71, 655–679.

CHAUDHURI, S., P. K. GOLDBERG, AND P. JIA (2006): "Estimating the effects of global patent protection in pharmaceuticals: A case study of Quinolones in India," *American Economic Review*, 96, 1477–1514.

CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): "Estimation and confidence regions for parameter sets in econometric models," *Econometrica*, 75, 1243–1284.

CILIBERTO, F. AND E. TAMER (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77, 1791–1828.

DAGSVIK, J. K. AND A. KARLSTRÖM (2005): "Compensating variation and Hicksian choice probabilities in random utility models that are nonlinear in income," *Review of Economic Studies*, 72, 57–76.

DEATON, A. AND J. MUELLBAUER (1980): "An almost ideal demand system," *American Economic Review*, 70, 312–326.

DUBE, J.-P., J. T. FOX, AND C.-L. SU (2012): "Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation," *Econometrica*, 80, 2231–2267.

DUBÉ, J.-P., A. HORTAÇSU, AND J. JOO (2020): "Random-coefficients Logit demand estimation with zero-valued market shares," *Working Paper*.

FAN, Y. (2013): "Ownership consolidation of product characteristics: A study of the US daily newspaper market," *American Economic Review*, 103, 1598–1628.

FREYBERGER, J. (2015): "Asymptotic theory for differentiated products demand models with many markets," *Journal of Econometrics*, 185, 162–181.

GANDHI, A., Z. LU, AND X. SHI (2019): "Estimating demand for differentiated products with zeroes in market share data," Unpublished Manuscript.

GOLDBERG, P. K. (1995): "Product differentiation and oligopoly in international markets: The case of the U.S. automobile industry," *Econometrica*, 63, 891–951.

GOLDBERG, P. K. AND F. VERBOVEN (2001): "The evolution of price dispersion in the European car market," *Review of Economic Studies*, 68, 811–848.

HANSEN, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

HAUSMAN, J., G. LEONARD, AND J. D. ZONA (1994): "Competitive analysis with differentciated products," *Annales D'Economie et de Statistique*, 34, 159–180.

HAUSMAN, J. A. AND W. K. NEWEY (1995): "Nonparametric estimation of exact consumers surplus and deadweight loss," *Econometrica*, 63, 1445–1476.

——— (2016): "Individual heterogeneity and average welfare," *Econometrica*, 84, 1225–1248.

HENDEL, I. AND A. NEVO (2006a): "Measuring the implications of sales and consumer inventory behavior," *Econometrica*, 74, 1637–1673.

——— (2006b): "Sales and consumer inventory," *RAND Journal of Economics*, 37, 543–561.

HERRIGES, J. A. AND C. L. KLING (1999): "Nonlinear income effects in random utility models," *Review of Economics and Statistics*, 81, 62–72.

HURWICZ, L. AND H. UZAWA (1971): *On the integrability of demand functions*, Harcourt Brace Jovanovich, chap. 6, 114–148, Preferences, Utility, and Demand.

IMBENS, G. W. AND C. F. MANSKI (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72, 1845–1857.

LEE, J. AND K. SEO (2015): "A computationally fast estimator for random coefficients logit demand models using aggregate data," *RAND Journal of Economics*, 46, 86–102.

MANSKI, C. F. AND E. TAMER (2002): "Inference on regressions with interval data on a regressor or outcome," *Econometrica*, 70, 519–546.

MCFADDEN, D. (1981): *Econometric models of probabilistic choice*, The MIT Press, chap. 5, 198–272, Structural analysis of discrete data with econometric applications.

NEVO, A. (2000): "A practitioner's guide to estimation of random-coefficients logit models of demand," *Journal of Economics and Management Strategy*, 9, 513–548.

——— (2001): "Measuring market power in the ready-to-eat cereal industry," *Econometrica*, 69, 307–342.

NOCKE, V. AND N. SCHUTZ (2017): "Quasi-linear integrability," *Journal of Economic Theory*, 169, 603–628.

PETRIN, A. (2002): "Quantifying the benefits of new products: The case of the minivan," *Journal of Political Economy*, 110, 705–729.

REISS, P. C. AND F. A. WOLAK (2007): *Structural econometric modeling: Rationales and examples from industrial organization*, Elsevier, vol. 6A of *Handbook of Econometrics*, chap. 64, 4277–4415.

SMALL, K. A. AND H. S. ROSEN (1981): "Applied welfare economics with discrete choice models," *Econometrica*, 49, 105–130.

Chapter 4

Estimation of Discrete-Game Models with Cross-sectional Data

In this chapter, we study estimation of model parameters characterized by the pure-strategy Nash equilibria when cross-sectional data are available to the researcher. The models covered in this section do not involve dynamics, in the sense that the players are not forward looking at the time when the payoffs are realized.[1] We study the discrete games with complete and incomplete information, respectively. The literature has focused on the entry in the oligopolistic markets, which we focus on in our presentation as well.

As in other structural econometric models, estimation of model parameters of a game typically assumes the observed data are coming from a single equilibrium, which in this case is a pure-strategy Nash equilibrium. A common challenge is that the underlying game models often have multiple equilibria, and thus some measure to handle this problem must exist. Multiplicity of equilibria is problematic in the estimation of model parameters as well as in the model predictions and counterfactuals. The measures taken in the literature are equilibrium refinement, the equilibrium-selection mechanism, and set-identification approach.[2] We provide an overview of each of the measures in chapters 4 and 5.

4.1 Static Discrete Games with Complete Information

4.1.1 Simultaneous Entry

4.1.1.1 Setup and Equilibrium

Estimation of static discrete-game models with complete information was pioneered by Bresnahan and Reiss (1990, 1991a). Our presentation closely follows their works. We consider the simplest setup for illustration: one-shot simultaneous complete information entry/exit games played by two players, denoted by $i \in \{1, 2\}$. The reduced-form long-run equilibrium monopoly profit function in market j is

[1] The games we study in this chapter may include, for example, the subgame perfect Nash equilibria.

[2] In contrast to point identification.

$\mathbf{x}_{i,j}'\boldsymbol{\alpha}_i + \epsilon_{i,j}$, where $\epsilon_{i,j} \perp\!\!\!\perp \mathbf{x}_{i,j}$.[3] To that end, consider the following payoff matrix:

$$\begin{array}{ccc} & O_2 & E_2 \\ O_1 & (0,0) & \left(0, \mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 + \epsilon_{2,j}\right) \\ E_1 & \left(\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 + \epsilon_{1,j}, 0\right) & \left(\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 - \delta_1 + \epsilon_{1,j}, \mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 - \delta_2 + \epsilon_{2,j}\right) \end{array}.$$

If only one firm enters, the firm enjoys the monopoly profit $\mathbf{x}_{i,j}'\boldsymbol{\alpha}_i + \epsilon_{i,j}$. If both enter, they enjoy duopoly profit $\mathbf{x}_{i,j}'\boldsymbol{\alpha}_i - \delta_i + \epsilon_{i,j}$. $\epsilon_{i,j}$'s are not observable to the econometrician, but both players know the realization of $\left(\epsilon_{1,j}, \epsilon_{2,j}\right)$ before making the entry decision. We assume the econometrician knows $F\left(\epsilon_{1,j}, \epsilon_{2j} | \boldsymbol{\theta}\right)$, the joint cumulative distribution function of $\left(\epsilon_{1,j}, \epsilon_{2,j}\right)$ up to a finite-dimensional parameter $\boldsymbol{\theta}$. We only consider the pure-strategy Nash equilibria of the game. Note we define δ_i to be the difference between the monopoly profit and the duopoly profit of firm i. A natural restriction in the context of entry game is $\delta_i > 0$ for $i = 1, 2$.[4] The model parameters subject to estimation are $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2, \boldsymbol{\theta})$. We suppress $\boldsymbol{\theta}$ for the sake of notational convenience unless necessary.

We assume that the observed data are coming from the pure-strategy Nash equilibrium. The pure-strategy Nash equilibrium realization depends on the parameters and the realizations of $\epsilon_{i,j}$'s. Firm i enters market j if it earns a nonnegative profit, that is, $\mathbf{1}\left(\pi_{i,j} \geq 0\right)$. Hence, we can summarize the realizations of Nash equilibrium and its corresponding region of $\left(\epsilon_{1,j}, \epsilon_{2,j}\right)$ as follows:

- $\mathcal{R}_3 = \{(O_1, O_2) \text{ is observed}\} = \left\{\left(\epsilon_{1,j}, \epsilon_{2,j}\right) : \mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 + \epsilon_{1,j} < 0 \text{ and } \mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 + \epsilon_{2,j} < 0\right\}$
- $\mathcal{R}_1 = \{(E_1, E_2) \text{ is observed}\} = \left\{\left(\epsilon_{1,j}, \epsilon_{2,j}\right) : \mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 - \delta_1 + \epsilon_{1,j} > 0 \text{ and } \mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 - \delta_2 + \epsilon_{2,j} > 0\right\}$
- $\mathcal{R}_2 + \mathcal{R}_5 = \{(E_1, O_2) \text{ is observed}\} = \left\{\left(\epsilon_{1,j}, \epsilon_{2,j}\right) : \mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 + \epsilon_{1,j} > 0 \text{ and } \mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 - \delta_2 + \epsilon_{2,j} < 0\right\}$
- $\mathcal{R}_4 + \mathcal{R}_5 = \{(O_1, E_2) \text{ is observed}\} = \left\{\left(\epsilon_{1,j}, \epsilon_{2,j}\right) : \mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 - \delta_1 + \epsilon_{1,j} < 0 \text{ and } \mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 + \epsilon_{2,j} > 0\right\}$.

$\mathcal{R}_1, ..., \mathcal{R}_5$, the sets of possible realizations of $\left(\epsilon_{1,j}, \epsilon_{2,j}\right)$, depend on the parameter $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2)$. Figure 4.1.1 plots $\mathcal{R}_1, ..., \mathcal{R}_5$ on $\left(\epsilon_{1,j}, \epsilon_{2,j}\right)$ plane.

[3] The equilibrium formulation studied in this subsection can be extended to finite players with finite actions.

[4] Bresnahan and Reiss (1991a) note the possibility that either one or both δ_i can be negative. However, negative δ_i's are not realistic for a non-cooperative entry game, and thus the following literature is mostly focused only on positive δ_i's.

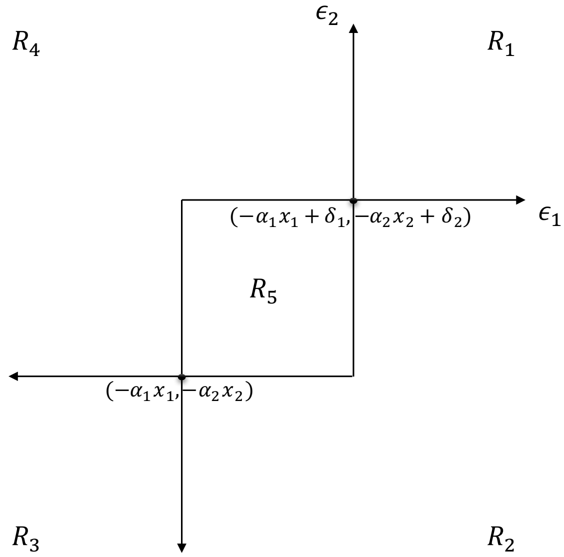

Figure 4.1.1: Equilibrium Strategy Plotted in $\left(\epsilon_{1,j}, \epsilon_{2,j}\right)$ Plane When $\delta_1 > 0$ and $\delta_2 > 0$

The overlapping region $\mathcal{R}_5$ is problematic in the prediction of the model outcomes and, in turn, in the identification and estimation of model parameters. In terms of the prediction, suppose we know the value of the model parameters $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2)$. If the realized value of $\left(\epsilon_{1,j}, \epsilon_{2,j}\right)$ is in $\mathcal{R}_5$, it can either yield the outcome (E_1, O_2) or (O_1, E_2). In other words, $\mathcal{R}_5$ can rationalize both (E_1, O_2) and (O_1, E_2) as an equilibrium outcome, so prediction becomes impossible in this region of $\left(\epsilon_{1,j}, \epsilon_{2,j}\right)$ realization. It becomes more problematic when we consider the identification and estimation of model parameters. Now suppose $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2)$ is unknown and we assume we calculate the action probabilities as usual, ignoring the overlapping region. The sum of the probabilities exceeds 1, and the resulting likelihood is not legitimate. This example nicely illustrates how a fully specified econometric model can be incomplete, in the sense that for a given $(\epsilon_{1,i}, \epsilon_{2,i})$ a range of parameters $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2)$ can rationalize the same outcome of the model. The literature has suggested a few different solutions to address this incompleteness problem, as we discuss below.

4.1.1.2 Incompleteness and Likelihoods

Our primary goal in the estimation is to derive the legitimate likelihoods of observing the data $\left\{\mathbf{x}_{i,j}, \mathbf{1}\left(\pi_{i,j} > 0\right)\right\}_{i\in\{1,2\}}$ The original solution suggested by Bresnahan and Reiss (1990, 1991a) is to exploit only the number of firms in a market. When (δ_1, δ_2) are constants, we can estimate the model parameters $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2, \boldsymbol{\theta})$ using MLE. The likelihoods of observing 0, 1, 2 firms are, respectively,

$$
\begin{aligned}
\mathcal{L}\left(0 \text{ Firm Enters} | \mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2, \boldsymbol{\theta}\right) &= \int_{\mathcal{R}_3} dF\left(\epsilon_{1,j}, \epsilon_{2,j} | \boldsymbol{\theta}\right) \\
\mathcal{L}\left(2 \text{ Firms Enter} | \mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2, \boldsymbol{\theta}\right) &= \int_{\mathcal{R}_1} dF\left(\epsilon_{1,j}, \epsilon_{2,j} | \boldsymbol{\theta}\right) \\
\mathcal{L}\left(1 \text{ Firm Enters} | \mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2, \boldsymbol{\theta}\right) &= 1 - \int_{\mathcal{R}_1} dF\left(\epsilon_{1,j}, \epsilon_{2,j} | \boldsymbol{\theta}\right) - \int_{\mathcal{R}_3} dF\left(\epsilon_{1,j}, \epsilon_{2,j} | \boldsymbol{\theta}\right).
\end{aligned}
$$

Bresnahan and Reiss (1990) further allow δ_1, δ_2 to be nonnegative random variables for more flexibility. To that end, they consider

$$\delta_{i,j} = g\left(\mathbf{z}_{i,j}'\gamma_i\right) + \eta_{i,j} \tag{4.1.1}$$

for $i = 1, 2$, where $g(\cdot)$ is a parametric nonnegative function and $\eta_{i,j}$ is a nonnegative random variable.

This classic approach has three shortcomings: (i) Using only the number of entrants in a market is throwing away some information from data, (ii) uniqueness in the number of entrants in the equilibrium does not always hold,[5] and (iii) generalizing it is difficult when more than two potential entrants exist.

Kooreman (1994) suggest an ad-hoc solution to resolve (i). The idea is to add a nuisance parameter, which is essentially the equilibrium-selection probability in region $\mathcal{R}_5$. As before, the likelihood of observing the samples (O_1, O_2), (E_1, E_2) can be written as

$$\begin{aligned}
\mathcal{L}\left(\{O_1, O_2\}|\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2, \boldsymbol{\theta}\right) &= \Pr\left(\epsilon_{1,j} < -\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1, \epsilon_{2,j} < -\mathbf{x}_{2,j}'\boldsymbol{\alpha}_2\right) \\
&= \int_{\mathcal{R}_3} dF\left(\epsilon_{1,j}, \epsilon_{2,j}|\boldsymbol{\theta}\right) && (4.1.2) \\
\mathcal{L}\left(\{E_1, E_2\}|\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2, \boldsymbol{\theta}\right) &= \Pr\left(\epsilon_{1,j} > -\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 + \delta_1, \epsilon_{2,j} < -\mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 + \delta_2\right) \\
&= \int_{\mathcal{R}_1} dF\left(\epsilon_{1,j}, \epsilon_{2,j}|\boldsymbol{\theta}\right). && (4.1.3)
\end{aligned}$$

For the likelihood of observing the $\{E_1, O_2\}$ and $\{O_1, E_2\}$ pair, we add a parameter $\lambda \in [0, 1]$ that can be jointly estimated. λ can be further parameterized as a function of covariates. The likelihood is legitimate now, which is

$$\begin{aligned}
\mathcal{L}\left(\{O_1, E_2\}|\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2, \boldsymbol{\theta}, \lambda\right) &= \Pr\left(\epsilon_{1,j} < -\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 + \delta_1, \epsilon_{2,j} > -\mathbf{x}_{2,j}'\boldsymbol{\alpha}_2\right) \\
&= \int_{\mathcal{R}_4} dF\left(\epsilon_{1,j}, \epsilon_{2,j}|\boldsymbol{\theta}\right) + \lambda \int_{\mathcal{R}_5} dF\left(\epsilon_{1,j}, \epsilon_{2,j}|\boldsymbol{\theta}\right) \\
\mathcal{L}\left(\{E_1, O_2\}|\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2, \boldsymbol{\theta}, \lambda\right) &= \Pr\left(\epsilon_{1,j} > -\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1, \epsilon_{2,j} < -\mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 + \delta_2\right) \\
&= \int_{\mathcal{R}_2} dF\left(\epsilon_{1,j}, \epsilon_{2,j}|\boldsymbol{\theta}\right) + (1-\lambda) \int_{\mathcal{R}_5} dF\left(\epsilon_{1,j}, \epsilon_{2,j}|\boldsymbol{\theta}\right).
\end{aligned}$$

The nuisance parameter λ is the conditional probability of choosing $\{O_1, E_2\}$ given that the $\left(\epsilon_{1,j}, \epsilon_{2,j}\right)$ realization is in $\mathcal{R}_5$. The idea to add a data-driven equilibrium-selection rule is further elaborated in Bajari, Hong, and Ryan (2010). They allow λ to be a function of the properties of the equilibrium itself: λ can be a function of indicators of pure-strategy equilibrium, Pareto-dominated equilibrium, and joint-payoff-maximizing equilibrium, coefficients of which are estimable from the data.

A modern approach to resolving the incompleteness of the underlying model is to exploit the moment inequalities, which leads to the bound estimates for the partially identified model parameters. Ciliberto and Tamer (2009) give a full treatment on the problem. Suppose we have observed a sample of $\{O_1, E_2\}$. Denote the conditional probability of observing $\{O_1, E_2\}$ given $\left(\mathbf{x}_{1,j}, \mathbf{x}_{2,j}\right)$ by $\Pr\left(\{O_1, E_2\}|\mathbf{x}_{1,j}, \mathbf{x}_{2,j}\right)$,[6]

[5] For example, consider when $\delta_1 < 0$ and $\delta_2 < 0$.

[6] Note the distinction of this conditional probability from the likelihoods defined above. The likelihoods are function of the

which is bounded as

$$\int_{\mathcal{R}_4} dF\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right) \leq \Pr\left(\{O_1,E_2\}\,|\mathbf{x}_{1,j},\mathbf{x}_{2,j}\right) \leq \int_{\mathcal{R}_4} dF\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right) + \int_{\mathcal{R}_5} dF\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right). \tag{4.1.4}$$

Similarly, $\Pr\left(\{E_1,O_2\}\,|\mathbf{x}_{1,j},\mathbf{x}_{2,j}\right)$ is bounded as

$$\int_{\mathcal{R}_2} dF\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right) \leq \Pr\left(\{E_1,O_2\}\,|\mathbf{x}_{1,j},\mathbf{x}_{2,j}\right) \leq \int_{\mathcal{R}_2} dF\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right) + \int_{\mathcal{R}_5} dF\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right). \tag{4.1.5}$$

We have two trivial bounds for $\Pr\left(\{O_1,O_2\}\,|\mathbf{x}_{1,j},\mathbf{x}_{2,j}\right)$ and $\Pr\left(\{E_1,E_2\}\,|\mathbf{x}_{1,j},\mathbf{x}_{2,j}\right)$ as

$$\int_{\mathcal{R}_3} dF\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right) \leq \Pr\left(\{O_1,O_2\}\,|\mathbf{x}_{1,j},\mathbf{x}_{2,j}\right) \leq \int_{\mathcal{R}_3} dF\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right) \tag{4.1.6}$$

$$\int_{\mathcal{R}_1} dF\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right) \leq \Pr\left(\{E_1,E_2\}\,|\mathbf{x}_{1,j},\mathbf{x}_{2,j}\right) \leq \int_{\mathcal{R}_1} dF\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right). \tag{4.1.7}$$

For simplicity of notation, let $\mathbf{y}$ denote the possible outcomes. Stack (4.1.4)-(4.1.7), and then denote the left-hand and the right-hand side of the inequalities by $\mathbf{h}^l\left(\mathbf{x}_{1,j},\mathbf{x}_{2,j};\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\delta_1,\delta_2,\boldsymbol{\theta}\right)$ and $\mathbf{h}^u\left(\mathbf{x}_{1,j},\mathbf{x}_{2,j};\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\delta_1,\delta_2,\boldsymbol{\theta}\right)$, respectively. Collecting (4.1.4)-(4.1.7) yields 2^2 inequalities:

$$\mathbf{h}^l\left(\mathbf{x}_{1,j},\mathbf{x}_{2,j};\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\delta_1,\delta_2,\boldsymbol{\theta}\right) \leq \mathbf{Pr}\left(\mathbf{y}|\mathbf{x}_{1,j},\mathbf{x}_{2,j}\right) \leq \mathbf{h}^u\left(\mathbf{x}_{1,j},\mathbf{x}_{2,j};\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\delta_1,\delta_2,\boldsymbol{\theta}\right). \tag{4.1.8}$$

[7] The parameter vector $(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\delta_1,\delta_2,\boldsymbol{\theta})$ satisfying (4.1.8) is not a singleton in general even if $F\left(\epsilon_{1,j},\epsilon_{2,j}|\boldsymbol{\theta}\right)$ has a well-behaved density. The parameter vector is only partially identified or set identified, as opposed to being point identified. Furthermore, as a result of being agnostic about the equilibrium selection, this approach to estimate only the bounds of the parameters may not be very informative in the model predictions and counterfactuals.

4.1.1.3 Estimation of the System Specified by Moment Inequalities

For the sake of notational simplicity, denote $\boldsymbol{\omega} := (\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\delta_1,\delta_2,\boldsymbol{\theta})$. (4.1.8) as the population moment inequalities. For estimation, we replace each term in the inequalities with its sample analogue. To that end, the term $\mathbf{Pr}\left(\mathbf{y}|\mathbf{x}_{1,j},\mathbf{x}_{2,j}\right)$ should be estimated first. Various nonparametric conditional cumulative distribution function estimators are available. Ciliberto and Tamer (2009) use the simple frequency estimator. More options can be found in, for example, Li and Racine (2007).

Let $\boldsymbol{\Omega}_0$ be the set such that for any $\boldsymbol{\omega} \in \boldsymbol{\Omega}_0$, (4.1.8) is satisfied for any $\left(\mathbf{x}_{1,j},\mathbf{x}_{2,j}\right)$ with probability 1. Such a set $\boldsymbol{\Omega}_0$ is called the identified set. The focus of the paper is to find a consistent set estimator $\hat{\boldsymbol{\Omega}}$ in the sense that $\hat{\boldsymbol{\Omega}} \subset \boldsymbol{\Omega}_0$ and $\boldsymbol{\Omega}_0 \subset \hat{\boldsymbol{\Omega}}$ with probability 1 as the sample size n grows to infinity. To that end,

parameters as well, whereas the conditional probabilities here, which are taken to be directly observable from data, are not a function of parameters.

[7] Although we focus on a two-player entry game, extension to a k-player entry game is straightforward. In general, we have 2^k inequalities.

they consider the following objective function:

$$Q_n(\boldsymbol{\omega}) := \frac{1}{n}\sum_{j=1}^{n}\left\{\left\|\min\left\{\mathbf{0}, \mathbf{Pr}\left(\mathbf{y}|\mathbf{x}_{1,j}, \mathbf{x}_{2,j}\right) - \mathbf{h}^{l}\left(\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\omega}\right)\right\}\right\| + \left\|\max\left\{\mathbf{0}, \mathbf{Pr}\left(\mathbf{y}|\mathbf{x}_{1,j}, \mathbf{x}_{2,j}\right) - \mathbf{h}^{u}\left(\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\omega}\right)\right\}\right\|\right\}, \tag{4.1.9}$$

where $\|\cdot\|$ is the Euclidian norm, and min and max operators inside $\|\cdot\|$ are taken elementwise. Notice each part of (4.1.9) penalizes the violation of the inequalities on only one side, and the objective function has some flat region where the inequalities in (4.1.8) are strict. By construction, it is clear that $Q(\boldsymbol{\omega}) \geq 0$ for any $\boldsymbol{\omega}$, and $Q(\boldsymbol{\omega}) = 0$ for $\boldsymbol{\omega} \in \boldsymbol{\Omega}_0$. Now, define the set estimator as

$$\hat{\boldsymbol{\Omega}} := \{\boldsymbol{\omega} \in \boldsymbol{\Omega} : nQ_n(\boldsymbol{\omega}) \leq \nu_n\},$$

where $\nu_n \to \infty$ and $\nu_n/n \to 0$. Chernozhukov, Hong, and Tamer (2007) show consistency and asymptotic properties of the set estimator $\hat{\boldsymbol{\Omega}}$ defined above.

The remaining part is an algorithm to construct $\mathbf{h}^{l}\left(\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\omega}\right)$ and $\mathbf{h}^{u}\left(\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\omega}\right)$. They suggest using simulation, which is similar to Berry (1992), which we study below. The key step is as follows: (i) Draw R numbers of $\left(\epsilon_{1,j}, \epsilon_{2,j}\right)$ for each j; (ii) compute the profits of each potential entrant; and (iii) when the equilibrium is unique for a simulation draw, the corresponding element of both $\mathbf{h}^{l}\left(\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\omega}\right)$ and $\mathbf{h}^{u}\left(\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\omega}\right)$ is set at 1. When the equilibrium is not unique, add 1 only on $\mathbf{h}^{u}\left(\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\omega}\right)$. For example, the upper bound estimate on the outcome probability $\Pr\left(\{E_1, E_1\}|\mathbf{x}_{1,j}, \mathbf{x}_{2,j}\right)$ is

$$\hat{h}^{u,(4)}\left(\mathbf{x}_{1,j}, \mathbf{x}_{2,j}; \boldsymbol{\omega}\right) = \frac{1}{R}\sum_{r=1}^{R}\mathbf{1}\left(\left\{\mathbf{x}'_{1,j}\boldsymbol{\alpha}_1 - \delta_1 + \epsilon^{r}_{1,j} > 0 \text{ and } \mathbf{x}'_{2,j}\boldsymbol{\alpha}_2 - \delta_2 + \epsilon^{r}_{2,j} > 0\right\}\right).$$

The simulation procedure is computationally very intensive, but it is parallelizable. Another computational challenge is that we cannot use the derivative-based optimizers due to the flat region of the objective function. Instead, the estimation algorithm should be essentially a grid search. Curse of dimensionality becomes a serious problem in a grid search.

4.1.2 Sequential Entry with More than Two Potential Entrants

The two-player discrete-game models studied in the previous subsection have a straightforward extension to multiple players. When the number of potential entrants is larger than two, the incompleteness problem becomes much more complicated. Some classical works, which we review in this subsection, rule out this potential multiple-equilibria problem by assuming sequential entry. We review Bresnahan and Reiss (1991a,b) and Berry (1992) here.

The basic idea of this equilibrium refinement is described in Bresnahan and Reiss (1991a). In a two-player entry game, when the entry is assumed to be sequential, the region $\mathcal{R}_5$ in Figure 4.1.1 only allows for the first mover's entry in the pure-strategy subgame perfect Nash equilibrium. The idea can readily be generalized to a game with more than two potential entrants.

Bresnahan and Reiss (1991b) assume the nonnegative-profit condition with sequential entry and homogeneous firms. In market j, n_j number of entrants are observed when $\pi_j(n_j) \geq 0$ and $\pi_j(n_j+1) < 0$ hold. The reduced-form profit function is

$$\begin{aligned}\pi_j(n_j) &= \left(\mathbf{q}_j'\boldsymbol{\lambda}\right)\left(\alpha_1 + \mathbf{x}_j'\boldsymbol{\beta} - \sum_{k=2}^{n_j}\alpha_k\right) - \left(\gamma_1 + \mathbf{w}_j'\delta + \sum_{k=2}^{n_j}\gamma_k\right) + \epsilon_j \\ &=: \bar{\pi}_j(n_j) + \epsilon_j,\end{aligned}$$

where $\mathbf{q}_j$ is demographics of the region in which firm j is located, $\mathbf{x}_j$ is the variable cost shifter, and $\mathbf{w}_j$ is the fixed cost shifter. The assumption on the error term is $\epsilon_j \sim i.i.d.\ \mathcal{N}(0,1)$. The model parameters to be estimated are $(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta})$. The i.i.d. normal assumption on the error term allows ordered Probit estimation. To that end, consider the likelihood of observing n_j firms:

$$\begin{aligned}\Pr\left(n_j \text{ firms observed}\right) &= \Pr\left(\pi_j(n_j) \geq 0, \pi_j(n_j+1) < 0\right) \\ &= \Pr\left(\bar{\pi}_j(n_j) + \epsilon_j \geq 0, \bar{\pi}_j(n_j+1) + \epsilon_j < 0\right) \\ &= \Pr\left(-\bar{\pi}_j(n_j) \leq \epsilon_j < -\bar{\pi}_j(n_j+1)\right) \\ &= \Phi\left(-\bar{\pi}_j(n_j+1)\right) - \Phi\left(-\bar{\pi}_j(n_j)\right).\end{aligned}$$

Maximum-likelihood estimation will yield the consistent, asymptotically normal estimators as usual.

Berry (1992) allows for firm-level heterogeneity in the fixed cost of the potential entrants. The model is again a complete-information oligopolistic entry game to a market. Only the pure-strategy equilibrium is considered. To that end, define the reduced-form profit function of a potential entrant i in market j as

$$\pi_{i,j}(n_j) = \mathbf{x}_j'\boldsymbol{\beta} - \delta\ln(n_j) + \mathbf{z}_{i,j}'\boldsymbol{\alpha} + \rho\epsilon_j + \sqrt{1-\rho^2}\epsilon_{i,j}, \tag{4.1.10}$$

where n_j is the equilibrium number of firms in market j, $\mathbf{x}_j$ is the market level variable profit shifter, $\mathbf{z}_{i,j}$ is the firm level fixed cost shifter. Market-specific heterogeneity is reflected into the term $\mathbf{z}_{i,j}'\boldsymbol{\alpha} + \sqrt{1-\rho^2}\epsilon_{i,j}$. Firm i enters market j if $\pi_{i,j}(n_j) > 0$.[8] It is assumed that ϵ_j's and $\epsilon_{i,j}$'s both follow i.i.d. standard normal distribution. Thus, within a market, the joint distribution of the composite error term $\mathbf{u}_j := \rho\epsilon_j + \sqrt{1-\rho^2}\boldsymbol{\epsilon}_j$ is

$$\begin{pmatrix} u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{N,j}\end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \vdots & \rho \\ \vdots & \cdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1\end{pmatrix}\right).$$

Furthermore, if we assume the firms enter in the order of profitability and operate with nonnegative profits, we can show that the equilibrium number of firms in a market is determined uniquely.

[8]Because Berry (1992) study the domestic airline market, a market is naturally defined by the city pair. Then, the potential entrant is defined by any airline that is operating in either of the endpoints. Identifying who is a potential entrant and who is not is generally a difficult problem.

The goal is to estimate the model parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \delta, \rho)$ in (4.1.10) consistently. However, even though we specify the model up to the distribution of error term, MLE is still practically infeasible because the likelihood of observing $1, 2, ..., N$ entrants should be computed for each datapoint. For each $\Pr\left(n_j^* = n | \mathbf{x}_j, \mathbf{z}_{i,j}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \delta, \rho\right)$, evaluation of approximately N times of integral is necessary, which is computationally expensive. Another difficult problem arises because n_j is endogenous, which is determined by the players' decision. Berry (1992) suggests GMM with simulated moment conditions. To implement GMM estimation, we must first specify the moment conditions. First, fix j and consider the equilibrium number of market entrants:

$$n_j^* = \sum_i \mathbf{1}\left(\pi_{i,j}\left(n_j^*\right) > 0 | \mathbf{x}_j, \mathbf{z}_{i,j}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \delta, \rho\right).$$

Taking expectation yields

$$n_j^* = \sum_i E_{u_{i,j}}\left[\mathbf{1}\left(\pi_{i,j}\left(n_j\right) > 0 | \mathbf{x}_j, \mathbf{z}_{i,j}; \boldsymbol{\beta}, \boldsymbol{\alpha}, \delta, \rho\right)\right]. \tag{4.1.11}$$

The corresponding moment condition is

$$E\left[n_j^* - \sum_i \Pr\left(\pi_{i,j}\left(n_j\right) > 0 | \mathbf{x}_j, \mathbf{z}_{i,j}\right) | \boldsymbol{\beta}, \boldsymbol{\alpha}, \delta, \rho\right] = 0. \tag{4.1.12}$$

The right-hand-side object of (4.1.11) is difficult to evaluate exactly. We employ a simulation approach, which is similar to BLP studied in Chapter 3. First, simulate many enough $\boldsymbol{\epsilon}_j$'s and store the simulated $\boldsymbol{\epsilon}_j$'s in the memory. Then, for each candidate $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \delta, \rho)$, calculate the simulation analogue of $\sum_i \Pr\left(\pi_{i,j}\left(n_j\right) > 0 | \mathbf{x}_j, \mathbf{z}_{i,j}\right)$. Iterate over $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \delta, \rho)$ to find the parameter values that minimize the GMM objective function corresponding to the moment condition (4.1.12).

The uniqueness of the equilibrium relies on the assumption that firms enter sequentially and the information is perfect.[9] We assume the firms enter in the order of profitability in Berry (1992), in which the actual sequence of entry is not observed. Note that the estimated model parameter may vary substantially if we assume a different order of entry, especially when the potential entrants are similar in terms of profitability. For further discussions on the assumed order of entry and robustness of the estimators, see Ellickson and Misra (2011).

4.1.3 Simultaneous Entry with Multiple Markets

Jia (2008) allows for the multiple location entry decisions by two firms. Consider two retail discount chains, Walmart and Kmart. Both chains decide whether to enter many markets simultaneously to maximize the chainwide profit. The profit functions of each chain store $i \in \{K, W\}$ at market j are given

[9] It can be shown that in an extensive-form game with perfect information and a finite number of nodes, a unique sequential equilibrium exists. See Theorem 4.7 of Myerson (1991).

by

$$\pi_{i,j}\left(\mathbf{d}_i, d_{-i,j}; \mathbf{x}_j, \mathbf{l}_j, \epsilon_j, \epsilon_{i,j}\right) = d_{i,j}\left(\mathbf{x}_j'\boldsymbol{\beta}_i + \delta_{i,-i}d_{-i,j} + \delta_{i,i}\sum_{k\neq j}\frac{d_{i,k}}{l_{j,k}} + \rho\epsilon_j + \sqrt{1-\rho^2}\epsilon_{i,j}\right),$$

where $d_{i,j} \in \{0,1\}$ is the endogenous (at equilibrium) entry-decision variable and $l_{j,k}$ is the physical distance between market j and k. The problem becomes complicated because each chain maximizes the profit over the sum over all the markets. To that end, the chainwide profit-maximization problem of chain i is

$$\max_{\left\{d_{i,1},d_{i,2},\ldots,d_{i,J}\right\}\in\{0,1\}^J}\sum_{j=1}^{J} d_{i,j}\left(\mathbf{x}_j'\boldsymbol{\beta}_i + \delta_{i,-i}d_{-i,j} + \delta_{i,i}\sum_{k\neq j}\frac{d_{i,k}}{l_{j,k}} + \rho\epsilon_j + \sqrt{1-\rho^2}\epsilon_{i,j}\right). \tag{4.1.13}$$

Solving this problem is complicated because (i) the number of decision variables 2^J increases exponentially over the number of markets, and (ii) the decision variable is discrete.

