Estimating Nonseparable Selection Models: A Functional Contraction Approach*

Fan Wu Yi Xin
California Institute of Technology

October 23, 2025

Abstract

We propose a novel method for estimating nonseparable selection models. We show that, given the selection rule and the observed selected outcome distribution, the potential outcome distribution can be characterized as the fixed point of an operator, which we prove to be a functional contraction. We propose a two-step semiparametric maximum likelihood estimator to estimate the selection model and the potential outcome distribution. The consistency and asymptotic normality of the estimator are established. Our approach performs well in Monte Carlo simulations and is applicable in a variety of empirical settings where only a selected sample of outcomes is observed. Examples include consumer demand models with only transaction prices, auctions with incomplete bid data, and Roy models with data on accepted wages.

Keywords: Sample Selection, Nonseparble Models, Functional Contraction, Potential Outcome Distribution, Semiparametric Estimation, Demand Estimation, Auction, Roy Models.
JEL Codes: C14, C24, C51, L11, D44, J31.

*Wu: Division of the Humanities and Social Sciences, California Institute of Technology, 1200 East California Blvd, MC 228-77, Pasadena, CA 91125. Email: fwu2@caltech.edu. Xin: Division of the Humanities and Social Sciences, California Institute of Technology, 1200 East California Blvd, MC 228-77, Pasadena, CA 91125. Email: yixin@caltech.edu. Financial support from the Ronald and Maxine Linde Institute of Economic and Management Sciences are gratefully acknowledged. We appreciate valuable discussions with Yingyao Hu, Yao Luo, Luciano Pomatto, Robert Sherman, and Omer Tamuz. We thank seminar and conference participants at Caltech, Umich, and USC.

1 Introduction

Sample selection issues arise when the data available for analysis is not representative of the entire population due to a selection process that systematically excludes certain observations. For example, in consumer demand studies, researchers often only have access to the transaction prices of chosen products, while the prices of non-selected products remain unobserved (Goldberg, 1996; Cicala, 2015; Crawford et al., 2018; Allen et al., 2019; Salz, 2022; Sagl, 2023; Cosconati et al., 2024). Similarly, in auctions, data may include only the winning bids (or certain order statistics), excluding all other submitted bids (Athey and Haile, 2002; Komarova, 2013; Guerre and Luo, 2019; Allen et al., 2024). Sample selection issues have long been recognized in labor market studies as well. For instance, wage data is typically available only for individuals who choose to work (Gronau, 1974; Heckman, 1974), and, in the original Roy model (Roy, 1951), which examines the occupational distribution of earnings, we observe earnings within an occupation only for those who self-select into working in that sector.

Observing only a selected sample of outcomes—such as prices, bids, or wages—presents significant challenges for estimating two key elements: (1) the model that governs the selection process, such as a consumer demand model, an auction's winning rule, or a labor force participation model; and (2) the distribution of outcomes *prior to selection*, often referred to as "potential outcomes" in the literature. Typically, it is assumed that potential outcomes are generated by an *outcome equation*, which depends on both observable characteristics and unobservable error terms. Flexibly estimating potential outcome distributions is crucial in many empirical contexts, such as analyzing price distributions to understand firms' pricing strategies and wage distributions to examine inequality.

The first solution to sample selection bias is to use maximum likelihood estimation, as in Heckman (1974) and Lee (1982, 1983), which relies heavily on distributional assumptions regarding the error terms. More commonly employed methods for sample selection models are two-step estimators proposed by Heckman (1976, 1979), which introduce a correction term to account for the non-random nature of the sample. A substantial body of theoretical work has been developed to relax the distributional assumptions in the two stages of the estimation procedure (Ahn and Powell, 1993; Andrews and Schafgans, 1998; Chen and Khan, 2003; Das et al., 2003; Newey, 2007, 2009; Chernozhukov et al., 2023). See also Vella (1998) for a comprehensive survey

on semi-parametric two-step estimation for selection models.

Our paper proposes a fundamentally different and novel approach to estimating selection models where the outcome equation is nonparametric and nonseparable in error terms. Rather than constructing a reduced-form bias correction term and controlling it in the outcome equation, we directly analyze how the selection model maps the potential outcome distributions to the distributions of selected outcomes and seek to *invert* the mapping. The key insight of our approach is that, given the selection model and potential outcome distributions across all alternatives, we can derive the likelihood of an outcome being selected. Conversely, if this selection likelihood *were* known, we could recover the potential outcome distributions from the observed outcome distributions. This two-way relationship characterizes a fixed-point problem. Building on this intuition, we construct an operator whose fixed point represents the potential outcome distributions and show that this operator is a functional contraction.

Formally, we consider a discrete choice problem in which each alternative is associated with a potential outcome distribution. A selection function maps a vector of realized potential outcomes to a probability distribution over the alternatives. For example, in the consumer demand setting, each alternative represents a product, and the potential outcome is the offered price, with the selection function micro-founded by the consumer's utility maximization problem. We allow the outcome equations to be fully nonparametric with nonseparable error terms and to vary flexibly across different alternatives. We assume that potential outcomes across different alternatives are conditionally independent given observables.

Given the selection function, we construct an operator whose fixed point is the potential outcome distributions. We establish sufficient conditions for it to be a functional contraction (Theorems 1 and 2). Proving contraction within a function space is challenging; to address this, we construct a metric in the same spirit as that in Thompson (1963). Our results imply that, given the selection function and the observed distributions of selected outcomes, we can nonparametrically recover the potential outcome distributions. Moreover, this identification result is *constructive*: starting with any initial guess for the potential outcome distributions, we iteratively apply the operator. As the number of iterations approaches infinity, this process converges to the potential outcome distributions associated with the selection function.

We propose a two-step semi-parametric maximum likelihood estimator for the

selection function, parameterized by a finite-dimensional parameter, and potential outcome distributions. In the first step, we obtain a nonparametric estimate of the selected outcome distribution directly from the data. Given this estimate, we use our contraction result to recover the potential outcome distributions for *any parameter* in the selection function. In the second step, we construct the model-implied choice probabilities and match them with the data moments. Once we have an estimator for the selection parameter, a plug-in estimator for the potential outcome distribution can be readily obtained.

We establish the consistency and asymptotic normality of the proposed estimator (Theorems 3 and 4). This is particularly challenging because the mapping from the potential to the selected outcome distributions does not have a closed form. We prove that this mapping is a homeomorphism, a key result in establishing consistency and asymptotic normality.

To examine the finite sample properties of our estimator, we conduct Monte Carlo simulations across various designs of the outcome equation. Our results show that the biases in our estimator are generally small, and the standard deviation decreases as the sample size increases across all simulation designs. Our nonparametric estimation of the potential outcome distributions outperforms the standard two-step method when the two-step method misspecifies the outcome equations. Notably, even when the selection function is misspecified by econometricians, our method performs robustly in estimating the potential outcome distributions.

Compared to the traditional two-step method, our approach offers several key advantages. First, we allow for fully nonparametric estimation of potential outcome distributions. Importantly, our approach accommodates nonseparable error terms in the outcome equation, allowing for fully heterogeneous effects of covariates on outcomes. Moreover, we impose no symmetry assumptions, allowing the potential outcome distributions to vary flexibly across alternatives. Unlike most selection correction approaches that focus on estimating conditional mean models (e.g., Das et al. (2003) and various other semi-parametric versions)[1], our goal is to recover the *entire outcome distribution* with a flexible specification. We correct for sample selection bias across the entire distribution of potential outcomes by examining how the bias is *sys-*

[1]These models restrict covariates to affecting only the location of the outcome distribution. A recent paper by Chernozhukov et al. (2023) proposes a semi-parametric generalization of the Heckman selection model which accommodates rich patterns of heterogeneity in the effects of covariates on outcomes and selection.

tematically generated by the selection model. More recently, Arellano and Bonhomme (2017) propose a method to correct for sample selection in quantile regression models; see also Newey (2007) and Fernández-Val et al. (2024) for recent developments in nonseparable sample selection models.

Second, our approach does not require an instrument to exogenously shift the choice probability, a typical requirement in the two-step method to avoid multicollinearity, nor does our approach rely on identification-at-infinity arguments. In practice, finding a suitable instrument can be quite challenging (see Vella (1998) for further discussion). d'Haultfoeuille and Maurel (2013a) and D'Haultfœuille et al. (2018) develop estimation methods for semiparametric sample selection models without an instrument or a large-support regressor, leveraging the independence-at-infinity assumption.

Our approach relies on an alternative assumption: conditional independence of potential outcomes given observables. This assumption is commonly invoked in auction models (e.g., independent private value auctions or mineral rights models)[2] and becomes more plausible when econometricians have access to a rich set of observables. In a binary selection model (e.g., the decision to work) where the potential outcome for one alternative is constant (e.g., the wage for not working is 0) or in censored regression models with a single observed dependent variable, our conditional independence assumption is trivially satisfied. We provide further discussion of this assumption in Section 2.1.

Finally, our method accommodates a flexible selection function, applicable to a variety of empirical settings, including consumer demand, multi-attribute auctions, and labor market decisions. The agent's utility in our model can depend on potential outcomes, observable characteristics, unobserved alternative-specific heterogeneity (such as product quality, compensating differentials, and other nonpecuniary factors), and random preference shocks. Incorporating nonpecuniary components into the selection model has proven essential in empirical studies (e.g., Heckman and Sedlacek, 1985; Berry, 1994; Berry et al., 1995) and has gained attention in recent theoretical research (Bayer et al., 2011; d'Haultfoeuille and Maurel, 2013b; Mourifie et al., 2020; Canay et al., 2024; Lee and Park, 2023).

Our method is applicable to a wide range of empirical applications. For example,

[2]See Athey and Haile (2007) for further discussion on the conditional independence assumption in auction models and potential testing approaches.

in a companion paper (Cosconati et al., 2024), we estimate consumer demand in the auto insurance market when only the transaction prices of selected insurance plans are observed. In this market, insurance companies employ risk-based pricing, leading to significant price variation across consumers. Our method enables nonparametric, firm-specific estimates of the offered price distribution, offering valuable insights into the heterogeneity of firms' pricing strategies and, ultimately, the precision of their risk-rating technology. In Section 6, we provide a more detailed discussion on applications to three empirical settings: consumer demand, auction models with incomplete bid information, and Roy models in labor economics, along with related literature.

The rest of the paper is organized as follows. Section 2 formally introduces our model, with an illustrative example provided at the end. Section 3 presents the main theoretical results. In Section 4, we describe the semi-parametric maximum likelihood estimator and its asymptotic properties. Section 5 reports the results of our Monte Carlo simulations, and Section 6 discusses various empirical applications. Finally, Section 7 concludes. All proofs are relegated to the appendix.

2 Model

In Sections 2–3, all analyses are conditional on observable characteristics x, which we omit to simplify notation. Throughout the paper, we use the consumer demand example to illustrate the main results and clarify ideas; however, the approach is broadly applicable to other selection models.

Consider a discrete choice problem. There is a finite set of alternatives $\mathcal{J} = \{1, \cdots, J\}$. Each alternative is associated with a price distribution. Let $G_j \in \Delta([\underline{p}_j, \overline{p}_j])$ represent the price distribution associated with alternative j, where $\Delta(Y)$ denotes the set of all cumulative distribution functions over a set $Y \subset \mathbb{R}$. We assume that $p_j \sim G_j$ are independently distributed across alternatives (conditional on x). The collection of G_j is denoted by $G = \prod_{j \in \mathcal{J}} G_j$. We refer to G as the *offered* price distribution.

A *selection function* is denoted by $f = (f_1, f_2, \cdots, f_J)$ where f_j maps the prices of alternatives $\boldsymbol{p} = (p_1, \cdots, p_J)$ to a strictly positive probability of selecting alternative

$j \in \mathcal{J}$.[3] We assume that the selection function is continuously differentiable,

$$f_j \in \mathscr{C}^1 \colon \prod_j [\underline{p}_j, \overline{p}_j] \to (0,1),$$

with $\sum_{j\in\mathcal{J}} f_j \leq 1$. Here, the inequality allows for the case with an outside option. The selection function is a primitive of the model. To provide a microfoundation, for example, f might be derived from a consumer's utility maximization problem as illustrated in Section 2.1.

Let $\boldsymbol{p}_{-j} = (p_1, \cdots, p_{j-1}, p_{j+1}, \cdots, p_J)$ denote the vector of prices excluding j's price. The probability of selecting j conditional on p_j is given by

$$Pr_j(p_j; G) = \int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k\neq j} dG_k(p_k), \tag{1}$$

where $Pr_j(\cdot; G)$ is a function defined on $[\underline{p}_j, \overline{p}_j]$. The assumption that prices are independent across different alternatives allows us to express the joint distribution of $\boldsymbol{p}_{-j}$ as the product of their individual marginal distribution functions.

Let $\tilde{G}_j \in \Delta([\underline{p}_j, \overline{p}_j])$ represent the price distribution conditional on selecting alternative j. We derive $\tilde{G}_j$ using Bayes' rule:

$$\tilde{G}_j(p) = \frac{\int_{\underline{p}_j}^{p} Pr_j(y; G) dG_j(y)}{\int_{\underline{p}_j}^{\overline{p}_j} Pr_j(y; G) dG_j(y)}. \tag{2}$$

Note that G_j and $\tilde{G}_j$ share the same support, as selection function f_j is strictly positive. Let $\tilde{G} = \prod_{j\in\mathcal{J}} \tilde{G}_j$ and we call $\tilde{G}$ *selected* price distribution. Equations (1) and (2) define a mapping from G to $\tilde{G}$. Let $F \colon \prod_j \Delta([\underline{p}_j, \overline{p}_j]) \to \prod_j \Delta([\underline{p}_j, \overline{p}_j])$ denote this mapping, i.e., $\tilde{G} = F(G)$.

In many empirical settings, researchers have access only to the selected price distribution. However, the key primitives of interest are often the offered price distribution. Our research question is how to recover the offered price distribution G from

[3] The assumption that the probability of selecting each alternative is strictly positive is analogous to the overlap assumption in the treatment effect literature, which requires each individual to have a positive probability of receiving each treatment level. This assumption is crucial for recovering the offered price distribution. To illustrate, consider a scenario where $f_j = 0$ whenever p_j falls within a certain subset of $[\underline{p}_j, \overline{p}_j]$. In this case, any p_j within that subset would not be observed in the data, making it impossible to identify G_j within that subset without introducing additional assumptions.

the observed selected price distribution $\tilde{G}$. Note that both G and $\tilde{G}$ are collections of J cumulative distribution functions. Therefore, the cardinality of unknowns and constraints are exactly the same in Equation (2) (assuming the selection function is known). Since a cumulative distribution function is an infinite-dimensional object, the key challenge is solving for a collection of infinite-dimensional objects entangled in a nonlinear system. We will explore this in detail in Section 3.

2.1 An Illustrative Example

We now present a simple example to illustrate the key assumptions of our model and compare them to the standard assumptions in the literature. Consider a consumer choosing between two products, $j = 1, 2$, to maximize her utility. The consumer's utility from product j is given by a scaler value:

$$u_j = \gamma p_j + \varepsilon_j, \tag{3}$$

where p_j represents the price of product j for this consumer, and ε_j represents an unobserved utility shock. We abstract from the possibility that the consumer's utility may depend on observable characteristics and unobserved product heterogeneity for this example. In this model, the price sensitivity parameter γ and the distribution of ε_j determine the selection function f. Let $\tilde{\varepsilon} = \varepsilon_1 - \varepsilon_2$ denote the error difference. If $\tilde{\varepsilon} \sim \mathcal{N}(0, 1)$, this represents a binary probit model, and the selection function for product 1 takes the following form:

$$f_1(p_1, p_2) = 1 - \Phi_{\mathcal{N}}\left(\gamma(p_2 - p_1)\right),$$

where $\Phi_{\mathcal{N}}$ denotes the CDF for standard normal distribution.

In this illustrative example, we consider a simple linear outcome equation with an additive error term. For each product $j = 1, 2$, the price is generated by the following equation:

$$p_j = x\beta_j + \eta_j, \tag{4}$$

where x represents observable characteristics, and η_j denotes a random shock, which, for simplicity, is assumed to be independent of x.

Suppose the econometrician observes the price of product 1 only when it is chosen by the consumer. We derive the conditional mean of p_1 given that it is observed:

$$\begin{aligned} E(p_1|x, u_1 > u_2) &= x\beta_1 + E(\eta_1|\gamma p_1 + \varepsilon_1 - (\gamma p_2 + \varepsilon_2) > 0) \\ &= x\beta_1 + E(\eta_1|x\underbrace{\gamma(\beta_1 - \beta_2)}_{\beta^*} + \underbrace{[\gamma(\eta_1 - \eta_2) + \tilde{\varepsilon}]}_{\text{composite error: } \varepsilon^*} > 0) \\ &= x\beta_1 + E(\eta_1|x\beta^* + \varepsilon^* > 0). \end{aligned} \tag{5}$$

The conditioning term $x\beta^* + \varepsilon^* > 0$ in Equation (5) represents the reduced-form selection model typically seen in the literature. Sample selection issue arises when η_1 and ε^* are correlated, so that $E(\eta_1|x\beta^* + \varepsilon^* > 0) \neq 0$. In the two-step estimation literature, researchers often impose assumptions on the joint distribution of $(\varepsilon^*, \eta_1, \eta_2)$. For example,

$$\begin{bmatrix} \varepsilon^* \\ \eta_1 \\ \eta_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \right).$$

We now take a closer look at the correlation between the composite error (ε^*) and the error in the outcome equation (η_1). Specifically,

$$\begin{aligned} cov(\varepsilon^*, \eta_1) &= cov(\gamma(\eta_1 - \eta_2) + \tilde{\varepsilon}, \eta_1) \\ &= \gamma var(\eta_1) - \gamma cov(\eta_1, \eta_2) + cov(\eta_1, \tilde{\varepsilon}). \end{aligned} \tag{6}$$

Equation (6) shows that the error term η_1 directly enters the composite error ε^*, implying that $cov(\varepsilon^*, \eta_1) \neq 0$ unless $\gamma = 0$. This correlation is *by construction* in selection models, as agents make decisions after observing the potential outcomes. Another common concern regarding selection bias arises from potential correlation between errors in the outcome equation (e.g., η_1) and those in the structural selection model (e.g., $\tilde{\varepsilon}$), as represented by the third term in Equation (6). For example, unobserved productivity factors may create correlation between a worker's willingness to work and their wage. Our model also accommodates this type of correlation.

The only assumption we impose is that the error terms in outcome equations across different alternatives are independent conditional on observables. This implies that $cov(\eta_1, \eta_2) = 0$ in Equation (6). In a simple binary model with only one dependent

variable of interest, such as Tobit Type 1 or Type 2, this assumption holds trivially. Heckman and Honore (1990) show that, under a strong log-normality assumption, the correlation structure between two outcome variables can be identified; however, this result does not hold more generally (see discussions in French and Taber, 2011). Due to the nature of the selection problem, the data include only the price of the selected alternative, while competing prices for unselected alternatives are not observed. If the prices of the two products tend to move together, we would not be able to observe this pattern. French and Taber (2011) point out that since the data provides only two one-dimensional price distributions, it is impossible to recover the full joint distribution of a two-dimensional object without imposing additional assumptions.

The conditional independence assumption is commonly employed in auction models, such as independent private value auctions or mineral rights models, where signals are assumed to be independent given the common value. This assumption is more plausible when econometricians have access to a rich set of observables. The conditional independence assumption essentially rules out the presence of a common unobserved factor, x^*, that introduces correlation between outcomes, even after conditioning on observables. When this assumption is not satisfied, the observed price distribution for each alternative, conditional on observable x, is a mixture of price distributions conditional on (x, x^*). We then need to first analyze this mixture model and use additional parametric structures or instruments to identify the selected price distributions conditional on (x, x^*). Techniques for this type of deconvolution problem have been studied in the literature (see the recent survey articles by Compiani and Kitamura, 2016; Hu, 2017) and are beyond the scope of this paper. We maintain the conditional independence assumption for the remainder of the paper.

Finally, we highlight several additional features that differentiate our model from the existing literature. First, our model allows the outcome equation to be fully flexible and nonparametrically specified as $p_j = h_j(x, \eta_j)$, where h_j is an unknown function that may be nonseparable in the error term. Our goal is to recover the entire distribution of p_j conditional on x, rather than only estimating the parameters in the conditional mean function, such as β_j in Equation (4). Importantly, we fully account for heterogeneity in the effects of covariates on outcomes. Second, our model does not require an instrument that exogenously shifts choices between alternatives and is excluded from the outcome equation—a critical requirement for identification and estimation in the two-step method. In other words, we allow the same set of

observables to enter both the outcome and selection equations. Moreover, we impose minimal assumptions on the selection function. It can accommodate nonparametric, nonseparable relationships between observable and unobserved errors, offering much greater flexibility than the utility specification in Equation (3); in fact, it *does not* even need to be derived from a utility maximization problem. Our framework also allows for alternative-specific unobserved heterogeneity, which is a desirable feature in many empirical contexts.

3 Main Results

Given the observed selected price distribution, we want to recover the offered price distribution. As the selected price distribution is derived from the offered price distribution through Bayes' rule in Equation (2), we can first invert Equation (2):

$$G_j(p_j) = \frac{\int_{\underline{p}_j}^{p_j} d\tilde{G}_j(p)/Pr_j(p;G)}{\int_{\underline{p}_j}^{\overline{p}_j} d\tilde{G}_j(p)/Pr_j(p;G)}.$$

Note that if we know $Pr_j(\cdot;G)$, the selection probability conditional on j's price, then we can easily recover the offered price distribution. However, $Pr_j(\cdot;G)$ also depends on the offered price distribution G which is unknown. A tentative solution is to plug in a conjecture Ψ of the offered price distribution to obtain a conjectured selection probability $Pr_j(\cdot;\Psi)$. Then the equation above can give us a new conjecture of the offered price distribution. This procedure of obtaining a new conjecture of offered price distribution from an old conjecture defines an operator $T\colon \prod_j \Delta([\underline{p}_j, \overline{p}_j]) \to \prod_j \Delta([\underline{p}_j, \overline{p}_j])$ as follows.

$$(T\Psi)_j(p_j) = \frac{\int_{\underline{p}_j}^{p_j} d\tilde{G}_j(p)/Pr_j(p;\Psi)}{\int_{\underline{p}_j}^{\overline{p}_j} d\tilde{G}_j(p)/Pr_j(p;\Psi)} \tag{7}$$

where $\Psi = (\Psi_1, \Psi_2, \cdots, \Psi_J) \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$.

Importantly, if the conjecture Ψ is correct, i.e., $\Psi = G$, the new conjecture $T\Psi$ will also equal G. Thus, the offered price distribution G a fixed point of the operator T.

The operator T is a contraction if there exists some real number $0 \leq \rho < 1$ such

that for all $\Psi, \Phi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$D(T\Psi, T\Phi) \leq \rho D(\Psi, \Phi),$$

given some metric D.[4] In the reminder of this section, we first construct the metric D and then characterize the modulus ρ. We discuss several special cases of our model at the end.