She simplifies the solution of the problem substantially, using the lattice structure of the choice variables. She uses the lattice fixed point theorem, which states that under certain conditions[10] the optimal decision $\left\{d_{i,1}^*, d_{i,2}^*, \ldots, d_{i,J}^*\right\}$ can be bounded elementwise. As an illustrative example, suppose $J = 3$ and the equilibrium entry decision is $\left\{d_{i,1}^*, d_{i,2}^*, d_{i,3}^*\right\} = \{0,1,0\}$. The lattice fixed-point theorem provides with an algorithm that gives the upper and lower bound for the equilibrium entry decision as, for example, $\{0,0,0\}$ and $\{0,1,1\}$. Then, the degree of freedom for numerical search being left is reduced by half.

The remaining estimation step is done by the method of simulated moments as in Berry (1992). For each candidate parameter, find the simulated profits as described above, and iterate over to minimize the objective function. The equilibrium-selection criteria are not the focus of this paper, and they are in a sense arbitrary. Jia chooses the equilibrium that best predicts the observed pattern of the chain store locations, and conducts some sensitivity analysis over different equilibrium-selection rules.

4.2 Static Discrete Games with Incomplete Information

4.2.1 Private Information and Bayesian Nash Equilibrium

The baseline discrete-game model pioneered by Bresnahan and Reiss (1991a) raises the fundamental problem of multiple equilibria. As we discussed above, the literature has developed four different solutions: (i) Aggregate the possible outcomes to the form that makes the model coherent; (ii) specify an equilibrium-selection rule as a nuisance parameter and estimate it from the data; (iii) abandon the point identification and exploit the moment inequalities, which essentially leads only to the set identification; and (iv) refine the equilibrium concepts.

Refining the Nash equilibria by introducing private information and employing the relevant equilibrium concept falls into (iv). The Bayesian Nash equilibrium, which requires that the belief probabilities

[10] The key condition is that $\left(\mathbf{x}_j'\boldsymbol{\beta}_i + \delta_{i,-i}d_{-i,j} + \delta_{i,i}\sum_{k\neq j}\frac{d_{i,k}}{l_{j,k}} + \rho\epsilon_j + \sqrt{1-\rho^2}\epsilon_{i,j}\right)$ is increasing in each $d_{i,k}$, that is, $\delta_{i,i} \geq 0$.

of the players line up with the strategies of each corresponding player, partly addresses the multiple-equilibria problem by refining the Nash equilibria. The Bayesian Nash equilibrium refinement, however, does not guarantee the uniqueness of the equilibrium. Although the degree of multiplicity shrinks substantially, the equilibrium-selection problem persists. In this section, we focus on the estimation of discrete games where the cross-sectional data are assumed to be generated from pure-strategy Bayesian Nash equilibria. Then, we discuss the recent developments on allowing more flexibility in the information structures.

4.2.1.1 Baseline Two-Player Simultaneous Entry Model with Additive i.i.d. Logit Errors

Consider the entry game studied in section 4.1.1. We keep restricting our attention to the two-player entry game for simplicity. Let $y_{i,j}$ be an indicator such that $y_{i,j} = 1$ if firm $i \in \{1,2\}$ entered market j, and $y_{i,j} = 0$ otherwise. When the realization of $(\epsilon_{1,j}, \epsilon_{2,j})$ was known to both firms, the following equations characterized the Nash equilibrium of this static entry game:

$$\begin{aligned} y_{1,j} &= \mathbf{1}\left(\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 - \delta_1 y_{2,j} + \epsilon_{1,j} \geq 0\right) \\ y_{2,j} &= \mathbf{1}\left(\mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 - \delta_2 y_{1,j} + \epsilon_{2,j} \geq 0\right). \end{aligned}$$

The key idea of Seim (2006) is to assume the realization of $\epsilon_{i,j}$ is the private information, and the distribution of $\epsilon_{i,j}$ is the common knowledge. Players know only the realization of own $\epsilon_{i,j}$'s. For tractability, we assume $\epsilon_{i,j}$ follows i.i.d. logistic distribution.[11]

The Bayesian Nash equilibrium is defined as the strategy profiles and belief systems of each player such that given (i) each player's posterior belief on other players' type[12] and (ii) other players' strategy, each player's strategy maximizes the expected payoff. Because we normalized the expected payoff of the option "exit" as zero, each player's strategy boils down to

$$\begin{aligned} y_{i,j} &= \mathbf{1}\left(\mathbf{x}_{i,j}'\boldsymbol{\alpha}_i - \delta_i E\left[y_{-i,j}|\mathbf{x}_{-i,j}\right] + \epsilon_{i,j} \geq 0\right) \\ &= \mathbf{1}\left(\mathbf{x}_{i,j}'\boldsymbol{\alpha}_i - \delta_i p_{-i,j} + \epsilon_{i,j} \geq 0\right), \end{aligned} \tag{4.2.1}$$

where $p_{-i,j} := E\left[y_{-i,j}|\mathbf{x}_{-i,j}\right] = \Pr\left(y_{-i,j} = 1|\mathbf{x}_{-i,j}\right)$. Because $\epsilon_{i,j}$'s are i.i.d., stacking (4.2.1) for $i = 1,2$ and taking expectations yields

$$p_{i,j} = 1 - F_\epsilon\left(-\mathbf{x}_{i,j}'\boldsymbol{\alpha}_i + \delta_i p_{-i,j}\right). \tag{4.2.2}$$

[11] Alternatively, we can assume

$$\begin{aligned} y_{1,j} &= \mathbf{1}\left(\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 - \delta_1 y_{2,j} + \epsilon_{1,j} \geq \epsilon_{1,j}'\right) \\ y_{2,j} &= \mathbf{1}\left(\mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 - \delta_2 y_{1,j} + \epsilon_{2,j} \geq \epsilon_{2,j}'\right), \end{aligned}$$

with $\epsilon_{i,j}$ and $\epsilon_{i,j}'$ both follow i.i.d. Type I extreme value distribution.

[12] In this context, a player's type is synonymous to the player's private information.

Invoking the logit error assumption, (4.2.2) becomes the set of equilibrium fixed-point equations as

$$p_{1,j} = \frac{\exp\left(\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 - \delta_1 p_{2,j}\right)}{1+\exp\left(\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 - \delta_1 p_{2,j}\right)} \tag{4.2.3}$$

$$p_{2,j} = \frac{\exp\left(\mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 - \delta_2 p_{1,j}\right)}{1+\exp\left(\mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 - \delta_2 p_{1,j}\right)}. \tag{4.2.4}$$

That is, the Bayesian Nash equilibrium data-generating assumption implies the probability of entry must satisfy the above fixed-point equations, that is, $p_{i,j} = \Pr\left(y_{i,j} = 1 | \mathbf{x}_{i,j}, \mathbf{x}_{-i,j}, p_{-i,j}; \boldsymbol{\omega}\right)$, where $\boldsymbol{\omega} := (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \delta_1, \delta_2)$.

4.2.1.2 Maximum-Likelihood Estimation

Again, the likelihood is

$$\begin{aligned}
&\mathcal{L}\left(\left\{y_{i,j}, \mathbf{x}_{i,j}\right\}_{i\in\{1,2\}, j\in\mathcal{J}}; \boldsymbol{\omega}\right) \\
= &\prod_{j\in\mathcal{J}}\prod_{i=1}^{2} \Pr\left(y_{i,j} = 1 | \mathbf{x}_{i,j}, \mathbf{x}_{-i,j}, p_{-i,j}; \boldsymbol{\omega}\right)^{y_{i,j}} \left\{1 - \Pr\left(y_{i,j} = 1 | \mathbf{x}_{i,j}, \mathbf{x}_{-i,j}, p_{-i,j}; \boldsymbol{\omega}\right)\right\}^{1-y_{i,j}}.
\end{aligned}$$

The problem in implementation is that the solution of (4.2.3)-(4.2.4) might not be unique. The number of different solutions is at most finitely many. That is, the model parameters may not be point identified, but the cardinality of the identified set is finite, as opposed to a positive Lebesgue measure set in $\mathbb{R}^2$. This reduction can be regarded as one of the results of the equilibrium refinement by imposing the specific information structure.

A more serious problem in the implementation of MLE is that the number of solutions depends on the candidate parameter values $\boldsymbol{\omega}$. The likelihood as a function of $\boldsymbol{\omega}$ may have kinks, which prevents one from using the derivative-based optimization routine. Vitorino (2012); Su (2014) converts the optimization problem to an equality-constrained one (MPEC), and addresses the multiplicity problem by trying many starting values. To this end, the problem can be restated as

$$\begin{aligned}
&\max_{\left\{p_{i,j}\right\}_{i\in\{1,2\}, j\in\mathcal{J}}, \boldsymbol{\omega}} \prod_{j\in\mathcal{J}}\prod_{i=1}^{2} p_{i,j}^{y_{i,j}} \left\{1 - p_{i,j}\right\}^{1-y_{i,j}} \\
&s.t. \quad p_{i,j} = \Pr\left(y_{i,j} = 1 | \mathbf{x}_{i,j}, \mathbf{x}_{-i,j}, p_{-i,j}; \boldsymbol{\omega}\right) \qquad i = 1,2 \quad j \in \mathcal{J}.
\end{aligned}$$

This constrained maximization problem is feasible because (4.2.3)-(4.2.4) are smooth in $\left(\boldsymbol{\omega}, p_{1,j}, p_{2,j}\right)$, and the Jacobian is of full rank. Su (2014) argues that when the Bayesian Nash equilibrium is not unique, the one that yields the highest likelihood value should be the data-generating equilibrium. This global maximum of the likelihood serves as an equilibrium-selection criterion.

4.2.1.3 Two-Stage GMM Estimation

Pesendorfer and Schmidt-Dengler (2008) suggest a two-stage GMM method, which is computationally lighter than the maximum-likelihood estimation because it does not need to exactly satisfy the fixed-point equation in the second-stage parameter estimation. Consider again (4.2.3)-(4.2.4). The key idea is to replace $\left(p_{1,j}, p_{2,j}\right)$ with its consistent estimators $\left(\hat{p}_{1,j}, \hat{p}_{2,j}\right)$, and then form the moment conditions

$$
\begin{aligned}
E\left[\hat{p}_{1,j} - \frac{\exp\left(\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 - \delta_1\hat{p}_{2,j}\right)}{1+\exp\left(\mathbf{x}_{1,j}'\boldsymbol{\alpha}_1 - \delta_1\hat{p}_{2,j}\right)}\right] &= 0 \\
E\left[\hat{p}_{2,j} - \frac{\exp\left(\mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 - \delta_2\hat{p}_{1,j}\right)}{1+\exp\left(\mathbf{x}_{2,j}'\boldsymbol{\alpha}_2 - \delta_2\hat{p}_{1,j}\right)}\right] &= 0
\end{aligned}
$$

for $j \in \mathcal{J}$. The second-stage estimation can be done relatively easily using the nonlinear GMM. Assuming the players' strategies corresponds to a single equilibrium, $\left(\hat{p}_{1,j}, \hat{p}_{2,j}\right)$ will converge to the true probabilities $\left(p_{1,j}, p_{2,j}\right)$ as the sample size grows to infinity.

The first-stage consistent estimator $\left(\hat{p}_{1,j}, \hat{p}_{2,j}\right)$ can be a simple frequency estimator, a Nadaraya-Watson-type smoothing estimator, or a partial maximum-likelihood estimator. Note the error of the first-stage estimates propagates to the second-stage parameter estimates, which means the asymptotic variances of the second-stage parameter estimates have to be corrected for this additional uncertainty.[13] Su (2014) reports that the finite-sample properties of the two-stage estimators can be worse than the full maximum-likelihood estimator.

4.2.2 Flexible Information Structures

The set of possible equilibria shrinks more as the assumption imposed on the information structures becomes stronger. What we have studied thus far is the two extremes; that is, the realization of $\boldsymbol{\epsilon}_{-i,j}$ is either fully known to player i or completely unknown to i. Furthermore, we assume $\epsilon_{i,j}$ is i.i.d. over i and j. Recent advances in the literature relax these extreme assumptions on the information structures.

Grieco (2014) introduces both private shocks and public shocks on the player's payoff functions. Grieco's analysis is based on the pure-strategy Bayesian Nash equilibrium, whereas the model parameters might be only partially identified due to the existence of public shocks in the payoff functions. Wan and Xu (2014); Liu, Vuong, and Xu (Forthcoming) study the estimation of discrete games with incomplete information when players' private information is possibly correlated, assuming the data are generated by a single Bayesian Nash equilibrium.

Aradillas-Lopez (2010) investigate semiparametric identification and estimation when the distribution of $\epsilon_{i,j}$'s are unknown to the researcher, assuming the data are generated from correlated equilibria developed by Aumann (1987). Magnolfi and Roncoroni (2017) employ the Bayes correlated equilibrium concept by Bergemann and Morris (2013, 2016), which is an extension of Aumann (1987) to games with

[13] A full treatment on this subject when the first-stage estimator is $\sqrt{n}$-consistent can be found in Newey and McFadden (1994).

incomplete information. In Magnolfi and Roncoroni's setup, player i receives a signal about the realization of $\epsilon_{-i,j}$, which can be either informative or uninformative. If the signal is not informative at all, the Bayes correlated equilibrium boils down to the Bayesian Nash equilibrium, and if it is fully informative, it boils down to the complete-information pure-strategy Nash equilibrium. They characterize the sharp identified set, and in their application, the set is small enough to be informative. Because the sharp identified set is much larger than what is characterized by the pure-strategy Nash equilibria, whether the informativeness can be generalized to other contexts remains an open question.

4.3 Discussion

We studied the estimation of model parameters of a discrete-game model with cross-sectional data, by assuming the observed data come from one or more of the Nash equilibria or their refinement. The fundamental issue in this literature is the incompleteness of the baseline economic model that yields the multiplicity of equilibria. If an econometrician wants to maintain the assumption that the data are generated by a single equilibrium, some equilibrium-selection criteria must exist. We reviewed that the equilibrium-selection criteria can be imposed in two ways: equilibrium refinement or the data-driven equilibrium-selection rule. One can choose to drop some information from the data, which was the original approach of Bresnahan and Reiss (1990, 1991a). The last possible choice is partial identification. This sophisticated method, however, often lacks the prediction, and thus, it may not be very useful to conduct policy counterfactuals.

We overviewed the estimation of complete-information-game models and incomplete-information-game models thus far. One can immediately question which is a more adequate concept. One possible argument is that complete-information-game models are more adequate for the markets that are under a long-run equilibrium.

Further discussions on this subject can be found in Berry and Tamer (2006); Berry and Reiss (2007). Berry and Tamer (2006) study the identification conditions for various oligopoly entry models under the presence of multiple equilibria, focusing more on mixed-strategy equilibria. Surveys on the estimation of discrete-game models with cross-sectional data can be found in Ellickson and Misra (2011); de Paula (2013).

References

ARADILLAS-LOPEZ, A. (2010): "Semiparametric estimation of a simultaneous game with incomplete information," *Journal of Econometrics*, 157, 409–431.

AUMANN, R. J. (1987): "Correlated equilibrium as an expression of Bayesian rationality," *Econometrica*, 55, 1–18.

BAJARI, P., H. HONG, AND S. P. RYAN (2010): "Identification and estimation of a discrete game of complete information," *Econometrica*, 78, 1529–1568.

BERGEMANN, D. AND S. MORRIS (2013): "Robust predictions in games with incomplete information," *Econometrica*, 81, 1251–1308.

——— (2016): "Bayes correlated equilibrium and the comparison of information structures in games," *Theoretical Economics*, 11, 487–522.

BERRY, S. (1992): "Estimation of a model of entry in the airline industry," *Econometrica*, 60, 889–917.

BERRY, S. AND P. REISS (2007): *Empirical models of entry and market structure*, Elsevier, vol. 3 of *Handbook of Industrial Organization*, chap. 29, 1845–1886.

BERRY, S. AND E. TAMER (2006): *Identification in models of oligopoly entry*, Cambridge: Cambridge University Press, vol. 2 of *Advances in Economics and Econometrics, Ninth World Congress of the Econometric Society*, 46–85.

BRESNAHAN, T. F. AND P. C. REISS (1990): "Entry in monopoly markets," *The Review of Economic Studies*, 57, 531–553.

——— (1991a): "Empirical models of discrete games," *Journal of Econometrics*, 48, 57–81.

——— (1991b): "Entry and competition in concentrated markets," *The Journal of Political Economy*, 99, 977–1009.

CHERNOZHUKOV, V., H. HONG, AND E. TAMER (2007): "Estimation and confidence regions for parameter sets in econometric models," *Econometrica*, 75, 1243–1284.

CILIBERTO, F. AND E. TAMER (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77, 1791–1828.

DE PAULA, A. (2013): "Econometric analysis of games with multiple equilibria," *Annual Review of Economics*, 5, 107–131.

ELLICKSON, P. B. AND S. MISRA (2011): "Estimating discrete games," *Marketing Science*, 30, 997–1010.

EPSTEIN, L. G., H. KAIDO, AND K. SEO (2015): "Robust confidence regions for incomplete models," *Working Paper*.

GALICHON, A. AND M. HENRY (2011): "Set identification in models with multiple equilibria," *Review of Economic Studies*, 78, 1264–1298.

GRIECO, P. L. E. (2014): "Discrete games with flexible information structures: an application to local grocery markets," *RAND Journal of Economics*, 45, 303–340.

JIA, P. (2008): "What happens when Wal-mart comes to town: An empirical analysis of the discount retailing industry," *Econometrica*, 76, 1263–1316.

KOOREMAN, P. (1994): "Estimation of econometric models of some discrete games," *Journal of Applied Econometrics*, 9, 255–268.

LI, Q. AND J. S. RACINE (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

LIU, N., Q. VUONG, AND H. XU (Forthcoming): "Rationalization and identification of binary games with correlated types," *Journal of Econometrics*.

MAGNOLFI, L. AND C. RONCORONI (2017): "Estimation of discrete games with weak assumptions on information," *Working Paper*.

MANZEL, K. (Forthcoming): "Inference for games with many players," *Review of Economic Studies*.

MAZZEO, M. J. (2002): "Product choice and oligopoly market structure," *RAND Journal of Economics*, 33, 221–242.

MYERSON, R. B. (1991): *Game Theory: Analysis of Conflict*, Harvard University Press.

NEWEY, W. K. AND D. MCFADDEN (1994): *Large sample estimation and hypothesis testing*, Amsterdam: Elsevier, vol. 4 of *Handbook of Econometrics*, chap. 36, 2111–2245.

PESENDORFER, M. AND P. SCHMIDT-DENGLER (2008): "Asymptotic least squares estimators for dynamic games," *Review of Economic Studies*, 75, 901–928.

SEIM, K. (2006): "An empirical model of firm entry with endogenous product-type choices," *RAND Journal of Economics*, 37, 619–640.

SU, C.-L. (2014): "Estimating discrete-choice games of incomplete information: Simple static examples," *Quantitative Marketing and Economics*, 12, 167–207.

SWEETING, A. (2009): "The strategic timing incentives of commercial radio stations: An empirical analysis using multiple equilibria," *RAND Journal of Economics*, 40, 710–742.

TAMER, E. (2003): "Incomplete simultaneous discrete response model with multiple equilibria," *Review of Economic Studies*, 70, 147–165.

VITORINO, M. A. (2012): "Empirical entry games with complementarities: An application to the shopping center entry," *Journal of Marketing Research*, 49, 175–191.

WAN, Y. AND H. XU (2014): "Semiparametric identification of binary decision games of incomplete information with correlated private signals," *Journal of Econometrics*, 182, 235–246.

Chapter 5

Estimation of Dynamic Discrete-Game Models

In this chapter, we study the estimation of dynamic game models. We first briefly discuss the general framework of Markov-perfect industry dynamics developed by Ericson and Pakes (1995), and move on to the methods to estimate the model parameters of a dynamic game under Markov state-transition probabilities and Markov decision rules. Estimation frameworks of the dynamic games under Markov assumptions are pioneered by Aguirregabiria and Mira (2007); Bajari, Benkard, and Levin (2007); Pakes, Ostrovsky, and Berry (2007); Pesendorfer and Schmidt-Dengler (2008). We go over their key ideas in this chapter.

The dynamic-game model-estimation frameworks presented in this chapter are extensions of single-agent dynamics we studied in section 2.3. In a sense, single-agent dynamics can be regarded as a game against nature. In all of the dynamic games we study in this section, each player has private information. Therefore, the games could be regarded as an extension of the static games with incomplete information that we studied in section 4.2.

5.1 Industry Dynamics in an Oligopolistic Market and the Markov Perfect Equilibrium

Ericson and Pakes (1995) develop a theoretical framework of industry dynamics, which is general enough to encompass most of the empirical frameworks used in the literature thus far. We review only the key features of the model here, leaving out all the technical details. We assume all the regularity conditions required for the Markov perfect Nash equilibria hold.

5.1.1 Setup

Consider the ex-ante homogeneous firms. Firms only differ by productivity level ω ($\in \Omega \subseteq \mathbb{Z}$), where $\mathbb{Z}$ denotes the set of integers. ω represents the producer-specific heterogeneity. ω^0, the state at the entry, is drawn from a set of states Ω^e ($\subseteq \Omega$). For the entrants, only the distribution $\Pr\left(\omega^0\right)$ over Ω^e is common prior knowledge. Potential entrants must pay the sunk cost of entry x^e in order to draw the initial level of productivity ω^0.

Industry structure is summarized by a measure $\mathbf{s}$, which gives a count of the number of firms at each ω. $\mathbf{s}$ can be regarded as the summary measure of the competitiveness in the industry. Note that firms' prospects are completely summarized by the $(\omega, \mathbf{s})$ pair. Each period, firms make investment decision x with the knowledge of $(\omega, \mathbf{s})$ and state-space evolution probability.

Firms have common salvage value ϕ, which can be interpreted as the opportunity cost of being in the industry. In every period, a stochastic negative shift occurs in all incumbent firms' ω. One explanation could be the progress outside the industry. The negative shifts are large enough that all firms will exit eventually absent any investment.

Let $R(\omega, \mathbf{s}; x) = A(\omega, \mathbf{s}) - c(\omega)x$ be the per-period net profit function. $A(\omega, \mathbf{s})$ is the per-period gross profit, x is the period investment, and $c(\omega)$ is the unit cost of investment that increases by ω. The state-transition rule is $\omega' = \omega + \tau + \eta$. τ, the stochastic outcome of investment decision, is $\tau \sim \pi(\omega, x)$. η is the stochastic negative shift of ω. $x = 0$ implies $\tau = 0$. Notice the future states are stochastically affected only by the current investment.

5.1.2 Value Functions, Firm Dynamics, and Markov Perfect Nash Equilibria

Incumbent firms choose either to operate or to scrap the firm with getting ϕ. The incumbent firms' value function is as below:

$$V(\omega, \mathbf{s}) = \max\left[\sup_{x\geq 0}\left\{R(\omega, \mathbf{s}; x) + \beta\sum_{\eta'}\sum_{\mathbf{s}'}\sum_{\omega'} V(\omega', \mathbf{s}')\, p(\omega'|\omega, x, \eta)\, q_{\omega}(\hat{\mathbf{s}}|\mathbf{s}, \eta')\, p_{\eta'}\right\}, \phi\right]. \tag{5.1.1}$$

$p_{\eta'}$ is the stochastic shift and $q_{\omega}(\mathbf{s}'|\mathbf{s}, \eta')$ produces the belief about the evolution of $\mathbf{s}$. The entrant's productivity evolution $\pi^e(\omega^0 - \eta'; 0, 0)$ is defined by $\pi^e(\omega^0 - \eta'; 0, 0) := p^e(\omega'|\omega, 0, \eta')$. At the beginning of each period, the entrant firm pays a sunk cost of entry x^e to enter the market, draws the initial level of productivity ω^0, and gets nothing in the period of entry. The entrant firms' value function is

$$V^e(\mathbf{s}, m) = \beta\sum_{\eta'}\sum_{\mathbf{s}'}\sum_{\hat{\omega}_m}\sum_{\omega^0} V\left(\omega^0, \mathbf{s}' + \mathbf{e}_{\omega^0} + \sum_{j=1}^{m-1}\mathbf{e}_{\omega_j^0}\right)\pi^e(\omega^0 - \eta'; 0, 0)\prod_{j=1}^{m-1}\pi^e\left(\omega_j^0 - \eta'; 0, 0\right) q^0(\mathbf{s}'|\mathbf{s}, \eta')\, p_{\eta'}, \tag{5.1.2}$$

where the number of entrants $m(\mathbf{s})$ is endogenously given by

$$m(\mathbf{s}) = \min\left\{0, \arg\min_{m\in\mathbb{Z}_+}\left\{V^e(\mathbf{s}, m) > x_m^e, V^e(\mathbf{s}, m+1) \leq x_{m+1}^e\right\}\right\}. \tag{5.1.3}$$

$\mathbf{e}_{\omega^0}$ is the unit vector with 1 in ω^0-th place. (5.1.1)-(5.1.3) summarize the entire dynamics of this economy.

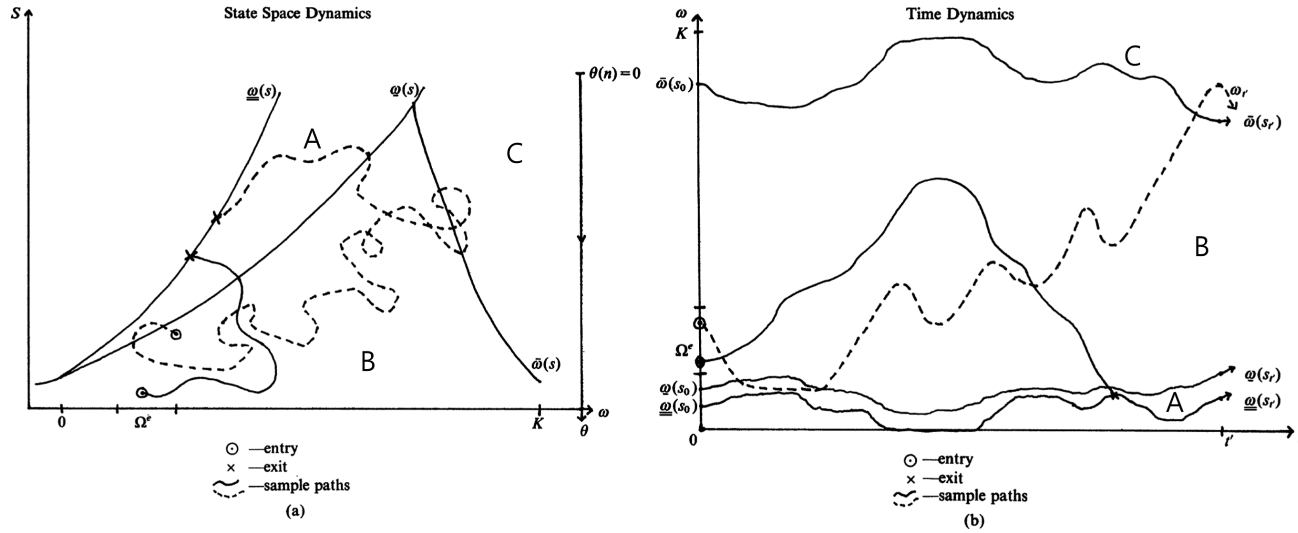

Figure 5.1.1: Figure 1 of Ericson and Pakes (1995)

Figure 5.1.1 depicts some simulated firm dynamics. In both panels, $\underline{\underline{\omega}}(\mathbf{s})$ denotes the threshold for exit, and $\underline{\omega}(\mathbf{s})$ and $\bar{\omega}(\mathbf{s})$ are the no-investment threshold. A denotes the no-investment region due to too-low productivity, C denotes the no-investment region due to a too-high cost. In B, investment occurs to improve the chance of moving on to the higher states in the future. The set Ω^e is also depicted, which is a subset of all possible states Ω.

Markov-perfect Nash equilibrium is defined by (i) the set of beliefs of each player regarding other players' unobservables, (ii) a Markov strategy profile[1] that maps the set of beliefs and the set of states to the set of own actions, and (iii) each player's belief coinciding with other players' actual action probability and the actual state-transition probability. Doraszelski and Satterthwaite (2010) prove the existence of the Markov perfect Nash equilibrium in a setup similar to this one. Existence of the equilibrium allows for empirical frameworks that employ or simplify this setup, to assume the data come from a Markov perfect equilibrium. Most empirical frameworks focus on (i) the algorithms finding the equilibria, with (ii) the equilibrium-selection rules.

Weintraub, Benkard, and van Roy (2008) develop a new equilibrium concept, the oblivious equilibrium, in a setup similar to Ericson and Pakes. In the oblivious equilibrium, players' strategy depends only on their own states and the long-run average industry state. The authors claim it is easier to compute than the Markov perfect Nash equilibrium. They further show the oblivious equilibrium is a limiting approximation to the Markov perfect Nash equilibrium under certain conditions.

[1]In the sense that the mapping depends only on the current states and the current beliefs. Note that without the Markov requirement, the equilibrium concept boils down to the perfect Bayesian Nash equilibrium.

5.2 Estimation Frameworks of Dynamic Discrete Games

5.2.1 Empirical Framework

Let $i\,(\in \mathcal{I} = \{1, 2, ..., I\})$ denote a firm that plays the repeated game in periods $t = 0, 1, 2, ..., \infty$. Let $a_{i,t}\,(\in \mathcal{J} = \{1, 2, ..., J\})$ be the action that firm i takes in period t. Let $\mathbf{s}_{i,t} := (x_{i,t}, \boldsymbol{\epsilon}_{i,t})$ be the state vector that has an observed component $x_{i,t}\,(\in \mathcal{X} = \{1, 2, ..., X\})$ and an unobserved component $\boldsymbol{\epsilon}_{i,t} \equiv \left(\epsilon_{i,j,t}\right)_{j\in\mathcal{J}}$. The sets $\mathcal{I}, \mathcal{J}, \mathcal{X}$[2] are taken as fixed. For tractability, we assume $\epsilon_{i,j,t}$ follows the mean-zero i.i.d. Type I extreme value distribution throughout the section unless noted otherwise.

Let $u_i\left(a_{i,t}, \mathbf{s}_{i,t}, \mathbf{a}_{-i,t}, \mathbf{s}_{-i,t}; \boldsymbol{\theta}\right) \equiv u_i\left(\mathbf{a}_t, \mathbf{s}_t; \boldsymbol{\theta}\right)$ be agent i's per-period utility function. Note agent i's utility is affected not only by her own state and action, but also by other players' state and action. Denote $\boldsymbol{\sigma}\left(\mathbf{s}_t\right) = \left(\sigma_i\left(\mathbf{x}_t, \boldsymbol{\epsilon}_{i,t}\right)\right)_{i\in\mathcal{I}}$ by the stationary Markov pure strategy of the players.[3] We impose the following assumptions, which are analogous to DDC(1)-DDC(3) in section 2.3.1. Although these assumptions are not the most general ones required, the assumptions encompass most empirical applications.

DDG(1) (Additive Separability of Preferences) Let $u_i\left(a_{i,t} = j, \mathbf{s}_{i,t}, \mathbf{a}_{-i,t}, \mathbf{s}_{-i,t}; \boldsymbol{\theta}\right)$ be the utility of agent i choosing action j at state $\mathbf{s}_t$. Then,

$$u_i\left(a_{i,t} = j, \mathbf{s}_{i,t}, \mathbf{a}_{-i,t}, \mathbf{s}_{-i,t}; \boldsymbol{\theta}\right) = \bar{u}_{i,j}\left(x_{i,t}, \mathbf{s}_{-i,t}, \mathbf{a}_{-i,t}; \boldsymbol{\theta}\right) + \epsilon_{i,j,t}.$$

DDG(2) ($\boldsymbol{\epsilon}_{i,t}$ is i.i.d. over t) $\forall \boldsymbol{\epsilon}_{i,t}\left(\in \mathbb{R}^J\right)$,

$$\Pr\left(\boldsymbol{\epsilon}_{i,t+1} | \boldsymbol{\epsilon}_{i,t}\right) = \Pr\left(\boldsymbol{\epsilon}_{i,t+1}\right).$$

DDG(3) ($\boldsymbol{\epsilon}_{i,t}$ is i.i.d. over i) $\forall \boldsymbol{\epsilon}_{i,t}\left(\in \mathbb{R}^J\right)$,

$$\Pr\left(\boldsymbol{\epsilon}_t\right) = \prod_{i\in\mathcal{I}} \Pr\left(\boldsymbol{\epsilon}_{i,t}\right).$$

DDG(4) (Conditional Independence of State-Transition Probabilities) Given $(\mathbf{x}_t, \mathbf{a}_t)$ observed,

$$\begin{aligned}
\Pr\left(\mathbf{x}_{t+1}, \boldsymbol{\epsilon}_{t+1} | \mathbf{x}_t, \boldsymbol{\epsilon}_t, \mathbf{a}_t; \boldsymbol{\varphi}\right) &\equiv \Pr\left(\boldsymbol{\epsilon}_{t+1} | \mathbf{x}_{t+1}, \mathbf{x}_t, \boldsymbol{\epsilon}_t, \mathbf{a}_t; \boldsymbol{\varphi}\right) \Pr\left(\mathbf{x}_{t+1} | \mathbf{x}_t, \boldsymbol{\epsilon}_t, \mathbf{a}_t; \boldsymbol{\varphi}\right) \\
&= \Pr\left(\boldsymbol{\epsilon}_{t+1}\right) \Pr\left(\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{a}_t; \boldsymbol{\varphi}\right).
\end{aligned}$$

DDG(2) and DDG(3) establish the i.i.d. of unobservable states over t and over i. DDG(4) establishes the

[2] We allow some players in $\mathcal{I}$ to be potentially inactive.

[3] We note here an important distinction in the notations. When the utility function $u_i\left(a_{i,t}, \mathbf{s}_{i,t}, \sigma_{-i}\left(\mathbf{s}_t\right), \mathbf{s}_{-i,t}; \boldsymbol{\theta}\right)$ or the value function $v_i\left(\boldsymbol{\sigma}\left(\mathbf{s}_t\right), \mathbf{s}_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right)$ is conditioned on other players' strategy, it denotes the expected utility or expected value over other players' actions. That is,

$$u_i\left(a_{i,t}, \mathbf{s}_{i,t}, \sigma_{-i}\left(\mathbf{s}_t\right), \mathbf{s}_{-i,t}; \boldsymbol{\theta}\right) := \int u_i\left(a_{i,t}, \mathbf{s}_{i,t}, \mathbf{a}_{-i,t}, \mathbf{s}_{-i,t}; \boldsymbol{\theta}\right) d\Pr\left(\mathbf{a}_{-i,t}\right).$$

However, player i's own strategy $\sigma_i\left(\mathbf{x}_t, \boldsymbol{\epsilon}_{i,t}\right)$ and own action a_i can be used interchangeably because they are observationally equivalent when players employ pure-strategies.

Markov state-transition probability such that the state transition is affected directly only by the current observed states and actions. As in the single-agent problem, the unobservable states $\boldsymbol{\epsilon}_t$ can affect the future states $\mathbf{x}_{t+1}$ only via the current action $\mathbf{a}_t$.

At each period t, each player's discounted sum of utility is

$$E_t\left[\sum_{s=t}^{\infty}\beta^{s-t}u_i\left(a_{i,s},\mathbf{s}_{i,s},\mathbf{a}_{-i,s},\mathbf{s}_{-i,s};\boldsymbol{\theta}\right)\mid\left\{\mathbf{x}_l\right\}_{l=1}^{t},\left\{\boldsymbol{\epsilon}_{i,l}\right\}_{l=1}^{t}\right],\tag{5.2.1}$$

where $\beta\in(0,1)$ is the discount factor. At each period t, each player observes the history of observed state variables $\{x_{i,l}\}_{l=1}^{t}$ for all the players and the history of own unobserved state variables $\{\boldsymbol{\epsilon}_{i,l}\}_{l=1}^{t}$, and then makes a decision $a_{i,t}$. Assuming DDG(1)-DDG(4) and that all the agents employ the stationary Markov decision rule $\boldsymbol{\sigma}(\mathbf{s}_t)=(\sigma_i(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}))_{i\in\mathcal{I}}$, the dynamic utility-maximization problem associated with (5.2.1) can be written recursively as follows:

$$\begin{aligned}&v_i\left(\boldsymbol{\sigma}\left(\mathbf{s}_t\right),\mathbf{s}_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\\=&\max_{\sigma_i\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right)}\left\{u_i\left(\sigma_i\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right),\boldsymbol{\sigma}_{-i}\left(\mathbf{s}_t\right),\mathbf{s}_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)+\beta\int v_i\left(\boldsymbol{\sigma}\left(\mathbf{s}_{t+1}\right),\mathbf{s}_{t+1};\boldsymbol{\theta},\boldsymbol{\varphi}\right)d\Pr\left(\mathbf{x}_{t+1}|\mathbf{x}_t,\boldsymbol{\sigma}\left(\mathbf{s}_t\right);\boldsymbol{\varphi}\right)\right\}.\end{aligned}$$

Analogous to the single-agent problem, we define the choice-specific value function $\bar{v}_{i,j}\left(\mathbf{x}_t;\boldsymbol{\sigma}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ and the integrated Bellman equation $\bar{\bar{v}}_i\left(\sigma_i\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right),\mathbf{x}_t;\boldsymbol{\sigma}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ as

$$\begin{aligned}&\bar{v}_{i,j}\left(\mathbf{x}_t;\boldsymbol{\sigma}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)\\:=\;&\bar{u}_{i,j}\left(\mathbf{x}_t;\boldsymbol{\sigma}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)+\beta\int v_i\left(\boldsymbol{\sigma}\left(\mathbf{s}_{t+1}\right),\mathbf{s}_{t+1};\boldsymbol{\theta},\boldsymbol{\varphi}\right)d\Pr\left(\mathbf{x}_{t+1}|\mathbf{x}_t,\boldsymbol{\sigma}\left(\mathbf{s}_t\right);\boldsymbol{\varphi}\right)\\&\bar{\bar{v}}_i\left(\sigma_i\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right),\mathbf{x}_t;\boldsymbol{\sigma}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)\\:=\;&E\left[v_i\left(\boldsymbol{\sigma}\left(\mathbf{s}_t\right),\mathbf{s}_t;\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right].\end{aligned}$$

Notice that $\bar{v}_{i,j}\left(\mathbf{x}_t;\boldsymbol{\sigma}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ and $\bar{\bar{v}}_i\left(\sigma_i\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right),\mathbf{x}_t;\boldsymbol{\sigma}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ are conditioned on other players' strategy $\boldsymbol{\sigma}_{-i}\left(\mathbf{s}_t\right)$.

The pure-strategy stationary Markov perfect Nash equilibrium $\boldsymbol{\sigma}^*\left(\mathbf{s}_t\right)$ is defined by the set of strategy profiles $\sigma_i\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right)$ such that for all $i\in\mathcal{I}$,

$$\sigma_i^*\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right)=\arg\max_{j\in\mathcal{J}}\left\{\bar{v}_{i,j}\left(\mathbf{x}_t;\boldsymbol{\sigma}_{-i}^*\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)+\epsilon_{i,j,t}\right\}.\tag{5.2.2}$$

As is common in other Nash equilibrium concepts, the Markov perfect Nash equilibrium requires each player's strategy to be the best response for other players' strategy.[4] As shown by Doraszelski and Satterthwaite (2010); Pesendorfer and Schmidt-Dengler (2008), the Markov perfect Nash equilibrium exists in the setup under some regularity conditions.

[4] To be precise, the strategy $\sigma_i^*\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right)$ returns the best-response action given player i's belief (in terms of the expectations) on the unobservables and on other players' strategies. Then, at the Markov perfect Nash equilibrium, the belief should coincide with the actual conditional choice probabilities of other players. The Markov perfect Nash equilibrium is a direct extension of the Bayesian Nash equilibrium to the infinite-horizon dynamic games under the Markov assumptions.

Existence of the equilibrium does not tell us much about the multiplicity of the equilibrium. When the data are generated by different equilibria that the econometrician cannot distinguish, estimation and policy counterfactuals might be pointless. All the estimation methods studied below in this chapter get around the problem by simply assuming the data come from a single pure-strategy Markov perfect equilibrium. Such an assumption may be more suitable if the econometrician uses the data of a single time series with fixed players.

5.2.2 Two-Sage Forward-Simulation Estimation

Bajari, Benkard, and Levin (2007) develop the two-stage estimation procedure, which extends the idea of forward simulation originally developed by Hotz, Miller, Sanders, and Smith (1994) in the context of a single agent's dynamic problem. In the first stage, the choice-specific value function $\bar{v}_{i,j}\left(\mathbf{x}_t;\boldsymbol{\sigma}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ and the state-transition probabilities $\Pr\left(\mathbf{x}_{t+1}|\mathbf{x}_t,\boldsymbol{\sigma}\left(\mathbf{s}_t\right);\boldsymbol{\varphi}\right)$ are estimated, and then the value functions $v_i\left(\sigma_i^*\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right),\mathbf{s}_t;\boldsymbol{\sigma}_{-i}^*\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)$ are simulated. In the second stage, the model parameter $\boldsymbol{\theta}$ is estimated.

Our presentation follows the reverse order; we first illustrate how the model parameter $\boldsymbol{\theta}$ can be estimated in the second stage. Then, we present how the policy functions and value functions can be estimated and simulated in the first stage.

5.2.2.1 Second-Stage Parameter Estimation

Let $\sigma_i\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t};\boldsymbol{\varphi}\right)$ be parametrized by a finite dimensional parameter vector $\boldsymbol{\varphi}$. Let $\hat{\boldsymbol{\varphi}}$ be the preliminary $\sqrt{N}$-consistent estimator of the parameter $\boldsymbol{\varphi}$.[5] If $\boldsymbol{\sigma}^*\left(\mathbf{s}_t\right)$ is an equilibrium strategy profile, it must be satisfied that

$$v_i\left(\sigma_i^*\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right),\mathbf{s}_t;\boldsymbol{\sigma}_{-i}^*\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)\geq v_i\left(\sigma_i'\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right),\mathbf{s}_t;\boldsymbol{\sigma}_{-i}^*\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right) \tag{5.2.3}$$

for all $i\in\mathcal{I}$ and $\sigma_i'\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right)$. (5.2.3) are the equilibrium inequalities that rationalize the strategy profile $\boldsymbol{\sigma}^*\left(\mathbf{s}_t\right)$. Let $\boldsymbol{\Theta}_0\left(\boldsymbol{\sigma}^*,\hat{\boldsymbol{\varphi}}\right)$ be the set of possible $\boldsymbol{\theta}$ such that given $\hat{\boldsymbol{\varphi}}$, (5.2.3) is satisfied for all possible $\left(\mathbf{s}_t,i,\sigma_i'\right)$. Suppose the set $\boldsymbol{\Theta}_0\left(\boldsymbol{\sigma}^*,\hat{\boldsymbol{\varphi}}\right)$ is a singleton.[6] The true parameter value $\boldsymbol{\theta}_0=\boldsymbol{\Theta}_0$ must satisfy that

$$\min_{\boldsymbol{\theta}} E\left[\left(\min\left\{v_i\left(\sigma_i^*\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right),\mathbf{s}_t;\boldsymbol{\sigma}_{-i}^*\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)-v_i\left(\sigma_i'\left(\mathbf{x}_t,\boldsymbol{\epsilon}_{i,t}\right),\mathbf{s}_t;\boldsymbol{\sigma}_{-i}^*\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right),0\right\}\right)^2\right]=0, \tag{5.2.4}$$

where the expectation is taken over $\left(\mathbf{s}_t,i,\sigma_i'\right)$ combination. Notice the term inside the expectation penalizes the violation of the inequality (5.2.3), which is similar to Ciliberto and Tamer (2009), studied in the preceding chapter. The authors suggest minimizing the simulation analogue of (5.2.4)

$$\begin{aligned}\frac{1}{N_s}\sum_{k_s=1}^{N_s}\Big(\min\Big\{&\hat{v}_{i,k_s}\left(\sigma_{i,k_s}^*\left(\mathbf{x}_{t,k_s},\boldsymbol{\epsilon}_{i,t,k_s}\right),\mathbf{s}_{t,k_s};\boldsymbol{\sigma}_{-i,k_s}^*\left(\mathbf{s}_{t,k_s}\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)\\&-\hat{v}_{i,k_s}\left(\sigma_{i,k_s}'\left(\mathbf{x}_{t,k_s},\boldsymbol{\epsilon}_{i,t,k_s}\right),\mathbf{s}_{t,k_s};\boldsymbol{\sigma}_{-i,k_s}^*\left(\mathbf{s}_{t,k_s}\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right),0\Big\}\Big)^2,\end{aligned} \tag{5.2.5}$$

[5] The state-transition probability $\Pr\left(\mathbf{x}_{t+1}|\mathbf{x}_t,\boldsymbol{\sigma}\left(\mathbf{s}_t\right);\boldsymbol{\varphi}\right)$ and its parameter $\hat{\boldsymbol{\varphi}}$ can be easily estimated using, for example, MLE or simple frequency estimator, depending on the specification.

[6] The set $\boldsymbol{\Theta}_0\left(\boldsymbol{\sigma}^*,\hat{\boldsymbol{\varphi}}\right)$ may not be a singleton. In such a case, bound estimation can be used to recover the confidence region. For details on the estimation and inference when $\boldsymbol{\theta}_0$ is not point identified, see Bajari, Benkard, and Levin (2007).

where k_s denotes the index for a randomly sampled $(\mathbf{s}_t, i, \sigma_i')$ combination.

One possibility for choosing $\sigma_i'(\mathbf{x}_t, \boldsymbol{\epsilon}_{i,t})$ is to give a slight perturbation on the estimated optimal policy, that is, $\sigma_i'(\mathbf{x}_t, \boldsymbol{\epsilon}_{i,t}) = \sigma_i^*(\mathbf{x}_t, \boldsymbol{\epsilon}_{i,t}) + \eta$. The authors note the choice of $\sigma_i'(\mathbf{x}_t, \boldsymbol{\epsilon}_{i,t})$ only affects the efficiency of the estimator, not the consistency. The estimator $\hat{\boldsymbol{\theta}}$ minimizing (5.2.5) is shown to be $\sqrt{N}$-consistent and asymptotically normal under suitable regularity conditions.

5.2.2.2 First-Stage Value-Function Estimation

The first stage is a combination of Hotz and Miller (1993); Hotz, Miller, Sanders, and Smith (1994). Consider the Hotz-Miller inversion:

$$\bar{v}_{i,j}\left(\mathbf{x}_t; \boldsymbol{\sigma}_{-i}^*(\mathbf{s}_t), \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}\right) - \bar{v}_{i,J}\left(\mathbf{x}_t; \boldsymbol{\sigma}_{-i}^*(\mathbf{s}_t), \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}\right) = \ln\left(\Pr\left(a_{i,j}|\mathbf{x}_t\right)\right) - \ln\left(\Pr\left(a_{i,J}|\mathbf{x}_t\right)\right). \tag{5.2.6}$$

[7] Because the sample analogue of the right-hand side is directly observed or can be easily estimated, we can obtain an estimate of all the choice-specific value differences for all the possible observed states $\mathbf{x}_t$. These difference estimates are used to find the optimal $\sigma_i(\mathbf{x}_t, \boldsymbol{\epsilon}_{i,t})$ during the forward simulation, where at each simulated path, $\epsilon_{i,j,t}$ is taken as known.

The remaining model primitives are the value-function estimates for each candidate parameter $\boldsymbol{\theta}$, with $\hat{\boldsymbol{\varphi}}$ fixed. Recall the definition of $v_i(\boldsymbol{\sigma}^*(\mathbf{s}_t), \mathbf{s}_t; \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}})$:

$$v_i(\boldsymbol{\sigma}^*(\mathbf{s}_t), \mathbf{s}_t; \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}) = E\left[\sum_{t=0}^{\infty} \beta^t u_i(a_{i,t}, \mathbf{s}_{i,t}, \mathbf{a}_{-i,t}, \mathbf{s}_{-i,t}; \boldsymbol{\theta})\,|\mathbf{x}_0, \boldsymbol{\epsilon}_{i,0}\right]. \tag{5.2.7}$$

We use forward simulation to compute $v_i(\boldsymbol{\sigma}^*(\mathbf{s}_t), \mathbf{s}_t; \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}})$. We impose the equilibrium assumption or perturbation assumption during the forward simulation. We list the steps below.

1. Fix the candidate parameter $\boldsymbol{\theta}$. Start at a randomly drawn state $\mathbf{x}_t$. Draw $\boldsymbol{\epsilon}_{i,t}$ for each $i \in \mathcal{I}$.

2. Compare the values of all $\bar{v}_{i,j}\left(\mathbf{x}_t; \boldsymbol{\sigma}_{-i}^*(\mathbf{s}_t), \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}\right) + \epsilon_{i,j,t}$ for all i and j pairs, which are obtained from (5.2.6), to find the optimal action $a_{i,t} \in \mathcal{J}$.

3. Compute $u_i(a_{i,t}, \mathbf{s}_{i,t}, \mathbf{a}_{-i,t}, \mathbf{s}_{-i,t}; \boldsymbol{\theta})$ for the set of optimal actions.

4. Draw a new state $(\mathbf{x}_{t+1}, \boldsymbol{\epsilon}_{t+1})$ using the estimates of $\Pr(\mathbf{x}_{t+1}|\mathbf{x}_t, \boldsymbol{\sigma}^*(\mathbf{s}_t); \hat{\boldsymbol{\varphi}})$ given the set of optimal actions.

5. Iterate over 1-4 until $\sum_{t=0}^{\infty} \beta^t u_i(a_{i,t}, \mathbf{s}_{i,t}, \mathbf{a}_{-i,t}, \mathbf{s}_{-i,t}; \boldsymbol{\theta})$ converges for all $i \in \mathcal{I}$.

6. Upon convergence, save the converged value in memory.

7. Repeat 1-6 many times, and then take the mean for each i, respectively, which will construct the simulation analogue of $v_i(\boldsymbol{\sigma}^*(\mathbf{s}_t), \mathbf{s}_t; \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}})$ for each candidate parameter $\boldsymbol{\theta}$.

[7] Notice that we invoke the Type I extreme value assumption here to simplify the inversion. The inversion, however, works for other distributions.

If the utility function $u_i\left(a_{i,t}, \mathbf{s}_{i,t}, \mathbf{a}_{-i,t}, \mathbf{s}_{-i,t}; \boldsymbol{\theta}\right)$ is linear in $\boldsymbol{\theta}$, that is,

$$u_i\left(a_{i,t}, \mathbf{s}_{i,t}, \mathbf{a}_{-i,t}, \mathbf{s}_{-i,t}; \boldsymbol{\theta}\right) = h_i\left(a_{i,t}, \mathbf{s}_{i,t}, \mathbf{a}_{-i,t}, \mathbf{s}_{-i,t}\right) \cdot \boldsymbol{\theta}$$

for some h_i, then iterating 1-4 only once will be sufficient to obtain the values of $v_i\left(\boldsymbol{\sigma}^*\left(\mathbf{s}_t\right), \mathbf{s}_t; \boldsymbol{\theta}, \hat{\boldsymbol{\varphi}}\right)$; that is, step 5 is not necessary, but step 7 is.

5.2.3 Estimation under Fixed-Point Equilibrium Characterization in a Probability Space

5.2.3.1 Characterization of the Markov perfect Nash Equilibrium as a Fixed Point of Conditional Choice Probabilities

Recall the condition for the Markov perfect Nash equilibrium (5.2.2): For each i,

$$\max_{j\in\mathcal{J}}\left\{\bar{v}_{i,j}\left(\mathbf{x}_t; \boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right), \boldsymbol{\theta}, \boldsymbol{\varphi}\right) + \epsilon_{i,j,t}\right\}.$$

Define $\bar{\bar{v}}_i\left(\mathbf{x}_t; \boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right), \boldsymbol{\theta}, \boldsymbol{\varphi}\right)$ as

$$\bar{\bar{v}}_i\left(\mathbf{x}_t; \boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right), \boldsymbol{\theta}, \boldsymbol{\varphi}\right) := \int \max_{j\in\mathcal{J}}\left\{\bar{v}_{i,j}\left(\mathbf{x}_t; \boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right), \boldsymbol{\theta}, \boldsymbol{\varphi}\right) + \epsilon_{i,j,t}\right\} \Pr\left(d\boldsymbol{\epsilon}_{i,t}\right). \tag{5.2.8}$$

Assume all the players employ the equilibrium strategy. As with the single-agent problem, we invert (5.2.8) in terms of the state-transition probability and the conditional choice probability. To that end, at an equilibrium, we have

$$\bar{\bar{\mathbf{v}}}_i\left(\mathbf{x}_t; \boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right), \boldsymbol{\theta}, \boldsymbol{\varphi}\right) = \left[\mathbf{I} - \beta\mathbf{P}\left(\mathbf{x}_{t+1}|\mathbf{x}_t; \boldsymbol{\sigma}^*\left(\mathbf{s}_t\right), \boldsymbol{\theta}, \boldsymbol{\varphi}\right)\right]^{-1} \mathbf{w}_i\left(\boldsymbol{\sigma}^*\left(\mathbf{s}_t\right), \mathbf{x}_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right) \tag{5.2.9}$$

for each $i \in \mathcal{I}$. Note each element of $\mathbf{w}_i\left(\sigma^*\left(\mathbf{s}_t\right), \mathbf{x}_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right)$ consists of

$$\mathbf{p}\left(a_{i,t}|\sigma^*_{-i}\left(\mathbf{s}_t\right), \mathbf{x}_t; \boldsymbol{\theta}, \boldsymbol{\varphi}\right)'\left(\bar{\mathbf{u}}_i\left(\mathbf{x}_t; \sigma^*_{-i}\left(\mathbf{s}_t\right), \boldsymbol{\theta}\right) - \ln\mathbf{p}\left(a_{i,t}|\mathbf{x}_t; \sigma^*_{-i}\left(\mathbf{s}_t\right), \boldsymbol{\theta}, \boldsymbol{\varphi}\right)\right).$$

8

Combining

$$\Pr\left(a_{i,t} = j|\mathbf{x}_t; \sigma^*_{-i}\left(\mathbf{s}_t\right), \boldsymbol{\theta}, \boldsymbol{\varphi}\right) = \frac{\exp\left(\bar{v}_{i,j}\left(\mathbf{x}_t; \sigma^*_{-i}\left(\mathbf{s}_t\right), \boldsymbol{\theta}, \boldsymbol{\varphi}\right)\right)}{\sum_{k\in\mathcal{J}}\exp\left(\bar{v}_{i,k}\left(\mathbf{x}_t; \sigma^*_{-i}\left(\mathbf{s}_t\right), \boldsymbol{\theta}, \boldsymbol{\varphi}\right)\right)} \tag{5.2.10}$$

[8] The dimensionality is such that we vectorized all the possible observed states. Both sides of (5.2.9) will be $X \times I$ dimensional vectors. (5.2.9) coincides with (2.3.24) otherwise.

and

$$
\begin{aligned}
& \bar{v}_{i,j}\left(\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right) \\
= \; & \bar{u}_{i,j}\left(\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right) \\
& +\beta \sum_{\mathbf{x}'\in\mathcal{X}^{\mathcal{I}}} \bar{\bar{\mathbf{v}}}_i\left(\mathbf{x}_{t+1}=\mathbf{x}';\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_{t+1}\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)\Pr\left(\mathbf{x}_{t+1}=\mathbf{x}'|\mathbf{x}_t,\boldsymbol{\sigma}^*\left(\mathbf{s}_t\right);\boldsymbol{\varphi}\right)
\end{aligned} \tag{5.2.11}
$$

with (5.2.9), we obtain the matrix-form fixed-point equations in the probability space:

$$
\mathbf{P}\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)=\Psi\left(\mathbf{P}\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\boldsymbol{\varphi}\right)\right). \tag{5.2.12}
$$

[9] Aguirregabiria and Mira (2007); Pesendorfer and Schmidt-Dengler (2008) show the fixed-point mapping (5.2.12) characterizes the Markov perfect Nash equilibrium of this dynamic game. To be specific, it is shown that (5.2.12) is the necessary and sufficient condition for the Markov perfect Nash equilibrium in this setup. Existence of the equilibrium follows by Brouwer's fixed-point theorem. But for a given $(\boldsymbol{\theta},\boldsymbol{\varphi})$, multiple sets of conditional choice probabilities that satisfy (5.2.12) may be possible.

5.2.3.2 Nested Pseudo-Likelihood Estimation

Aguirregabiria and Mira (2007) suggest a nested pseudo-likelihood (NPL) estimation method, which is essentially an extension of Aguirregabiria and Mira (2002). We assume that $\hat{\boldsymbol{\varphi}}$, the consistent estimator for $\boldsymbol{\varphi}$, can be obtained separately. The estimation algorithm proceeds in two stages as follows. Begin from the initial guess $\mathbf{P}^0\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)$. The first stage is to obtain

$$
\hat{\boldsymbol{\theta}}^K=\arg\max_{\boldsymbol{\theta}}\prod_i\mathcal{L}_i\left(\mathbf{P}^K\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)\right), \tag{5.2.13}
$$

where $\mathcal{L}_i\left(\cdot\right)$ denotes the likelihood induced by the argument and i denotes the data. The second stage is to update $\mathbf{P}\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)$ using the right-hand side of (5.2.9), and then plugging back into (5.2.11) and (5.2.10) sequentially. To that end, the K-stage pseudo-likelihood is updated as

$$
\mathbf{P}^{K+1}\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)=\Psi\left(\mathbf{P}^K\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\hat{\boldsymbol{\theta}}^K,\hat{\boldsymbol{\varphi}}\right)\right). \tag{5.2.14}
$$

Iterating over K until convergence of $\left(\hat{\boldsymbol{\theta}}^K,\mathbf{P}^K\right)$ will yield the nested pseudo-likelihood estimator, and the corresponding fixed point $\mathbf{P}^\infty\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\hat{\boldsymbol{\theta}}^\infty,\hat{\boldsymbol{\varphi}}\right)$.

When $K=0$, that is, when the econometrician does not iterate over K, the resulting estimator is called the two-step pseudo-maximum likelihood (PML) estimator. Given that the initial guess $\mathbf{P}^0\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)$ is a consistent estimator, two-step PML estimators are consistent and asymptotically normal, but the finite sample properties can be worse than nested pseudo-likelihood estimators.

Multiple set of fixed points $\left(\hat{\boldsymbol{\theta}}^\infty,\mathbf{P}^\infty\right)$ may exist that satisfy (5.2.14). In such a case, Aguirregabiria

[9] To be precise, the operator $\Psi\left(\cdot\right)$ should have $(\boldsymbol{\theta},\boldsymbol{\varphi})$ as its argument as well. For brevity, we suppress it unless otherwise noted.

and Mira (2007) asserts that the one that maximizes the right-hand side of (5.2.13) as the estimator $\hat{\boldsymbol{\theta}}^{\infty}$. $\hat{\boldsymbol{\theta}}^{\infty}$ is consistent only when it attains the global maximum of the pseudo-likelihood function. Thus, Aguirregabiria and Mira suggest the econometrician try different $\mathbf{P}^{0}\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^{*}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)$'s to find "all" the possible convergence points of $\hat{\boldsymbol{\theta}}^{\infty}$, and then compare the pseudo-likelihood values.

However, such an algorithm may never converge to the true data-generating fixed points, regardless of the choice of the initial values. That is, although the nested pseudo-likelihood estimator is well defined, the algorithm suggested by Aguirregabiria and Mira may fail to converge to the true parameter. Pesendorfer and Schmidt-Dengler (2010) show the nested pseudo-likelihood iteration always converges, but only to the stable fixed points. Thus, if the data come from an unstable equilibrium, the pseudo-likelihood iteration may fail to converge to the equilibrium that generates the data.

Kasahara and Shimotsu (2012) further show the nested pseudo-likelihood algorithm converges only when the fixed-point mapping has the local contraction property. They suggest an alternative updating algorithm in (5.2.14) as follows:

$$\mathbf{P}^{K+1}\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^{*}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)=\left\{\Psi\left(\mathbf{P}^{K}\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^{*}_{-i}\left(\mathbf{s}_t\right),\hat{\boldsymbol{\theta}}^{K},\hat{\boldsymbol{\varphi}}\right)\right)\right\}^{\lambda}*\left\{\mathbf{P}^{K}\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^{*}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)\right\}^{1-\lambda}$$

for a small $\lambda\in(0,1)$, where $*$ denotes the elementwise product and the exponents are taken elementwise. Optimal[10] λ cannot be known to the econometrician ex ante, because it is a function of the model primitives. All the fixed points satisfying the local contraction property is more likely when λ is small. The tradeoff of a small λ is that it will require more iterations until convergence.

5.2.3.3 Constrained Full Maximum Likelihood Estimation

Egesdal, Lai, and Su (2015) suggest the full maximum likelihood estimation, the possibility of which is also briefly mentioned by Aguirregabiria and Mira (2007). They consider the following constrained optimization problem:

$$\max_{\boldsymbol{\theta},\left\{\bar{\bar{\mathbf{v}}}_i\left(\mathbf{x}_t;\boldsymbol{\sigma}^{*}_{-i}(\mathbf{s}_t),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)\right\}_{i\in\mathcal{I}},\left\{\mathbf{P}\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^{*}_{-i}(\mathbf{s}_t),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)\right\}_{i\in\mathcal{I}}}\prod_i\mathcal{L}_i\left(\mathbf{P}\left(a_{i,t}|\mathbf{x}_t;\boldsymbol{\sigma}^{*}_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)\right),$$

such that (5.2.12), (5.2.11), (5.2.10), and (5.2.9) hold. In this constrained maximization problem, the objective function and all the constraints are smooth functions, so the derivative-based optimizer can find the local maxima as usual. When multiple local maxima exist, the researcher should try many starting values and try to find the global maximum. They report that this full maximum likelihood performs better than the two-step PML or NPL estimators.

5.2.3.4 Asymptotic Least-Squares Estimators

Pesendorfer and Schmidt-Dengler (2008) take the fixed-point equilibrium characterization (5.2.12) as a

[10] It is optimal in the sense that it is a required minimum to guarantee the local contraction property to every fixed point.

moment condition. Taking the expectation on (5.2.12) yields

$$E\left[\text{vec}\left\{\mathbf{P}\left(a_{i,t}|\mathbf{x}_t;\sigma^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)\right\}-\text{vec}\left\{\Psi\left(\mathbf{P}\left(a_{i,t}|\mathbf{x}_t;\sigma^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)\right)\right\}\right]=\mathbf{0}.$$

To calculate the fixed-point iteration

$$\Psi\left(\mathbf{P}\left(a_{i,t}|\mathbf{x}_t;\sigma^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right);\sigma^*\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right),$$

the state-transition probability $\mathbf{P}\left(\mathbf{x}_{t+1}|\mathbf{x}_t;\sigma^*\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)$ and the conditional choice probabilities $\mathbf{P}\left(a_{i,t}|\mathbf{x}_t;\sigma^*_{-i}\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)$ have to be taken as the argument of $\Psi\left(\cdot;\sigma^*\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)$. They suggest a few possibilities. One of them is to use a consistent estimator $\left\{\hat{\mathbf{P}}\left(a_{i,t}|\mathbf{x}_t\right),\hat{\mathbf{P}}\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right)\right\}$ such as simple frequency, and take it as data. Another possibility is to use

$$E\left[\text{vec}\left\{\mathbf{1}\left(a_{i,t}|\mathbf{x}_t\right)\right\}-\text{vec}\left\{\Psi\left(\hat{\mathbf{P}}\left(a_{i,t}|\mathbf{x}_t\right);\sigma^*\left(\mathbf{s}_t\right),\boldsymbol{\theta},\hat{\boldsymbol{\varphi}}\right)\right\}\right]=\mathbf{0}.$$

Either way will yield the consistent estimates, but the asymptotic variance will be different. They derive the asymptotically efficient weighting matrix. The remaining estimation step is as usual: Impose the corresponding moment conditions and implement the nonlinear GMM.

The two-step estimators of Bajari, Benkard, and Levin (2007); Pakes, Ostrovsky, and Berry (2007) can be characterized as a variant of the asymptotic least-squares estimators—as using different conditional choice probabilities, different state-transition probabilities, or different weighting matrices in the nonlinear GMM estimation. An advantage of using these two-stage estimators is that they place less computational burden than the estimation methods that iterate over the fixed-point equation. A possible downside is that the two-stage estimators may suffer from a large finite-sample bias.

5.2.4 Further Issues and Discussion

5.2.4.1 (Under)-Identification

Underidentification in dynamic games becomes much more serious than in the single-agent problem. Pesendorfer and Schmidt-Dengler (2008) show the degree of underidentification increases exponentially with the number of players in the game. The outline of the argument is as below. Each agent has J possible actions. When I players are in the game, the combination of possible beliefs on the actions of all the players are $I \times J^I$. But the dimension of the observed conditional choice probabilities are at most $I \times J$. Thus, the degree of underidentification $I \times J^I - I \times J$ increases exponentially with I.[11] One possible restriction that Pesendorfer and Schmidt-Dengler suggest is to impose the independence of the per-period utility of each player in the game.

Bajari, Chernozhukov, Hong, and Nekipelov (2015) suggest imposing the exclusion restriction on per-period utility, i.e., some of the state variables are excluded from player i's per-period choice specific

[11] The argument is not exact, because we do not consider the number of free equations, which is $J-1$.

value function, which implies

$$\bar{v}_{i,j}\left(\mathbf{x}_t;\sigma^*_{-i}\left(\mathbf{s}_t\right)\right)=\bar{v}_{i,j}\left(\mathbf{x}_{i,t};\sigma^*_{-i}\left(\mathbf{s}_t\right)\right)$$

[12] where $\mathbf{x}_t\equiv\left(\mathbf{x}_{i,t},\mathbf{x}_{-i,t}\right)$. Bajari, Chernozhukov, Hong, and Nekipelov show the nonparametric identification of $\bar{u}_{i,j}\left(\mathbf{x}_t;\sigma^*_{-i}\left(\mathbf{s}_t\right)\right)$ can be achieved if (i) the conditional distribution of $\mathbf{x}_{-i,t}$ given $\mathbf{x}_{i,t}$ has more support points than J, and (ii) the following J^{I-1} equations are linearly independent:

$$\bar{u}_{i,j}\left(\mathbf{x}_t;\sigma^*_{-i}\left(\mathbf{s}_t\right)\right)=\sum_{\mathbf{a}_{-i}\in\mathcal{J}^{I-1}}\Pr\left(\mathbf{a}_{-i}|\mathbf{x}_t\right)\bar{u}_{i,j}\left(\mathbf{x}_{i,t};\mathbf{a}_{-i,t}\right).$$

Aguirregabiria and Magesan (2015) extend this identification result to the circumstances where players' beliefs are not in equilibrium.

5.2.4.2 Unobserved Heterogeneity

All the estimation frameworks for the dynamic-game models that we studied thus far commonly assume no private information other than the i.i.d. idiosyncratic shocks $\boldsymbol{\epsilon}_t$. Such shocks are unobservable both to the other players in the game and to the econometrician. In that sense, the econometrician has the same degree of information as the players. This assumption can often be inadequate. Recently, attempts have been made to incorporate the unobserved heterogeneity, which is possibly serially correlated, into the model.

Igami and Yang (2017) estimate the dynamic entry model of hamburger chains, which accommodates the unobserved heterogeneity, combining the estimation procedures suggested by Bajari, Benkard, and Levin (2007); Arcidiacono and Miller (2011), building upon the identification results provided by Kasahara and Shimotsu (2009). They consider the time-invariant unobserved market types, where some markets can be more lucrative than others, which leads to some markets being more attractive for entry than others. The method is essentially a combination of Arcidiacono and Miller (2011) and Bajari, Benkard, and Levin (2007). Kasahara and Shimotsu discretize the type of markets, estimate the conditional probability of each market belonging to a specific type, then compute the market-type-specific conditional choice probability of entry. The rest of the estimation procedure follows Bajari, Benkard, and Levin. Igami and Yang find that ignoring the unobserved market heterogeneity attenuates the effect of competition.

5.2.4.3 Dynamic Discrete Games with Equilibrium Concepts Other than MPE

Fershtman and Pakes (2012) develop a setup and framework that can accommodate the unobserved actions and states in dynamic games. It is based on a new dynamic equilibrium concept that they introduce, which is called the *experience based equilibrium*.[13] In the Markov perfect equilibrium-based frameworks,

[12] $(\boldsymbol{\theta},\boldsymbol{\varphi})$ are suppressed because we are considering the *nonparametric* identification problem. That is, we do not assume here that the per-period utility function is parametrized by $\boldsymbol{\theta}$.

[13] The idea of such an equilibrium concept is first outlined in Pakes et al. (2007).

accommodating the unobserved actions or states is difficult, mainly because for each t, each player's belief should coincide with the actual conditional choice probabilities and the actual state-transition probabilities. An experience-based equilibrium relaxes this requirement. It only requires that the beliefs should coincide with the conditional choice probabilities only at the limiting recurrent states. In the experience-based equilibrium, each agent maximizes the sum of the expected discounted value, conditional on the limiting information sets that are observed to the corresponding players when the game is repeated infinitely many times. All the players in the game learn about other players' best responses as the game is repeated.

Experience-based equilibrium is a relaxation of the Markov perfect Nash equilibrium. Because the Markov perfect Nash equilibrium exists in such a setup, the existence of experience-based equilibrium is guaranteed. Fershtman and Pakes provide an algorithm to calculate the experience-based equilibria under unobserved actions and/or states, and then conduct a simulation study. The framework is well suited to the environment where the learning of the players is involved, because it naturally involves the learning process of the players to calculate the equilibrium. Asker, Fershtman, Jeon, and Pakes (2016); Doraszelski, Lewis, and Pakes (Forthcoming) apply the experience-based equilibrium framework to collusion and entry to the UK electricity system, respectively.

Igami (2017, Forthcoming) assumes the finite-horizon sequential entry/exit game with nonstationarity, in which the firms are forward looking because of the investment or innovation decisions. He estimates the corresponding dynamic game assuming the firms employ the perfect Bayesian Nash equilibrium strategies.

References

AGUIRREGABIRIA, V. AND C.-Y. HO (2012): "A dynamic oligopoly game of the US airline industry: Estimation and policy experiments," *Journal of Econometrics*, 168, 156–173.

AGUIRREGABIRIA, V. AND A. MAGESAN (2015): "Identification and estimation of dynamic games when players' beliefs are not in equilibrium," *Working Paper*.

AGUIRREGABIRIA, V. AND P. MIRA (2007): "Sequential estimation of dynamic discrete games," *Econometrica*, 75, 1–53.

——— (2010): "Dynamic discrete choice structural models: A survey," *Journal of Econometrics*, 156, 38–67.

AGUIRREGABIRIA, V., P. MIRA, AND H. ROMAN (2007): "An estimable dynamic model of entry, exit, and growth in oligopoly retail markets," *American Economic Review: Papers and Proceedings*, 97, 449–454.

AGUIRREGABIRIA, V. AND N. SUZUKI (2014): "Identification and counterfactuals in dynamic models of market entry and exit," *Quantitative Marketing and Economics*, 12, 267–304.

ARCIDIACONO, P. AND R. A. MILLER (2011): "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity," *Econometrica*, 79, 1823–1867.