3.1 Constructing the Metric

We begin by defining a metric in the set of all cumulative distribution functions for alternative j. Let Ψ_j and Φ_j denote two probability measures in $\Delta([\underline{p}_j, \overline{p}_j])$. Recall that two probability measures Ψ_j and Φ_j are equivalent, denoted $\Psi_j \sim \Phi_j$, if they are absolutely continuous with respect to each other. When $\Psi_j \sim \Phi_j$, the Radon-Nikodym derivative,

$$\frac{d\Psi_j}{d\Phi_j} \colon [\underline{p}_j, \overline{p}_j] \to (0, \infty),$$

exists, as guaranteed by the Radon-Nikodym Theorem. If both Ψ_j and Φ_j have continuous densities, the Radon-Nikodym derivative simplifies to the ratio of densities:

$$\frac{d\Psi_j}{d\Phi_j}(p) = \frac{\Psi_j'(p)}{\Phi_j'(p)}.$$

Note that

$$\Psi_j = \Phi_j \quad \Leftrightarrow \quad \frac{d\Psi_j}{d\Phi_j}(p) = 1 \qquad \Phi_j\text{-a.e.}$$

In the space $\Delta([\underline{p}_j, \overline{p}_j])$, we define a metric $d \colon \Delta([\underline{p}_j, \overline{p}_j]) \times \Delta([\underline{p}_j, \overline{p}_j]) \to [0, +\infty]$ to simplify the analysis.[5]

$$d(\Psi_j, \Phi_j) = \begin{cases} \ln \operatorname{ess\,sup}_{y \in [\underline{p}_j, \overline{p}_j]} \frac{d\Psi_j}{d\Phi_j}(y) + \ln \operatorname{ess\,sup}_{y \in [\underline{p}_j, \overline{p}_j]} \frac{d\Phi_j}{d\Psi_j}(y), & \text{if } \Psi_j \sim \Phi_j, \\ +\infty & \text{otherwise.} \end{cases}$$

[4] We adopt the convention that $+\infty$ and $+\infty$ are not comparable, but $c < +\infty$ for any $c \in \mathbb{R}_+$.

[5] This metric is a variant of the Thompson metric (Thompson, 1963). The Thompson metric between two functions $s, q \in \mathbb{R}^Y$ is

$$d_{Thompson}(s, q) = \max\{\ln \sup \frac{s(y)}{q(y)}, \ln \sup \frac{q(y)}{s(y)}\}.$$

Given our operator T in Equation (7), for all $\Psi_j, \Phi_j \in \Delta([\underline{p}_j, \overline{p}_j])$,

$$(T\Psi)_j \sim \tilde{G}_j \sim (T\Phi)_j.$$

Thus,

$$d((T\Psi)_j, (T\Phi)_j) = \ln \operatorname*{ess\,sup}_{p_j} \frac{d(T\Psi)_j}{d(T\Phi)_j}(p_j) + \ln \operatorname*{ess\,sup}_{p_j} \frac{d(T\Phi)_j}{d(T\Psi)_j}(p_j).$$

The observed selected price distribution $\tilde{G}_j$ appears in both $(T\Psi)_j$ and $(T\Phi)_j$. As a result, $\tilde{G}_j$ cancels out in the distance above. Moreover, the denominator in our operator is a normalizing factor, which is also canceled out after we take the sum of log ratios. Consequently, the distance between $(T\Psi)_j$ and $(T\Phi)_j$ only relies on the ratio between selection probabilities:

$$d((T\Psi)_j, (T\Phi)_j) \leq \sup_{p_j} \ln \frac{Pr_j(p_j; \Psi)}{Pr_j(p_j; \Phi)} + \sup_{p_j} \ln \frac{Pr_j(p_j; \Phi)}{Pr_j(p_j; \Psi)},$$

where equality holds when $\tilde{G}_j$ admits full support on $[\underline{p}_j, \overline{p}_j]$.

Next, we define a metric in the space $\prod_j \Delta([\underline{p}_j, \overline{p}_j])$ by taking the maximum distance among all alternatives:

$$D(\Psi, \Phi) = \max_{j \in \mathcal{J}} d(\Psi_j, \Phi_j)$$

for any $\Psi, \Phi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$. From now on, we work with the metric space $(\prod_j \Delta([\underline{p}_j, \overline{p}_j]), D)$.

3.2 Functional Contraction

For $j \in \mathcal{J}$, we define the *maximum semi-elasticity difference* as

$$M_j = \sup_{p_j, \boldsymbol{p}_{-j}, \boldsymbol{p}'_{-j}} \left| \frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} - \frac{\partial \ln f_j(p_j, \boldsymbol{p}'_{-j})}{\partial p_j} \right|. \tag{8}$$

The quantity $\frac{\partial \ln f_j}{\partial p_j}$ measures how sensitive the log of the choice probability changes with respective to the price, and therefore represents the semi-elasticity. Let

$$\rho = \frac{J-1}{4} \max_{j \in \mathcal{J}} (\overline{p}_j - \underline{p}_j) M_j.$$

Theorem 1. *If $\rho < 1$, the operator T is a contraction with modulus less than ρ.*

Proof. See Appendix B.1. □

By the Banach fixed point theorem, whenever $\rho < 1$, any selected distribution $\tilde{G}$ corresponds to a unique offered distribution G. Theorem 1 implies that we can nonparametrically identify the potential outcome distributions G from the observed selected outcome distribution $\tilde{G}$, given the selection function f. Moreover, this result provides a constructive method for solving G. Take any $\Psi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$, by Theorem 1,

$$D(T^n\Psi, G) = D(T^n\Psi, TG) \leq \rho D(T^{n-1}\Psi, G) \leq \rho^{n-1} D(T\Psi, G),$$

where $D(T\Psi, G)$ is finite. This implies

$$\lim_{n\to\infty} D(T^n\Psi, G) = 0.$$

$$\lim_{n\to\infty} T^n\Psi = G.$$

Thus, we can simply take an initial guess for the potential outcome distributions and iteratively apply the operator. As the number of iterations approaches infinity, this process converges to the potential outcome distributions associated with the selection function.

Note that the condition of Theorem 1 is a joint constraint on the selection function and the price range. The bound on the modulus, ρ, consists of the product between the number of alternatives, the price range $\overline{p}_j - \underline{p}_j$, and the maximum semi-elasticity difference.[6] Our condition requires this product to be small. If we expand the support $[\underline{p}_j, \overline{p}_j]$ to $[\underline{p}'_j, \overline{p}'_j]$ where

$$\underline{p}'_j < \underline{p}_j < \overline{p}_j < \overline{p}'_j$$

with $\tilde{G}$ unchanged, ρ becomes weakly larger, which implies now it is more difficult for the operator T to contract. This comparison is intuitive. The larger domain $\prod_{k\neq j} \Delta([\underline{p}_k, \overline{p}_k]) \times \Delta([\underline{p}'_j, \overline{p}'_j])$ nests more collections of probability measures, making it more challenging to control $\frac{D(T\Psi, T\Phi)}{D(\Psi, \Phi)}$ for all Ψ and Φ in this domain.

[6] Note that by definition ρ is unitless. Changing the unit of price does not affect ρ. Moreover, the bound ρ is relatively tight: there exist selection functions for which the supremum is arbitrarily close ρ.

To understand the maximum semi-elasticity difference M_j in the modulus ρ, consider an extreme case where the choice probabilities do not vary with prices at all, indicating perfectly inelastic demand. In this scenario, there is effectively no selection and the offered price distribution coincides with the selected price distribution. The modulus equals 0 and we obtain the fixed point immediately.

It may be a concern that a large number of alternatives J would result in a large modulus. However, we show that a large number of alternatives could lead to a small maximum semi-elasticity difference. For example, consider the multinomial logit model, arguably the most popular model for discrete choices due to its analytical form and ease of estimation:

$$f_j(p_1, \cdots, p_J) = \frac{\exp(\gamma p_j)}{\sum_{k=1}^{J} \exp(\gamma p_k)}, \tag{9}$$

where γ represents the consumer's price sensitivity. We derive the semi-elasticity for the logit model,

$$\frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} = \gamma(1 - f_j(\boldsymbol{p})).$$

When J is large, the choice probability for each alternative tends to be small, so that the log derivative is approximately equal to γ. As a result, the maximum semi-elasticity difference is close to 0.

The crux and the bulk of the proof for Theorem 1 is to provide a bound on the ratio

$$\sup_{\Psi, \Phi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])} \frac{D(T\Psi, T\Phi)}{D(\Psi, \Phi)}.$$

This is difficult as the domain of the supreme, $\prod_j \Delta([\underline{p}_j, \overline{p}_j])$, is a large space. For instance, if $J = 10$, the supreme is over 20 functions. In the proof of this theorem in Appendix B.1, we employ a technique called a change of measure, also know as the tilted measure, and combine it with insights from transportation problem. In Appendix A, we connect our contraction result with quantal response equilibria (McKelvey and Palfrey, 1995).

3.3 Special Cases

Thus far, we have not imposed any structure on the selection function. For a general selection function, we have to take the supreme over $\boldsymbol{p}_{-j}, \boldsymbol{p}'_{-j}$ to compute the

maximum semi-elasticity difference. Now we impose an assumption on the selection function to determine where the supreme is attained.

Assumption 1 (Log Supermodularity)**.** For all $j \in \mathcal{J}$ and $p_j \in [\underline{p}_j, \overline{p}_j]$, $\frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j}$ is weakly increasing in each p_k with $k \neq j$.

Given log supermodularity, the maximum semi-elasticity difference is attained at the boundary,

$$M_j = \sup_{p_j} \left| \frac{\partial \ln f_j(p_j, \overline{\boldsymbol{p}}_{-j})}{\partial p_j} - \frac{\partial \ln f_j(p_j, \underline{\boldsymbol{p}}_{-j})}{\partial p_j} \right|.$$

What is left in the definition of maximum semi-elasticity difference is the supreme over p_j. It turns out that we can use $\overline{p}_j - \underline{p}_j$ in the definition of ρ to eliminate the supreme over p_j and give a tighter bound. The result is as follows,

$$\rho^* = \frac{J-1}{4} \max_{j \in \mathcal{J}} [\ln f_j(\overline{\boldsymbol{p}}) - \ln f_j(\underline{p}_j, \overline{\boldsymbol{p}}_{-j}) - \ln f_j(\overline{p}_j, \underline{\boldsymbol{p}}_{-j}) + \ln f_j(\underline{\boldsymbol{p}})].$$

Theorem 2. *Suppose that Assumption 1 holds. If $\rho^* < 1$, the operator T is a contraction with modulus less than ρ^*.*

Proof. See Appendix B.2. □

Under Assumption 1, the modulus ρ^* takes a much simpler form and is straightforward to compute. The log-supermodularity assumption holds in models widely adopted by empirical researchers. For example, the multinomial logit model satisfies Assumption 1. Another example is the binary probit model we describe in Section 2.1. The log-supermodularity condition in Assumption 1 holds for the binary probit model and Theorem 2 applies.[7] However, Assumption 1 may not hold for probit models with three or more alternatives; in such cases, the more general results in Theorem 1 can be applied.

[7] To see this, we compute the log derivative for the binary probit model:

$$\frac{\partial \ln f_1(p_1, p_2)}{\partial p_1} = \frac{\gamma \phi_{\mathcal{N}}(\Delta)}{1 - \Phi_{\mathcal{N}}(\Delta)},$$

$$\frac{\partial^2 \ln f_1(p_1, p_2)}{\partial p_1 \partial p_2} = \gamma^2 \frac{d}{d\Delta} \left[\frac{\phi_{\mathcal{N}}(\Delta)}{1 - \Phi_{\mathcal{N}}(\Delta)} \right],$$

where $\Delta = \gamma(p_2 - p_1)$ and the term in the square bracket is known as the hazard rate or inverse Mills ratio. As Gaussian satisfies increasing hazard rate (Baricz, 2008), the log-supermodularity condition in Assumption 1 holds.