ASKER, J., C. FERSHTMAN, J. JEON, AND A. PAKES (2016): "The competitive effects of information sharing," *Working Paper*.

BAJARI, P., C. L. BENKARD, AND J. LEVIN (2007): "Estimating dynamic models of imperfect competition," *Econometrica*, 75, 1331–1370.

BAJARI, P., V. CHERNOZHUKOV, H. HONG, AND D. NEKIPELOV (2015): "Identification and efficient semiparametric estimation of a dynamic discrete game," *NBER Working Paper*, no. 21125.

BHASKAR, V., G. J. MAILATH, AND S. MORRIS (2013): "A foundation for Markov equilibria in sequential games with finite social memory," *Review of Economic Studies*, 80, 925–948.

CILIBERTO, F. AND E. TAMER (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77, 1791–1828.

DORASZELSKI, U., G. LEWIS, AND A. PAKES (Forthcoming): "Just starting out: Learning and equilibrium in a new market," *American Economic Review*.

DORASZELSKI, U. AND A. PAKES (2007): *A framework for applied dynamic analysis in IO*, Elsevier, vol. 3 of *Handbook of Industrial Organization*, chap. 30, 1887–1966.

DORASZELSKI, U. AND M. SATTERTHWAITE (2010): "Computable Markov-perfect industry dynamics," *RAND Journal of Economics*, 41, 215–243.

EGESDAL, M., Z. LAI, AND C.-L. SU (2015): "Estimating dynamic discrete-choice games of incomplete information," *Quantitative Economics*, 6, 567–597.

ERICSON, R. AND A. PAKES (1995): "Markov-perfect industry dynamics: A framework for empirical work," *Review or Economic Studies*, 62, 53–82.

FERSHTMAN, C. AND A. PAKES (2012): "Dynamic games with asymmetric information: A framework for empirical work," *Quarterly Journal of Economics*, 127, 1611–1661.

HOTZ, V. J. AND R. A. MILLER (1993): "Conditional choice probabilities and the estimation of dynamic models," *Review of Economic Studies*, 60, 497–529.

HOTZ, V. J., R. A. MILLER, S. SANDERS, AND J. SMITH (1994): "A simulation estimator for dynamic models of discrete choice," *Review of Economic Studies*, 61, 265–289.

IGAMI, M. (2017): "Estimating the innovator's dilemma: Structural analysis of creative destruction in the hard distk drive industry, 1981-1998," *Journal of Political Economy*, 125, 798–847.

——— (Forthcoming): "Industry dynamics of offshoring: The case of hard disk drives," *American Economic Journal: Microeconomics*.

IGAMI, M. AND N. YANG (2017): "Unobserved heterogeneity in dynamic games: Cannibalization and preemptive entry of hamburger chains in Canada," *Quantitative Economics*, 7, 483–521.

KASAHARA, H. AND K. SHIMOTSU (2009): "Nonparametric identification of finite mixture models of dynamic discrete choices," *Econometrica*, 77, 135–175.

——— (2012): "Sequential estimation of structural models with a fixed point constraint," *Econometrica*, 80, 2303–2319.

MASKIN, E. AND J. TIROLE (1988): "A theory of dynamic oligopoly, I: Overview and quantity competition with large fixed costs," *Econometrica*, 56, 549–569.

——— (2001): "Markov perfect equilibrium: I. Observable actions," *Journal of Economic Theory*, 100, 191–219.

PAKES, A., M. OSTROVSKY, AND S. BERRY (2007): "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)," *RAND Journal of Economics*, 38, 373–399.

PESENDORFER, M. AND P. SCHMIDT-DENGLER (2008): "Asymptotic least squares estimators for dynamic games," *Review of Economic Studies*, 75, 901–928.

——— (2010): "Sequential estimation of dynamic discrete games: A comment," *Econometrica*, 78, 833–842.

WEINTRAUB, G. Y., C. L. BENKARD, AND B. VAN ROY (2008): "Markov perfect industry dynamics with many firms," *Econometrica*, 76, 1375–1411.

Chapter 6

Empirical Frameworks of Consumer Search

Search is a major modeling framework reflecting friction in the information of an economic agent. In this chapter, we briefly overview the historical development of theoretical frameworks on optimal consumer search and price dispersion. Next, we study empirical frameworks that combine data with the predictions from the search models such as optimal search sequence or search size.

Search models commonly derive the optimal size or degree of search under different setups. When the theory models of search is to be combined with empirics, availability of data becomes a primary concern. In an ideal world, an econometrician should be able to observe individual-level data of (i) consumers' thought process regarding the search-mode decision (simultaneous / sequential / mixed), (ii) the set (or even better, sequence) of searched alternatives, and (iii) the purchase decisions. Such detailed individual-level search data are not easily available, and earlier works developed empirical frameworks to identify and estimate the search-cost distribution out of (iv) market data on the observed price dispersion. In this section, we review the empirical frameworks on consumer search combining different degrees of data availability with theoretic models of consumer search. We note that to the best of our knowledge, none of the existing empirical works to date incorporate (i) directly.

6.1 Utility Specification and Some Preliminary Results

6.1.1 Utility Specification in Consumer Search Models

We focus on the price search of consumers within the context of a more general utility search. Consumers search over discrete alternatives to maximize the utility net of search cost. Let $u_{i,j}$ be consumer i's utility of purchasing $j \in \mathcal{J}$. The utility is given by

$$u_{i,j} := u\left(-p_j, \mathbf{x}_j, \xi_j, \epsilon_{i,j}\right). \tag{6.1.1}$$

The utility function above takes price p_j, observed product characteristics $\mathbf{x}_j$, unobserved product characteristics ξ_j, and idiosyncratic utility shock $\epsilon_{i,j}$ as its argument. We put the negative sign on the price term p_j to emphasize that the utility is decreasing in the prices.

The utility specification (6.1.1) is fairly general, encompassing most empirical setups that have been employed in the literature thus far. For example, if consumers search only for the prices for an identical product, the utility function is simply, for some large y,

$$u_{i,j} = y - p_j.$$

This way, the problem of finding an alternative that yields the best utility can be directly converted into the problem of finding an alternative that yields the lowest price. In the context of differentiated-products markets, the search procedure can be thought of as consumers locking in to the characteristics first, and then search for the prices (e.g., all the classical search models are covered in section 6.2, Honka (2014); Honka and Chintagunta (2017)). Notice that a one-unit increase of p_j cannot be compensated by any change of $\mathbf{x}_j$. The other extreme would be the linear utility, that is,

$$u_{i,j} := -\alpha p_j + \mathbf{x}_j' \boldsymbol{\beta} + \xi_j + \epsilon_{i,j},$$

where a price increase can be compensated linearly by an increase/decrease of $\mathbf{x}_j$ or ξ_j.

Throughout sections 6.1 and 6.2, we assume the product is homogeneous; therefore, consumers search only for the lowest price unless otherwise noted. Consumers want to find a seller that offers the lowest price for the item, but search is costly. Consumers therefore optimize over the size and/or order of search to maximize the expected benefit net of search costs. Note that the results and implications derived in sections 6.1 and 6.2 apply directly to the utility search context where consumers search in order to maximize the utility net of search costs.

6.1.2 Some Preliminary Results on Stochastic Dominance

For a quick reference, we present some equivalent statements for the first-order and second-order stochastic dominance without proof. The following results are used extensively in the search literature.

Proposition 6.1.1. *(First-Order Stochastic Dominance) Let $F(\cdot)$ and $G(\cdot)$ be two continuous cumulative distribution functions on $[0,\infty)$. The following statements are equivalent.*
(i) $F(\cdot)$ first-order stochastically dominates $G(\cdot)$.
(ii) $\forall x \in [0,\infty)$, $F(x) \leq G(x)$.
(iii) For any nondecreasing function $u : \mathbb{R}^+ \to \mathbb{R}^+$, for any Borel subset $\mathcal{B}$ of $[0,\infty)$,

$$\int_{\mathcal{B}} u(x)\, dF(x) \geq \int_{\mathcal{B}} u(x)\, dG(x).$$

Proposition 6.1.2. *(Second-Order Stochastic Dominance) Let $F(\cdot)$ and $G(\cdot)$ be two continuous cumulative distribution functions on $[0,\infty)$. The following statements are equivalent.*
(i) $F(\cdot)$ second-order stochastically dominates $G(\cdot)$.
(ii) $\forall x_0 \in [0,\infty)$,

$$\int_0^{x_0} F(x)\, dx \leq \int_0^{x_0} G(x)\, dx.$$

(iii) For any nondecreasing concave function $u : \mathbb{R}^+ \to \mathbb{R}^+$, *for any Borel subset* $\mathcal{B}$ *of* $[0, \infty)$,

$$\int_{\mathcal{B}} u(x)\, dF(x) \geq \int_{\mathcal{B}} u(x)\, dG(x).$$

A few remarks are in order. First, Proposition 6.1.1 (iii) implies that if $F(\cdot)$ first-order stochastically dominates $G(\cdot)$, then $E_F[x_j] \geq E_G[x_j]$. However, $E_F[x_j] \geq E_G[x_j]$ does not tell anything about first-order stochastic dominance. Second, first-order stochastic dominance implies second-order stochastic dominance, but not vice versa. This is straightforward by comparing (iii) of Propositions 6.1.1 with 6.1.2. Third, if $E_F[x_j] = E_G[x_j]$, $F(\cdot)$ second-order stochastically dominates $G(\cdot)$ if and only if $G(\cdot)$ is a mean-preserving spread of $F(\cdot)$. Note that if $G(\cdot)$ is a mean-preserving spread of $F(\cdot)$, then $Var_G(x_j) \geq Var_F(x_j)$, but not vice versa. Comparing only variance says little about the mean-preserving spread.

6.2 Classical Search-Theoretic Models: Sequential Search and Simultaneous Search

We review classical consumer search-theoretic models in this section. Classical models of consumer search, dating back to Stigler (1961), are motivated by observed price dispersion of a product even when the product is identical. It was considered a major challenge to the classical Walrasian demand model's law-of-one-price prediction. Our purpose here is not to give a comprehensive review of the search theory literature; rather, we intend to give a concise overview of the extent to which they are employed in recent empirical literature.

The classical consumer search models use two major search schemes: sequential search and simultaneous search.[1] In sequential search, consumers draw one price at each period and decide whether to keep searching for one more or to stop searching. The sequential search problem boils down to the optimal stopping problem. In simultaneous search, consumers commit to the search set ex ante before engaging in the actual search. Then, they draw all the prices in the search set, collect the prices, and then choose an alternative that has the lowest price.

The baseline assumptions common to the search models that we study in this section are as follows:

AS(1) Each consumer has a unit inelastic demand.

AS(2) A continuum of consumers have the identical choke price θ.

AS(3) Consumers know their own search-cost function.

AS(4) Consumers know the common price distribution $F(p)$, which is exogenous, nondegenerate, and homogeneous over $j\ (\in \mathcal{J})$.[2]

AS(5) Each price draw is independent.[3]

[1] Chronologically simultaneous search was developed first. However, we study sequential search first for illustrative purposes.

[2] It implies the price distribution $F(p)$ is not indexed by j, which we relax in section 6.2.2.2.

[3] An independent price draw is closely related to "no-learning."

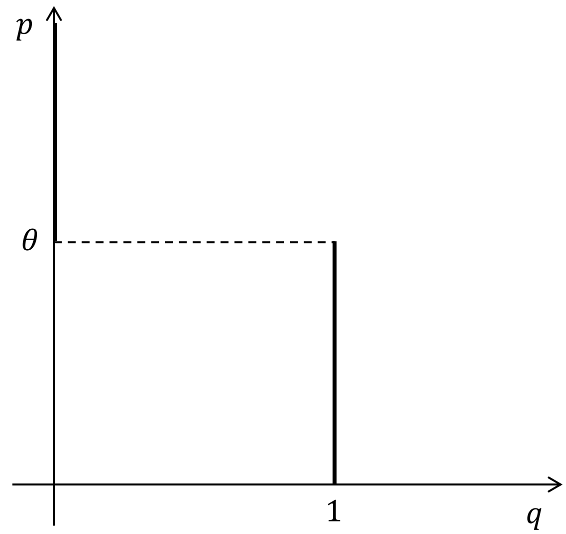

Figure 6.2.1: Unit Inelastic Individual Demand Curve

The shape of consumers' common unit inelastic demand, AS(1) and AS(2) combined, is depicted in Figure 6.2.1. AS(3) and AS(4) are about the prior knowledge of the consumers. The actual price distribution is considered to be prior knowledge of consumers, and no learning component is involved. We assume the price distribution is a priori knowledge of the consumers. AS(5) implies no spillover of information in search. We stick to this assumption because the majority of empirical works to date do not seem to incorporate search with learning.

6.2.1 Sequential Search

6.2.1.1 Sequential Search with Homogeneous Ex-ante Price Distribution and Perfect Recall

The sequential search framework originates from McCall (1970)'s wage-search problem, which was later adapted to the consumers' price-search context. In sequential search, consumers decide whether to stop or keep searching after each price draw. In addition to AS(1)-AS(5), we impose the following three assumptions for the sequential search.

AS(SEQ1) At the beginning of each period t, consumers draw one price out of J stores, and then decide whether to stop searching or keep searching for one more.

AS(SEQ2) Consumers' constant discount factor is $\delta\ (\leq 1)$ and the constant marginal cost of searching one more price is $c\ (> 0)$. δ and c are common across $\mathcal{J}$.

AS(SEQ3) Consumers can revisit any of the previously searched stores costlessly.

AS(SEQ1) is the sequential search assumption, AS(SEQ2) specifies consumers' time preference and search cost, and AS(SEQ3) is the perfect recall assumption.

Let p_t^B be the minimum price searched thus far at period t, that is,

$$p_t^B := \min_{1\leq s\leq t} \{p_s\}.$$

At each period, consumers can collect p_t^B for sure, or try to search one more store by paying the marginal cost of search c. If consumers are searching once more, the search result p_{t+1} would either be such that $p_{t+1} \geq p_t^B$ or $p_{t+1} < p_t^B$. Consumers will stick to p_t^B in the former case. To formalize, the period t decision problem becomes

$$\begin{aligned} &\min\left\{p_t^B, c + \delta p_t^B \Pr\left(p_{t+1} > p_t^B\right) + \delta \Pr\left(p_{t+1} \leq p_t^B\right) E\left[p_{t+1} | p_{t+1} \leq p_t^B\right]\right\} \\ =&\min\left\{p_t^B, c + \delta p_t^B \int_{p_t^B}^{\bar{p}} dF(p) + \delta \int_0^{p_t^B} p dF(p)\right\}. \end{aligned} \tag{6.2.1}$$

We claim the solution to the problem (6.2.1) has a reservation-price structure. That is, a unique price level p^* exists such that consumers accept p_t^B if $p_t^B \leq p^*$ and reject p_t^B if $p_t^B > p^*$. To show the existence and uniqueness of the solution for the reservation price, let us subtract $\delta p_t^B \int_0^{\bar{p}} dF(p)$ from each term inside the min operator, which will not change the solution of the original problem. After rearranging, (6.2.1) simplifies as

$$\min\left\{p_t^B (1-\delta), c + \delta \int_0^{p_t^B} \left(p - p_t^B\right) dF(p)\right\}.$$

Subtracting $\delta \int_0^{p_t^B} \left(p - p_t^B\right) dF(p)$ and rearranging yields

$$\min\left\{p_t^B (1-\delta) + \delta \int_0^{p_t^B} \left(p_t^B - p\right) dF(p), c\right\}. \tag{6.2.2}$$

Now, denote

$$\psi\left(p_t^B\right) := p_t^B (1-\delta) + \delta \int_0^{p_t^B} \left(p_t^B - p\right) dF(p). \tag{6.2.3}$$

It can be easily verified that $\psi(\cdot)$ is continuous, monotonically increasing,[4] $\lim_{p_t^B \to 0+} \psi\left(p_t^B\right) = 0$, and $\lim_{p_t^B \to +\infty} \psi\left(p_t^B\right) = \infty$. Because $c > 0$, $\psi\left(p_t^B\right)$ crosses c just once. Hence, a unique solution for the problem (6.2.2) exists, and thus for the equivalent problem (6.2.1) as well. We define p^* implicitly as the solution for

$$\psi(p^*) = c, \tag{6.2.4}$$

[4]Under some regularity conditions,

$$\frac{\partial}{\partial p_B}\psi(p_B) = (1-\delta) + \frac{\partial}{\partial p_B}\delta \int_0^{p_B} (p_B - p) dF(p),$$

and it can be easily verified that the second part $\frac{\partial}{\partial p_B}\delta \int_0^{p_B} (p_B - p) dF(p)$ is simply $\delta F(p_B)$ by using the Leibniz integral rule. Then, it is also straightforward that $\frac{\partial^2}{\partial p_B^2}\psi(p_B) > 0$, thereby implying $\psi(\cdot)$ is convex.

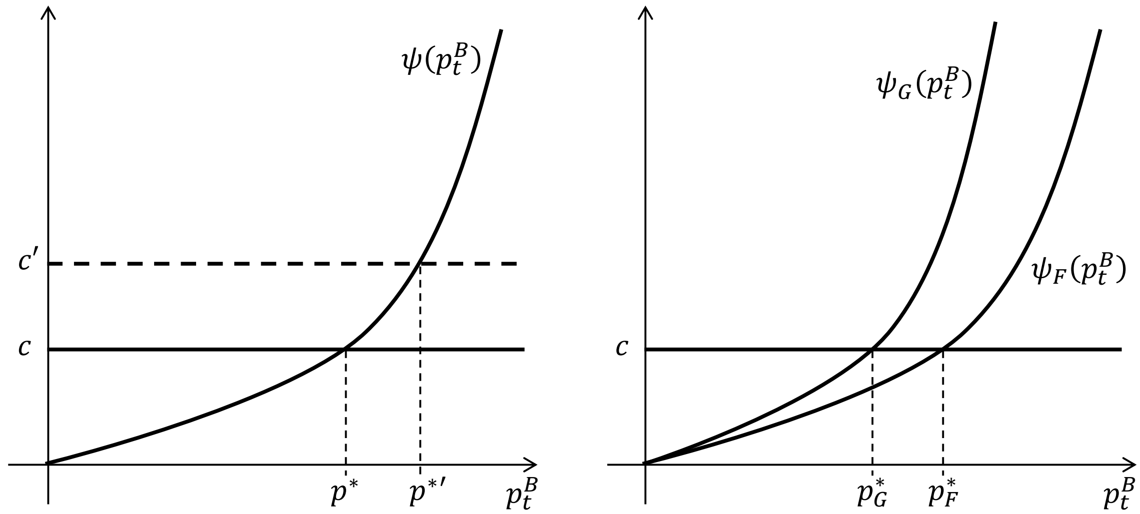

Figure 6.2.2: Determination of the Reservation Price p^*

which is well defined as we have just shown. We refer to p^* as the reservation price henceforth.

The left panel of Figure 6.2.2 illustrates how p^* is implicitly determined. When the consumer has $p_t^B < p^*$ in hand, collecting the current p_t^B and quitting is better than drawing one more price in the next period. The consumer will therefore choose the first option in the choice problem (6.2.1). When the consumer has $p_t^B > p^*$ in hand, drawing one more price is better than just collecting p_t^B and quitting, so the consumer will choose the second option in (6.2.1). Notice that a higher marginal cost of search c' will induce a higher reservation price, and therefore, consumers will accept higher price draws then before. Such a prediction is consistent with the intuition that consumers want to cease the search earlier because of the increased search cost. Another important implication is that consumers will accept the first price draw that is below p^* because of this reservation-price structure and homogeneity of the ex-ante price distribution $F(\cdot)$. That is, recall never occurs in the setup when the ex-ante price distribution is homogeneous.

We further examine what happens to the reservation price when the underlying price distribution $F(\cdot)$ changes in a specific way: first-order and second-order stochastic dominance. Suppose a price distribution $G(\cdot)$ exists such that $F(\cdot)$ first-order or second-order stochastically dominates $G(\cdot)$. Then, for any $p_t^B > 0$, because $\left(p - p_t^B\right)$ is an increasing and (weakly) concave function in p, the following holds by Propositions 6.1.1 and 6.1.2:

$$\int_0^{p_t^B} \left(p - p_t^B\right) dF(p) \geq \int_0^{p_t^B} \left(p - p_t^B\right) dG(p). \tag{6.2.5}$$

Now let us index $\psi(\cdot)$ by F and G, respectively, and then subtract:

$$\begin{aligned}\psi_F\left(p_t^B\right)-\psi_G\left(p_t^B\right) &= -\delta\left\{\int_0^{p_t^B}\left(p-p_t^B\right)dF(p)-\int_0^{p_t^B}\left(p-p_t^B\right)dG(p)\right\}\\ &\leq 0.\end{aligned}$$

The above inequality implies $\psi_F(\cdot)$ will be always be below $\psi_G(\cdot)$ at the same p_t^B value. As a result, the associated p_G^* would be lower than p_F^*. Determination of p^* with respect to the ex-ante price-distribution change is depicted in the right panel of Figure 6.2.2.

Recall from Propositions 6.1.1 and 6.1.2 in section 6.1.2 that $E_F\left[x_j\right]\geq E_G\left[x_j\right]$ when (6.2.5) holds. Furthermore, if the equality holds so $E_F\left[x_j\right]=E_G\left[x_j\right]$, then $Var_G\left(x_j\right)\geq Var_F\left(x_j\right)$. An immediate result is that a "better" ex-ante distribution[5] in the sense of first-order stochastic dominance will decrease the reservation price. Intuitively, consumers will have higher standards on the draws from $G(\cdot)$ than from $F(\cdot)$. If the expectations of the two ex-ante price distributionsare identical, a "riskier" distribution will decrease the reservation price. All else being equal, consumers have "more to gain by searching" when an option is riskier. These intuitions will carry over when the ex-ante price distribution is heterogeneous.

6.2.1.2 An Excursus: Sequential Search with Homogeneous Ex-ante Price Distribution and No Recall

AS(SEQ3) allows for a perfect recall. However, consumers never return to a previously visited store to collect p_s for $s<t$ in the problem structure presented above, where the price distribution is homogeneous across alternatives. One may conjecture even more: The reservation price p^* might be identical regardless of whether recall is allowed, and we show here that such a conjecture is indeed true.[6] AS(SEQ3) is now replaced by the following:

AS(SEQ3') Consumers cannot revisit any of the previously searched stores.

When recall is not allowed, consumers must either take the current price draw p_t or abandon p_t to search for one more price. To formalize, consumers solve

$$\begin{aligned}&\min\left\{p_t, c+\delta\Pr\left(p_{t+1}\leq p_t\right)E\left[p_{t+1}|p_{t+1}\leq p_t\right]\right\}\\ =&\min\left\{p_t, c+\delta\int_0^{p_t}pdF(p)\right\}.\end{aligned}\tag{6.2.6}$$

[5] Better means the chance of low price realization is higher. In this case, $G(\cdot)$ is better than $F(\cdot)$.

[6] This is the original wage-search model studied by McCall (1970).

Subtracting $\delta \int_0^{p_t} pdF(p)$ and rearranging yields

$$\begin{aligned}
&\min\left\{(1-\delta)\,p_t+\delta p_t-\delta\int_0^{p_t} pdF(p)\,,c\right\}\\
=&\min\left\{(1-\delta)\,p_t+\delta\int_0^{p_t} p_t dF(p)-\delta\int_0^{p_t} pdF(p)\,,c\right\}\\
=&\min\left\{(1-\delta)\,p_t+\delta\int_0^{p_t}(p_t-p)\,dF(p)\,,c\right\},
\end{aligned}$$

which boils down to the identical decision problem as in (6.2.2). It is trivial that the decision rule is also identical.

6.2.1.3 Sequential Search with Heterogeneous Ex-ante Price Distributions

In this subsection, we relax the homogeneity assumptions AS(4) and AS(SEQ2), keeping AS(1), (2), (3), (5), (SEQ1), and (SEQ3) intact. The relaxed assumptions are as follows.

AS(4') Consumers know the price distribution $F_j(p)$, which is exogenous and nondegenerate. $F_j(p)$ may differ over $j\,(\in\mathcal{J})$.

AS(SEQ2') Consumers' constant discount factor is $\delta_j\,(<1)$ and the constant marginal cost of searching one more price is $c_j\,(>0)$. δ_j and c_j may differ over $j\,(\in\mathcal{J})$.

The price distribution, marginal cost of search, and the discount factor are all alternative specific. The alternative-specific discount factor may be interpreted as some alternatives taking more time for the price to be discovered than others. Because alternatives are heterogeneous, consumers' sequential decision problem entails an extra component: which alternative to search first. Weitzman (1979) proved the optimal decision rule in this setup follows three simple criteria using the reservation prices: ordering, stopping, and choice rule.

Before presenting the criteria, recall how the reservation-price structure was determined in the homogeneous case in (6.2.2). In this heterogeneous alternative case, (6.2.2) is converted to

$$\min\left\{p_B\left(1-\delta_j\right)+\delta_j\int_0^{p_B}(p_B-p)\,dF_j(p)\,,c_j\right\}, \tag{6.2.7}$$

where p_B is the lowest price in hand. Notice the reservation price p_j^* itself, which is determined implicitly by equating the two terms in (6.2.7), is alternative specific. p_j^* can be thought of as a function of $\left(F_j(\cdot)\,,\delta_j,c_j\right)$, and it is not affected by the order of which alternative is searched first. We present Weitzman's optimal decision rule below, which corresponds to the ordering, stopping, and choice rule, respectively.

WS(1) Calculate the reservation price p_j^* for $j=1,...,J$. Without losing generality, order $\mathcal{J}$ such that $p_1^*<p_2^*<\cdots<p_J^*$. Search from alternative 1 to J sequentially in the increasing order.

WS(2) Stop at j if the actual price draw $p_j<p_{j+1}^*$; that is, stop if the actual price draw is lower than all the reservation prices of the unsearched alternatives.

WS(3) Collect the alternative with the lowest price among those searched.

We briefly discussed in section 6.2.1.1 that consumers' higher standards on the ex-ante price distributions will boil down to the form of a lower reservation price in the sequential search models with homogeneous ex-ante price distributions. When the ex-ante price distribution is heterogeneous over alternatives, the ordering rule WS(1) states that consumers draw prices first from the alternatives that "better outcomes are expected." The rule seems intuitive, but the proof is rather technical, so we do not go over it here.

Allowing or not allowing for perfect recall is no longer innocuous. That is, consumers may return to a foregone alternative to collect p_s for $s < t$. To see why, consider the example in Table 6.1 with four alternatives. The consumer stops at $t = 2$, returns to the first alternative, and collects 25. The search ceases at $t = 2$ even if the current best draw is not smaller than alternative 2's reservation price. The reason is that when the consumer proceeds to $t = 3$, the best draw before drawing the price of store 3 is already smaller than 30. Waiting for one more period is not optical for the consumer, especially because the discount factor is smaller than 1. Lastly, we note that Salop (1973) shows a similar implication to Weitzman (1979), namely, that the reservation price increases over the order of draw even when recall is not allowed.

Order of Draw	0	1	2	3	4
Reservation Price	–	10	20	30	40
Actual Draw	–	25	28	–	–
Current Best Draw	∞	25	25	–	–

Table 6.1: An Example of Recall When Alternatives Are Ex-ante Heterogeneous

6.2.2 Simultaneous Search

6.2.2.1 Simultaneous Search with Homogeneous Ex-ante Price Distributions

Simultaneous search, often referred to as fixed-sample search or nonsequential search, was pioneered by Stigler (1961). We assume the following in addition to AS(1)-AS(5):

AS(SIM1) Consumers choose and commit to the search set $\mathcal{S}$ ($\subset \mathcal{J}$) ex ante before engaging in search.

AS(SIM2) Marginal cost of search c is constant and common across consumers.

Notice that choosing the search set $\mathcal{S}$ ex-ante is equivalent to choosing $n := |\mathcal{S}|$, the size of the search set, when the price distribution is homogeneous.

Let $F(p)$ be the prior distribution that has a compact support $[0, \bar{p}]$ and a density $f(p)$.[7] Before engaging in search, consumers choose the size of price draws n and commits to it. Consumers choose n to minimize the expected minimum price plus the total cost of search. The marginal cost of search c is

[7] Extension of the argument to the domain $[0, \infty)$ is straightforward.

constant. Consumers solve

$$\begin{aligned} n^* &= \arg\min_n \left\{ E\left[\min_{1\le k\le n} p_k\right] + cn \right\} \\ &= \arg\min_n \left\{ \int_0^{\bar{p}} np\left(1-F(p)\right)^{n-1} f(p)\,dp + cn \right\}. \end{aligned} \tag{6.2.8}$$

Equality (6.2.8) follows from the fact that the density of the minimum-order statistics among n i.i.d. draws is given by $np\left(1-F(p)\right)^{n-1} f(p)$.[8]

It is immediate that increasing $n$involves a tradeoff: The expected minimum price decreases while the cost of search increases. To see the fact more clearly, let us convert the cost-minimization problem (6.2.8) as the net benefit maximization problem by using a negative sign and replacing the minimum operator with the maximum operator as follows:

$$\begin{aligned} n^* &= \arg\max_n \left\{ -E\left[\min_{1\le k\le n} p_k\right] - cn \right\} \\ &= \arg\max_n \left\{ E\left[-\min_{1\le k\le n} p_k\right] - cn \right\} \\ &= \arg\max_n \left\{ E\left[\max_{1\le k\le n} \{-p_k\}\right] - cn \right\} \\ &= \arg\max_n \left\{ E\left[\max_{1\le k\le n} u_k\right] - cn \right\}, \end{aligned} \tag{6.2.9}$$

where we use the utility specification $u_k = y - p_k$ with location adjustment y in (6.2.9), which will not change the solution of the original problem. Denote M_n by the benefit of drawing n prices:

$$\begin{aligned} M_n &:= -E\left[\min_{1\le k\le n} p_k\right] \\ &= -\int_0^{\bar{p}} np\left(1-F(p)\right)^{n-1} f(p)\,dp \\ &= -\int_0^{\bar{p}} \left(1-F(p)\right)^n dp. \end{aligned} \tag{6.2.10}$$

[8] Because

$$\begin{aligned} \Pr\left(p_1 > p, p_2 > p, ..., p_n > p\right) &= \prod_{k=1}^{n} \Pr\left(p_k > p\right) \\ &= \left(1-F(p)\right)^n, \end{aligned}$$

the cumulative distribution function of the minimum price is $1-\left(1-F(p)\right)^n$, with density $n\left(1-F(p)\right)^{n-1} f(p)$. Similarly, the cumulative distribution function of the maximum price is simply $\left(F(p)\right)^n$, with density $n\left(F(p)\right)^{n-1} f(p)$.

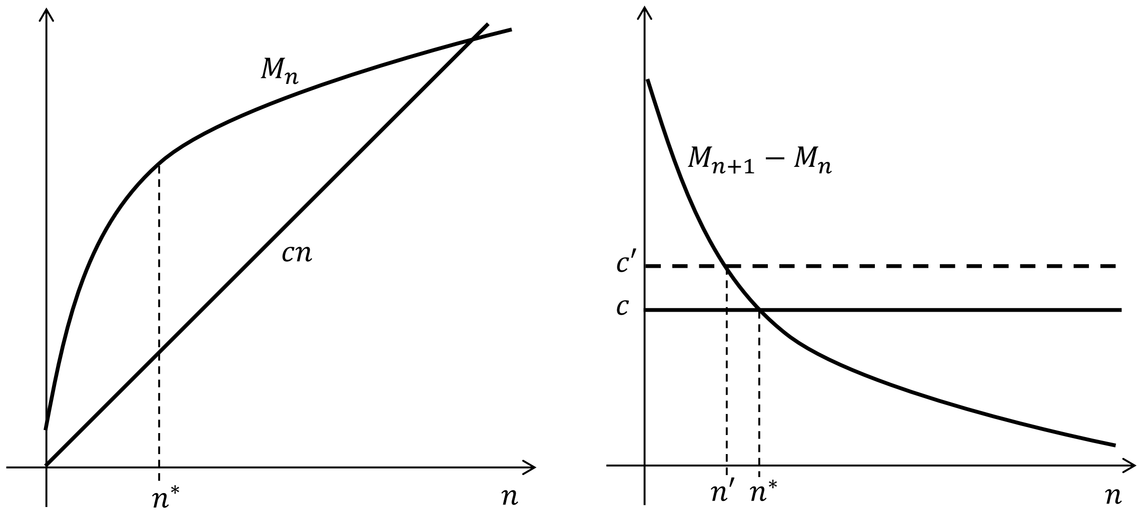

Figure 6.2.3: Determination of the Optimal Search Size in the Stigler's Model

[9] Then, the marginal benefit of drawing one more price is

$$M_{n+1} - M_n = \int_0^{\bar{p}} (1 - F(p))^n F(p)\, dp. \tag{6.2.11}$$

The right-hand side of (6.2.11) is monotonically decreasing, converging to zero as $n \to \infty$.

The search cost, on the other hand, is linearly increasing over n.[10] Figure 6.2.3 depicts how the optimal search is determined. The left panel illustrates the original problem (6.2.8),[11] and the right panel shows the fact that the optimal search size n^* is determined at the level at which the marginal benefit of search equals the marginal cost of search. When the marginal cost of search c increases, the optimal search size n^* decreases.

---

[9]The following lemma is used:

Lemma 6.2.1. *Let $G_{(n)}(\cdot)$ be a cumulative distribution function with a density $g_{(n)}(\cdot)$ that has a support $[0, \bar{p}]$. Then,*

$$\int_0^{\bar{p}} p g_{(n)}(p)\, dp = \int_0^{\bar{p}} \left(1 - G_{(n)}(p)\right) dp.$$

Proof. It is straightforward from integration by parts:

$$\int uv' = uv - \int u'v$$

,and letting $u = 1 - G_{(n)}(p)$ and $v = p$. □

[10]Search cost does not need to be linear over n; it only has to be convex over n to draw the same implications.

[11]We assumethe benefit of searching for at least one alternative is high enough in drawing the left panel, that is, $y \gg 0$ in the utility specification $u_j = y - p_j$. The assumption can be understood as a positive location adjusted to M_n, which will not change the solution of the original problem as long as y is sufficiently large. This assumption $y \gg 0$ is often replaced with the assumption that the first search is free.

6.2.2.2 Simultaneous Search with Heterogeneous Ex-ante Price Distributions

Vishwanath (1992); Chade and Smith (2005, 2006) study generalized versions of the simultaneous search problem where each store may have a different ex-ante distribution of prices. That is, assumptions AS(1), (2), (3), (5), and (SIM) remain intact, but AS(4) is changed as the following.

AS(4') Consumers know the price distribution $F_j(p)$, which is exogenous and nondegenerate. $F_j(p)$ may differ over $j\,(\in \mathcal{J})$.