The simple form of ρ^* allows us to easily check when $\rho^* < 1$. For instance, consider the multinomial logit model (9) with two alternatives. If two alternatives are symmetric, i.e., $\underline{p}_1 = \underline{p}_2$ and $\overline{p}_1 = \overline{p}_2$, then a price range as large as $\overline{p}_1 - \underline{p}_1 = 5.4\frac{1}{\gamma}$ can allow for $\rho^* < 1$. If the two alternatives are more asymmetric, then the price range can be larger. For instance, if $\overline{p}_2 - \underline{p}_2 = \overline{p}_1 - \underline{p}_1$ and $\underline{p}_2 - \underline{p}_1 = 3\frac{1}{\gamma}$, then price range $7.1\frac{1}{\gamma}$ still allows $\rho^* < 1$. The analytic form of ρ^* might lead one to expect that f_j must be sufficiently bounded away from 0, which excludes applications where some alternative has a very small market share. However, this intuition is incorrect. In the two alternative example, a very small market share stems from a large asymmetry, which actually allows for a larger price range. We can also check what happens with more alternatives. With three symmetric alternatives, the price range can be up to $3.4\frac{1}{\gamma}$. Although the semi-elasticity difference decreases, it is outweighed by the increases in $\frac{J-1}{4}$. Importantly, note that we are checking a sufficient condition. Even if this condition is violated, the operator may still be a contraction.

To summarize, our contraction results provide a novel method for identifying the potential outcome distribution from the observed selected outcome distribution, given any selection function f—whether parametric or nonparametric, and regardless of whether it is microfounded in a utility maximization problem. Moreover, the identification is constructive: starting with an initial guess, iterative application of the operator converges to the potential outcome distributions associated with the selection function. These theoretical results are essential for estimating the selection function and potential outcome distributions, which will be discussed in the next section.

4 Estimation

Building on the theoretical results in Section 3, we now turn to the estimation of the model's primitives. We begin by discussing the estimation of the offered price distribution G when the selection function f is known, followed by the more complex case where both f and G must be jointly estimated.

In the data, for each individual i, we observe their choice, characteristics, and the price of the selected product. Let $y_{ij} = 1$ if j is chosen by i, and 0 otherwise. Since the alternatives are exclusive, $\sum_{j=1}^{J} y_{ij} = 1$. Let x_{ij} represent a vector of observable characteristics. We define $y_i = (y_{i1}, \cdots, y_{iJ})'$ and $x_i = (x_{i1}', \cdots, x_{iJ}')' \in X$. The observed

selected prices in the data enable us to estimate $\tilde{G}$ using standard nonparametric methods. Let $\hat{G}$ denote the estimate of $\tilde{G}$, and $\hat{G}(x)$ denote the estimate conditional on observable x.

4.1 Estimation with a Known Selection Function

In Section 3, we show that for a given selection function f, the offered price distribution G can be uniquely determined from the selected price distribution $\tilde{G}$, as the number of iterations of the operator T defined in Equation (7) goes to infinity. In practice, however, econometricians typically do not observe the true selected price distribution $\tilde{G}$, but rather an estimate $\hat{G}$, which is subject to sampling errors. Moreover, when iterating the operator T to obtain the offered price distribution G, the process stops after a finite number of iterations m. Therefore, our estimation of G contains these two sources of error.

Let $T^m_{\hat{G}}\Psi$ denote our estimator for G, using the estimated selected price distribution $\hat{G}$ and initiating the operator iteration with $\Psi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$. The distance between our estimator and the true G is bounded by the sum of sampling errors from a finite sample size and the approximation errors from finite iterations, as shown in the following triangular inequality.

$$D(G, T^m_{\hat{G}}\Psi) \leq \underbrace{D(G, F^{-1}(\hat{G}))}_{\text{finite sample size}} + \underbrace{D(F^{-1}(\hat{G}), T^m_{\hat{G}}\Psi)}_{\text{finite iteration}},$$

where F^{-1} denotes the inverse of F. Recall that F is the mapping from G to $\tilde{G}$ defined in Equations (1) and (2). The inverse mapping, F^{-1}, maps $\tilde{G}$ back to G. Theorem 1 guarantees that we can obtain G from $\tilde{G}$ by iterating the operator T an infinite number of times.

We first focus on the sampling error $D(G, F^{-1}(\hat{G}))$. The next proposition shows that this error goes to zero as $\hat{G}$ converges to $\tilde{G}$.

Proposition 1. *Suppose $\rho < 1$. The mapping F is a homeomorphism. Moreover, both F and F^{-1} are Lipschitz continuous, with Lipschitz constants $1 + \rho$ and $\frac{1}{1-\rho}$, respectively.*

Proof. See Appendix B.3. □

Since F is a homeomorphism, the inverse F^{-1} is well-defined and $G = F^{-1}(\tilde{G})$. Since F^{-1} is continuous, we have

$$F^{-1}(\hat{G}) \xrightarrow{p} F^{-1}(\tilde{G}) \quad \text{as} \quad \hat{G} \xrightarrow{p} \tilde{G}.$$

Moreover, as F^{-1} is Lipschitz continuous, $F^{-1}(\hat{G})$ converges to G at the same rate as $\hat{G}$ converges to $\tilde{G}$.

We now analyze the approximation error $D(T^m_{\hat{G}}\Psi, F^{-1}(\hat{G}))$ due to the finite number of iterations. Note that this error term tends to 0 at speed ρ^m. Thus, if ρ^m decays faster than the convergence rate of $\hat{G}$ to $\tilde{G}$, then $T^m_{\hat{G}}\Psi$ converges to G at the same rate as $\hat{G}$ converges to $\tilde{G}$. We let $m(n)$ express the dependence of the number of iterations on the sample size. The following result summarizes the discussion above.

Corollary 1. *Suppose that $\hat{G} \xrightarrow{p} \tilde{G}$ at a polynomial rate of n^k with $k > 0$. If*

$$\liminf_{n\to+\infty} \frac{m(n)}{\ln n} > k(\ln(1/\rho))^{-1},$$

then $T^{m(n)}_{\hat{G}}\Psi \xrightarrow{p} G$ at rate n^k.

For instance, if the support of $\tilde{G}$ is finite, $\hat{G} \to \tilde{G}$ at rate $\sqrt{n}$. If

$$\lim_{n\to\infty} \rho^{m(n)}\sqrt{n} = 0 \quad \text{or} \quad \liminf_{n\to+\infty} \frac{m(n)}{\ln n} > \frac{1}{2}(\ln(1/\rho))^{-1},$$

$T^m_{\hat{G}}\Psi$ converges to G at rate $\sqrt{n}$.

4.2 Estimation with an Unknown Selection Function

We now consider the case where the selection function f is unknown to econometricians and we jointly estimate f and G. As discussed in Section 3, given any selection function f, whether parametric or nonparametric, our contraction results provide a straightforward method for recovering the potential outcome distribution from the observed selected outcome distribution. This step utilizes all the information contained in the selected outcome distribution. To further identify and estimate the selection function f, we must leverage additional data, specifically the "market share" of each alternative.

The dimensionality of market shares determines how flexibly we can estimate f. For example, if market shares are observed conditional on continuously distributed covariates, it is possible to estimate a semiparametric single-index model (Ichimura, 1993; Klein and Spady, 1993) for the selection function f. While allowing for a semiparametric or nonparametric selection function is theoretically possible, implementing it would be highly complex and data-intensive. In most empirical settings, market shares are observed conditional on discrete values of covariates. We therefore focus on the case where the selection function is parametrically specified in the estimation.[8]

We assume that the selection function f is derived from a standard multinomial choice model with an indirect utility given by

$$u_{ij} = v_j(p_{ij}, x_{ij}, \varepsilon_{ij}; \theta),$$

where v_j is a known function parametrized by a finite-dimensional parameter θ; p_{ij} denotes the offered price of alternative j for individual i; the vector of unobserved error terms $\varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{iJ})$ is jointly distributed according to a known distribution. Note that our framework fully allows that the unobserved error term enters the utility function in a nonseparable way. The individual chooses an alternative to maximize utility, and the selection function f is captured by the parameter θ. Let θ_0 denote the true parameter. For example, one commonly used specification is as follows:

$$u_{ij} = \gamma p_{ij} + x_{ij}'\beta + \xi_j + \varepsilon_{ij}, \quad j = 1, 2, \cdots, J,$$

where ξ_j represents a scalar-valued unobserved characteristic of alternative j. In this example, $\theta = (\gamma, \beta, \boldsymbol{\xi})$, where $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_J)$.

We estimate the parameter θ in the selection function by matching the model-implied choice probabilities to those observed in the data. Specifically, for an individual with observable characteristic x_i, the probability of choosing alternative j is given by the following equation:

$$Prob_j(x; \theta, \hat{G}, m) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta) d\big(T^m_{\hat{G}(x),\theta}\Psi\big)(\boldsymbol{p}), \tag{10}$$

[8] In our Monte Carlo simulations, we consider a scenario where the selection function is misspecified by the econometrician. We find that our estimates of the potential outcome distributions remain quite robust even when the selection function is misspecified.

where $T^m_{\hat{G}(x),\theta}\Psi$ represents the estimated offered price distribution after iterating the operator T for m steps, starting with the initial value Ψ. The operator is constructed using the estimated selected price distribution conditional on x, denoted by $\hat{G}(x)$, and the selection function parameterized by θ. Note that θ affects the choice probabilities both directly through the selection function and indirectly through the estimated offered price distribution.

Let $z_i = \{x_i, y_i\}$. Given an i.i.d. sample of $\{z_i\}_{i=1}^n$ and a first-step nonparametric estimator $\hat{G}(x)$, we propose a semiparametric maximum likelihood estimator for θ:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \hat{Q}_n(\theta), \tag{11}$$

where

$$\hat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij} \ln Prob_j(x_i; \theta, \hat{G}, m(n)). \tag{12}$$

Once $\hat{\theta}$ is obtained, a plug-in estimator for G is given by $T^{m(n)}_{\hat{G},\hat{\theta}}\Psi$.

4.3 Consistency and Asymptotic Normality

In this section, we show that the estimator defined in Equation (11) is consistent and asymptotically normal. We maintain the previous assumptions on the selection function: $f_j \in \mathscr{C}^1$: $\prod_j [\underline{p}_j, \overline{p}_j] \to (0,1)$. The additional technical conditions required for the consistency of $\hat{\theta}$ are as follows.

Assumption 2. (i) The space Θ of parameter θ is compact; (ii) for each $x \in X$, the selection function $f(\boldsymbol{p}; x, \theta)$ is jointly continuous in θ and $\boldsymbol{p}$; (iii) the condition in Theorem 1 holds for all $\theta \in \Theta$, that is, $\sup_{\theta\in\Theta} \rho(\theta) \leq \bar{\rho} < 1$ for some $\bar{\rho}$; (iv) the number of iterations $m(n) \to \infty$; (v) $\hat{G} \xrightarrow{p} \tilde{G}$.