Consumers choose the set to search $\mathcal{S}\,(\subset \mathcal{J})$ ex ante, draw the prices of all the alternatives in $\mathcal{S}$, and then choose the best draw from $\mathcal{S}$. The cost function of search $c(\cdot)$ is assumed to depend only on the cardinality of $\mathcal{S}$. Because $F_j(p)$ differs by j, the problem of choosing $\mathcal{S}$ now entails two distinct factors: choosing the composition and size of $\mathcal{S}$. Chade and Smith consider both (i) choosing $n^* \equiv |\mathcal{S}|$ first and then the composition of $\mathcal{S}$, and (ii) choosing $|\mathcal{S}|$ and the composition of $\mathcal{S}$ simultaneously. (i) and (ii) can be achieved simply by adjusting the cost function $c(|\mathcal{S}|)$. In case (i),

$$c(|\mathcal{S}|) = \begin{cases} 0 & \text{if } |\mathcal{S}| \le n^* \\ \infty & \text{otherwise} \end{cases},$$

and in case (ii),

$$c(|\mathcal{S}|) = c\,|\mathcal{S}|\,.$$

Now let us formalize the problem. Consumers solve

$$\begin{aligned} \mathcal{S}^* &= \arg\min_{\mathcal{S}\subset\mathcal{J}} E\left[\min_{k\in\mathcal{S}}\{p_k\} + c(|\mathcal{S}|)\right] \\ &= \arg\min_{\mathcal{S}\subset\mathcal{J}} \left\{E\left[\min_{k\in\mathcal{S}}\{p_k\}\right] + c(|\mathcal{S}|)\right\} \end{aligned} \tag{6.2.12}$$

to find the optimal search set $\mathcal{S}^*$. Consumers' search problem to find $\mathcal{S}^*$ entails two intertwined choice problems: (i) the size and (ii) composition of $\mathcal{S}^*$. Simple formulas such as in equation (6.2.8) do not exist, because now all the price distributions are indexed by j. In principle, consumers have to consider every possible subset of $\mathcal{J}$ and then compare each possible subset to find the optimal set $\mathcal{S}^*$. The cardinality of the set to be compared is $2^{|\mathcal{J}|}$. This brute-force comparison is inefficient, and the computing time explodes because the number of comparisons increas exponentially over $|\mathcal{J}|$.

Chade and Smith (2005, 2006) provide a simple recursive algorithm that has a polynomial computing time and the sufficient conditions under which the algorithm finds the optimal set $\mathcal{S}^*$. We first describe the algorithm, which is called the Marginal Improvement Algorithm (MIA). Let $\mathcal{S}_t$ denote the optimal search set at stage t. At stage t,[12]

MIA(1) Choose l_{t+1} such that it improves the expected price the most, that is,

$$l_{t+1} = \arg\min_{l\in(\mathcal{J}\setminus\mathcal{S}_t)} E\left[\min_{k\in(l\cup\mathcal{S}_t)}\{p_k\}\right].$$

[12] We assume $c(1)$ is low enough, or u in $u_k := u - p_k$ is high enough, to ensure consumers engage in searching.

MIA(2) Stop if the marginal benefit is smaller than the marginal cost. That is, if

$$E\left[\min_{k\in\mathcal{S}_t}\{p_k\}\right]-E\left[\min_{k\in(l_{t+1}\cup\mathcal{S}_t)}\{p_k\}\right]<c\,(t+1)-c\,(t)\,,$$

ompare $E\left[\min_{k\in\mathcal{S}_t}\{p_k\}+c\left(|\mathcal{S}_t|\right)\right]$ with $E\left[\min_{k\in\mathcal{S}_{t+1}}\{p_k\}+c\left(|\mathcal{S}_{t+1}|\right)\right]$, and take $\mathcal{S}^*$ as the set that returns a smaller value.[13]

MIA(3) Otherwise, take $\mathcal{S}_{t+1}:=\mathcal{S}_t\cup l_{t+1}$ and go back to MIA(1).

The algorithm resembles Weitzman (1979)'s rules in sequential search particularly because it sorts the alternative set $\mathcal{J}$ ex ante. The crucial difference is, however, the ordering criteria: Weitzman's rule sorts the alternatives by the reservation price, whereas the MIA sorts the alternatives by the order of improving the expected minimum price. Now we state some sufficient conditions for the MIA to be optimal.

Theorem 6.2.1. *(Sufficient Conditions for the Optimality of the MIA) MIA(1)-MIA(3) find the search set $\mathcal{S}^*$ that solves (6.2.12) if either of the following is met:*

(i) (First-Order or Quasi-Second-Order Stochastic Dominance) $\mathcal{J}$ can be ordered in a way that for each $j=1,...,J-1$, $\forall r\in[0,\bar{p}]$ where $[0,\bar{p}]$ contains the union of the supports of $\{F_j(\cdot)\}$,

$$\int_r^{\bar{p}}F_j\left(p\right)dp\leq\int_r^{\bar{p}}F_{j+1}\left(p\right)dp \tag{6.2.13}$$

with strict inequality at $r=0$.

(ii) (Binary Success) For each $j\in\mathcal{J}$, the cumulative distribution function $F_j(p)$ has the following form:

$$F_j\left(p\right)=\begin{cases}0 & \text{if } p<p_j^0\\ \alpha & \text{if } p_j^0\leq p<\bar{p}\\ 1 & \text{if } p>\bar{p}\end{cases}$$

with $0<\alpha\leq 1$ and $p_j^0\in(0,\bar{p})$.

Notice that first-order stochastic dominance over j is sufficient for (6.2.13), in which case part (i) states that MIA(1)-MIA(3) finds the optimal search set $\mathcal{S}^*$. Part (ii) states that if the price draw has a gamble structure such that for some positive probability the draw is the highest possible price (failure),[14] MIA(1)-MIA(3) find the optimal search set $\mathcal{S}^*$.

MIA(1)-MIA(3) reduce the order of the number of comparisons from $2^{|\mathcal{J}|}$ to $(|\mathcal{J}|-1)+...+1$, by introducing the ex-ante optimal ordering of $\mathcal{J}$. To see why, notice that MIA(1) compares $E\left[\min_{k\in\mathcal{J}}\{p_k\}\right]$ for each k at $t=1$, and therefore, the maximum number of comparisons is $(|\mathcal{J}|-1)$. The same logic applies downward for $t=2,...,J$. An important implication of Theorem 6.2.1 is that when first-order

[13] This comparison is necessary because t can take only integer values.
[14] This is the case when $\alpha<1$, where the failure probability is $1-\alpha$.

or second-order stochastic dominance of the underlying price distribution can be assumed, one can simplify the search algorithm substantially. We study some examples of empirical applications in section 6.4. As a last remark, similar to the homogeneous price-distribution case of Stigler (1961), it is immediate from MIA(2) that an upward shift/pivot of the cost function $c(\cdot)$ will decrease the size of $\mathcal{S}^*$.

6.3 Price Dispersion in the Market Equilibrium and Search Cost Identification With Price Data

6.3.1 Critiques on Classical Consumer Search Models as Explanations of the Observed Price Dispersion

We studied the sequential and simultaneous search models in sections 6.2.1 and 6.2.2, respectively. Stigler (1961)'s original motivation for developing the search model was to explain the observed, persistent price dispersion of even an identical product. Roughly speaking, by introducing the search friction, the simple search models we studied above might successfully generate an elastic downward-sloping nondegenerate market demand function from a collection of homogeneous degenerate demand functions.[15] One of the common core assumptions imposed in the search models was that the price distribution $F(\cdot)$ is exogenously given to be a nondegenerate, well-behaving distribution.[16] The supplier-side problem from which $F(\cdot)$ is generated, however, is left unresolved: an observed price dispersion should not exist if it is optimal for the suppliers to sell at the identical price. It turns out that it is indeed optimal for the suppliers to set a single price in the simple search models we studied above, which leads to a "paradox." Here, we provide a brief overview of the arguments for why such a paradox arises, and discuss what adjustments to the models can possibly resolve the paradox.

This paradox was raised by Diamond (1971); Rothschild (1973). They argue that firms' optimal pricing has no price dispersion if consumers are searching, and therefore, the assumption that $F(\cdot)$ is a nondegenerate distribution leads to a contradiction in the equilibrium. We describe the essential logic of their critiques in a simplified static setup where firms engage in price competition to maximize their own profits. Let $q(p_j)$ and $c(p_j)$ be the demand and cost function that firm j faces. Let

$$\pi(p_j) := p_j q(p_j) - c(p_j)$$

be the profit function of firm j. Supposea single joint profit-maximizing price level $\bar{p} > 0$ exists such that for any $p_j \neq \bar{p}$, $\pi(p_j) < \pi(\bar{p})$.

The argument for consumers engaging in simultaneous search is as follows. Consumers have the same c, the same $F(\cdot)$, and therefore the same optimal search size n^*. Let p_j be the price that firm j

[15]A prominent example of a homogeneous degenerate demand function is the unit-inelastic demand curve illustrated in Figure 6.2.1.

[16]Weitzman (1979); Chade and Smith (2005, 2006), who allow $F(\cdot)$ to be heterogeneous across suppliers, do not mention much about how the heterogeneous price distribution can be sustained as an equilibrium behavior.

charges. A consumer will purchase from the firm if j is in the searched set $\mathcal{S}$[17] and

$$p_j \leq \min_{k \in \mathcal{S}} \{p_k\}.$$

Let I be the measure of consumers in the market. The ex-ante (expected) demand curve that a firm faces is

$$\begin{aligned} q(p_j) &= I \Pr\left(\{j \in \mathcal{S}\} \cap \left\{p_j \leq \min_{k \in \mathcal{S}} \{p_k\}\right\}\right) \\ &= I \Pr(j \in \mathcal{S}) \Pr\left(p_j \leq \min_{k \in \mathcal{S}} \{p_k\} \middle| j \in \mathcal{S}\right) \\ &= I \frac{n^*}{|\mathcal{J}|} \left(1 - F(p_j)\right)^{n^*-1}. \end{aligned} \tag{6.3.1}$$

[18] Knowing this demand curve, firms will charge the price that maximizes expected profit. Because all the firms will charge the same price, we should see no persistent equilibrium price dispersion.[19]

The argument for consumers engaging in sequential search is as follows. Because every consumer has the same cost of marginal search c, the associated reservation price p^* is identical across consumers. Note that p^* is strictly in between 0 and $\bar{p}$. Consider any firm that charges $p < p^*$. The firm will have an incentive to charge $p + \epsilon$, where $0 < \epsilon < c$. Consumers will still accept $p + \epsilon$ because the search cost c is higher than ϵ. The same logic yields that $p + \epsilon$, $p + 2\epsilon$, and so on cannot be the equilibrium price. Now that all the firms will charge p^*, consumers' new reservation price should shift up to somewhere between p^* and $\bar{p}$. On the other hand, no firm will have an incentive to charge any price higher than $\bar{p}$. Repeating this logic, it turns out that the only possible price at the equilibrium is $\bar{p}$, which is the joint profit-maximizing price. The existence of the homogeneous search cost c in the sequential search actually induces a stable collusion.

The key factor causing this "paradox" is the homogeneity of the suppliers. To resolve this paradox to allow for an equilibrium price dispersion within the consumer search context, the following is usually necessary: (i) a downward-sloping market demand curve with search friction and (ii) an upward-sloping

[17] Recall that $|\mathcal{S}| = n^*$.

[18] Derivation of (6.3.1) is as follows:

$$\begin{aligned} \Pr\left(p_j \leq \min_{k \in \mathcal{S}} \{p_k\} \middle| j \in \mathcal{S}\right) &= \Pr\left(\min_{k \in \{\mathcal{S} \setminus j\}} \{p_k\} > p_j \middle| j \in \mathcal{S}\right) \\ &= \prod_{k \in (\mathcal{S} \setminus j)} \Pr\left(p_k > p_j\right) \\ &= \prod_{k \in (\mathcal{S} \setminus j)} \left(1 - \Pr\left(p_k \leq p_j\right)\right) \\ &= \left(1 - F\left(p_j\right)\right)^{n^*-1}. \end{aligned} \tag{6.3.2}$$

Note that we invoked the assumption of independent draws in (6.3.2).

[19] Furthermore, the price that the firms charge will not equal the Bertrand price in general, because an ϵ deviation from the optimal price does not make the firm lose all of its demand. Exactly what price is optimal depends on the shape of $F(\cdot)$. However, investigating the property of $F(\cdot)$ is not informative in this context, because in equilibrium, the assumption that $F(\cdot)$ is a nondegenerate distribution is contradictory anyways.

supply curve. We illustrate a few different ways the literature uses to derive the equilibrium price dispersion.

6.3.2 Equilibrium Price-Dispersion Models and Search-Cost Distribution Identification Using Market-level Price Data

In this subsection, we study models that explain the sustained price dispersion in a market equilibrium. Price dispersion can be an equilibrium outcome if (i) different firms charge different prices and (ii) firms do not have incentives to change the current pricing strategy. As mentioned previously, firm-side heterogeneity is required for a price dispersion to be sustained as a market equilibrium. The models we study in this section show the existence of a price-dispersion equilibrium along with one or multiple single price equilibria. Specific assumptions that lead to a price-dispersion equilibrium differ by individual model.

Note a few characteristics about the equilibrium price-dispersion models. First, consider equilibrium selection. The theory models we study prove the existence of a price-dispersion equilibrium under different setups, but they provide little information about when and how the price-dispersion equilibrium is realized. The second characteristic concerns dynamics and stationarity. Consumer search models, especially sequential search models, by nature entail dynamics. In the context of dynamics, the equilibrium concepts employed throughout this subsection assume stationarity, where the same market (downward-sloping) demand and (upward-sloping) supply functions replicate over time.

Throughout the subsection, we denote $F_p(\cdot)$ by the distribution of prices charged by firms, $F_r(\cdot)$ by the distribution of firms' marginal cost or production, $G_c(\cdot)$ by the distribution of consumers' marginal cost of search, and $G_{p^*}(\cdot)$ by the distribution of consumers' reservation price. Lowercase $f_p(\cdot), f_r(\cdot), g_c(\cdot), g_{p^*}(\cdot)$ denote their densities. All the models studied in this subsection assume the heterogeneity of consumers' search costs, which ensures $G_c(\cdot)$ to be a nondegenerate distribution. The world from which the data are generated is assumed to be in a stationary price-dispersion equilibrium. Therefore, the distributions $F_p(\cdot), F_r(\cdot), G_c(\cdot), G_{p^*}(\cdot)$ are stationary distributions of the respective object in a strict sense.

6.3.2.1 Sequential Search of Consumers, Downward-Sloping Individual Demand, Heterogeneous Marginal Cost of Firms

Consider the sequential search framework. Reinganum (1979) relaxes the unit inelastic demand assumption of consumers and cost-function homogeneity of suppliers. For consumers, Reinganum introduces a downward-sloping individual demand curve when the price is smaller than the reservation price. To this end, AS(1) and AS(2) are replaced with:

AS(R1) Each consumer has a downward-sloping individual demand function given by $q(p)$, where $q(p)$ has the constant own-price elasticity $\varepsilon < -1$.

AS(R2) A continuum of consumers exists with the identical demand function. The measure of consumers in the market is I.

Consumers engage in sequential search. When $p_j < p^*$, the demand function for firm j when it is searched is given by:

$$q(p_j) = \begin{cases} q(p_j) & \text{if } p_j \le p^* \\ 0 & \text{if } p_j > p^* \end{cases}.$$

Because firms are ex-ante homogeneous, the share of consumers who sample each firm is identical for each firm.[20] The equilibrium share of consumers that each firm being sampled is denoted by λ. Accordingly, firms' profits are given by

$$\pi(p_j; r_j) = \begin{cases} \lambda I (p_j - r_j) q(p_j) & \text{if } p_j \le p^* \\ 0 & \text{if } p_j > p^* \end{cases}.$$

Firms' marginal costs are exogenously given, distributed over $[\underline{r}, \bar{r}]$, with cumulative distribution function $F_r(\cdot)$. Each firm maximizes the profit $\pi(p_j; r_j)$ taking other firms' strategy as given.

Reinganum shows the following $F_p(\cdot)$ can be sustained as the price-dispersed equilibrium price distribution:

$$F_p(r) = \begin{cases} F_r\left(r\frac{1+\varepsilon}{\varepsilon}\right) & \text{for } p < p^* \\ 1 & \text{for } p = p^*. \end{cases}$$

The equilibrium range of price is $\left[\underline{r}\left(\frac{\varepsilon}{1+\varepsilon}\right), p^*\right]$. Notice that $F_p(p)$ has a mass point at $p = p^*$, which is problematic for an empirical application. Models that we review below suggest equilibria with nondegenerate price distributions over their supports, and therefore have been employed as bases of recent empirical works.

6.3.2.2 Simultaneous Search with Reservation Price, Homogeneous Marginal Cost of Firms

Burdett and Judd (1983) establish the existence of a price-dispersion equilibrium when homogeneous consumers engage in simultaneous search and consumers have common reservation price p^*. If all the revealed prices are higher than p^*, consumers search again. Building upon Burdett and Judd's price-dispersion equilibrium, Hong and Shum (2006)[21] develop a nonparametric identification result of consumers' search-cost distribution from the observed price-distribution data. Assumptions AS(1)-(5) remain intact, but AS(SIM1)-(SIM2) are modified as follows.

AS(HS1) Consumers choose and commit to the search set $\mathcal{S}_n\ (\subset \mathcal{J})$ ex ante before engaging in search. However, if all the searched prices are higher than the reservation price, consumers search again from the beginning.

AS(HS2) Consumers' marginal cost of search c_n[22] may differ across consumers.

[20] Because $F_p(\cdot)$ that a firm faces is homogeneous across firms.

[21] In the first part of their paper.

[22] The reason why c is indexed by the search size n, not simply i, becomes clear below. This setup in which search costs are heterogeneous across consumers is an extension made by Hong and Shum. Burdett and Judd's original setup is such that consumers' search cost is homogeneous, and they show that a price-dispersion equilibrium, where consumers search up to two

Identification of Search-Cost Distribution from Observed Price Dispersion An econometrician observes $\bar{k}$ distinct prices $\{p_k\}_{k=1}^{\bar{k}}$[23] and its (equilibrium) distribution $\left\{F_p(p_k)\right\}_{k=1}^{\bar{k}}$. The goal of search-cost distribution identification is to find a mapping $\Phi : \left(\left\{p_k, F_p(p_k)\right\}_{k=1}^{\bar{k}}\right) \to \left(\left\{c_n, G_c(c_n)\right\}_{n=1}^{\bar{n}}\right)$, a mapping from the observed (equilibrium) price dispersion to the search-cost distribution. As an identification condition, we assume $\bar{n} \leq \bar{k} - 1$. In practice, during the estimation, $\bar{n}$ is treated as a tuning parameter set a priori by an econometrician, because it is assumed that the search size is not directly observed.

Each consumer, whose optimal search size is n, has consumer-specific heterogeneous search cost c_n,[24] whereas the ex-ante price distribution $F_p(\cdot)$ is common across consumers. Recall equations (6.2.10) and (6.2.11) that specify the (marginal) benefit of search:

$$\begin{aligned} M_n &= -E\left[\min_{1\leq l\leq n}\{p_l\}\right] \\ &= -\int_0^{\bar{p}} \left(1 - F_p(p)\right)^n dp \\ M_{n+1} - M_n &= \int_0^{\bar{p}} \left(1 - F_p(p)\right)^n F_p(p)\, dp \\ &=: \Delta_n. \end{aligned} \tag{6.3.3}$$

We have shown Δ_n is monotonically decreasing over n in section 6.2.1. Furthermore, the optimality condition for the consumers states that

$$\Delta_n = c_n;$$

that is, the indifferent consumers' marginal cost of search equals the marginal benefit. Consumers with $c_i > \Delta_1$ will draw only one price, $\Delta_2 \leq c_i < \Delta_1$ two prices, and so on.[25] Constructing the mapping Φ proceeds in three stages as we describe below.

Stage 1. The first stage is to construct the mapping from the proportion of consumers who searched for n prices $\{\phi_n\}_{n=1}^{\bar{n}}$ to the consumers' search-cost distribution $\{G_c(\Delta_n)\}_{n=1}^{\bar{n}}$. By construction, we have a set of equations as follows:

$$\begin{aligned} \phi_1 &= \Pr(c_i > \Delta_1) = 1 - G_c(\Delta_1) \\ \phi_2 &= \Pr(\Delta_2 \leq c_i < \Delta_1) = G_c(\Delta_1) - G_c(\Delta_2) \\ &\vdots \\ \phi_{\bar{n}-1} &= \Pr(\Delta_{\bar{n}-1} \leq c_i < \Delta_{\bar{n}-2}) = G_c(\Delta_{\bar{n}-2}) - G_c(\Delta_{\bar{n}-1}) \\ \phi_{\bar{n}} &= \Pr(\Delta_{\bar{n}} \leq c_i < \Delta_{\bar{n}-1}) = G_c(\Delta_{\bar{n}-1}) - G_c(\Delta_{\bar{n}}). \end{aligned} \tag{6.3.4}$$

prices, can exist.

[23]Without loss of generality, $\{p_k\}_{k=1}^{\bar{k}}$ is indexed in ascending order $p_1 < p_2 < ... < p_{\bar{k}}$. If a tied price is observed, one can give ϵ-perturbation on either of the observations.

[24]The setup of heterogeneous search costs across consumers is an extension made by Hong and Shum. The original setup of Burdett and Judd's original setup is such that consumers' search cost is homogeneous, and they showed that a price-dispersion equilibrium, where consumers search up to two prices, can exist.

[25]Assuming that either the first search is free or $u_j = y - p_j$ with $y \gg 0$.

ϕ_n is an intermediate parameter that is not directly observable to the econometrician in this setup. An additional step is required to link $\{\phi_n\}_{n=1}^{\bar{n}}$ with $\{p_k, F_p(p_k)\}_{k=1}^{\bar{k}}$.

Stage 2. The second stage is to construct the mapping from $\{p_k, F(p_k)\}_{k=1}^{\bar{k}}$ to $\{\phi_n\}_{n=1}^{\bar{n}}$. To that end, Hong and Shum consider the equilibrium pricing equation of Burdett and Judd (1983). Let r be the firms' common marginal cost. Assume the market is in the price-dispersion equilibrium. Burdett and Judd show that the following pricing equation should hold in the equilibrium:

$$\begin{aligned}(p_{\bar{k}} - r)\,\phi_1 &= (p_k - r)\left[\sum_{n=1}^{\bar{n}} \phi_n n \left(1 - F_p(p_k)\right)^{n-1}\right] \qquad \text{for } k = 1, ..., \bar{k}-1 \\ \phi_{\bar{n}} &= 1 - \sum_{n=1}^{\bar{n}-1} \phi_n, \end{aligned} \tag{6.3.5}$$

where $0 < \phi_1 < 1$ and $p^* > r$. Because we assumed $\bar{n} \leq \bar{k} - 1$, we have $\bar{k}$ equations for $\bar{n}$ unknowns, which allows us to solve for $\left\{r, \{\phi_n\}_{n=1}^{\bar{n}-1}\right\}$. Combining stages 1 and 2, we have a mapping from $\{p_k, F_p(p_k)\}_{k=1}^{\bar{k}}$ to $\{G(\Delta_n)\}_{n=1}^{\bar{n}}$ via $\left\{r, \{\phi_n\}_{n=1}^{\bar{n}-1}\right\}$. The only remaining part that is missing is $\{\Delta_n\}_{n=1}^{\bar{n}}$.

Stage 3. The last stage is to construct $\{c_n\}_{n=1}^{\bar{n}} = \{\Delta_n\}_{n=1}^{\bar{n}}$ from the right-hand side of (6.3.3) and their differences over n. Because the prices and their distribution $F_p(\cdot)$ are observed, the calculation of M_n, the expected minimum price by drawing n prices, is straightforward. Calculating sequentially from $n = 1$ to $\bar{n}$, we have $\{\Delta_n\}_{n=1}^{\bar{n}}$, and thus, $\{c_n\}_{n=1}^{\bar{n}}$.

Estimation of Search-Cost Distribution Let $\hat{F}_p(p) := \frac{1}{\bar{k}}\sum_{k=1}^{\bar{k}} \mathbf{1}(p_k \leq p)$ be the empirical distribution function of the price. We treat $\hat{F}_p(p)$ as a discrete distribution with support $\{p_k\}_{k=1}^{\bar{k}}$. The data consist of $\{p_k, \hat{F}_p(p_k)\}_{k=1}^{\bar{k}}$, where $F_p(p_1) = 0$ and $F_p(p_{\bar{k}}) = 1$. The mapping we just constructed, $\Phi : \left(\{p_k, F_p(p_k)\}_{k=1}^{\bar{k}}\right) \to \left(\{c_n, G_c(c_n)\}_{n=1}^{\bar{n}}\right)$, has $\bar{k}$ observations[26] for $\bar{n}+1$ unknowns, where $\bar{k} \geq \bar{n}+1$ by construction. The estimation procedure proceeds backward from Stage 3 to 1.[27]

For Stage 3, replacing $F_p(p)$ with its empirical distribution $\hat{F}_p(p_k)$ and plugging back in to (6.3.3) yields, for $n = 1, 2, ..., \bar{n}$,

$$\begin{aligned}\hat{\Delta}_n &= \sum_{j=1}^{\bar{k}} \left\{1 - \frac{1}{\bar{k}}\sum_{k=1}^{\bar{k}} \mathbf{1}(p_k \leq p_j)\right\}^n \left\{\frac{1}{\bar{k}}\sum_{k=1}^{\bar{k}} \mathbf{1}(p_k \leq p_j)\right\} \\ &=: \hat{c}_n.\end{aligned}$$

We estimate $(\hat{r}, \hat{\phi}_1, ..., \hat{\phi}_{\bar{n}})$, which corresponds to Stage 2. Abbreviate $\hat{\boldsymbol{\theta}} := (\hat{r}, \hat{\phi}_1, ..., \hat{\phi}_{\bar{n}})$, and write the

[26]One observation per one equation.
[27]For the asymptotics, we assumet $\bar{k}$ grows to infinity while $\bar{n}$ is fixed.

empirical counterpart of (6.3.5) as

$$(p_{\bar{k}} - r)\hat{\phi}_1 = (p_k - r)\left[\sum_{n=1}^{\bar{n}} \hat{\phi}_n n \left(1 - \left[\frac{1}{\bar{k}} \sum_{k=1}^{\bar{k}} \mathbf{1}(p_k \leq p_j)\right]\right)^{n-1}\right] \qquad \text{for } k = 1, ..., \bar{k} - 1 \tag{6.3.6}$$

$$\hat{\phi}_{\bar{n}} = 1 - \sum_{n=1}^{\bar{n}-1} \hat{\phi}_n. \tag{6.3.7}$$

By using the fact that $F_p(p_1) = \hat{F}_p(p_1) = 0$, evaluate (6.3.6) at p_1 to obtain

$$r = \frac{p_1\left(\sum_{n=1}^{\bar{n}} \hat{\phi}_n n\right) - p_{\bar{k}}\hat{\phi}_1}{\left(\sum_{n=1}^{\bar{n}} \hat{\phi}_n n\right) - \hat{\phi}_1}. \tag{6.3.8}$$

Plugging (6.3.8) back into each k of (6.3.6) eliminates firms' common marginal cost r. Then, we can use (6.3.6) and (6.3.7) as a moment condition, to proceed with the GMM or empirical likelihood. Hong and Shum use the empirical likelihood and provide the corresponding asymptotics.

The remaining step of estimation, which corresponds to Stage 1, is to use the one-to-one mapping (6.3.4) sequentially from top to bottom, which maps $\{\hat{\phi}_n\}_{n=1}^{\bar{n}}$ to $\{\hat{G}_c(\hat{c}_n)\}_{n=1}^{\bar{n}}$.

6.3.2.3 Sequential Search of Consumers, Homogeneous Marginal Cost of Firms

Identification Argument of Consumers' Reservation-Price Distribution from Observed Price Dispersion Rob (1985) establishes price-dispersion equilibrium by employing Nash-Stackelberg equilibrium concept between a continuum of profit-maximizing firms with zero marginal cost[28] when consumers search sequentially. Hong and Shum (2006)[29] establish an argument to identify the search-cost distribution, building upon Rob's setup. AS(SEQ2) is modified as

AS(RS2) Consumers' constant discount factor is $\delta\ (< 1)$ and the constant marginal cost of searching for one more price is $c_k\ (> 0)$. c_k follows a nondegenerate distribution $G_c(\cdot)$ that has a density $g_c(\cdot)$.

Note the only difference from section 6.2.1.1 other than introducing the search-cost distribution $G_c(\cdot)$ is homogeneity of the discount factor δ. All the results and implications from section 6.2.1.1 carry over. In particular, recall the indifference equation that determines consumers' reservation price:

$$c_k = p_k^*(1-\delta) + \delta \int_0^{p_k^*} (p_k^* - p)\, dF_p(p). \tag{6.3.9}$$

Note the reservation price p_k^* is indexed by k because it is determined as an implicit function of c_k. We showed in section 6.2.1 that for each given $c_k > 0$, (6.3.9) has a unique solution p_k^*. Furthermore, p_k^* is a (strictly) increasing function of c_k provided that $F_p(\cdot)$ is well behaving. This construes a one-to-one mapping from consumers' reservation price to consumers' marginal cost of search, and therefore, the

[28] It can be easily extended to a positive, constant-marginal-cost case.

[29] In the second part of their paper.

remaining task is to construct a mapping between the distribution of the observed prices $\{p_k\}_{k=1}^{\bar{k}}$ and consumers' reservation price $\{p_k^*\}_{k=1}^{\bar{k}}$.

Let $[p_1, p_{\bar{k}}]$ be the support of $F_p(\cdot)$, and let $\{p_k\}_{k=1}^{\bar{k}}$ be the observed price data.[30] It is immediate that $F_p(p_1) = 0$ and $F_p(p_{\bar{k}}) = 1$. Let r be the firms' common, constant marginal cost. In equilibrium, each participating firm in the market must not have an incentive to deviate to charge a different price. Rob's key equilibrium assumption is the nonnegative equal-profit condition over the support of the equilibrium price distribution $F_p(\cdot)$. That is, for any observed price p_k ($\neq p_{\bar{k}}$),

$$(p_{\bar{k}} - r)\, q(p_{\bar{k}}) = (p_k - r)\, q(p_k) \tag{6.3.10}$$

should hold, which gives us $\bar{k} - 1$ effective equations for $k = 1, ..., \bar{k} - 1$. Furthermore, using the fact that consumers will employ the reservation-price strategy in sequential search, the demand function $q(p_{\bar{k}})$ is defined such that

$$\begin{aligned} q(p) &= I \Pr(p^* > p) \\ &= I\left(1 - G_{p^*}(p)\right), \end{aligned} \tag{6.3.11}$$

where I is the measure of consumers. Plugging (6.3.10) back, (6.3.10) can be transformed to the form of consumers' reservation-price distribution $G_{p^*}(\cdot)$ as follows:

$$(p_{\bar{k}} - r)\left(1 - G_{p^*}(p_{\bar{k}})\right) = (p_k - r)\left(1 - G_{p^*}(p_k)\right). \tag{6.3.12}$$

Parametric Estimation of Search-Cost Distribution $\{G_{p^*}(p_k)\}_{k=1}^{\bar{k}}$, which is the distribution of the reservation price of the consumers, is an intermediate object that will be eliminated eventually. A non-parametric estimation as in section 6.3.2.2, however, is not feasible here, because we have only $\bar{k} - 1$ equations for $\bar{k}$ objects.[31] Therefore, Hong and Shum proceed with parametrizing $G_c(\cdot)$ as a gamma distribution to derive the likelihood and construct a mapping between $G_c(\cdot)$ and $G_{p^*}(\cdot)$. Let $\boldsymbol{\theta}$ be the finite-dimensional parameter vector that dictates the behavior of the distribution $G_c(\cdot;\boldsymbol{\theta})$. The goal is to write down the likelihood of observing $\{p_k\}_{k=1}^{\bar{k}}$ in terms of $g_c(\cdot;\boldsymbol{\theta})$.

Determination of α We solve for $\alpha := 1 - G_{p^*}(p_{\bar{k}})$ first. As an initial condition, we impose $\Pr(p^* \leq p_1) = G_{p^*}(p_1) = 0$, implying that

$$\begin{aligned} 1 - G_{p^*}(p_{\bar{k}}) &= \frac{(p_1 - r)}{(p_{\bar{k}} - r)} \\ &= 1 - \Pr(p^* \leq p_{\bar{k}}) \\ &=: \alpha. \end{aligned} \tag{6.3.13}$$

[30] In the corner case when the c_k is too large, the reservation price is taken to be $\min\{p_k^*, p_{\bar{k}}\}$.

[31] Even though we set $G_{p^*}(p_1) = 0$ so we have $\bar{k} - 1$ effective equations for $\bar{k} - 1$ objects, it is still insufficient, because the number of objects to be estimated grows at the same rate as the number of equations. For asymptotics, we need the number of equations to grow at least a faster rate than the number of parameters to be estimated.

Rearranging, we have

$$\alpha \left(p_{\bar{k}} - r\right) = p_1 - r. \tag{6.3.14}$$

Plugging (6.3.13) back into (6.3.12) and rearranging, we have

$$G_{p^*}\left(p_k\right) = 1 - \alpha \frac{\left(p_k - r\right)}{\left(p_{\bar{k}} - r\right)}. \tag{6.3.15}$$

Deriving the Expression of $F_p\left(\cdot;\boldsymbol{\theta}\right)$ in Terms of $g_c\left(\cdot;\boldsymbol{\theta}\right)$ Denoting $\tau_k := G_{p^*}\left(p_k\right)$ by the the quantile of p_k. (6.3.12) becomes

$$\left(p_{\bar{k}} - r\right)\alpha = \left(G_{p^*}^{-1}\left(\tau_k\right) - r\right)\left(1 - \tau_k\right),$$

which yields

$$G_{p^*}^{-1}\left(\tau_k\right) = \alpha \frac{\left(p_{\bar{k}} - r\right)}{\left(1 - \tau_k\right)} + r \tag{6.3.16}$$

for $k = 1, ..., \bar{k} - 1$. Then, by the reservation-price equation,

$$G_c^{-1}\left(\tau_k;\boldsymbol{\theta}\right) = \int_{p_{k_1}}^{G_{p^*}^{-1}\left(\tau_k\right)} F_p\left(p\right) dp.$$

Differentiating with respect to τ_k yields

$$\frac{\partial G_c^{-1}\left(\tau_k;\boldsymbol{\theta}\right)}{\partial \tau_k} = \alpha \frac{\left(p_{\bar{k}} - r\right)}{\left(1 - \tau_k\right)^2} F_p\left(\alpha \frac{\left(p_{\bar{k}} - r\right)}{\left(1 - \tau_k\right)} + r\right). \tag{6.3.17}$$

Let $\iota\left(\tau_k;\boldsymbol{\theta}\right)$ be the τ_k'th quantile of $G_c\left(\cdot;\boldsymbol{\theta}\right)$ so that $G_c\left(\iota\left(\tau_k;\boldsymbol{\theta}\right);\boldsymbol{\theta}\right) = \tau_k$. By the inverse function theorem,

$$\frac{\partial G_c^{-1}\left(\tau_k;\boldsymbol{\theta}\right)}{\partial \tau_k} = \frac{1}{g_c\left(\iota\left(\tau_k;\boldsymbol{\theta}\right);\boldsymbol{\theta}\right)}. \tag{6.3.18}$$

Combining 6.3.17 and 6.3.18 yields

$$\frac{1}{g_c\left(\iota\left(\tau_k;\boldsymbol{\theta}\right);\boldsymbol{\theta}\right)} = \alpha \frac{\left(p_{\bar{k}} - r\right)}{\left(1 - \tau_k\right)^2} F_p\left(\alpha \frac{\left(p_{\bar{k}} - r\right)}{\left(1 - \tau_k\right)} + r\right).$$

Substituting 6.3.15 and 6.3.16 to eliminate $\tau_k \equiv G_{p^*}\left(p_k\right)$, we finally have

$$F_p\left(p_k;\boldsymbol{\theta}\right) = \frac{\alpha\left(p_{\bar{k}} - r\right)}{\left(p_k - r\right)^2 g_c\left(\iota\left(1 - \alpha\frac{\left(p_k - r\right)}{\left(p_{\bar{k}} - r\right)};\boldsymbol{\theta}\right);\boldsymbol{\theta}\right)}. \tag{6.3.19}$$

Notice that $F_p\left(p_k;\boldsymbol{\theta}\right)$ inherits $\boldsymbol{\theta}$ from $g_c\left(\cdot;\boldsymbol{\theta}\right)$.

Determination of r Because $F_p(p_{\bar{k}};\boldsymbol{\theta}) = 1$ has to be satisfied, combining (6.3.14) with (6.3.19), we get

$$1 \equiv F_p(p_{\bar{k}};\boldsymbol{\theta}) = \frac{(p_1 - r)}{(p_{\bar{k}} - r)^2 g_c\left(\iota\left(1 - \frac{(p_1 - r)}{(p_{\bar{k}} - r)};\boldsymbol{\theta}\right);\boldsymbol{\theta}\right)},$$

which determines r.

Parametrization and MLE Estimation Lastly, taking the first-order derivative on (6.3.19) with respect to p_k gives $f_p(p_k;\boldsymbol{\theta})$. The likelihood of observing the entire price data is simply

$$\prod_{k=1}^{\bar{k}} f_p(p_k;\boldsymbol{\theta}).$$

By parametrizing $g_c(\cdot;\boldsymbol{\theta})$ as a gamma density, plugging back to (6.3.19) and running the maximum likelihood estimation will give a consistent estimate for $\boldsymbol{\theta}$. Hong and Shum note they employ a gamma distribution due to restrictions that Rob's equilibrium condition implies.[32]

6.3.2.4 Sequential Search of Consumers, Heterogeneous Marginal Cost of Firms

Carlson and McAfee (1983) derive equilibrium price dispersion from consumers searching sequentially without recall. Consumers have heterogeneous search costs and firms have heterogeneous marginal costs of production. Based on Carlson and McAfee, Hortaçsu and Syverson (2004) develop the identification results and estimation method of consumers' search-cost distribution. One key advance Hortaçsu and Syverson made is that they utilize the quantity data $\{q_k\}_{k=1}^{\bar{k}}$ in addition to the price-distribution data for identification. Hortaçsu and Syverson's setup is the utility search with linear utility specification, whereas we stick to the price-search setup in the following illustration for simplicity.

Carlson and McAfee; Hortaçsu and Syverson assume the price distribution has the discrete support $\{p_k\}_{k=1}^{\bar{k}}$.[33] The price distribution $F_p(\cdot)$ can be summarized as a probability mass function that has the following form:

$$f_p(p) = \begin{cases} \rho_k & p = p_1, ..., p_{\bar{k}} \\ 0 & \text{otherwise} \end{cases}, \tag{6.3.20}$$

with $\rho_k > 0$ and $\sum_{k=1}^{\bar{k}} \rho_k = 1$. The price distribution $\{p_k, \rho_k\}_{k=1}^{\bar{k}}$ is taken to be the prior knowledge of the consumers. ρ_k is interpreted as the ex-ante probability of consumers' sample price p_k.

Derivation of Downward-Sloping Market Demand from Search-Cost Distribution

[32] Namely, the density has to be decreasing over the support.

[33] As before, p_k is ordered in a way that $p_1 < p_2 < ... < p_{\bar{k}}$ without loss of generality.

Identification of Cutoff Search Costs Recall the consumers' problem (6.2.6) from section 6.2.1.2, where we assume the discount factor $\delta = 1$, and consumer i already has p_k in her hand:

$$\min\left\{p_k, c_i + \int_0^{p_k} p dF_p(p)\right\} = \min\left\{\int_0^{p_k}(p_k - p)\, dF_p(p), c_i\right\}.$$

The indifference condition of this problem states that consumers should stop searching if the marginal cost of searching exceeds the marginal benefit, that is,

$$\begin{aligned} c_i &> \int_0^{p_k}(p_k - p)\, dF_p(p) \\ &=: \psi_k. \end{aligned} \tag{6.3.21}$$

By combining (6.3.21) with (6.3.20), the discrete version of (6.3.21) applies. The marginal benefit of searching for one more price ψ_k is

$$\psi_k = \sum_{k'=1}^{k} \rho_k (p_k - p_{k'}). \tag{6.3.22}$$

ψ_k is simply the sum of the marginal benefit of drawing $p_{k'}$ for each $k' \le k$, weighted by its sampling probability $\rho_{k'}$. It is straightforward that $\psi_1 = 0$,[34] and $\{\psi_k\}_{k=2}^{\bar{k}}$ can be identified given that $\{\rho_k\}_{k=1}^{\bar{k}}$ can be either parametrically or nonparametrically estimated.[35] This procedure identifies the cutoff value of $\{\psi_k\}_{k=1}^{\bar{k}}$, and thus that of c_i.

Identification of Downward-Sloping Market Demand Function from Search-Cost Distribution Individual consumer's reservation price p_i^* is a monotonically increasing implicit function of c_i.[36] Furthermore, a consumer will buy from anywhere provided that the visited firm offers a price less than her reservation price p_i^*. Thus, the cutoff rule gives a downward-sloping demand function.

Our goal now is to write down the demand function for each firm k, $\{q_k\}_{k=1}^{\bar{k}}$, in terms of consumers' search-cost distribution $G_c(\cdot)$. We begin from the most expensive product $\bar{k}$. For product $\bar{k}$, only the consumer with $c_i \ge \psi_{\bar{k}}$ will buy. Demand for firm $\bar{k}$ is therefore the measure of consumers that have a search cost higher than $\psi_{\bar{k}}$ multiplied by its sampling probability $\rho_{\bar{k}}$:

$$q_{\bar{k}} = I\rho_{\bar{k}}\left(1 - G_c\left(\psi_{\bar{k}}\right)\right),$$

[34] Intuitively if a consumer already knows she has the lowest possible price, searching for one more price provides no benefit.

[35] Hortaçsu and Syverson (2004) assume the following functional form during the estimation:

$$\rho_j = \frac{z_j^\alpha}{\sum_{k=1}^{\bar{k}} z_k^\alpha},$$

where z_j is an index of product-level observables that influence the probability a firm is sampled.

[36] Recall Figure 6.2.2 of Section 6.2.1.1.

where I is the mass of consumers. Demand for product $\bar{k}-1$ is composed of two terms as follows:

$$\begin{aligned} q_{\bar{k}-1} &= I\left\{\rho_{\bar{k}-1}\left(1-G_c\left(\psi_{\bar{k}}\right)\right)+\frac{\rho_{\bar{k}-1}}{1-\rho_{\bar{k}}}\left(G_c\left(\psi_{\bar{k}}\right)-G_c\left(\psi_{\bar{k}-1}\right)\right)\right\} \\ &= I\rho_{\bar{k}-1}\left\{1+\frac{\rho_{\bar{k}}}{1-\rho_{\bar{k}}}G_c\left(\psi_{\bar{k}}\right)-\frac{1}{1-\rho_{\bar{k}}}G_c\left(\psi_{\bar{k}-1}\right)\right\}. \end{aligned} \tag{6.3.23}$$

The first term of (6.3.23), $\rho_{\bar{k}-1}\left(1-G_c\left(\psi_{\bar{k}}\right)\right)$, is the probability that consumers with search cost higher than $\psi_{\bar{k}}$ has sampled $\bar{k}-1$ times. The second term is the probability that consumers with search cost between $\psi_{\bar{k}-1}$ and $\psi_{\bar{k}}$ has sampled $\bar{k}-1$ times, after sampling only $\bar{k}$ in the previous periods. Notice the stationary probability of drawing $\bar{k}$ in the previous periods, $\left(1-\rho_{\bar{k}}\right)^{-1}=1+\rho_{\bar{k}}+\rho_{\bar{k}}^2+...$, is multiplied by in the second term. Analogous calculations yield the expression for q_j for $1\leq j\leq \bar{k}-2$:

$$\begin{aligned} q_j &= I\rho_j\left\{\left(1-G_c\left(\psi_{\bar{k}}\right)\right)+\frac{1}{1-\rho_{\bar{k}}}\left(G_c\left(\psi_{\bar{k}}\right)-G_c\left(\psi_{\bar{k}-1}\right)\right)+...+\frac{1}{1-\rho_{\bar{k}}-...\rho_{j+1}}\left(G_c\left(\psi_{j+1}\right)-G_c\left(\psi_j\right)\right)\right\} \\ &= I\rho_j\left\{1+\frac{\rho_{\bar{k}}}{1-\rho_{\bar{k}}}G_c\left(\psi_{\bar{k}}\right)+\frac{\rho_{\bar{k}-1}}{\left(1-\rho_{\bar{k}}\right)\left(1-\rho_{\bar{k}}-\rho_{\bar{k}-1}\right)}G_c\left(\psi_{\bar{k}-1}\right)\right. \\ &\quad \left.+\sum_{k=\bar{k}-2}^{j+1}\frac{\rho_k}{\left(1-\rho_{\bar{k}}-...-\rho_{k+1}\right)\left(1-\rho_{\bar{k}}-...-\rho_k\right)}G_c\left(\psi_k\right)-\frac{1}{\left(1-\rho_{\bar{k}}-...-\rho_{j+1}\right)}G_c\left(\psi_j\right)\right\}. \end{aligned} \tag{6.3.24}$$

It can be easily shown that $\frac{dq_j}{dp_j}<0$.

We assume an econometrician has the market share data $\{q_k\}_{k=1}^{\bar{k}}$, and therefore, $I=1$. The system (6.3.24) then gives $\bar{k}-1$ equations for $\bar{k}-1$ objects, namely, $\{G_c\left(\psi_k\right)\}_{k=2}^{\bar{k}}$. Combined with the fact that search cost cannot be negative, $G_c\left(\psi_1\right)=G_c\left(0\right)=0$, we have all the necessary cutoff points for the search-cost distribution. Another possibility to proceed is to parametrize the distribution $G_c\left(\cdot;\boldsymbol{\theta}\right)$ to a family of distributions and estimate $\boldsymbol{\theta}$ that governs $G_c\left(\cdot;\boldsymbol{\theta}\right)$.

Supply Side and Equilibrium One of the sufficient condition, although it is not exhaustive, for the existence of the persistent price dispersion of a homogeneous product from Carlson and McAfee is that $g_c\left(\cdot\right)$ is nondecreasing over its support. Hortaçsu and Syverson do not impose or test the condition, because their setup is the utility search in a differentiated-products market. In the context of utility search with heterogeneous alternatives, the existence of the persistent price dispersion is trivial. Instead, Hortaçsu and Syverson utilized additional conditions from the suppliers' pricing equation to estimate the levels of search cost, and then to derive the slope of the demand function, $\frac{\partial q_j}{\partial p_j}$. We omit the part of suppliers' pricing equation in estimating the level of search costs in the context of utility search.

6.4 Empirical Frameworks with Search-Set Data or Its Proxy

In this section, we study empirical frameworks of consumer search that make use of more granular data than the market-level price-quantity data. The goal is to write down the likelihood under the relevant search assumptions with the available data. Ideally, an econometrician should be able to observe (i) all the utility orderings of the alternatives, (ii) search sequence, and (iii) the choice. Because observing all the components of this "ideal" data is virtually impossible, we study how the lack of some component is reflected in writing down the likelihood.

The models studied in this section are based on the utility search in a differentiated-products market. Nevertheless, all the results and implications of the optimal search rules on price search apply, which we studied in section 6.2.1.

6.4.1 Empirical Frameworks of Sequential Search

Kim et al. (2010, 2017) develop an empirical framework of consumer search based on Weitzman (1979)'s theoretic setup on sequential search. Consumers search sequentially on heterogeneous alternatives with recall, to find the alternative with the highest utility among the choice set $\mathcal{J}_i$.

6.4.1.1 Setup and Basic Results

Let the alternative-specific utility of consumer i choosing $j \in \mathcal{J}_i$ be

$$\begin{aligned} u_{i,j} &:= \bar{u}_{i,j}\left(\boldsymbol{\beta}_i, \alpha_i\right) + \epsilon_{i,j} \\ &= \mathbf{x}_j' \boldsymbol{\beta}_i - \alpha_i p_j + \epsilon_{i,j}, \end{aligned} \tag{6.4.1}$$

where the parameter vectors are assumed to follow

$$\begin{aligned} &\boldsymbol{\beta}_i \sim \mathcal{N}\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\beta}\right) \\ &\ln\left(\alpha_i\right) \sim \mathcal{N}\left(\alpha, \sigma_{\alpha}^2\right) \\ &\epsilon_{i,j} \sim \mathcal{N}\left(0, \sigma_{i,j}^2\right). \end{aligned} \tag{6.4.2}$$

In the original setup of Kim et al. (2010), each consumer knows her own $(\boldsymbol{\beta}_i, \alpha_i)$, and the goal of search is to resolve the uncertainty of $\epsilon_{i,j}$. If alternative i is searched, no uncertainty remains in the utility term $u_{i,j}$. The econometrician, however, does not observe $(\boldsymbol{\beta}_i, \alpha_i)$, and estimating them is part of the goal of estimation. $\sigma_{i,j}$ can be heterogeneous across individuals, which reflects the difference between "more knowledgable" and "less knowledgable" consumers and/or alternatives. When $\sigma_{i,j}$ is heterogeneous across individuals or alternatives, it will be further parametrized as following some parametric distribution.

The baseline assumption is that consumers search sequentially over heterogeneous alternatives. Weitzman (1979)'s rules, which we studied in 6.2.1.3, apply here. The only difference is that now consumers try to find an alternative that yields the maximum utility not the minimum prices, but the principle of the rules are the same. For each alternative $k \in \mathcal{J}_i$, the reservation utility $u_{i,k}^*$ can be calculated implicitly by the following equation:

$$c_{i,k} = \int_{u_{i,k}^*}^{\bar{u}_{i,k}} \left(u_{i,k} - u_{i,k}^*\right) dF_{\epsilon_{i,j}}\left(\epsilon_{i,j}\right).$$

Let us first sort $\mathcal{J}_i$ in the descending order of $u^*_{i,k}$ (i.e., higher reservation utility has lower index) without losing generality. Recall the rule proceeds in three stages: (i) (Ordering) Sort the alternatives in the descending order of the reservation utility $u^*_{i,k}$ and search sequentially from the order made in the previous step. (ii) (Stopping) At each stage $t \in \mathcal{J}_i$, compare $\max_{1\le k\le t}\{u_{i,k}\}$ with $u^*_{i,t+1}$ and stop if $\max_{1\le k\le t}\{u_{i,k}\} > u^*_{i,t+1}$, and define $\mathcal{S}_i\,(\subset \mathcal{J}_i)$[37] as the searched set thus far. (iii) (Choice) If stopped, collect the alternative j such that $j = \arg\max_{k\in\mathcal{S}_i}\{u_{i,k}\}$.

The individual-product-specific search cost $c_{i,j}$ is further parametrized as follows:

$$\begin{aligned} c_{i,j} &\propto \exp\left(\mathbf{w}'_{i,j}\gamma_i\right) \\ \gamma_i &\sim \mathcal{N}\left(\gamma, \boldsymbol{\Sigma}_\gamma\right). \end{aligned} \tag{6.4.3}$$

$\mathbf{w}_{i,j}$ is the individual and/or product-specific search-cost shifter,[38] which contains 1, and therefore the coefficient of 1 will represent the mean search cost over individuals. Under the model assumptions with given parametrizations (6.4.1), (6.4.2), and (6.4.3), it can be shown that the reservation utility has the following closed-form solution:

$$\begin{aligned} u^*_{i,j} &= \bar{u}_{i,j}\left(\boldsymbol{\beta}_i, \alpha_i\right) + \sigma_{i,j}\zeta\left(\frac{c_{i,j}\left(\gamma_i\right)}{\sigma_{i,j}}\right) \\ &= \mathbf{x}'_j\boldsymbol{\beta}_i - \alpha_i p_j + \sigma_{i,j}\zeta\left(\frac{c_{i,j}\left(\gamma_i\right)}{\sigma_{i,j}}\right), \end{aligned}$$

where $\zeta(t)$ is a function defined implicitly by the following equation:

$$t = \left(1 - \Phi(\zeta)\right)\left(\frac{\phi(\zeta)}{1-\Phi(\zeta)} - \zeta\right).$$

A straightforward application of the implicit function theorem shows $\zeta(t)$ is a monotone function, strictly decreasing over its domain.

Let $\boldsymbol{\theta} := \left(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \alpha, \sigma_\alpha, \gamma, \boldsymbol{\Sigma}_\gamma, \left\{\sigma_{i,j}\right\}_{i\in\mathcal{I}, j\in\mathcal{J}_i}\right)$ be the vector of model parameters. The goal of the estimation is now to estimate $\boldsymbol{\theta}$ consistently with the given data. Availability of data items substantially restricts the path that an econometrician can take, and we study a few paths that can be taken or have already been taken in the literature in what follows.

6.4.1.2 Ideal-Case Likelihood When Individual-level Search Sequence and Choice Are Observed

Let us consider an ideal-case likelihood of observing the order of search set $\mathcal{S}_i = \{1, 2, ..., S_i\}$, out of the available alternatives $\mathcal{J}_i = \{1, 2, ..., S_i, S_i + 1, ..., J_i\}$. Again, without loss of generality, $\mathcal{S}_i$ and $\mathcal{J}_i$ are ordered according to the reservation utility so the consumer searches from index 1 first and then sequentially moves on to the next index, stops at S_i, and collects j.

[37] Notice $\mathcal{S}_i$ has the same order as $\mathcal{J}_i$.

[38] Ursu (Forthcoming), notably, uses the randomized ranking of the alternatives in $\mathbf{w}_{i,j}$. By utilizing the experimental variation on the ranking determination, she removes the endogeneity when the ranking is determined.

The unconditional probability of observing the $\mathcal{J}_i$ and the choice $j \in \mathcal{S}_i$ is such that

$$\begin{aligned}
&\Pr\left(\{j \in \mathcal{S}_i \text{ Chosen}\} \cap \{\text{Stopped at } S_i\} \cap \{\mathcal{J}_i \text{ Ordered}\}; \boldsymbol{\theta}\right) \\
=& \Pr\left(j \in \mathcal{S}_i \text{ Chosen} | \{\text{Stopped at } S_i\} \cap \{\mathcal{J}_i \text{ Ordered}\}; \boldsymbol{\theta}\right) \\
&\times \Pr\left(\text{Stopped at } S_i | \mathcal{J}_i \text{ Ordered}; \boldsymbol{\theta}\right) \\
&\times \Pr\left(\mathcal{J}_i \text{ Ordered}; \boldsymbol{\theta}\right). \qquad (6.4.4)
\end{aligned}$$

Here, "$\mathcal{J}_i$ Ordered" is shorthand for $\mathcal{J}_i$ being ordered as $\{1, 2, ..., S_i, S_i + 1, ..., J_i\}$. The equality (6.4.4) is obtained simply by repeatedly applying the law of total probability.

Ordering Rule In this ideal case, the search sequence of each consumer $\mathcal{S}_i = \{1, 2, ..., S_i\}$ is observed, with the actual choice $j \in \mathcal{S}_i$. The amount of actual content of this S_i-tuple is $\frac{1}{2}S_i(S_i - 1)$, because it contains all the pairwise comparisons. Furthermore, $S_i(J_i - S_i)$ observations exist because for each $k \in \mathcal{S}_i$ and $m \in \mathcal{J}_i \backslash \mathcal{S}_i$, the ordering should be such that $k < m$.

According to Weitzman's ordering rule, we have the following three restrictions (i)-(iii). Each restriction is for $\mathcal{S}_i$, between $\mathcal{S}_i$ and $\mathcal{J}_i \backslash \mathcal{S}_i$, and for $\mathcal{J}_i$, respectively. The restrictions are as follows. (i) For each $k, l \in \mathcal{S}_i$ such that $k < l$, $u^*_{i,k} > u^*_{i,l}$ holds. (ii) For each $k \in \mathcal{S}_i$, $m \in \mathcal{J}_i \backslash \mathcal{S}_i$, $u^*_{i,k} > u^*_{i,m}$ holds. (iii) For each $m, n \in \mathcal{J}_i \backslash \mathcal{S}_i$ such that $m < n$, $u^*_{i,m} > u^*_{i,n}$ holds.

The ordering problem is deterministic for the consumers, but it is stochastic for the econometrician. From the econometrician's perspective, the source of stochasticity is the uncertainty in $(\boldsymbol{\beta}_i, \alpha_i, \gamma_i, \sigma_{i,j})$. The ideal-case likelihood of observing an ordered sequence $\mathcal{J}_i$, which is infeasible to implement because the order of $\mathcal{J}_i \backslash \mathcal{S}_i$ is never observed, is

$$\begin{aligned}
\Pr\left(\mathcal{J}_i \text{ Ordered}; \boldsymbol{\theta}\right) = &\left\{ \prod_{\{k,l \in \mathcal{S}_i : k<l\}} \Pr\left(u^*_{i,k} - u^*_{i,l} > 0; \boldsymbol{\theta}\right) \right\} \\
&\times \left\{ \prod_{\{k \in \mathcal{S}_i, m \in \mathcal{J}_i \backslash \mathcal{S}_i\}} \Pr\left(u^*_{i,k} - u^*_{i,m} > 0; \boldsymbol{\theta}\right) \right\} \\
&\times \left\{ \prod_{\{m,n \in \mathcal{J}_i \backslash \mathcal{S}_i : m<n\}} \Pr\left(u^*_{i,m} - u^*_{i,n} > 0; \boldsymbol{\theta}\right) \right\}.
\end{aligned}$$

Instead, in an ideal dataset, the search order within $\mathcal{S}_i$ can be observed. Observing $\mathcal{S}_i$ out of $\mathcal{J}_i$ is

$$\begin{aligned}
\Pr\left(\mathcal{S}_i \left(\subset \mathcal{J}_i\right) \text{ Ordered}; \boldsymbol{\theta}\right) = &\left\{ \prod_{\{k,l \in \mathcal{S}_i : k<l\}} \Pr\left(u^*_{i,k} - u^*_{i,l} > 0; \boldsymbol{\theta}\right) \right\} \\
&\times \left\{ \prod_{\{k \in \mathcal{S}_i, m \in \mathcal{J}_i \backslash \mathcal{S}_i\}} \Pr\left(u^*_{i,k} - u^*_{i,m} > 0; \boldsymbol{\theta}\right) \right\}. \qquad (6.4.5)
\end{aligned}$$

To back out $\Pr\left(\mathcal{J}_i \text{ Ordered}; \boldsymbol{\theta}\right)$ using $\Pr\left(\mathcal{S}_i \left(\subset \mathcal{J}_i\right) \text{ Ordered}; \boldsymbol{\theta}\right)$, in principle, one can construct all the possible orders of $\mathcal{J}_i \backslash \mathcal{S}_i$ and then use the probability-weighted sum to derive the expression $\Pr\left(\mathcal{J}_i \text{ Ordered}; \boldsymbol{\theta}\right)$

from only $\Pr\left(\mathcal{S}_i \left(\subset \mathcal{J}_i\right) \text{ Ordered}; \boldsymbol{\theta}\right)$. In practice, calculating all the permutation orders can be further simplified by using the simulation methods.

Stopping Rule If the consumer decided to stop at S_i, the following two conditions must hold: (i) $\max_{1\leq k\leq t}\{u_{i,k}\} \leq u^*_{i,t+1}$ for all the previous periods $t\,(< S_i)$, and (ii) $\max_{1\leq k\leq S_i}\{u_{i,k}\} \geq u^*_{i,S_i+1}$ for the period S_i. Combined with the fact $u_{i,j} = \max_{1\leq k\leq S_i}\{u_{i,k}\}$, the (infeasible) conditional probability of the consumer stopped at S_i conditional on the order $\mathcal{J}_i$ is

$$\Pr\left(\text{Stopped at } S_i | \mathcal{J}_i \text{ Ordered}; \boldsymbol{\theta}\right) = \Pr\left(\left\{\bigcap_{t=1}^{S_i-1}\left\{\max_{1\leq k\leq t}\{u_{i,k}\} < u^*_{i,t+1}\right\}\right\} \cap \left\{u_{i,j} \geq u^*_{i,S_i+1}\right\}; \boldsymbol{\theta}\right).$$

However, because the order of $\mathcal{J}_i \backslash \mathcal{S}_i$ is never observed, any alternative in $\mathcal{J}_i \backslash \mathcal{S}_i$ can be in $S_i + 1$'th reservation utility order. Thus, the feasible version would be

$$\begin{aligned}
&\Pr\left(\text{Stopped at } S_i | \mathcal{J}_i \text{ Ordered}; \boldsymbol{\theta}\right)\\
&= \Pr\left(\left\{\bigcap_{t=1}^{S_i-1}\left\{\max_{1\leq k\leq t}\{u_{i,k}\} < u^*_{i,t+1}\right\}\right\} \cap \left\{\max_{1\leq k\leq S_i}\{u_{i,k}\} \geq u^*_{i,S_i+1}\right\}; \boldsymbol{\theta}\right)\\
&= \sum_{l\in\{\mathcal{J}_i\backslash\mathcal{S}_i\}} \Pr\left(u_{i,l} = \max_{S_i+1\leq m\leq J_i}\{u_{i,m}\}; \boldsymbol{\theta}\right) \Pr\left(\left\{\bigcap_{t=1}^{S_i-1}\left\{\max_{1\leq k\leq t}\{u_{i,k}\} < u^*_{i,t+1}\right\}\right\} \cap \left\{u_{i,j} \geq u^*_{i,l}\right\}; \boldsymbol{\theta}\right).
\end{aligned} \tag{6.4.6}$$

In practice, one can use simulation methods to calculate the probability $\Pr\left(u_{i,l} = \max_{S_i+1\leq m\leq J_i}\{u_{i,m}\}\right)$.

Choice Rule When the searched set $\mathcal{S}_i$ and the chosen alternative $j \in \mathcal{S}_i$ are observed, the choice rule states that for the chosen item j,

$$u_{i,j} = \max_{k\in\mathcal{S}_i}\{u_{i,k}\}$$

should hold.

The conditional choice probability of j out of $\mathcal{S}_i$, conditioned on the two events $\{\text{Stopped at } S_i\}$ and $\{\mathcal{J}_i \text{ Ordered}\}$, is

$$\begin{aligned}
\Pr\left(j \in \mathcal{S}_i \text{ Chosen} | \{\text{Stopped at } S_i\} \cap \{\mathcal{J}_i \text{ Ordered}\}; \boldsymbol{\theta}\right) &= \Pr\left(\left\{k \in \mathcal{S}_i : u_{i,j} \geq \max_{k\in\mathcal{S}_i}\{u_{i,k}\}\right\}; \boldsymbol{\theta}\right)\\
&= \Pr\left(\left\{k \in \mathcal{S}_i : u_{i,j} \geq u_{i,k}\right\}; \boldsymbol{\theta}\right).
\end{aligned}$$

Estimation of Parametrized Search-Cost Distribution Using Ordered Individual-level Search-Set Data Chen and Yao (2016) utilize a clickstream dataset, which is the closest possible to the ideal case in the literature. Although the essence is the same, their likelihood representation is slightly different from (6.4.4), combining the events $\{\text{Stopped at } S_i\} \cap \{\mathcal{J}_i \text{ Ordered}\}$ in one expression.

Estimation of Parametrized Search-Cost Distribution Using Unordered Individual-level Search-Set Data Honka and Chintagunta (2017); Ursu (Forthcoming) use individual-level unordered search-set data to identify and estimate the parametrized search-cost distribution. All other components of the likelihood are identical to what is illustrated in section 6.4.1.2. Because the search order even in $\mathcal{S}_i$ is not observed, it makes the permutation problems in (6.4.5) and (6.4.6) more complicated than when $\mathcal{S}_i$ is fully observed.

6.4.1.3 Estimation of Parametrized Search-Cost Distribution Using Market-level View-Rank Data

Commonality Index Data and Construction of the Model Counterpart Kim et al. (2010) use Amazon's pairwise conditional view-rank data that record the view rank of all other alternatives $\mathcal{J}\backslash j$ for each alternative j.[39] Amazon's commonality index between product j and k is defined as

$$CI_{jk} := \frac{n_{jk}}{\sqrt{n_j}\sqrt{n_k}}, \tag{6.4.7}$$

where n_{jk} is the number of consumers who viewed j and k in the same session, n_j and n_k only j and k in the same session, respectively. The commonality index is defined to reflect the viewing preferences in such a way that for each j, if (j,k) is viewed more than (j,l), then $CI_{jk} > CI_{jl}$, and vice versa. The estimation proceeds with constructing a model counterpart of (6.4.7), and then implementing either moment matching or maximum likelihood.

Weitzman (1979)'s optimal stopping rule dictates that the inclusion probability of product j in the searched set is

$$\begin{aligned}\Pr\left(\{j \in \mathcal{S}_i\};\boldsymbol{\theta}\right) &= \Pr\left(\max_{k\in\{1,2,\ldots,j-1\}}\{u_{i,k}\} < u^*_{i,j};\boldsymbol{\theta}\right)\\ &= \prod_{k=1}^{j-1} F_{\epsilon_{i,j}}\left(u^*_{i,j} - u_{i,k};\boldsymbol{\theta}\right).\end{aligned}$$

Note that $\boldsymbol{\theta} := \left(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \alpha, \sigma_\alpha, \gamma, \boldsymbol{\Sigma}_\gamma, \left\{\sigma_{i,j}\right\}_{i\in\mathcal{I}, j\in\mathcal{J}_i}\right)$ contains all the utility parameters and search-cost parameters. $\Pr\left(\{1 \in \mathcal{S}_i\};\boldsymbol{\theta}\right) = 1$ by the assumption that consumers search at least once. In other words, the possibility of an outside option is not allowed here. Then, it is trivial to show that for any pair (j, j') such that $j < j'$,

$$\Pr\left(\{j \in \mathcal{S}_i\};\boldsymbol{\theta}\right) > \Pr\left(\{j' \in \mathcal{S}_i\};\boldsymbol{\theta}\right)$$

and

$$\begin{aligned}\Pr\left(\{j \in \mathcal{S}_i\} \cap \{j' \in \mathcal{S}_i\};\boldsymbol{\theta}\right) &= \min\left\{\Pr\left(\{j \in \mathcal{S}_i\};\boldsymbol{\theta}\right), \Pr\left(\{j' \in \mathcal{S}_i\};\boldsymbol{\theta}\right)\right\}\\ &= \Pr\left(\{j' \in \mathcal{S}_i\};\boldsymbol{\theta}\right)\end{aligned}$$

hold. Also note that

[39] Notice the individual index i is taken out because only market-level data are observed.

$$\Pr\left(\mathcal{S}_i = \{1, ..., S_i\}; \boldsymbol{\theta}\right) = \Pr\left(\{S_i \in \mathcal{S}_i\}; \boldsymbol{\theta}\right) - \Pr\left(\{(S_i + 1) \in \mathcal{S}_i\}; \boldsymbol{\theta}\right).$$

[40] The model-side $\widehat{CI}_{jk}(\boldsymbol{\theta})$ is defined as the following:

$$\widehat{CI}_{jk}(\boldsymbol{\theta}) := \frac{\hat{n}_{jk}(\boldsymbol{\theta})}{\sqrt{\hat{n}_j(\boldsymbol{\theta})}\sqrt{\hat{n}_k(\boldsymbol{\theta})}}, \tag{6.4.8}$$

where $\hat{n}_j(\boldsymbol{\theta}) = \sum_{i\in\mathcal{I}} \Pr\left(\{j \in \mathcal{S}_i\}; \boldsymbol{\theta}\right)$ and $\hat{n}_{jk}(\boldsymbol{\theta}) = \sum_{i\in\mathcal{I}} \min\left\{\Pr\left(\{j \in \mathcal{S}_i\}; \boldsymbol{\theta}\right), \Pr\left(\{k \in \mathcal{S}_i\}; \boldsymbol{\theta}\right)\right\}$.

Estimation: Moment Matching or MLE The remaining step of estimation is either moment matching or maximum-likelihood estimation. Moment matching can be carried out by minimizing the squared difference of the model commonality index (6.4.8) and data (6.4.7):

$$\min_{\boldsymbol{\theta}} \sum_{j,k\in\mathcal{J}} \left\{\widehat{CI}_{jk}(\boldsymbol{\theta}) - CI_{jk}\right\}^2.$$

Maximum-likelihood estimation can be carried out by further assuming the parametric distribution of ϵ_{jk}, that is,

$$CI_{jk} = \widehat{CI}_{jk}(\boldsymbol{\theta}) + \epsilon_{jk} \qquad \epsilon_{jk} \sim i.i.d.\,\mathcal{N}\left(0, \frac{\nu^2}{2}\right).$$

Then, the building block for the likelihood becomes

$$\Pr\left(\widehat{CI}_{jk}(\boldsymbol{\theta}) - \widehat{CI}_{jl}(\boldsymbol{\theta}); \boldsymbol{\theta}, \nu\right) = \Phi\left(\frac{1}{\nu}\left(\widehat{CI}_{jk}(\boldsymbol{\theta}) - \widehat{CI}_{jl}(\boldsymbol{\theta})\right)\right).$$

Therefore, the maximum-likelihood estimation problem is as follows:

$$\max_{(\boldsymbol{\theta},\nu)} \prod_{j,k,l\in\mathcal{J}} \Pr\left(\widehat{CI}_{jk}(\boldsymbol{\theta}) - \widehat{CI}_{jl}(\boldsymbol{\theta}); \boldsymbol{\theta}, \nu\right)^{\iota_{j,kl}},$$

where $\iota_{j,kl} := \mathbf{1}\left(\{CI_{jk} - CI_{jl} > 0\}\right)$.

6.4.1.4 Remarks and Further Readings

Kim et al. (2017); Moraga-González et al. (2017) derive additional results on the building blocks of the likelihood under different scenarios on data availability and under different distributional assumptions. Kim et al. rely on the Gaussian assumption, and Moraga-González et al. rely on the T1EV assumption and augment the sequential search likelihood on Berry et al. (1995)'s demand-estimation framework.

[40] Recall that $\mathcal{S}_i$ is sorted by the decreasing order of the reservation utility.

6.4.2 Empirical Frameworks of Simultaneous Search

6.4.2.1 Setup and Data

Honka (2014) combines the simultaneous search and sorting rule based on Chade and Smith (2005, 2006) with consideration-set data. The context is consumers choosing automobile insurance, and some of the setups below are specific to her context. She connects the simultaneous search to the consideration-set formation. The data are an unordered individual-level consideration set without search sequence, in which case it is natural to consider simultaneous search rather than sequential search. Consumers are assumed to choose characteristics first, lock in to the characteristics, and then search to reveal the uncertainty of only the prices. In other words, elasticity of substitution between the price component and characteristics component is zero.

Consumer i's utility of choosing alternative j is specified as

$$\begin{aligned} u_{i,j} &:= \bar{u}_{i,j}\left(\boldsymbol{\beta}, \alpha, \eta_{i,j}\right) + \epsilon_{i,j} \\ &= \mathbf{x}'_{i,j}\boldsymbol{\beta} - \alpha p_{i,j} + \eta_{i,j} + \gamma_i \mathbf{1}\left(i \text{ chose } j \text{ at } t-1\right) + \epsilon_{i,j}, \end{aligned}$$

where

$$\begin{aligned} \epsilon_{i,j} &\sim i.i.d.\ T1EV\left(0,1\right) \\ \boldsymbol{\eta}_i &\sim \mathcal{N}\left(\boldsymbol{\eta}, \boldsymbol{\Sigma}_{\eta}\right) \\ \gamma_i &= \gamma_0 + \mathbf{w}'_i \gamma_1. \end{aligned}$$

$\mathbf{x}_{i,j}$ contains (consumer-specific) attributes, demographics, psychographic factors, and region fixed effects that are common across j. Note the region fixed effects drop out in the (within-region) conditional choice decision, but they may play a role in the search and consideration decisions. $p_{i,j}$ is the quoted price, and we assume $p_{i,j} \sim T1EV\left(\mu_{i,j}, \sigma_p\right)$, where σ_p is constant across alternatives. $\boldsymbol{\eta}_i$ captures consumer-specific brand intercepts. The term $\mathbf{1}\left(i \text{ chose } j \text{ at } t-1\right)$ reflects the inertia. $\mathbf{w}_i$ contains demographics, psychographic factors, customer satisfaction with the previous insurer, and therefore, γ_i captures the interaction between the consumer characteristics with the inertia.