Assumption 3 (Identification)**.** There does not exist $\theta' \in \Theta$, $\theta' \neq \theta_0$, offered price distributions G, $G' \in \left(\prod_j \Delta([\underline{p}_j, \overline{p}_j])\right)^X$ such that for all $j \in \mathcal{J}$ and $x \in X$

$$F(G(x); \theta_0) = F(G'(x); \theta'),$$

$$\int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta_0) dG(x)(\boldsymbol{p}) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta') dG'(x)(\boldsymbol{p}).$$

Assumption 2 (i) and (ii) are standard regularity conditions. Assumption 2 (iii) ensures that for all $\theta \in \Theta$, the operator T is a contraction. Assumption 2 (iv) requires that the number of iterations m tends to infinity, but it does not impose any restrictions on the rate at which m approaches infinity. Assumption 2 (v) ensures that our first-step estimator $\hat{G}$ is consistent. Assumption 3 imposes the identification condition, which requires that there does not exist another parameter that can yield the same selected price distribution and choice probabilities.

Theorem 3 (Consistency)**.** *Under Assumptions 2 and 3,* $\hat{\theta} \xrightarrow{p} \theta_0$, $T^{m(n)}_{\hat{G},\hat{\theta}}\Psi \xrightarrow{p} G$.

Proof. See Appendix B.3. □

Proving this theorem turns out to be challenging. We cannot rely on the standard consistency arguments for maximum likelihood estimators, as $\hat{Q}_n(\theta)$ is not a sample average. Since all data points are already used to estimate $\hat{G}$, each term in $\hat{Q}_n(\theta)$ depends on the entire dataset. Moreover, the number of iterations depends on the sample size n.

To prove consistency, we invoke the fundamental consistency theorem for extremum estimators (Theorem 2.1 in Newey and McFadden (1994)). We construct the true population objective function as follows:

$$Q_0(\theta) = \mathbb{E}_x \sum_{j=1}^{J} \left(\int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta_0) dG(x)(\boldsymbol{p}) \right) \ln \left(Prob^*_j(x; \theta, \tilde{G}) \right),$$

where $\int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta_0) dG(x)(\boldsymbol{p})$ represents the true probability of selecting alternative j conditional on x; and

$$Prob^*_j(x; \theta, \tilde{G}) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; x, \theta) dF^{-1}(\tilde{G}(x), \theta)(\boldsymbol{p}). \tag{13}$$

Equation (13) represents the model-implied choice probability for alternative j conditional on x, given the model parameter θ, the true selected price distribution $\tilde{G}$, and as the number of iterations goes to infinity. By the identification condition in Assumption 3, Q_0 is uniquely maximized at θ_0.

Similarly to Section 4.1, there are two sources of error in the sample objective function $\hat{Q}_n(\theta)$ when approximating the true population objective function $Q_0(\theta)$: (1) sampling error, and (2) errors resulting from the finite number of iterations of the

operator T. To focus on the sampling error, we construct the following intermediate objective function where the number of iterations m in Equation (12) goes to infinity:

$$\hat{Q}_n^*(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij} \ln\left(Prob_j^*(x_i;\theta,\hat{G})\right).$$

We use the homeomorphism in Proposition 1 to show that $\hat{Q}_n^*$ converges pointwise to Q_0 in probability. We then prove that $\hat{Q}_n^*$ is equicontinuous, which ensures its uniform convergence to Q_0. Lastly, we show that $\hat{Q}_n$ converges uniformly in probability to $\hat{Q}_n^*$ as the number of iterations approaches infinity, which implies that $\hat{Q}_n$ converges uniformly in probability to Q_0, a key to establishing the consistency result. Further details of each step can be found in Appendix B.3.

Next, we show that the estimator defined in Equation (11) is asymptotically normal. Motivated by our discussion above, we first study the behavior of the estimator when m tends to infinity for each n. Let

$$\hat{\theta}^* = \arg\max_{\theta} \hat{Q}_n^*(\theta),$$

$$\mathfrak{g}^*(z_i;\theta,\hat{G}) = \nabla_\theta\left(\sum_{j=1}^{J} y_{ij} \ln Prob_j^*(x_i;\theta,\hat{G})\right),$$

where ∇_θ denote the gradient operator with respect to θ. The estimator $\hat{\theta}^*$ solves the first-order condition

$$\frac{1}{n}\sum_{i=1}^{n} \mathfrak{g}^*(z_i;\hat{\theta}^*,\hat{G}) = 0.$$

Proving the asymptotic normality of a semiparametric two-step estimator typically requires a first-order expansion around the nonparametric estimator (see Theorem 8.1 in Newey and McFadden (1994)). In our case, this involves expanding the equation above around $\hat{G}$. A standard argument would apply if $\hat{G}$ entered directly into Equation (13). However, it enters through F^{-1}, for which we lack an analytic form. As a result, continuing to work with an infinite-dimensional distribution $\tilde{G}$ becomes extremely challenging.

To make the analysis tractable, we assume that the support of $\tilde{G}$ is finite. This assumption is practically innocuous, as nonparametric estimators are always represented as finite-dimensional vectors in numerical applications. For instance, in con-

sumer demand estimation, $\tilde{G}$ represents a distribution over prices, which are measured in discrete units (e.g., cents), so this assumption is reasonable.

Assumption 4. (i) $\text{supp}(\tilde{G})$ is finite. (ii) θ_0 is in the interior of Θ. (iii) f is twice continuously differentiable in θ. (iv) $\mathbb{E}\nabla_\theta \mathfrak{g}^*(z;\theta_0,\tilde{G})$ is nonsingular. (v) The number of iterations satisfies $\liminf_{n\to+\infty} \frac{m(n)}{\ln n} > \frac{1}{2}(\ln(1/\bar{\rho}))^{-1}$.

Assumption 4(ii)–(iv) are standard regularity conditions. Assumption 2–4(iv) ensure that the estimator $\hat{\theta}^*$ is asymptotically normal. Assumption 4(v) requires that the number of iterations increases rapidly enough for the error introduced by finite iterations to become negligible compared to the error of $\hat{\theta}^*$. Particularly, it guarantees $\sqrt{n}(\hat{\theta} - \hat{\theta}^*) \xrightarrow{p} 0$, which gives us the next result.

Theorem 4 (Asymptotic Normality)**.** *Suppose that Assumption 2, 3, and 4 hold. Then $\hat{\theta}$ is asymptotically normal and $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$.*[9] *$T^{m(n)}_{\hat{G},\hat{\theta}}\Psi$ converges to G in probability at rate $\sqrt{n}$.*

Proof. See Appendix B.4. □

5 Monte Carlo Simulations

To examine how our estimator for θ and the offered price distribution may perform in practice, we conduct a Monte Carlo simulation experiment with $J = 2$. The utility individual i derives from the two alterantives are specified as follows:

$$u_{i1} = -\gamma \log(p_{i1}) + \xi_1 + \beta x_{i1} + \varepsilon_i,$$
$$u_{i2} = -\gamma \log(p_{i2}) + \xi_2,$$

where p_{ij} and ξ_j are, respectively, the offered price and unobserved heterogeneity for alternative j; $x_{i1} \in \{0, 1\}$ is a binary observable with $Pr(x_{i1} = 1) = 0.5$ that shifts individual i's choice probabilities; and $\varepsilon_i \sim N(0, 1)$ is the error term. Throughout the simulation exercises, we set the utility parameters as follows: $\gamma = 1$, $\xi_1 = 0$, $\xi_2 = 1$, $\beta = 0.5$. Let $y_i \in \{1, 2\}$ denote the choice of individual i.

We consider five data generating processes for the offered prices. Let x_{i2} denote the observable characteristic of individual i that enters the pricing equation. For

[9] See the analytical form of V in the proof of Lemma 7.

simplicity, we also restrict x_{i2} to take binary values from $\{0, 1\}$, with $Pr(x_{i2} = 1) = 0.7$.

DGP 1: $\log(p_{ij}) = \delta_{0j} + \delta_j x_{i2} + \eta_{ij}$, where $\delta_{01} = 0.2, \delta_1 = 0.5, \eta_{i1} \sim N(0, 0.1)$, $\delta_{02} = 0.1, \delta_2 = 1, \eta_{i2} \sim N(0, 0.2)$.

DGP 2: $\log(p_{ij}) = \delta_{0j} + \delta_j x_{i2} + \eta_{ij}$, where $\delta_{01} = 0.2, \delta_1 = 0.5, \eta_{i1} \sim EV(0, 0.1)$, $\delta_{02} = 0.1, \delta_2 = 1, \eta_{i2} \sim EV(0, 0.2)$.

DGP 3: $\log(p_{ij}) = (\delta_{0j} + \delta_j x_{i2})(1 + \eta_{ij})$, where $\delta_{01} = 0.2, \delta_1 = 0.5, \eta_{i1} \sim N(0, 0.1)$, $\delta_{02} = 0.1, \delta_2 = 1, \eta_{i2} \sim N(0, 0.3)$.

DGP 4: $\log(p_{ij}) = \exp\left((\delta_{0j} + \delta_j x_{i2})(1 + \eta_{ij})\right)$, where $\delta_{01} = 0.2, \delta_1 = 0.1, \eta_{i1} \sim N(0, 0.1)$, $\delta_{02} = 0.1, \delta_2 = 0.3, \eta_{i2} \sim N(0, 0.2)$.

DGP 5: $\log(p_{ij}) = (\delta_{0j} + \delta_j x_{i2})(1 + \eta_{ij})^{-1}$, where $\delta_{01} = 0.2, \delta_1 = 0.1, \eta_{i1} \sim N(0, 0.1)$, $\delta_{02} = 0.1, \delta_2 = 0.3, \eta_{i2} \sim N(0, 0.2)$.

In DGP 1, the error term in the pricing equation is additively separable and follows a normal distribution, which is commonly assumed in empirical applications. DGP 2 assumes instead that the error term follows an extreme value distribution, while in DGP 3, we relax the homoskedasticity assumption. Finally, DGPs 4 and 5 consider scenarios where the pricing function takes a nonseparable form.[10]

For each DGP, we simulate offered prices and individual choices, and assume that the econometricians observe $(y_i, x_{i1}, x_{i2}, p_i)$, where p_i is the price of the chosen alternative. We then apply our method from Section 4 to estimate the parameters of the selection function, i.e., $\theta = (\gamma, \xi_2, \beta)$ with ξ_1 normalized to 0, along with the offered price distribution for each alternative.[11] For comparison, we employ the classic two-step method, assuming that the pricing equations are linearly separable, with an error term that is independent of x_{i2} and normally distributed. Under this assumption, the two-step method misspecifies the pricing equation under DGPs 2–5. For each design, we run 500 simulations of 1000 and 5000 observations.

[10] Although all the offer price distributions admit unbounded support, in simulation we shall assume that the realized price range coincides with the true price range. Given a large sample size, the realized price range supports almost all the probability mass of the offered price distribution. Later we show that the estimation of the offered price distribution performs well.

[11] We estimate the cumulative distribution function of prices at 300 grid points.

We report Monte Carlo biases, standard deviations, and root mean squared errors for θ using our method in the first three columns of Table 1. For the cumulative distribution functions of log(price), we compute the integrated squared biases and integrated mean squared errors, as shown in the first two columns of Table 2. These results are based on a sample size of $N = 1000$. The results for $N = 5000$ are provided in Tables 3–4 in Appendix C. Our estimator performs well in finite samples across all DGPs we consider. The biases of our estimator are small, and the standard deviation decreases as the sample size increases in all simulation designs.