Consumers get a quote from the previous insurer for free. We assume consumers know the characteristics vector $\mathbf{x}_{i,j}$ without cost.[41] The only variable subject to searching is the price, which is a setup adopted from Mehta et al. (2003). Consumers observe $E_p\left[u_{i,j}\right]$, given by

$$E_p\left[u_{i,j}\right] = \mathbf{x}'_{i,j}\boldsymbol{\beta} - \alpha \mu_{i,j} + \eta_{i,j} + \gamma_i \mathbf{1}\left(i \text{ chose } j \text{ at } t-1\right) + \epsilon_{i,j},$$

whereas $\left(\epsilon_{i,j}, \eta_{i,j}\right)$ is not directly observable to the econometrician.

[41] This assumption might only be applicable to some restricted contexts such as insurance-premium searching where the majority of consumers lock in to their coverage first, and then search only for the prices.

6.4.2.2 Marginal Improvement Algorithm and the Likelihood

A specific form of the first-order stochastic dominance among the (belief) distribution of $p_{i,j}$ across j is assumed to implement Chade and Smith (2005, 2006)'s MIA. The price-belief distribution $p_{i,j} \sim T1EV\left(\mu_{i,j}, \sigma_p\right)$ has the same variance over alternatives, but different means. This assumption of price belief follows exactly same shape of distribution allows Honka to calculate the optimal order of the alternative set $\mathcal{J}_i$ by using the MIA that we studied in section 6.2.2.2. By the virtue of the MIA, the derivation of the likelihood below resembles what we studied in section 6.4.1 for the sequential search.

The search and choice rule is as follows. Let 0 denote the previous insurer. Consumers, as well as the econometrician, sort $\mathcal{J}_i$ in the descending order of $E_p\left[u_{i,j}\right]$.[42] Without loss of generality, let $\mathcal{J}_i \backslash 0 = \{1, ..., J_i\}$ follow this order. Let $\mathcal{S}_i$ contain the previous insurer as index 0.[43] Consumers know the realization of $\left\{\epsilon_{i,j}\right\}_{j \in \mathcal{J}_i}$ already.

Ordering Rule The ordering rule, combined with the data availability, contains the following information:

$$\min_{k \in \mathcal{S}_i} \left\{E_p\left[u_{i,k}\right]\right\} \geq \max_{k \in \mathcal{J}_i \backslash \mathcal{S}_i} \left\{E_p\left[u_{i,k}\right]\right\}. \tag{6.4.9}$$

Using the extreme-value assumption on the price distributions, the likelihood calculation becomes much simpler.

Stopping Rule Recall the net benefit of searching the set $\mathcal{S}_i = \{0, 1, ..., S_i\}$ is given by

$$\Gamma\left(S_i\right) := E_p\left[\max_{k \in \mathcal{S}_i} \left\{u_{i,k}\right\}\right] - c_i S_i.$$

Furthermore, for the observed search set $\mathcal{S}_i$, for all $1 \leq s \leq J_i$,

$$\Gamma\left(S_i\right) \geq \Gamma\left(s\right) \tag{6.4.10}$$

should hold.

Choice Rule Conditional on $\mathcal{S}_i$ being observed, the choice rule is as usual

$$j = \arg\max_{k \in \mathcal{S}_i} u_{i,k}. \tag{6.4.11}$$

Likelihood Now let us combine the events of ordering, stopping, and choice rule with the probability of observing the unordered consideration-set data with the choice to derive the likelihood. To simplify notation, denote $\boldsymbol{\theta} := \left(\boldsymbol{\beta}, \alpha, \boldsymbol{\eta}, \boldsymbol{\Sigma}_{\eta}, \gamma_0, \gamma_1\right)$ by the set of model parameters to be estimated. The event

[42] The search rule described below is, in fact, slightly simpler than the original MIA, because it is assumed that the distribution $p_{i,j}$ follows the same shape. If the shape of the distribution that $p_{i,j}$ follows is heterogeneous, comparing only the mean is no longer sufficient, because what matters in the MIA is the tail behavior of the price distribution.

[43] Inclusion of the "free" option reflects the specific context that previous insurer sends the quote for free.

$\{\text{Ordering } \mathcal{J}_i\} \cap \{\text{Stopping at } \mathcal{S}_i\}$ is the event of observing $\mathcal{S}_i$ ($\subset \mathcal{J}_i$). The likelihood of observing $\mathcal{S}_i$ and the choice $j \in \mathcal{S}_i$ is

$$\begin{aligned}
&\Pr\left(\{\text{Ordering } \mathcal{J}_i\} \cap \{\text{Stopping at } \mathcal{S}_i\} \cap \{\text{Choose } j \in \mathcal{S}_i\}\,; \boldsymbol{\theta}\right) \\
&= \Pr\left(\{\text{Ordering } \mathcal{J}_i\} \cap \{\text{Stopping at } \mathcal{S}_i\}\,; \boldsymbol{\theta}\right) \Pr\left(\text{Choose } j \in \mathcal{S}_i \middle| \{\text{Ordering } \mathcal{J}_i\} \cap \{\text{Stopping at } \mathcal{S}_i\}\,; \boldsymbol{\theta}\right),
\end{aligned}$$

where

$$\begin{aligned}
&\Pr\left(\{\text{Ordering } \mathcal{J}_i\} \cap \{\text{Stopping at } \mathcal{S}_i\}\,; \boldsymbol{\theta}\right) \\
&= \Pr\left(\left\{\min_{k \in \mathcal{S}_i} \{E_p[u_{i,k}]\} \geq \max_{k \in \mathcal{J}_i \setminus \mathcal{S}_i} \{E_p[u_{i,k}]\}\right\} \cap \{1 \leq s \leq J_i : \Gamma(S_i) \geq \Gamma(s)\}\,; \boldsymbol{\theta}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\Pr\left(\text{Choose } j \in \mathcal{S}_i \middle| \{\text{Ordering } \mathcal{J}_i\} \cap \{\text{Stopping at } \mathcal{S}_i\}\,; \boldsymbol{\theta}\right) \\
&= \Pr\left(\left\{0 \leq k \leq S_i : u_{i,j} \geq u_{i,k}\right\}; \boldsymbol{\theta}\right).
\end{aligned}$$

$\left(\epsilon_{i,j}, \eta_{i,j}\right)_{i \in \mathcal{I}, j \in \mathcal{J}_i}$ are observable to the consumers but not to the econometrician. Only their distributions are parametrized. The maximum-likelihood problem solves the following to take into account the variables unobservable to the econometrician:

$$\begin{aligned}
\arg\max_{\boldsymbol{\theta}} \prod_{i \in \mathcal{I}} \int \Bigg[\prod_{\mathcal{S}_i \subset \mathcal{J}_i} \prod_{j \in \mathcal{S}_i} &\Pr\left(\{\text{Ordering } \mathcal{J}_i\} \cap \{\text{Stopping at } \mathcal{S}_i\}\,; \boldsymbol{\theta}\right)^{\mathbf{1}(\mathcal{S}_i \text{ Observed})} \\
&\times \Pr\left(\text{Choose } j \in \mathcal{S}_i \middle| \{\text{Ordering } \mathcal{J}_i\} \cap \{\text{Stopping at } \mathcal{S}_i\}\,; \boldsymbol{\theta}\right)^{\mathbf{1}(j \in \mathcal{S}_i \text{ Chosen})} \Bigg] d\Pr\left(\epsilon_{i,j}, \eta_{i,j}; \boldsymbol{\theta}\right).
\end{aligned}$$

Note $\prod_{\mathcal{S}_i \subset \mathcal{J}_i}$ means all the possible combinations of subset $\mathcal{S}_i$ of $\mathcal{J}_i$.

6.4.3 Testing the Modes of Search and Further Readings

6.4.3.1 Sequential or Simultaneous?

We presented the theoretical and empirical framework of consumer search under two large assumptions on the consumers' thought process: sequential search and simultaneous search. In the beginning of this chapter, we mentioned that little research to date has taken account of the consumers' thought process on the search modes. However, if *all* other assumptions hold for the respective search modes, sequential search and simultaneous search will yield different patterns on the observed search sequence or searched set, allowing for an econometrician to test whether consumers search sequentially or simultaneously. We provide a brief overview of the key idea of De los Santos et al. (2012); Honka and Chintagunta (2017) below, who attempted to test the modes of search using individual-level data. Both De los Santos et al.; Honka and Chintagunta conclude that consumers' search pattern is closer to simultaneous search.

De los Santos et al. (2012) use ComScore (partial) sequence data of browsing and purchase. The following patterns from the data are shown, which can be considered a "reduced-form test" on search modes. First, recall occurs often, suggesting the benchmark model of sequential search that assumes the homogeneous ex-ante price distribution is rejected. Second, even when allowing for the ex-ante heterogeneity over price distributions, search size does not depend on the revealed prices. That is, if consumers are searching sequentially, consumers searching once are more likely to have found a relatively low price, and consumers searching twice are likely to have found a relatively high price in the first search. This pattern is not discovered in their data.

Honka and Chintagunta (2017) use the unordered consideration-set data from Honka (2014), and suggest examining the mean of the prices over the consideration set. If consumers employ simultaneous search strategy, the mean price over the consideration set would be equal to the mean price over population. On the other hand, if consumers employ the sequential search strategy, the mean price over the searched set can be different from the population average, because consumers adaptively decide whether to stop searching or not at each period t. That is, they argue that the mean price over the searched set is lower than the mean over the population when consumers search sequentially. Honka and Chintagunta show that the prediction from the sequential search holds true for the the subset of consumers who ended up searching only once.

A possible common criticism of these comparisons is that whether the violation of the pattern in the sequential search is originated from consumers' search mode or from other assumptions that are imposed in the search models. Consumers might employ a mixed search strategy, or learning and spillover on the price distribution of unsearched alternatives might occur. Many more assumptions are involved in deriving the predictions from the sequential/simultaneous search models we studied above, which are the arena for future research as more granular and/or experimental data are becomes available.

6.4.3.2 Further Readings

Three recent literature reviews have been published on the subject. Baye et al. (2006) give an overview in the perspective of equilibrium price dispersion and its sustainability. Ratchford (2009)'s review focuses more on the empirical evidence of search. A recent review by Honka et al. (2018) focuses on the empirical frameworks and on the relation between search and consideration-set models.

References

ANDERSON, S. P. AND R. RENAULT (1999): "Pricing, product diversity, and search costs: A Bertrand-Chamberlin-Diamond model," *RAND Journal of Economics*, 30, 719–735.

——— (2000): "Consumer information and firm pricing: Negative externalities from improved information," *International Economic Review*, 41, 721–742.

ATHEY, S. AND G. IMBENS (2011): "Position auction with consumer search," *Quarterly Journal of Economics*, 126, 1213–1270.

BAYE, M. R., J. MORGAN, AND P. SCHOLTEN (2006): *Information, search, and price dispersion*, Boston, MA: Elsevier B.V., vol. 1 of *Economics and Information Systems*, chap. 6, 323–375, 1st ed. ed.

BENABOU, R. (1993): "Search market equilibrium, bilateral heterogeneity, and repeat purchases," *Journal of Economic Theory*, 60, 140–158.

BRAVERMAN, A. (1980): "Consumer search and alternative market equilibria," *Review of Economic Studies*, 47, 487–502.

BRONNENBERG, B. J., J. B. KIM, AND C. F. MELA (2016): "Zooming in on choice: How do consumers search for cameras online?" *Marketing Science*, 35, 693–712.

BROWN, J. R. AND A. GOOLSBEE (2002): "Does the internet make markets more competitive? Evidence from the life insurance industry," *Journal of Political Economy*, 110, 481–507.

BURDETT, K. AND K. L. JUDD (1983): "Equilibrium price dispersion," *Econometrica*, 51, 955–969.

BUTTERS, G. R. (1977a): "Equilibrium distributions of sales and advertising prices," *Review of Economic Studies*, 44, 465–491.

——— (1977b): "Market allocation through search: Equilibrium adjustment and price dispersion - A comment," *Journal of Economic Theory*, 15, 225–227.

CARLSON, J. A. AND R. P. MCAFEE (1983): "Discrete equilibrium price dispersion," *Journal of Political Economy*, 91, 480–493.

CHADE, H. AND L. SMITH (2005): "Simultaneous search," *Working Paper*.

——— (2006): "Simultaneous search," *Econometrica*, 74, 1293–1307.

CHEN, Y. AND S. YAO (2016): "Sequential search with refinement: Model and application with click-stream data," *Management Science*, 63, 4345–4365.

DE LOS SANTOS, B., A. HORTAÇSU, AND M. R. WILDENBEEST (2012): "Testing models of consumer search using data on web browsing and purchasing behavior," *American Economic Review*, 102, 2955–2980.

——— (2017): “Search with learning for differentiated products: Evidence from e-commerce,” *Journal of Business and Economic Statistics*, 35, 626–641.

DIAMOND, P. A. (1971): “A model of price adjustment,” *Journal of Economic Theory*, 3, 156–168.

DIXIT, A. K. AND J. E. STIGLITZ (1977): “Monopolistic competition and optimum product diversity,” *American Economic Review*, 67, 297–308.

ELLISON, G. AND S. F. ELLISON (2009): “Search, obfuscation, and price elasticities on the internet,” *Econometrica*, 77, 427–452.

GOEREE, M. S. (2008): “Limited information and advertising in the U.S. personal computer industry,” *Econometrica*, 76, 1017–1074.

GOLDMANIS, M., A. HORTAÇSU, C. SYVERSON, AND ÖNSEL EMRE (2009): “E-commerce and the market structure of retail industries,” *Economic Journal*, 120, 651–682.

HONG, H. AND M. SHUM (2006): “Using price distributions to estimate search costs,” *RAND Journal of Economics*, 37, 257–275.

HONKA, E. (2014): “Quantifying search and switching costs in the US auto insurance industry,” *RAND Journal of Economics*, 45, 847–884.

HONKA, E. AND P. CHINTAGUNTA (2017): “Simultaneous or sequential? Search strategies in the U.S. auto insurance industry,” *Marketing Science*, 36, 21–42.

HONKA, E., A. HORTAÇSU, AND M. A. VITORINO (2017): “Advertising, consumer awareness, and choice: evidence from the U.S. banking industry,” *RAND Journal of Economics*, 48, 611–646.

HONKA, E., A. HORTAÇSU, AND M. WILDENBEEST (2018): “Empirical search and consideration sets,” *Working Paper*.

HORTAÇSU, A. AND C. SYVERSON (2004): “Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds,” *Quarterly Journal of Economics*, 119, 403–456.

IOANNIDES, Y. M. (1975): “Market allocation through search: Equilibrium adjustment and price dispersion,” *Journal of Economic Theory*, 11, 247–262.

KIM, J. B., P. ALBUQUERQUE, AND B. J. BRONNENBERG (2010): “Online demand under limited consumer search,” *Marketing Science*, 29, 1001–1023.

——— (2011): “Mapping online consumer search,” *Journal of Marketing Research*, 48, 13–27.

——— (2017): “The Probit choice model under sequential search with an application to online retailing,” *Marketing Science*, 63, 3911–3929.

KOULAYEV, S. (2014): "Search for differentiated products: Identification and estimation," *RAND Journal of Economics*, 45, 553–575.

MACMINN, R. D. (1980): "Search and market equilibrium," *Journal of Political Economy*, 88, 308–327.

MCCALL, J. J. (1970): "Economics of information and job search," *Quarterly Journal of Eocnomics*, 84, 113–126.

MEHTA, N., S. RAJIV, AND K. SRINIVASAN (2003): "Price uncertainty and consumer search: A structural model of consideration set formulation," *Marketing Science*, 22, 58–84.

MORAGA-GONZÁLEZ, J. L., Z. SÁNDOR, AND M. R. WILDENBEEST (2017): "Consumer search and prices in the automobile market," *Working Paper*.

MORGAN, P. AND R. MANNING (1985): "Optimal search," *Econometrica*, 53, 923–944.

PIRES, T. (2016): "Costly search and consideration sets in storable goods markets," *Quantitative Marketing and Economics*, 14, 157–193.

RATCHFORD, B. T. (2009): *Consumer Search Behavior and Its Effect on Markets*, Foundations and Trends in Marketing, Hanover, MA: now Publishers Inc.

RAUH, M. T. (2009): "Strategic complementarities and search market equilibrium," *Games and Economic Behavior*, 66, 959–978.

REINGANUM, J. F. (1979): "A simple model of equilibrium price dispersion," *Journal of Political Economy*, 87, 851–858.

ROB, R. (1985): "Equilibrium price distributions," *Review of Economic Studies*, 52, 487–504.

ROTHSCHILD, M. (1973): "Models of market organization with imperfect information: A survey," *Journal of Political Economy*, 81, 1283–1308.

——— (1974): "Searching for the lowest price when the distribution of prices is unknown," *Journal of Political Economy*, 82, 689–711.

SALOP, S. AND J. STIGLITZ (1977): "Bargains and ripoffs: A model of monopolistically competitive price dispersion," *Review of Economic Studies*, 44, 493–510.

SALOP, S. C. (1973): "Systematic job search and uneomplyment," *Review of Economic Studies*, 40, 191–201.

SEILER, S. (2013): "The impact of search costs on consumer behavior: A dynamic approach," *Quantitative Marketing and Economics*, 11, 155–203.

STAHL II, D. O. (1989): "Oligopolistic pricing with sequential consumer search," *American Economic Review*, 79, 700–712.

STIGLER, G. J. (1961): "The economics of information," *Journal of Political Economy*, 69, 213–225.

URSU, R. M. (Forthcoming): "The power or rankings: Quantifying the effect of rankings on online consumer search and purchase decisions," *Marketing Science*.

VARIAN, H. R. (1980): "A model of sales," *American Economic Review*, 70, 651–659.

VISHWANATH, T. (1992): "Parallel search for the best alternative," *Economic Theory*, 2, 495–507.

WEITZMAN, M. L. (1979): "Optimal search for the best alternative," *Econometrica*, 47, 641–654.

WOLINSKY, A. (1986): "True monopolistic competition as a result of imperfect information," *Quarterly Journal of Economics*, 101, 493–512.

Appendix A

Review on Basic Estimation Methods

In this chapter, we review some basic characteristics of the maximum-likelihood estimator, generalized method of moments estimator, and some simulation-based estimators. The focus is to give a concise summary of the asymptotics, namely, consistency results and asymptotic normality. We recommend the readers consult relevant econometrics textbooks and handbook chapters for a full treatment on the materials overviewed here.

A.1 Maximum-Likelihood Estimation

Maximum-likelihood estimation is the primary way of estimating nonlinear models when the model is fully specified up to a set of finite-dimensional model parameters. Provided that the econometric model is correctly specified, the maximum-likelihood estimator is consistent and asymptotically efficient.

A.1.1 Definition

We begin from presenting the definition of the maximum-likelihood estimator. Throughout the section, we assume the parameter space $\Theta \subset \mathbb{R}^p$, and the data $\mathbf{x}_i \in \mathbb{R}^k$. Also, we assume all the regularity conditions for the differentiation under the integral sign are satisfied.

Definition. (Maximum-Likelihood Estimator) Let $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ be the given data. Let $\mathbf{X} = (\mathbf{x}_i)_{i=1}^n \in \mathbb{R}^{n\times k}$. Suppose $\mathcal{L}(\boldsymbol{\theta}|\cdot)$ is the joint density function corresponding to the true data-generating process. Then,

$$\hat{\boldsymbol{\theta}}_{MLE} := \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{X}).$$

Proposition A.1.1. *(MLE for i.i.d. Data) Let $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ be the given data that are mutually independent and identically distributed. Let $f(\mathbf{x}_i|\boldsymbol{\theta})$ be the marginal density function corresponding to the true data-generating process. Then,*

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} f(\mathbf{x}_i|\boldsymbol{\theta}).$$

Corollary. *(MLE for i.i.d. Data Using Log-Likelihood) Let* $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ *be the given data that are mutually independent and identically distributed. Let* $f(\mathbf{x}_i|\boldsymbol{\theta})$ *be the marginal density function corresponding to the true data-generating process. Then,*

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln f(\mathbf{x}_i|\boldsymbol{\theta}).$$

Definition. (Score Function) Let $f(.|\boldsymbol{\theta})$ be the density function of the true data-generating process. A score function $\mathbf{s}: \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^k$ is defined by

$$\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}_i) := \frac{\partial \ln f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Remark. We sometimes abbreviate $\mathbf{s}_i(\boldsymbol{\theta}) = \mathbf{s}(\boldsymbol{\theta}|\mathbf{x}_i)$.

Proposition A.1.2. *(Expectation of Score) Let* $f(.|\boldsymbol{\theta})$ *be the density function of the true data-generating process. Let* $\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}_i)$ *be the corresponding score function. Then,* $E_{\boldsymbol{\theta}_0}[\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}_i)] = \mathbf{0}$.

Proof.

$$\begin{aligned} E_{\boldsymbol{\theta}_0}[\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}_i)] &= \int \frac{\partial \ln f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} f(\mathbf{x}_i|\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} d\mathbf{x}_i \\ &= \int \left[\frac{1}{f(\mathbf{x}_i|\boldsymbol{\theta})} \frac{\partial f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} f(\mathbf{x}_i|\boldsymbol{\theta}_0)\, d\mathbf{x}_i \\ &= \left[\frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}_i|\boldsymbol{\theta})\, d\mathbf{x}_i\right]\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} 1 \\ &= \mathbf{0}. \end{aligned}$$

[1] □

Notice this proposition asserts that MLE can be viewed as a subset of GMM; the moment condition that MLE imposes is $E_{\boldsymbol{\theta}_0}[\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}_i)] = \mathbf{0}$.

Proposition A.1.3. *(MLE with Differentiability of the Log-Likelihood Function) Suppose* $\forall \boldsymbol{\theta} \in \Theta$, $\ln f(\mathbf{x}_i|\boldsymbol{\theta})$ *exists. Then,* $\hat{\boldsymbol{\theta}}_{MLE}$ *solves* $\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(\boldsymbol{\theta}|\mathbf{x}_i) = \mathbf{0}$.

Remark. The above proposition provides us with a justification for the first-derivative method.

Definition. (Information Matrix) Let $\mathbf{s}(\boldsymbol{\theta}|\mathbf{x})$ be the score function for an i.i.d. data-generating process F. The Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ is defined by

$$\mathbf{I}(\boldsymbol{\theta}) := E_{\boldsymbol{\theta}_0}\left[\{\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}_i)\}\{\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}_i)\}'\right].$$

[1] Note. Although we have evaluated this equation at the true $\boldsymbol{\theta}_0$, it holds for any $\boldsymbol{\theta} \in \Theta$.

Theorem A.1.1. *(Information Equality) Let $\mathbf{s}(\boldsymbol{\theta}|\mathbf{x})$ be the score function for an i.i.d. data-generating process F. Then, under some regularity conditions,*

$$\begin{aligned}
\mathbf{I}(\boldsymbol{\theta}) &= -E_{\boldsymbol{\theta}_0}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{s}(\boldsymbol{\theta}|\mathbf{x}_i)\right] \\
&= -E_{\boldsymbol{\theta}_0}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\ln f(\mathbf{x}_i|\boldsymbol{\theta})\right] \\
&\equiv -\mathbf{H}(\boldsymbol{\theta}).
\end{aligned}$$

A.1.2 Consistency and Asymptotic Efficiency

Theorem A.1.2. *(Cramer-Rao Lower Bound) Let $\mathcal{L}(\mathbf{x}_1,...,\mathbf{x}_n|\boldsymbol{\theta})$ be the likelihood of $\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$ given $\boldsymbol{\theta}$. Let $\tilde{\boldsymbol{\theta}}\left(\{\mathbf{x}_i\}_{i=1}^n\right)$ be an unbiased estimator of $\boldsymbol{\theta}_0$. Then,*

$$Var_{\boldsymbol{\theta}_0}\left(\tilde{\boldsymbol{\theta}}\right) \geq \left[\mathbf{I}(\boldsymbol{\theta})\right]^{-1}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0},$$

i.e., $Var_{\theta_0}\left(\tilde{\boldsymbol{\theta}}\right) - \left[\mathbf{I}(\boldsymbol{\theta}_0)\right]^{-1}$ is positive semidefinite.

Proof. Let $\tilde{\boldsymbol{\theta}}$ be an unbiased estimator of $\boldsymbol{\theta}_0$, so that $E_{\boldsymbol{\theta}_0}\left[\tilde{\boldsymbol{\theta}}\right] = \boldsymbol{\theta}_0$. We want to show that

$$\begin{aligned}
Var_{\theta_0}\left(\tilde{\boldsymbol{\theta}}\right)\cdot\mathbf{I}(\boldsymbol{\theta}_0) &\equiv Var_{\boldsymbol{\theta}_0}\left(\tilde{\boldsymbol{\theta}}\right)\cdot Var_{\boldsymbol{\theta}_0}\left(\mathbf{s}(\boldsymbol{\theta}_0|\mathbf{x}_i)\right) && \text{(A.1.1)} \\
&\geq \mathbf{I}_k, && \text{(A.1.2)}
\end{aligned}$$

where $\mathbf{I}_k$ is an $k \times k$ identity matrix.

The Cauchy-Schwartz inequality asserts that $Var_{\boldsymbol{\theta}_0}\left(\tilde{\boldsymbol{\theta}}\right)\cdot\mathbf{I}(\boldsymbol{\theta}_0) \geq \left\{Cov_{\boldsymbol{\theta}_0}\left(\tilde{\boldsymbol{\theta}},\mathbf{s}(\boldsymbol{\theta}_0|\mathbf{x}_i)\right)\right\}^2$. Because

$$\begin{aligned}
Cov_{\boldsymbol{\theta}_0}\left(\tilde{\boldsymbol{\theta}},\mathbf{s}(\boldsymbol{\theta}_0|\mathbf{x}_i)\right) &= \int\tilde{\boldsymbol{\theta}}\left(\frac{1}{f(\mathbf{x}_i|\boldsymbol{\theta}_0)}\frac{\partial f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)f(\mathbf{x}_i|\boldsymbol{\theta}_0)\,d\mathbf{x}_i \\
&= \int\tilde{\boldsymbol{\theta}}\left(\frac{\partial f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)d\mathbf{x}_i \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\int\tilde{\boldsymbol{\theta}}f(\mathbf{x}_i|\boldsymbol{\theta}_0)\,d\mathbf{x}_i \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}E_{\boldsymbol{\theta}_0}\left[\tilde{\boldsymbol{\theta}}\right] \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\boldsymbol{\theta}_0 \\
&= \mathbf{I}_k.
\end{aligned}$$

Thus, we have established the inequality (A.1.2). $\square$

For the sake of notational simplicity, we define

$$\begin{aligned} Q_0(\boldsymbol{\theta}) &:= E_{\boldsymbol{\theta}_0}[\ln f(\mathbf{x}_i|\boldsymbol{\theta})] \\ &= \int [\ln f(\mathbf{x}_i|\boldsymbol{\theta})] f(\mathbf{x}_i|\boldsymbol{\theta}_0)\, d\mathbf{x}_i \\ \hat{Q}_n(\boldsymbol{\theta}) &:= \frac{1}{n}\sum_{i=1}^{n} \ln f(\mathbf{x}_i|\boldsymbol{\theta}). \end{aligned}$$

We assume the following:

MLE(1) The support of the joint distribution of $\mathbf{x}_i$ does not depend on the parameter value $\boldsymbol{\theta} \in \Theta$.

MLE(2) $\hat{Q}_n(\boldsymbol{\theta}) \Rightarrow_p Q_0(\boldsymbol{\theta})$ uniformly over Θ.[2]

MLE(3) If $Q_0(\boldsymbol{\theta})$ attains the maximum over $\boldsymbol{\theta} \in \Theta$, the maximizer is unique.

MLE(4) Θ is compact.

MLE(5) $Q_0(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$.

Remark. The pointwise convergence of $\forall \boldsymbol{\theta} \in \Theta$, $\hat{Q}(\boldsymbol{\theta}) \rightarrow_p Q_0(\boldsymbol{\theta})$ is guaranteed by the weak law of large numbers and the continuous mapping theorem. However, as we shall see, we need a stricter condition than the pointwise convergence to show the consistency of MLE.

Theorem A.1.3. *(Consistency and Asymptotic Efficiency of MLE) Let $\hat{\boldsymbol{\theta}}_{MLE}$ be a maximum-likelihood estimator for a true parameter $\boldsymbol{\theta}_0$. Then, under Assumption MLE(1)-(5),*
(i) $\hat{\boldsymbol{\theta}}_{MLE} \rightarrow_p \boldsymbol{\theta}_0$ as $n \rightarrow \infty$.
(ii) $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0\right) \rightarrow_d \mathcal{N}\left(\mathbf{0}, [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}\right)$.
Furthermore, $n\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0\right)\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0\right)' \rightarrow_p [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$ as $n \rightarrow \infty$, so $\hat{\boldsymbol{\theta}}_{MLE}$ is asymptotically efficient.

Proof. (i) We first show that $\forall \tilde{\boldsymbol{\theta}} \neq \boldsymbol{\theta}_0$, $Q(\tilde{\boldsymbol{\theta}}) - Q(\boldsymbol{\theta}_0) \leq 0$.

$$\begin{aligned} Q(\tilde{\boldsymbol{\theta}}) - Q(\boldsymbol{\theta}_0) &= E_{\boldsymbol{\theta}_0}\left[\ln f\left(\mathbf{x}_i|\tilde{\boldsymbol{\theta}}\right)\right] - E_{\boldsymbol{\theta}_0}[\ln f(\mathbf{x}_i|\boldsymbol{\theta}_0)] \\ &= \int \ln \frac{f\left(\mathbf{x}_i|\tilde{\boldsymbol{\theta}}\right)}{f(\mathbf{x}_i|\boldsymbol{\theta}_0)} f(\mathbf{x}_i|\boldsymbol{\theta}_0)\, d\mathbf{x}_i \\ &\leq \ln \int \frac{f\left(\mathbf{x}_i|\tilde{\boldsymbol{\theta}}\right)}{f(\mathbf{x}_i|\boldsymbol{\theta}_0)} f(\mathbf{x}_i|\boldsymbol{\theta}_0)\, d\mathbf{x}_i \\ &= \ln \int f\left(\mathbf{x}_i|\tilde{\boldsymbol{\theta}}\right) d\mathbf{x}_i \\ &= 0. \end{aligned}$$

The third inequality is by the Jensen's inequality, using the fact that $\ln(.)$ is a concave function. $E_{\theta_0}[\ln f(\mathbf{x}_i|\theta)]$ attains its maximum at true $\boldsymbol{\theta}_0$. Because MLE(2) asserts that

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\|\hat{Q}_n(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})\right\| \rightarrow_p 0,$$

[2]We can replace this assumption with the bounded $Q(\boldsymbol{\theta})$.

which yields

$$\arg\max_{\boldsymbol{\theta}\in\Theta} \hat{Q}_n(\boldsymbol{\theta}) \equiv \hat{\boldsymbol{\theta}}_{MLE} \rightarrow_p \boldsymbol{\theta}_0.$$

This establishes the consistency.

(ii) Recall the definition of MLE, that MLE solves

$$\frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial \ln f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MLE}} = \mathbf{0}.$$

The intermediate value theorem asserts that $\exists \boldsymbol{\theta}^* \in \left[\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_{MLE}\right]$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left.\frac{\partial \ln f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \left.\frac{\partial^2 \ln f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right|_{\boldsymbol{\theta}^*\in\left[\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}_{MLE}\right]}\left(\hat{\boldsymbol{\theta}}_{MLE}-\boldsymbol{\theta}_0\right)\right] = \mathbf{0}.$$

Rearranging, we obtain

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{MLE}-\boldsymbol{\theta}_0 &= -\frac{\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial \ln f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}}{\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \ln f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right|_{\boldsymbol{\theta}^*\in\left[\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}_{MLE}\right]}} \\ &\rightarrow_p \mathbf{0}.\end{aligned}$$

[3]

To use the central limit theorem, multiplying $\sqrt{n}$ and sending $n \rightarrow \infty$, we have

$$\begin{aligned}-\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left.\frac{\partial \ln f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &\rightarrow_d \mathcal{N}\left(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_0)\right) \\ \frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \ln f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right|_{\boldsymbol{\theta}^*\in\left[\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}_{MLE}\right]} &\rightarrow_p -E\left[\left.\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\ln f(\mathbf{x}_i|\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right] \\ &= -\mathbf{I}(\boldsymbol{\theta}_0).\end{aligned}$$

The last equality is by the information equality. Therefore, the continuous mapping theorem asserts that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{MLE}-\boldsymbol{\theta}_0\right) \rightarrow_d \mathcal{N}\left(\mathbf{0}, \left[\mathbf{I}(\boldsymbol{\theta}_0)\right]^{-1}\right),$$

which attains the Cramer-Rao efficiency bound. □

Remark. Note that if the model is misspecified as $g(\mathbf{x}_i|\boldsymbol{\theta})$ but the true data-generating process is $f(\mathbf{x}_i|\boldsymbol{\theta})$,

[3]Note. $\frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial \ln f(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \rightarrow_p \frac{1}{n}\sum_{i=1}^{n} E\left[\mathbf{s}(\mathbf{x}_i|\boldsymbol{\theta}_0)\right] = \mathbf{0}.$

then $\hat{\boldsymbol{\theta}}_{MLE}$ is not a consistent estimator of θ_0, because

$$\begin{aligned} Q_0(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}_0}[\ln g(\mathbf{x}_i|\boldsymbol{\theta})] \\ &= \int \ln g(\mathbf{x}_i|\boldsymbol{\theta}) f(\mathbf{x}_i|\boldsymbol{\theta}_0)\, d\mathbf{x}_i, \end{aligned}$$

which is not necessarily maximized at $\boldsymbol{\theta}_0$.

However, the quasi-MLE estimator is consistent if $g(\mathbf{x}_i|\boldsymbol{\theta})$ and $f(\mathbf{x}_i|\boldsymbol{\theta})$ both belong to a linear exponential family. The Q-MLE estimator minimizes the following Kullback-Leiber information criterion:

$$I(g, f; \boldsymbol{\theta}, \boldsymbol{\theta}_0) = \int \ln\left[\frac{g(\mathbf{x}_i|\boldsymbol{\theta})}{f(\mathbf{x}_i|\boldsymbol{\theta}_0)}\right] f(\mathbf{x}_i|\boldsymbol{\theta}_0)\, d\mathbf{x}_i.$$

Example A.1.1. (Sources of Stochasticity) Suppose a researcher estimates the intertemporal elasticity of substitution to test the life-cycle consumption-saving hypothesis using MLE. Let $\boldsymbol{\theta}$ be the intertemporal elasticity of substitution, let $\mathbf{x}_i$ be the wage process, and let y_i be savings. We have obtained a dataset on $(\mathbf{x}_i, y_i)$ to fit the model:

$$y_i = g(\mathbf{x}_i, \boldsymbol{\theta}) + u_i.$$

Four sources of stochasticity are possible here: (i) heterogeneity in the structural form $g_i(\boldsymbol{\theta})$; (ii) the dependence of true $\mathbf{x}$ on some other factor such as unexpected transfer income, wealth, and so on; (iii) optimization failure, which is often observed in the lab experiments; and (iv) measurement error on $\mathbf{x}_i$. In many cases, we let all the stochasticity comes from (iv) just to consistently estimate the parameters. However, if it is not the case, a significant bias in the estimates can be involved.

A.2 Generalized Method of Moments

Structural econometric models often boil down to a set of moment conditions, especially when the researcher does not specify the distribution of unobservables. In such a case, the generalized method of moments by Hansen (1982) provides an integrated solution.

A.2.1 Motivation and Setup

Let $\{\mathbf{x}_t\}_{t=1}^N \in \mathbb{R}^n$ be the observed data. We want to figure out $\boldsymbol{\beta}_{(k\times 1)} \in \mathbb{P} \subset \mathbb{R}^k$, where $\mathbb{P}$ is the parameter space. We have a $\mathbf{f} : \mathbb{P} \times \mathbb{R}^n \to \mathbb{R}^r$ such that

$$E[\mathbf{f}(\mathbf{x}, \boldsymbol{\beta})]\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathbf{0} \tag{A.2.1}$$

if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ is the true parameter. The system of equations (A.2.1) has k unknowns and r equations. (A.2.1) is called the moment condition.

We assume $r \geq k$, so the system (A.2.1) is identified. It is immediate that we can formulate the sample analogue of (A.2.1) defined by $\mathbf{g}_N : \mathbb{P} \times \mathbb{R}^n \to \mathbb{R}^r$ as

$$\mathbf{g}_N(\boldsymbol{\beta}) := \frac{1}{N}\sum_{t=1}^{N} \mathbf{f}(\mathbf{x}_t, \boldsymbol{\beta}) = \mathbf{0}. \tag{A.2.2}$$

If $r > k$, there might be no solution for (A.2.2) even if (A.2.1) has a solution.

We consider using a selection matrix $\mathbf{A}_{(k\times r)}$ to reduce the k dimensional equations to the r dimension as

$$\mathbf{A}_{(k\times r)}\mathbf{g}_N(\boldsymbol{\beta}) = \mathbf{0} \tag{A.2.3}$$

and denote $\mathbf{b}_N$ by the solution for the equation (A.2.3). The remaining task is to figure out the good choice of $\mathbf{A}$. As one changes the specification of $\mathbf{A}$, the estimator $\mathbf{b}_N$ changes accordingly. Thus, the entire family of estimator $\{\mathbf{b}_N\}$ can only be pinned down by specifying $\mathbf{A}$. Thus, *the solution* $\mathbf{b}_N$ *is indexed by* $\mathbf{A}$.

A.2.2 Efficiency Bound

A.2.2.1 Limit Approximations for $\boldsymbol{\beta}_0$

In this section, we want to establish the efficiency bound for the estimator $\mathbf{b}_N$ of $\boldsymbol{\beta}_0$. Let $\mathbf{g}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{t=1}^{N} \mathbf{f}(X_t, \boldsymbol{\beta})$. We begin from the following assumptions.

GMM(1) The underlying data-generating process $\{\mathbf{x}_t\}_{t=1}^{N}$ is stationary and ergodic.[4]

GMM(2) (Central Limit Theorem) $\sqrt{N}\mathbf{g}_N(\boldsymbol{\beta}_0) \to_d \mathcal{N}(0, \mathbf{V})$.[5]

GMM(3) (Consistency of $\mathbf{b}_N$) $\mathbf{b}_N \to_{a.s.} \boldsymbol{\beta}_0$.[6]

GMM(4) (Differentiability of $\mathbf{f}$) $\mathbf{f}(\mathbf{x}_t, \boldsymbol{\beta})$ is continuously differentiable in $\boldsymbol{\beta}$.

[4]Ergodicity can be relaxed.

[5]We impose this central limit theorem as an assumption to rule out the case where the data-generating process $\{\mathbf{x}_t\}$ may be more strongly correlated than the assumption for the central limit theorem to hold. In other words, we want to rule out a strong dependence of the underlying process.

[6]We assume $\mathbf{b}_N \to_{a.s.} \boldsymbol{\beta}_0$ pointwise. To derive the consistency of $\mathbf{b}_N$, it is insufficient just to ensure the pointwise convergence of the following:

$$\mathbf{g}_N(\boldsymbol{\beta}) \equiv \frac{1}{N}\sum_{t=1}^{N} \mathbf{f}(\mathbf{x}_t, \boldsymbol{\beta}) \to_{a.s} E[\mathbf{f}(\mathbf{x}_t, \boldsymbol{\beta})].$$

Instead, we should impose two additional restrictions.

1. For a compact parameter space $\mathbb{P}$,
$$\mathbf{g}_N(.) \to_{a.s.} E[\mathbf{f}(\mathbf{x}_t, .)]$$
uniformly. This is the uniform convergence-type restriction.
2. $\exists \epsilon > 0$ such that
$$E\left[\sup_{\|\beta-\tilde{\beta}\|<\epsilon, \tilde{\beta}} \left\|\mathbf{f}(\mathbf{x}_t, \boldsymbol{\beta}) - \mathbf{f}(\mathbf{x}_t, \tilde{\boldsymbol{\beta}})\right\|\right] < \infty.$$
Note that because $\mathbf{f}(\mathbf{x}_t, \boldsymbol{\beta})$ is a random function, we have integrated it out. Under this condition, the minimizer $\mathbf{b}_N$ of $\mathbf{g}_N(\boldsymbol{\beta})' \mathbf{W}\mathbf{g}_N(\boldsymbol{\beta})$ can be shown to converge almost surely to the true parameter.

GMM(5) (Differentiability of $\mathbf{f}$ at $\boldsymbol{\beta}_0$) $\exists\epsilon > 0$ such that $\mathcal{B}_\epsilon(\boldsymbol{\beta}_0) \subset \mathbb{P}$. That is, $\boldsymbol{\beta}_0$ is located in the interior of $\mathbb{P}$.[7]

GMM(6) (Smoothness of $\mathbf{f}$ around $\boldsymbol{\beta}_0$) $E\left[\sup_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|<\epsilon}\left|\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{f}(\mathbf{x}_t,\boldsymbol{\beta}) - \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{f}(\mathbf{x}_t,\boldsymbol{\beta}_0)\right|\right] < \infty$.

Throughout the section, denote $\mathbf{D}_N := \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{g}_N(\mathbf{b}_N)$ and $\mathbf{D} := E\left[\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{f}(\mathbf{x}_t,\boldsymbol{\beta}_0)\right]$.[8] The following lemma is immediate from the assumptions.

Lemma A.2.1. *(Consistency of Derivative Matrix)* $\mathbf{D}_N \to_{a.s.} \mathbf{D}$.

Proof. We have

$$\begin{aligned}
\mathbf{D}_N &= \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{g}_N(\mathbf{b}_N) \\
&= \frac{1}{N}\sum_{t=1}^{N}\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{f}(\mathbf{x}_t,\mathbf{b}_N) \\
&\to_{a.s.} E\left[\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{f}(\mathbf{x}_t,\boldsymbol{\beta}_0)\right] \\
&= \mathbf{D}.
\end{aligned}$$

The convergence follows by GMM(3), (4), (6), and the continuous mapping theorem that $\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{f}$ is continuous as well. $\square$

Lemma A.2.2. *(Mean Value Approximation) Under GMM(3)-(6),*

$$\mathbf{g}_N(\mathbf{b}_N) \backsimeq \mathbf{g}_N(\boldsymbol{\beta}_0) + \mathbf{D}_N(\mathbf{b}_N - \boldsymbol{\beta}_0). \tag{A.2.4}$$

The above lemmas combined with GMM(2) yields the following theorem.

Theorem A.2.1. *(Central Limit Theorem for a GMM Estimator) Let $\mathbf{b}_N$ be a GMM estimator indexed by the selection matrix $\mathbf{A}$. Then,*

$$\sqrt{N}(\mathbf{b}_N - \boldsymbol{\beta}_0) \to_d \mathcal{N}(0, Cov(\mathbf{A})), \tag{A.2.5}$$

where $Cov(\mathbf{A}) = (\mathbf{AD})^{-1}\mathbf{AVA}'(\mathbf{AD})'^{-1}$.

Proof. Let $\mathbf{A}_{(k\times r)}$ be a selection matrix. From (A.2.3), we know $\mathbf{b}_N$ is a solution for the equation $\mathbf{A}_{(k\times r)}\mathbf{g}_N(\boldsymbol{\beta}) = \mathbf{0}$, so that

$$\mathbf{A}_{(k\times r)}\mathbf{g}_N(\mathbf{b}_N) = \mathbf{0}$$

holds by definition.[9]

Recall (A.2.4) and suppose N is large. Multiplying $\mathbf{A}_{(r\times k)}$ yields

$$\mathbf{A}\mathbf{g}_N(\boldsymbol{\beta}_0) + \mathbf{AD}(\mathbf{b}_N - \boldsymbol{\beta}_0) = \mathbf{A}\mathbf{g}_N(\mathbf{b}_N) \equiv \mathbf{0}.$$

[7] This condition is required for the existence of the derivatives in vicinity of $\boldsymbol{\beta}_0$.
[8] $\mathbf{D}$ and $\mathbf{D}_N$ are $r \times k$ matrices.
[9] Notice that $\mathbf{g}_N(\mathbf{b}_N)$ may not be $\mathbf{0}$.

So we have the equality

$$\mathbf{A}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right) = -\mathbf{A}\mathbf{D}\left(\mathbf{b}_N - \boldsymbol{\beta}_0\right).$$

Multiplying $\sqrt{N}$ yields

$$\mathbf{A}\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right) = -\mathbf{A}\mathbf{D}\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}_0\right).$$

By rearranging, we have

$$\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}_0\right) = -\left(\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{A}\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right). \tag{A.2.6}$$

Notice that by A2, $\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right) \to_d \mathcal{N}\left(0, \mathbf{V}\right)$. So we have that

$$\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}_0\right) \to_d \mathcal{N}\left(0, \left(\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{A}\mathbf{V}\mathbf{A}'\left(\mathbf{A}\mathbf{D}\right)'^{-1}\right).$$

[10] □

A.2.2.2 Asymptotic Efficiency Bound

The theorem that asserts the GMM estimator is asymptotically normally distributed leads our interests to the efficiency bound for the asymptotic variance of $\mathbf{b}_N$ and the way to achieve the bound, by selecting an adequate $\mathbf{A}$. We begin from showing that the asymptotic covariance matrix is unique up to invertible transformations.

Lemma A.2.3. *(GMM Estimator and Its Asymptotic Variance Is Unique up to Invertible Linear Transformation) Let $\mathbf{B}_{(k\times k)}$ be invertible. Consider $\mathbf{BA}$ as our new selection-matrix choice. Then,*

(i) $\mathbf{b}_N$ solves $\mathbf{A}_{(k\times r)}\mathbf{g}_N\left(\boldsymbol{\beta}\right) = \mathbf{0}$ if and only if the same $\mathbf{b}_N$ solves $\mathbf{B}_{(k\times k)}\mathbf{A}_{(k\times r)}\mathbf{g}_N\left(\boldsymbol{\beta}\right) = \mathbf{0}$.

(ii) $Cov\left(\mathbf{A}\right) = Cov\left(\mathbf{BA}\right)$.

Proof. (i) It is immediate by the definition of an invertible matrix.

(ii) Consider (A.2.6) again, replacing $\mathbf{A}$ with $\mathbf{BA}$:

$$\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}_0\right) = -\left(\mathbf{B}\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{B}\mathbf{A}\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right)$$

$$\begin{aligned}\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}_0\right) &\to_d \mathcal{N}\left(\mathbf{0}, \left(\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{B}^{-1}\mathbf{B}\mathbf{A}\mathbf{V}\mathbf{A}'\mathbf{B}'\mathbf{B}'^{-1}\left(\mathbf{A}\mathbf{D}\right)'^{-1}\right)\\ &=_d \mathcal{N}\left(\mathbf{0}, \left(\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{A}\mathbf{V}\mathbf{A}'\left(\mathbf{A}\mathbf{D}\right)'^{-1}\right).\end{aligned}$$

[10]Multiplying $\mathbf{D}$ yields

$$\sqrt{N}\mathbf{D}\left(\mathbf{b}_N - \boldsymbol{\beta}_0\right) = -\mathbf{D}\left(\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{A}\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right).$$

Using (A.2.4) again, we obtain

$$\begin{aligned}\sqrt{N}\mathbf{g}_N\left(\mathbf{b}_N\right) &= \left[\mathbf{I} - \mathbf{D}\left(\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{A}\right]\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right)\\ &\to_d \left[\mathbf{I} - \mathbf{D}\left(\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{A}\right]\mathcal{N}\left(0, \mathbf{V}\right).\end{aligned}$$

So we find that $\left[\mathbf{I} - \mathbf{D}\left(\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{A}\right]$ is the adjustment term. Premultiplying $\mathbf{A}$ yields the both side $\mathbf{0}$, which is trivial by assumption.

So the asymptotic covariance matrix is the same. □

By virtue of the above lemma, we can pick $\mathbf{B}$ in a convenient way. Consider $\mathbf{B} = (\mathbf{AD})^{-1}$ so that $\mathbf{AD} = \mathbf{I}$. The asymptotic covariance matrix reduces to $\mathbf{AVA}'$ under the choice of $\mathbf{B}$. Thus, hereafter, without loss of generality we can assume $Cov(\mathbf{A}) = \mathbf{AVA}'$. The following theorem establishes the choice of $\mathbf{A}$ under which the lower bound of the asymptotic covariance matrix is achieved.

Theorem A.2.2. *(Asymptotic Efficiency Bound of GMM Estimator) Suppose $\mathbf{V}$ defined in GMM(2) is nonsingular. Let $\mathbf{A}^* = \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{V}^{-1}$.*[11] *Then, $\mathbf{A}^*$ achieves the asymptotic efficiency bound of GMM estimator $\mathbf{b}_N$ with the asymptotic covariance matrix $Cov(\mathbf{A}^*) = \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}$, that is,*

$$\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}_0\right) \to_d \mathcal{N}\left(\mathbf{0}, \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}\right).$$

Furthermore, $\forall \mathbf{A}_{(k\times r)}$, $Cov(\mathbf{A}) - Cov(\mathbf{A}^*)$ is positive semidefinite.

Proof. Let $\mathbf{A}_{(k\times r)}$ be a selection matrix such that $\mathbf{AD} = \mathbf{I}$.

Guess $\mathbf{A}^* = \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{V}^{-1}$. It is immediate that

$$Cov(\mathbf{A}^*) = \mathbf{A}^*\mathbf{V}(\mathbf{A}^*)' = \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{D}\left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1} = \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}.$$

We have

$$\begin{aligned}
\mathbf{AV}(\mathbf{A}^*)' &= \mathbf{AVV}^{-1}\mathbf{D}\left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1} \\
&= \mathbf{AD}\left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1} \\
&= \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}.
\end{aligned}$$

The last equality is by our choice of $\mathbf{A}$ such that $\mathbf{AD} = \mathbf{I}$.

Now consider

$$\begin{aligned}
Cov(\mathbf{A}) - Cov(\mathbf{A}^*) &= \mathbf{AVA}' - \mathbf{A}^*\mathbf{V}(\mathbf{A}^*)' \\
&= \mathbf{AVA}' - \left(\mathbf{DV}^{-1}\mathbf{D}'\right)^{-1} \\
&= \mathbf{AVA}' - \left(\mathbf{DV}^{-1}\mathbf{D}'\right)^{-1} + \left(\mathbf{DV}^{-1}\mathbf{D}'\right)^{-1} - \left(\mathbf{DV}^{-1}\mathbf{D}'\right)^{-1} \\
&= \mathbf{AVA}' - \mathbf{A}^*\mathbf{VA}' - \mathbf{AV}(\mathbf{A}^*)' + \mathbf{A}^*\mathbf{V}(\mathbf{A}^*)' \\
&= (\mathbf{A} - \mathbf{A}^*)\mathbf{V}(\mathbf{A} - \mathbf{A}^*)' \\
&= \left[(\mathbf{A} - \mathbf{A}^*)\sqrt{\mathbf{V}}\right]\left[(\mathbf{A} - \mathbf{A}^*)\sqrt{\mathbf{V}}\right]',
\end{aligned}$$

[11] Because $\left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}$ is invertible, $\mathbf{A}^*$ can be replaced by $\mathbf{D}'\mathbf{V}^{-1}$ in practice, but not in this proof, because we want to exploit the feature $\mathbf{AD} = \mathbf{I}$.

which is a positive semidefinite matrix, because $\mathbf{V}$ is a positive definite matrix because $\mathbf{V}$ is nonsingular. □

Remark. Because $\mathbf{A}^* = \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{V}^{-1}$ and $\left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}$ is just some nonsingular matrix, the bound is attained if

$$\mathbf{A}^* = \mathbf{B}\mathbf{D}'\mathbf{V}^{-1} \tag{A.2.7}$$

for some $\mathbf{B}$.

Remark. Notice the similarity of the proof structure with the Gauss-Markov theorem.

The results up to now can be summarized by the following four basic approximation equations:

$$\sqrt{N}\left(\mathbf{b}_N - \boldsymbol{\beta}_0\right) \approx -\left(\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{A}\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right) \tag{A.2.8}$$

$$\sqrt{N}\mathbf{g}_N\left(\mathbf{b}_N\right) \approx \left[\mathbf{I} - \mathbf{D}\left(\mathbf{A}\mathbf{D}\right)^{-1}\mathbf{A}\right]\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right) \tag{A.2.9}$$

$$\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right) \to_d \mathcal{N}\left(0, \mathbf{V}\right) \tag{A.2.10}$$

$$Cov\left(\mathbf{A}^*\right) = \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}. \tag{A.2.11}$$

A.2.3 Tests of Overidentifying Restrictions

This subsection provides the asymptotic test on the validity of the moment conditions using the efficient selection matrix.

A.2.3.1 Using the Selection Matrix $\mathbf{A}^*$

We had that the efficiency bound is attained when we set $\mathbf{A}^*_{(k\times r)} = \mathbf{B}\mathbf{D}'\mathbf{V}^{-1}$. Set $\mathbf{A}^* = \left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{V}^{-1}$ and factor $\mathbf{V}^{-1}_{(r\times r)} = \boldsymbol{\Lambda}'\boldsymbol{\Lambda}$ so that $\mathbf{V}_{(r\times r)} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}'^{-1}$.

Lemma A.2.4.

$$\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right) \to_d \mathcal{N}\left(0, \mathbf{I}_r\right)$$

Proof. It is immediate from (A.2.10) and our construction of $\boldsymbol{\Lambda} = \sqrt{\mathbf{V}^{-1}}$. □

Now plugging back our construction of A^* in (A.2.9) yields

$$\sqrt{N}\mathbf{g}_N\left(\mathbf{b}_N\right) \approx \left[\mathbf{I} - \mathbf{D}\left(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{V}^{-1}\right]\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right). \tag{A.2.12}$$

Premultiplying by $\boldsymbol{\Lambda}$ and substituting back $\mathbf{V}^{-1} = \boldsymbol{\Lambda}'\boldsymbol{\Lambda}$, we have that

$$\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N\left(\mathbf{b}_N\right) \approx \boldsymbol{\Lambda}\left[\mathbf{I} - \mathbf{D}\left(\mathbf{D}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\mathbf{D}\right)^{-1}\mathbf{D}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right]\sqrt{N}\mathbf{g}_N\left(\boldsymbol{\beta}_0\right). \tag{A.2.13}$$

Defining $\boldsymbol{\Delta}_{(r\times k)} := \boldsymbol{\Lambda}_{(r\times r)}\mathbf{D}_{(r\times k)}$. (A.2.13) yields

$$\begin{aligned}\sqrt{N}\boldsymbol{\Lambda}g_N(\mathbf{b}_N) &\approx \left[\boldsymbol{\Lambda}-\boldsymbol{\Lambda}\mathbf{D}\left(\mathbf{D}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\mathbf{D}\right)^{-1}\mathbf{D}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right]\sqrt{N}\mathbf{g}_N(\boldsymbol{\beta}_0) && \text{(A.2.14)}\\ &= \left[\boldsymbol{\Lambda}-\boldsymbol{\Delta}\left(\boldsymbol{\Delta}'\boldsymbol{\Delta}\right)^{-1}\boldsymbol{\Delta}'\boldsymbol{\Lambda}\right]\sqrt{N}\mathbf{g}_N(\boldsymbol{\beta}_0) && \text{(A.2.15)}\\ &= \left[\mathbf{I}_r-\boldsymbol{\Delta}\left(\boldsymbol{\Delta}'\boldsymbol{\Delta}\right)^{-1}\boldsymbol{\Delta}'\right]\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0). && \text{(A.2.16)}\end{aligned}$$

It is straightforward that

$$\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\mathbf{b}_N) \to_d N\left(0, \mathbf{I}_r-\boldsymbol{\Delta}\left(\boldsymbol{\Delta}'\boldsymbol{\Delta}\right)^{-1}\boldsymbol{\Delta}'\right).$$

Lemma A.2.5. *(Properties of $\mathbf{J}_1$ and $\mathbf{J}_2$) Define $\mathbf{J}_1 = \boldsymbol{\Delta}\left(\boldsymbol{\Delta}'\boldsymbol{\Delta}\right)^{-1}\boldsymbol{\Delta}'$ and $\mathbf{J}_2 = \mathbf{I}-\boldsymbol{\Delta}\left(\boldsymbol{\Delta}'\boldsymbol{\Delta}\right)^{-1}\boldsymbol{\Delta}'$.*

(i) $\mathbf{J}_1$ and $\mathbf{J}_2$ are symmetric and idempotent.

(ii) $\mathbf{J}_1+\mathbf{J}_2 = \mathbf{I}_r$.

(iii) $\mathbf{J}_1\mathbf{J}_2 = \mathbf{O}$.

Proof. Basic linear algebra yields the result. □

We are ready to characterize the limiting distribution for our overidentification test. The following theorem illustrates the limiting distribution of our squared objective function.

Theorem A.2.3. *(Asymptotic Distributions of the GMM Test Statistics) Let $\mathbf{J}_1, \mathbf{J}_2$ be as defined above. Let $\mathbf{g}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{t=1}^{N}\mathbf{f}(\mathbf{x}_t,\boldsymbol{\beta})$. Let $\mathbf{A}^* = \mathbf{D}'\mathbf{V}^{-1}$ be an efficient selection matrix. Suppose $\mathbf{b}_N$ is the GMM estimator thatsatisfies $\mathbf{A}^*\mathbf{g}_N(\mathbf{b}_N) = \mathbf{0}$. Then,*

(i) $N\mathbf{g}_N(\boldsymbol{\beta}_0)'\mathbf{V}^{-1}\mathbf{g}_N(\boldsymbol{\beta}_0) \to_d \chi^2_r$.

(ii) $N\mathbf{g}_N(\mathbf{b}_N)'\mathbf{V}^{-1}\mathbf{g}_N(\mathbf{b}_N) \to_d \chi^2_{r-k}$.

(iii) $\begin{pmatrix}\mathbf{J}_1\\ \mathbf{J}_2\end{pmatrix}\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0) \to_d \mathcal{N}\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{J}_1 & \mathbf{O}\\ \mathbf{O} & \mathbf{J}_2\end{pmatrix}\right)$.

Proof. (i)

$$N\mathbf{g}_N(\boldsymbol{\beta}_0)'\mathbf{V}^{-1}\mathbf{g}_N(\boldsymbol{\beta}_0) = \left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right]'\left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right], \tag{A.2.17}$$

and we have shown that $\left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right] \to \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$. Thus, we find the test statistic converges in distribution to χ^2_r.

(ii) We begin from (A.2.17).

$$\left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right]'\left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right] = \left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right]'\mathbf{J}_1\left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right] + \left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right]'\mathbf{J}_2\left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right]$$

by the previous lemma that $\mathbf{J}_1+\mathbf{J}_2 = \mathbf{I}_r$. From (A.2.12) and (A.2.16) we have:

$$\left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\mathbf{b}_N)\right]'\left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\mathbf{b}_N)\right] \approx \left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right]'\mathbf{J}_2\left[\sqrt{N}\boldsymbol{\Lambda}\mathbf{g}_N(\boldsymbol{\beta}_0)\right]$$

we have the desired convergence.

(iii) It is straightforward by the previous results and the fact that $\mathbf{J}_1\mathbf{J}_2 = \mathbf{O}$. □

Remark. Notice the important difference between (i) and (ii) on the degree of freedom. When $\mathbf{b}_N$ is plugged in, the degree of freedom becomes $r-k$, while the actual $\boldsymbol{\beta}_0$ returns the degree of freedom r. The proof of the theorem was simple because of our selection of $\mathbf{A}^* = \mathbf{D}'\mathbf{V}^{-1}$.

In practice, $N\mathbf{g}_N(\mathbf{b}_N)'\mathbf{V}^{-1}\mathbf{g}_N(\mathbf{b}_N) \to_d \chi^2_{r-k}$ can be used for the overspecification test. If the test statistic is large, it can be considered as evidence that the model is somehow misspecified. That is, either the moment condition is wrong or some other restrictions are wrong. However, the literature has also documented that the test seems to reject overidentifying restrictions too often in practice.

A.2.4 Quadratic-Form Minimization and Implementation

We can think of the GMM estimator in a different way. Let $\mathbf{W}_{(r\times r)}$ be a weighting matrix. Consider the minimization problem:

$$\min_{\boldsymbol{\beta}\in\mathbb{P}} \frac{1}{2}\mathbf{g}_N(\boldsymbol{\beta})'\mathbf{W}\mathbf{g}_N(\boldsymbol{\beta}). \tag{A.2.18}$$

The first-order condition is

$$\left[\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{g}_N(\boldsymbol{\beta})'\mathbf{W}\right]\mathbf{g}_N(\boldsymbol{\beta}) = \mathbf{0}. \tag{A.2.19}$$

Let $\mathbf{b}_N$ denote the solution for (A.2.19). Notice $\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{g}_N(\boldsymbol{\beta}) \to_{a.s.} \mathbf{D}$. Thus, we find the efficient choice for $\mathbf{W}$ is $\mathbf{V}^{-1}$.

Let K_N be the minimized objective function for (A.2.19) where $\mathbf{W} = \mathbf{V}^{-1}$, that is, $K_N := \frac{1}{2}\mathbf{g}_N(\mathbf{b}_N)'\mathbf{V}^{-1}\mathbf{g}_N(\mathbf{b}_N)$. Then,

$$N\mathbf{g}_N(\mathbf{b}_N)'\mathbf{V}^{-1}\mathbf{g}_N(\mathbf{b}_N) \to_d \chi^2_{r-k} \tag{A.2.20}$$

as before.

In practice, one uses the consistent estimator of $\mathbf{V}^{-1}$ to find the estimator that solves (A.2.18). One way is do the estimation in two stages or, iteratively, begin from $\mathbf{W} = \mathbf{I}$. An alternative is the continuous-updating GMM, which uses the sample analogue of the following covariance matrix estimator:

$$\mathbf{V}(\boldsymbol{\beta}) = \lim_{N\to\infty} E\left[\frac{1}{\sqrt{N}}\left(\sum_{t=1}^{N}\mathbf{f}(\mathbf{x}_t,\boldsymbol{\beta}) - \frac{1}{N}\sum_{t=1}^{N}\mathbf{f}(\mathbf{x}_t,\boldsymbol{\beta})\right)\frac{1}{\sqrt{N}}\left(\sum_{t=1}^{N}\mathbf{f}(\mathbf{x}_t,\boldsymbol{\beta}) - \frac{1}{N}\sum_{t=1}^{N}\mathbf{f}(\mathbf{x}_t,\boldsymbol{\beta})\right)'\right],$$

which is a function of $\boldsymbol{\beta}$. Then, we can minimize the following:

$$\min_{\boldsymbol{\beta}\in\mathbb{P}} \frac{1}{2}\mathbf{g}_N(\boldsymbol{\beta})'(\mathbf{V}(\boldsymbol{\beta}))^{-1}\mathbf{g}_N(\boldsymbol{\beta}). \tag{A.2.21}$$

The solution $\mathbf{b}_N$ for the problem (A.2.21) also attains the efficiency bound.

A.3 Simulation-Based Estimation Methods

When a set of model parameters is point identified, most of the estimation method boils down either to MLE or GMM. As studied above, MLE is desirable when the likelihood is fully specified up to a finite-dimensional model parameter, by solving out a (structural) model. GMM is semiparametric, in the sense that the distribution of the error term can be left unspecified. However, even if a set of likelihoods or moment conditions can be derived from solving a model, they often do not have an analytic expression, or they may be practically impossible to evaluate. In some such instances, simulation-based methods can be useful.

A.3.1 A Starting Point: Glivenko-Cantelli Theorem

Let $\{x_i\}_{i=1}^n$ be an i.i.d. sequence of random variables with distribution $F(.)$ on $\mathbb{R}$. The empirical distribution function is a function of $x \in \mathbb{R}$ defined as follows.

Definition. (Empirical Distribution Function) Let $\{x_i\}_{i=1}^N \sim i.i.d.\ F(.)$. The empirical distribution function $\hat{F}(x)$ is defined by

$$\hat{F}_N(x) := \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(x_i \leq x).$$

For any given $x \in \mathbb{R}$, the strong law of large numbers asserts that $\hat{F}_N(x) \rightarrow_{a.s.} F(x)$ as $\hat{F}_N(x)$ is a form of sample mean.[12] The Glivenko-Cantelli theorem asserts more. $\hat{F}_N(x)$ is a reasonable estimate of $F(x)$ when both are viewed as functions of x.

Theorem A.3.1. *(Glivenko-Cantelli) Let $\{x_i\}_{i=1}^N$ be an i.i.d. sequence of random variables with distribution function $F(.)$ on $\mathbb{R}$. Then,*

$$\sup_{x\in\mathbb{R}} \left|\hat{F}_N(x) - F(x)\right| \rightarrow_{a.s.} 0.$$

McFadden (1989); Pakes and Pollard (1989) generalize this theorem to a more general class of empirical measures and empirical process. Then, they give the consistency and asymptotic normality of the simulated GMM estimators. The key condition other than the regularity conditions for the consistency and asymptotic normality of GMM is that the simulation analogue of the original moment condition converges fast enough to the true moment condition. In the following subsections, we illustrate the method of simulated moments and method of simulated likelihoods.

A.3.2 Method of Simulated Moments

Although the method of simulated moments can be used in a slightly more general setup, we consider the individual coefficients model in this subsection for the purpose of illustration. Consider a model,

[12] We only need to show that $E[|\mathbf{1}(x_i \leq x)|] < \infty$, which is trivial because $\mathbf{1}(x_i \leq x)$ is bounded and N is finite.

which yields the (unconditional) moment condition

$$
\begin{aligned}
E\left[g\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}_i\right)\right] &= \int\int g\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}_i\right) d\Pr\left(\boldsymbol{\theta}_i\right) d\Pr\left(y_j, \mathbf{x}_j\right) \\
&= 0,
\end{aligned}
$$

where $\Pr\left(\boldsymbol{\theta}_i\right) =_d \mathcal{N}\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$. The target model parameters are $\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$. The GMM estimation solves, given sample $\left\{\left(y_j, \mathbf{x}_j\right)\right\}_{j=1}^{J}$,

$$
\min_{\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}} \left\{\frac{1}{J}\sum_{j=1}^{J}\int g\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}_i\right) d\Pr\left(\boldsymbol{\theta}_i\right)\right\}^2 .
$$

Given a candidate parameter $\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$, it is practically infeasible to analytically evaluate

$$
\int g\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}_i\right) d\Pr\left(\boldsymbol{\theta}_i\right)
$$

when $\dim\left(\boldsymbol{\theta}\right)$ is larger than, say, 5. For first and second derivatives, the problem gets even more serious.

The key idea is to replace $\int g\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}_i\right) d\Pr\left(\boldsymbol{\theta}_i\right)$ with its simulation analogue. For each $\left(y_j, \mathbf{x}_j\right)$, consider the simulation analogue

$$
\frac{1}{N_s}\sum_{i=1}^{N_s} g\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}_i\right);
$$

that is, we simulate $\Pr\left(\boldsymbol{\theta}_i\right)$ and take the mean. A variant of Glivenko-Cantelli theorem asserts that $\frac{1}{N_s}\sum_{i=1}^{N_s} g\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}_i\right)$ is uniformly consistent to $\int g\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}_i\right) d\Pr\left(\boldsymbol{\theta}_i\right)$. An immediate limitation is that $\Pr\left(\boldsymbol{\theta}_i\right)$ has to be known up to the model parameters $\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$.

Let $\boldsymbol{\phi} = \operatorname{vec}\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$, and let $\boldsymbol{\phi}_0$ be the true parameter. Let $\hat{\boldsymbol{\phi}}_{GMM}$ and $\hat{\boldsymbol{\phi}}_{MSM}$ be GMM and MSM estimators, respectively. Let $\mathbf{V}_{GMM}$ be the GMM asymptotic variance. Pakes and Pollard (1989) show that for a fixed number N_s with a frequency simulator,

$$
\sqrt{J}\left(\hat{\boldsymbol{\phi}}_{MSM} - \boldsymbol{\phi}_0\right) \to_d \mathcal{N}\left(\mathbf{0}, \left(1 + \frac{1}{N_s}\right)\mathbf{V}_{GMM}\right).
$$

That is, the asymptotic variance is inflated by $\left(1 + \frac{1}{N_s}\right)$.

A.3.3 Maximum Simulated Likelihood

The intuition of maximum simulated likelihood is similar with method of simulated moments. Suppose the likelihood or part of the likelihood is defined in terms of integrals:

$$
f\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}\right) = \int h\left(y_j, \mathbf{x}_j, \mathbf{u}_i; \boldsymbol{\theta}\right) d\Pr\left(\mathbf{u}_j\right).
$$

Prominent examples include the multivariate normal $\mathbf{u}_i$'s, or the latent utility models, where the domain of integration is a strict subset of Euclidean space. The MLE objective function is

$$\sum_{j=1}^{J} \ln f\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}\right) = \sum_{j=1}^{J} \ln \left\{ \int h\left(y_j, \mathbf{x}_j, \mathbf{u}_i; \boldsymbol{\theta}\right) d \Pr\left(\mathbf{u}_i\right) \right\}.$$

With N_s simulated draws of $\mathbf{u}_i$, approximate the right-hand side using

$$\sum_{j=1}^{J} \ln f\left(y_j, \mathbf{x}_j; \boldsymbol{\theta}\right) = \sum_{j=1}^{J} \ln \left\{ \frac{1}{N_s} \sum_{i=1}^{N_s} h\left(y_j, \mathbf{x}_j, \mathbf{u}_i; \boldsymbol{\theta}\right) \right\}.$$

The asymptotics are slightly different. Let $\hat{\boldsymbol{\theta}}_{MSL}$ be the method of simulated likelihood estimator. Suppose $N_s = O\left(\sqrt{J} + \delta\right)$ for some $\delta > 0$. Let $\mathbf{I}\left(\boldsymbol{\theta}_0\right)$ be the Fisher information matrix. Then,

$$\sqrt{J}\left(\hat{\boldsymbol{\theta}}_{MSL} - \boldsymbol{\theta}_0\right) \to \mathcal{N}\left(\mathbf{0}, \mathbf{I}^{-1}\left(\boldsymbol{\theta}_0\right)\right).$$

That is, the method of simulated likelihood is asymptotically equivalent to MLE when the number of simulation draws increase at a rate faster than $\sqrt{J}$.

A.3.4 Implementation Algorithms

The following four steps are typical in implementing the algorithm: (i) Draw N_s random vectors, (ii) take the candidate parameter value, (iii) calculate the simulation analogue of the likelihood/moment condition, and (iv) iterate over (ii)-(iii) to find the parameter value that maximizes/minimizes the objective function. In practice, one should never draw N_s again at each iteration, in order to guarantee the numerical convergence.

Example A.3.1. (Berry, Levinsohn, and Pakes (1995)) $\Pr\left(\boldsymbol{\theta}_i\right) =_d \mathcal{N}\left(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$. First, draw N_s standard normal random vectors $\boldsymbol{\eta}_i$. For each candidate parameter value, $\boldsymbol{\theta}_i = \boldsymbol{\theta} + \sqrt{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}\boldsymbol{\eta}_i$, where $\left(\sqrt{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}\right)\left(\sqrt{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}}\right)' = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. Calculate the simulation analogue of the moment condition correspondingly. Then iterate over.

References

HANSEN, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

MCFADDEN, D. (1989): "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica*, 57, 995–1026.

NEWEY, W. K. AND D. MCFADDEN (1994): *Large sample estimation and hypothesis testing*, Amsterdam: Elsevier, vol. 4 of *Handbook of Econometrics*, chap. 36, 2111–2245.

PAKES, A. AND D. POLLARD (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica*, 57, 1027–1057.