Table 1: Simulation Results for Utility Parameters: $N = 1000$

	Functional Contraction			Two-Step Method		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
			DGP 1			
γ	-0.0075	0.1958	0.1957	0.0027	0.2126	0.2124
ξ_2	0.0021	0.0721	0.0721	0.0003	0.0980	0.0979
β	-0.0010	0.0906	0.0905	0.0016	0.0929	0.0929
			DGP 2			
γ	-0.0087	0.1990	0.1990	0.0196	0.2451	0.2457
ξ_2	0.0021	0.0728	0.0728	0.0183	0.1203	0.1215
β	-0.0049	0.0945	0.0946	0.0036	0.0960	0.0960
			DGP 3			
γ	-0.0254	0.1603	0.1621	0.1704	0.2398	0.2940
ξ_2	-0.0006	0.0702	0.0701	0.0097	0.0860	0.0864
β	-0.0045	0.0930	0.0930	-0.0023	0.0947	0.0946
			DGP 4			
γ	-0.0131	0.3485	0.3484	0.0368	0.3826	0.3840
ξ_2	-0.0016	0.0677	0.0676	-0.0044	0.0731	0.0731
β	-0.0045	0.0933	0.0933	-0.0040	0.0941	0.0941
			DGP 5			
γ	0.0551	0.9650	0.9656	0.1873	0.7830	0.8044
ξ_2	-0.0023	0.0671	0.0671	0.0047	0.0675	0.0676
β	-0.0050	0.0886	0.0886	-0.0050	0.0885	0.0886

Compared to the classic two-step method, our estimator outperforms the standard approach in DGPs 2–5. Because our method allows for nonparametric estimation of the offered price distributions, while the standard method misspecifies the pricing equation, we achieve significantly lower integrated squared bias and mean squared error for the cumulative distribution functions of log(price). This result can also be visualized in Figure 1, where we plot the true CDFs of log(price) for firms 1 and 2, alongside those obtained using our method and the two-step method.

Table 2: Simulation Results for CDF of log(Price): $N = 1000$

	Func. Contraction		Two-Step Method	
	IBias^2	IMSE	IBias^2	IMSE
		DGP 1		
$F_1(\cdot\|x_{i2}=0)$	0.0003	0.0029	0.0006	0.0211
$F_2(\cdot\|x_{i2}=0)$	0.0001	0.0006	0.0000	0.0016
$F_1(\cdot\|x_{i2}=1)$	0.0004	0.0032	0.0003	0.0125
$F_2(\cdot\|x_{i2}=1)$	0.0002	0.0013	0.0001	0.0042
		DGP 2		
$F_1(\cdot\|x_{i2}=0)$	0.0006	0.0032	0.0042	0.0269
$F_2(\cdot\|x_{i2}=0)$	0.0002	0.0006	0.0021	0.0040
$F_1(\cdot\|x_{i2}=1)$	0.0008	0.0037	0.0035	0.0177
$F_2(\cdot\|x_{i2}=1)$	0.0003	0.0014	0.0022	0.0070
		DGP 3		
$F_1(\cdot\|x_{i2}=0)$	0.0060	0.0086	0.0247	0.0501
$F_2(\cdot\|x_{i2}=0)$	0.0028	0.0032	0.0499	0.0525
$F_1(\cdot\|x_{i2}=1)$	0.0007	0.0033	0.0022	0.0119
$F_2(\cdot\|x_{i2}=1)$	0.0002	0.0013	0.0129	0.0170
		DGP 4		
$F_1(\cdot\|x_{i2}=0)$	0.0007	0.0033	0.0049	0.0304
$F_2(\cdot\|x_{i2}=0)$	0.0008	0.0012	0.0281	0.0303
$F_1(\cdot\|x_{i2}=1)$	0.0005	0.0046	0.0005	0.0161
$F_2(\cdot\|x_{i2}=1)$	0.0002	0.0011	0.0087	0.0112
		DGP 5		
$F_1(\cdot\|x_{i2}=0)$	0.0014	0.0034	0.0026	0.0226
$F_2(\cdot\|x_{i2}=0)$	0.0014	0.0018	0.0211	0.0234
$F_1(\cdot\|x_{i2}=1)$	0.0008	0.0058	0.0008	0.0192
$F_2(\cdot\|x_{i2}=1)$	0.0002	0.0011	0.0071	0.0086

Note: The IBias^2 of a function h is calculated as follows. Let $\hat{h}_r$ be the estimate of h from the r-th simulated dataset, and $\bar{h}(x) = \frac{1}{R}\sum_{r=1}^{R}\hat{h}_r(x)$ be the point-wise average over R simulations. The integrated squared bias is calculated by numerically integrating the point-wise squared bias $(\bar{h}(x) - h(x))^2$ over the distribution of x. The integrated MSE is computed in a similar way.

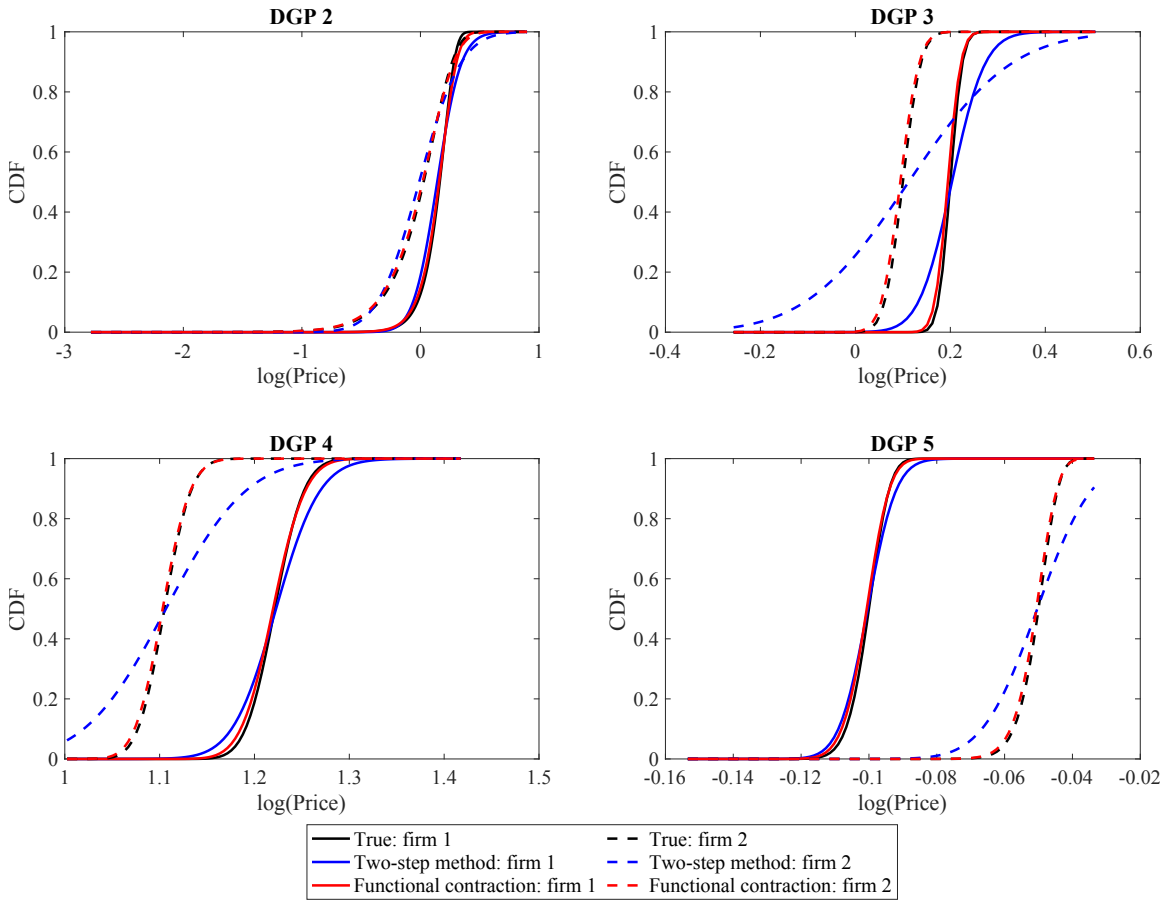

Figure 1: CDF of log(price) for firms 1 and 2 (conditional on $x = 0$). The black, blue, and red curves represent the true CDF, the CDF estimated using the two-step method, and the CDF estimated using the functional contraction method, respectively. Solid lines represent the CDFs for firm 1, while dashed lines represent those for firm 2.

For the two-step method, the misspecification of the pricing equation also creates a severe bias in estimating the parameters in the selection function. In particular, when the error term exhibits heteroskedasticity (DGP 3) or is nonseparable in the pricing equation (DGPs 4–5), the bias for the price sensitivity parameter γ is large and does not vanish as the sample size increases.

Another key advantage of our approach is that it does not require an instrument to exogenously shift the selection probability. It is well known in the literature that the two-step method is nearly unidentified when the same regressors are used in both the selection function and the outcome equation. This occurs because the inverse Mills ratio is approximately linear over a wide range of its argument. In our simulations, when the regressor in the pricing equation is discrete, the bias correction term becomes perfectly collinear with the regressor, rendering the two-step method infeasible without an excluded variable in the selection equation.

In contrast, our approach does not require an excluded variable in the selection equation. To illustrate this, we conduct a set of Monte Carlo simulations where the excluded variable x_{i1} is removed from the indirect utility, using the same five DGPs for log(price). The results for this specification are reported in Tables 5–6 in Appendix C. As shown, our estimator performs well in finite samples, even without an additional excluded variable to exogenously shift the selection probability. Our estimator consistently shows low bias across different DGPs and exhibits a decreasing standard deviation as the sample size increases.

Our method requires that the functional form of the selection function is known to econometricians. To assess the performance of our estimator when the selection function is misspecified, we conduct a series of Monte Carlo simulations. Specifically, we consider a scenario where the econometrician assumes that ε follows a logistic distribution, while it is actually generated from a normal distribution. In Tables 7–8 in Appendix C, we report the estimation results for the utility parameters and CDFs of log(price) under this misspecification. Although we observe a 7–8% bias in the utility parameters, our estimator for the offered price distributions performs well. The integrated squared bias and mean squared errors of the CDFs remain close to those in Table 2. This exercise suggests that our estimator for the offered price distributions is robust to misspecification of the selection function, a valuable feature in practice, especially when the econometrician lacks prior knowledge about the form of the selection function.

Finally, we briefly discuss how our functional contraction performs in practice. We compute the modulus ρ^* across all five simulation designs. Except for DGP 2—where the error term in the pricing equations is drawn from extreme value distributions, resulting in a wider price range—the modulus in all other cases is quite small (for example, $\rho^* = 0.37$ in DGP 1).[12] Consequently, our iteration process converges within 3–5 iterations. For DGP 2, although the modulus exceeds 1 ($\rho^* = 1.23$), the iteration process still performs well and converges to the same fixed point, even with different initial values. This is not surprising, as Theorems 1 and 2 provide only sufficient conditions for the contraction.

6 Applications

Our estimator introduced in Section 4 is broadly applicable to a variety of empirical settings. It effectively addresses the challenge of selection bias, which arises when only the outcomes of chosen alternatives are observed in the data. We impose no parametric restrictions on the potential outcome distribution and allow it to vary flexibly across alternatives. Moreover, the selection function in our model can incorporate alternative-specific unobserved heterogeneity and does not require an excluded variable, which is desirable in many empirical settings. In the following section, we discuss three types of empirical applications: consumer demand estimation, auctions with missing bids, and Roy models.

6.1 Consumer Demand

The first application of our method is the standard differentiated product demand estimation pioneered by Berry (1994) and Berry et al. (1995). In classic demand models, the price of a product is often assumed to be uniform across all consumers (e.g., the list price of a vehicle). But this assumption does not hold in contexts involving price discrimination or personalized pricing (Sagl, 2023; Buchholz et al., 2020; Dubé and Misra, 2023), discount negotiation (Goldberg, 1996; Allen et al., 2014), or

[12] The magnitude of the modulus depends heavily on the product of the price sensitivity parameter γ and price range. In our Monte Carlo simulations, γ is normalized to be 1. In DGP 1, a price range of approximately 2.5 leads to a small $\rho^* = 0.37$. In empirical applications, the price sensitivity parameter γ is around 10^{-3} (for example, see Cosconati et al., 2024). Then with a price range around 2500 euros, the modulus remains small.

risk-based pricing (Crawford et al., 2018; Cosconati et al., 2024). In these contexts, researchers can relatively easily gather data on the transaction prices consumers pay, but it is challenging to gain access to competing prices offered to consumers.

In a companion paper with coauthors (Cosconati et al., 2024), we apply our method to estimate demand and insurance companies' information technology in the auto insurance market, where only the transaction prices of selected insurance plans are observed. In this market, insurance companies employ risk-based pricing. For each consumer, an insurance company generates a noisy estimate of their risk type and prices accordingly. Our goal is to quantify the heterogeneity in insurers' information technology, as measured by the dispersion of their risk estimates. Since the shape of the offered price distribution reflects the distribution of risk estimates, allowing for flexible estimation of the offered price distribution is crucial.

We nonparametrically estimate each insurance company's offered price distribution using our functional contraction approach. In this application, we assume that the offered prices across different firms are independent, conditional on the consumer's true risk type, which is estimated using a panel of *ex-post* realized claim records over multiple years.

In Figure 2, we plot the nonparametrically estimated density functions for prices from several firms. These distributions vary significantly, with noticeable differences in mean, variance, and skewness, across firms, suggesting substantial heterogeneity in their information technology and pricing strategies. Building on this estimation, we further estimate the price sensitivity parameter, firm-specific unobserved heterogeneity (e.g., service quality or brand loyalty), and each firm's information precision. Our findings provides key insight for analyzing competition under various forms of supply-side heterogeneity in this market (Cosconati et al., 2024).

From a practical point of view, our iterative procedure to numerically solve for the offered price distributions given demand parameters is easy to implement and performs well in practice. In our empirical application using data from 11 insurers, the iterative algorithm converges very quickly, typically requiring only 6–7 iterations.

6.2 Auctions with Missing Bids

In certain auctions, not all bids are available, either due to the auction's structure or incomplete data. For instance, in Dutch auctions, only the winning bid is recorded,

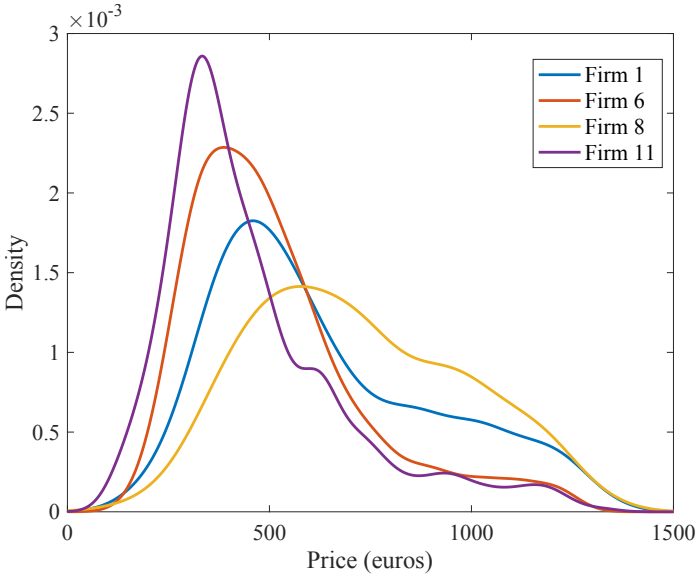

Figure 2: Estimated density functions

as the auction concludes as soon as the first bid is placed. Allen et al. (2024) study FDIC auctions for insolvent banks, where only the winning and the second-highest bids are recorded. Similarly, U.S. Forest Service timber auctions record only the top fourteen bids, while the Washington State Department of Transportation publishes only the three lowest bids for their highway procurement auctions.

The existing literature has shown that certain types of auction models can be identified using only winning bids or transaction prices. For example, Athey and Haile (2002) show that the symmetric IPV models are identified with the transaction price by exploiting a one-to-one mapping between an order statistic and its parent distribution. Komarova (2013) analyzes asymmetric second-price auctions where only the winning bids and the winner's identity are observed. A related result for generalized competing risks models can be found in Meilijson (1981). More recently, Guerre and Luo (2019) examine nonparametric identification of symmetric IPV first-price auctions with only winning bids, accounting for unobserved competition.

Our method is valuable for nonparametrically recovering the complete bid distribution and the auctioneer's scoring weights in multi-attribute auctions when the data contain only the winning bids and winner's identity, particularly in the presence of bidder asymmetry.[13] Auctions in many settings have used the scoring rule that

[13]Flexibly accommodating bidder asymmetries is known to be challenging in auction models (see discussions in the handbook chapter by Athey and Haile (2007)). Bidder asymmetries may arise

departs from the lowest bid criterion by accounting for quality differences (Asker and Cantillon, 2008; Lewis and Bajari, 2011; Nakabayashi, 2013; Yoganarasimhan, 2016; Takahashi, 2018; Krasnokutskaya et al., 2020; Allen et al., 2024). Our selection model is closely related to Krasnokutskaya et al. (2020), which employs a discrete choice framework with unknown, buyer-specific weights in the scoring rule. We allow the scoring rule to depend on both observed (x_{ij}) and unobserved bidder heterogeneity (ξ_j), with the error term (ε_{ij}) capturing uncertainty in the scoring rule.[14]

Beyond independent private value models, our method can be applied to certain common value auction models, such as the mineral rights model, where bidders' signals are assumed to be independent conditional on the common value. In these auctions, we can recover the bid distributions conditional on the ex-post realized common value.

6.3 Roy Models

Another important application of our method is estimating Roy models (Roy, 1951) in labor market contexts. Variants of the Roy model have been widely used in the literature to study decisions such as whether to continue schooling (Willis and Rosen, 1979), which occupation to pursue (Heckman and Sedlacek, 1985), whether to join a union (Lee, 1978), and whether to migrate (Borjas, 1987). Our selection model falls within the framework of "Generalized Roy Model", as defined by Heckman and Vytlacil (2007). We allow the utility that individual i gains from alternative j to depend not only on prices (or wages in labor market contexts) but also on non-pecuniary aspects of the alternative, either observable or unobservable to the econometrician. The comparison between our approach and standard two-step methods for estimating Roy models has already been discussed in the introduction; therefore, we do not reiterate it here.

from factors such as distance to the contract location (Flambard and Perrigne, 2006), information advantages (Hendricks and Porter, 1988; De Silva et al., 2009), varying risk attitudes (Campo, 2012), or strategic sophistication (Hortaçsu et al., 2019).

[14] Other recent papers that consider unknown weights in the scoring rule include Takahashi (2018) and Allen et al. (2024).

7 Conclusion

We introduce a novel method for estimating nonseparable selection models when only a selected sample of outcomes is observed. We show that potential outcome distributions can be nonparametrically identified from the observed distribution of selected outcomes, given a selection function. We achieve this by constructing an operator whose fixed point represents the potential outcome distributions and proving that this operator is a functional contraction. Building on this theoretical result, we propose a two-step semiparametric maximum likelihood estimator for both the selection function and potential outcome distributions. The consistency and asymptotic normality of the proposed estimator are established.

Our approach fundamentally differs from the classic two-step method for addressing sample selection bias. We allow the outcome equation to be fully nonparametric and nonseparable in error terms. Our goal is to recover the entire distribution of potential outcomes rather than focusing on specific moments or quantiles. In essence, we correct for sample selection bias across the entire distribution of potential outcomes by examining how the bias is *systematically* generated by the selection model. This approach allows for fully heterogeneous effects of covariates on outcomes, which is a crucial feature for empirical analysis, as discussed in Chernozhukov et al. (2023). Another key advantage of our approach is that it does not rely on instruments to exogenously shift selection probabilities, which are often challenging to find in empirical settings, or on identification-at-infinity arguments. Our approach also accommodates asymmetry in outcome distributions across alternatives and flexibly incorporates unobserved alternative-specific heterogeneity in the selection model.

We find that the proposed estimation strategy performs well in both simulations and real-world data applications (see our demand estimation using insurance market data in Cosconati et al. (2024)). Moreover, our approach is straightforward to implement and computationally efficient, making it highly appealing to empirical researchers. The estimator can be readily applied to a variety of empirical settings where only a selected sample of outcomes is observed, including consumer demand models with only transaction prices, auctions with incomplete bid data, and various selection models in labor economics. Our method is particularly valuable in applications where the entire distribution of outcomes is of interest.

References

Ahn, H. and Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1-2):3–29.

Allen, J., Clark, R., Hickman, B., and Richert, E. (2024). Resolving failed banks: Uncertainty, multiple bidding and auction design. *Review of Economic Studies*, 91(3):1201–1242.

Allen, J., Clark, R., and Houde, J.-F. (2014). Price dispersion in mortgage markets. *The Journal of Industrial Economics*, 62(3):377–416.

Allen, J., Clark, R., and Houde, J.-F. (2019). Search frictions and market power in negotiated-price markets. *Journal of Political Economy*, 127(4):1550–1598.

Andrews, D. W. and Schafgans, M. M. (1998). Semiparametric estimation of the intercept of a sample selection model. *The Review of Economic Studies*, 65(3):497–517.

Arellano, M. and Bonhomme, S. (2017). Quantile selection models with an application to understanding changes in wage inequality. *Econometrica*, 85(1):1–28.

Asker, J. and Cantillon, E. (2008). Properties of scoring auctions. *The RAND Journal of Economics*, 39(1):69–85.

Athey, S. and Haile, P. A. (2002). Identification of standard auction models. *Econometrica*, 70(6):2107–2140.

Athey, S. and Haile, P. A. (2007). Nonparametric approaches to auctions. *Handbook of econometrics*, 6:3847–3965.

Baricz, Á. (2008). Mills' ratio: Monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications*, 340(2):1362–1370.

Bayer, P., Khan, S., and Timmins, C. (2011). Nonparametric identification and estimation in a roy model with common nonpecuniary returns. *Journal of Business & Economic Statistics*, 29(2):201–215.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262.

Borjas, G. (1987). Self-selection and the earnings of immigrants. *American Economic Review*, 77:531–553.

Buchholz, N., Doval, L., Kastl, J., Matějka, F., and Salz, T. (2020). The value of time: Evidence from auctioned cab rides. *CEPR Discussion Paper No. DP14666*.

Campo, S. (2012). Risk aversion and asymmetry in procurement auctions: Identification, estimation and application to construction procurements. *Journal of Econometrics*, 168(1):96–107.

Canay, I. A., Mogstad, M., and Mountjoy, J. (2024). On the use of outcome tests for detecting bias in decision making. *Review of Economic Studies*, 91(4):2135–2167.

Chen, S. and Khan, S. (2003). Semiparametric estimation of a heteroskedastic sample selection model. *Econometric Theory*, 19(6):1040–1064.

Chernozhukov, V., Fernández-Val, I., and Luo, S. (2023). Distribution regression with sample selection and uk wage decomposition. Technical report, cemmap working paper.

Cicala, S. (2015). When does regulation distort costs? lessons from fuel procurement in us electricity generation. *American Economic Review*, 105(1):411–444.

Compiani, G. and Kitamura, Y. (2016). Using mixtures in econometric models: a brief review and some new results. *The Econometrics Journal*, 19(3):C95–C127.

Cosconati, M., Xin, Y., Wu, F., and Jin, Y. (2024). Competing under information heterogeneity: Evidence from auto insurance. *Working paper*.

Crawford, G. S., Pavanini, N., and Schivardi, F. (2018). Asymmetric information and imperfect competition in lending markets. *American Economic Review*, 108(7):1659–1701.

Das, M., Newey, W. K., and Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70(1):33–58.

De Silva, D. G., Kosmopoulou, G., and Lamarche, C. (2009). The effect of information on the bidding and survival of entrants in procurement auctions. *Journal of Public Economics*, 93(1-2):56–72.

Dubé, J.-P. and Misra, S. (2023). Personalized pricing and consumer welfare. *Journal of Political Economy*, 131(1):131–189.

d'Haultfoeuille, X. and Maurel, A. (2013a). Another look at the identification at infinity of sample selection models. *Econometric Theory*, 29(1):213–224.

d'Haultfoeuille, X. and Maurel, A. (2013b). Inference on an extended roy model, with an application to schooling decisions in france. *Journal of Econometrics*, 174(2):95–106.

D'Haultfœuille, X., Maurel, A., and Zhang, Y. (2018). Extremal quantile regressions for selection models and the black–white wage gap. *Journal of Econometrics*, 203(1):129–142.

Fernández-Val, I., van Vuuren, A., and Vella, F. (2024). Nonseparable sample selection models with censored selection rules. *Journal of Econometrics*, 240(2):105088.

Flambard, V. and Perrigne, I. (2006). Asymmetry in procurement auctions: Evidence from snow removal contracts. *The Economic Journal*, 116(514):1014–1036.

French, E. and Taber, C. (2011). Identification of models of the labor market. In *Handbook of labor economics*, volume 4, pages 537–617. Elsevier.

Goldberg, P. K. (1996). Dealer price discrimination in new car purchases: Evidence from the consumer expenditure survey. *Journal of Political Economy*, 104(3):622–654.

Gronau, R. (1974). Wage comparisons–a selectivity bias. *Journal of political Economy*, 82(6):1119–1143.

Guerre, E. and Luo, Y. (2019). Nonparametric identification of first-price auction with unobserved competition: A density discontinuity framework. *arXiv preprint arXiv:1908.05476*.

Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica: journal of the econometric society*, pages 679–694.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. NBER.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:153–161.

Heckman, J. J. and Honore, B. E. (1990). The empirical content of the roy model. *Econometrica: Journal of the Econometric Society*, pages 1121–1149.

Heckman, J. J. and Sedlacek, G. (1985). Heterogeneity, aggregation, and market wage functions: An empirical model of self-selection in the labor market. *Journal of political Economy*, 93(6):1077–1125.

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics*, 6:4779–4874.

Hendricks, K. and Porter, R. H. (1988). An empirical study of an auction with asymmetric information. *The American Economic Review*, pages 865–883.

Hortaçsu, A., Luco, F., Puller, S. L., and Zhu, D. (2019). Does strategic ability affect efficiency? evidence from electricity markets. *American Economic Review*, 109(12):4302–4342.

Hu, Y. (2017). The econometrics of unobservables–latent variable and measurement error models and their applications in empirical industrial organization and labor economics. *Manuscript in preparation*.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of econometrics*, 58(1-2):71–120.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421.

Komarova, T. (2013). A new approach to identifying generalized competing risks models with application to second-price auctions. *Quantitative Economics*, 4(2):269–328.

Krasnokutskaya, E., Song, K., and Tang, X. (2020). The role of quality in internet service markets. *Journal of Political Economy*, 128(1):75–117.

Lee, J. H. and Park, B. G. (2023). Nonparametric identification and estimation of the extended roy model. *Journal of Econometrics*, 235(2):1087–1113.

Lee, L.-F. (1978). Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International economic review*, pages 415–433.

Lee, L.-F. (1982). Some approaches to the correction of selectivity bias. *The Review of Economic Studies*, 49(3):355–372.

Lee, L.-F. (1983). Generalized econometric models with selectivity. *Econometrica: Journal of the Econometric Society*, pages 507–512.

Lewis, G. and Bajari, P. (2011). Procurement contracting with time incentives: Theory and evidence. *The Quarterly Journal of Economics*, 126(3):1173–1211.

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38.

Meilijson, I. (1981). Estimation of the lifetime distribution of the parts from the autopsy statistics of the machine. *Journal of Applied Probability*, 18(4):829–838.

Mourifie, I., Henry, M., and Meango, R. (2020). Sharp bounds and testability of a roy model of stem major choices. *Journal of Political Economy*, 128(8):3220–3283.

Nakabayashi, J. (2013). Small business set-asides in procurement auctions: An empirical analysis. *Journal of Public Economics*, 100:28–44.

Newey, W. K. (2007). Nonparametric continuous/discrete choice models. *International Economic Review*, 48(4):1429–1439.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal*, 12(suppl_1):S217–S229.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146.

Sagl, S. (2023). Dispersion, discrimination, and the price of your pickup. *Working paper*.

Salz, T. (2022). Intermediation and competition in search markets: An empirical case study. *Journal of Political Economy*, 130(2):310–345.

Takahashi, H. (2018). Strategic design under uncertain evaluations: Structural analysis of design-build auctions. *The RAND Journal of Economics*, 49(3):594–618.

Thompson, A. C. (1963). On certain contraction mappings in a partially ordered vector space. *Proceedings of the American Mathematical Society*, 14(3):438–443.

Vella, F. (1998). Estimating models with sample selection bias: a survey. *Journal of Human Resources*, pages 127–169.

Willis, R. J. and Rosen, S. (1979). Education and self-selection. *Journal of Political Economy*, 87(5, Part 2):S7–S36.

Yoganarasimhan, H. (2016). Estimation of beauty contest auctions. *Marketing Science*, 35(1):27–54.

A Connection to Quantal Response Equilibria

In this section, we connect our result to the quantal response equilibria (McKelvey and Palfrey, 1995).

Let us rename our variables. There is a set $\mathcal{J} = \{1, 2, \cdots, J\}$ of players. For each player $j \in \mathcal{J}$, there is a finite set $P_j = \{p_{j1}, p_{j2}, \cdots, p_{jn_j}\} \subset [\underline{p}_j, \overline{p}_j]$ consisting of n_j pure strategies. A payoff function $f\colon \prod_{j\in\mathcal{J}} P_j \to \Delta(\mathcal{J})$ assigns payoff f_j to player j. Let $g_j \in \Delta P_j$ denote player j' mixed strategy and $g = \prod_{j\in\mathcal{J}} g_j$. The player j's expected payoff for playing pure strategy p_j, given other players' strategy g_{-j}, is

$$Pr_j(p_j; g) = \int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k\neq j} g_k(p_k).$$

We define the quantal response operator $\mathbb{T}\colon \prod_j \Delta(P_j) \to \prod_j \Delta(P_j)$ by

$$(\mathbb{T}g)_j(p_j) = \frac{\exp(-\lambda Pr_j(p_j; g))}{\sum_{p_j\in P_j} \exp(-\lambda Pr_j(p_j; g))}.$$

In words, given the expected payoff $Pr_j(p_j; g)$, player j's probability of playing strategy p_j is proportional to $\exp(-\lambda Pr_j(p_j; g))$. Lemma 1 in McKelvey and Palfrey (1995) states that operator $\mathbb{T}$ is a contraction for a sufficiently small λ. This is intuitive as $\mathbb{T}$ sends probability measures to the center of the simplex when λ is small.

Note that our operator T is quite different. By definition,

$$(T\Psi)_j(p) = \frac{\int_{\underline{p}_j}^{p} d\tilde{G}_j(y)/Pr_j(y; \Psi)}{\int_{\underline{p}_j}^{\overline{p}_j} d\tilde{G}_j(y)/Pr_j(y; \Psi)}.$$

Given the expected probability Pr, to compute the new measure, each p_j is weighted by $d\tilde{G}(p_j)$, where $\tilde{G}$ can be any measure. This distinction complicates our problem. With the sup norm, McKelvey and Palfrey (1995) show that $\mathbb{T}$ is a contraction for sufficiently small λ. However, the presence of $\tilde{G}$ renders the sup norm not suitable for our task. Instead, our metric d is designed specifically to deal with $\tilde{G}$.

B Omitted Proofs

B.1 Proof of Theorem 1

Lemma 1. *For two probability measures $S, Q \in \Delta(Y)$, $\delta > 0$,*

$$\sup_{d(S,Q)\leq\delta} ||S - Q||_{TV} \leq \delta/2.$$

Proof of Lemma 1. We first consider the case where Y contains only two elements. Then we can identify S with $(p, 1-p)$ for some $p \in [0,1]$. We can pin down the Q that achieves the maximum $||S - Q||_{TV}$ under the constraint that $d(S,Q) \leq \delta$. At the maximum, this constraint is binding. Let $Q = (p - \epsilon, 1 - p + \epsilon)$. By $d(S,Q) = \delta$,

$$\ln \frac{p}{p-\epsilon} + \ln \frac{1-p+\epsilon}{1-p} = \delta. \tag{14}$$

We can solve for ϵ

$$\epsilon = \frac{p(1-p)(e^\delta - 1)}{p + (1-p)e^\delta}.$$

Plug this into the total variation norm

$$\frac{1}{2}||S - Q||_{TV} = \epsilon = (e^\delta - 1)[\frac{1}{1-p} + \frac{e^\delta}{p}]^{-1}.$$

Then we take sup over p. Note that $\frac{1}{1-p} + \frac{e^\delta}{p}$ as a function of p is convex and achieves a unique minimum at $p = \frac{e^{\delta/2}}{1+e^{\delta/2}}$. As a result,

$$\sup_{d(S,Q)\leq\delta} \frac{1}{2}||S - Q||_{TV} = \frac{(e^\delta - 1)}{(1 + e^{\delta/2})^2} = \frac{e^{\delta/2} - 1}{e^{\delta/2} + 1}.$$

To show $\sup_{d(S,Q)\leq\delta} ||S - Q||_{TV} \leq \delta/2$, it suffices to show that for all $\delta \geq 0$,

$$\frac{e^{\delta/2} - 1}{e^{\delta/2} + 1} \leq \delta/4$$

which holds true.[15] Note that the limiting case $\delta \to 0$, $p = \frac{1}{2}$, $\epsilon = \frac{\delta}{4}$ achieves this upper bound.

Now we prove this lemma for a general space Y and general CDF. For any $S, Q \in \Delta(Y)$ and $d(S,Q) \leq \delta$. Define two functions

$$P_Q(S,Q) = \int_{y \in Y : \frac{dS}{dQ}(y) \geq 1} dQ(y)$$

$$P_S(S,Q) = \int_{y \in Y : \frac{dS}{dQ}(y) \geq 1} dS(y).$$

Note that

$$\frac{P_S(S,Q)}{P_Q(S,Q)} \leq \operatorname*{ess\,sup}_{y \in Y} \frac{dS}{dQ}(y)$$

$$\frac{1 - P_Q(S,Q)}{1 - P_S(S,Q)} \leq \operatorname*{ess\,sup}_{y \in Y} \frac{dQ}{dS}(y)$$

which implies

$$\ln \frac{P_S(S,Q)}{P_Q(S,Q)} + \ln \frac{1 - P_Q(S,Q)}{1 - P_S(S,Q)} \leq \operatorname{ess\,sup} \ln \frac{dS}{dQ}(y) + \operatorname{ess\,sup} \ln \frac{dQ}{dS}(y) \leq \delta$$

since $d(S,Q) \leq \delta$. Observe that here $P_S(S,Q)$ faces the same constraint as p in the two-point support case in Equation (14). Thus, the total variation norm

$$||S - Q||_{TV} = 2[P_S(S,Q) - P_Q(S,Q)] \leq \delta/2.$$

□

[15] To see this,

$$\begin{aligned} &\frac{e^\delta - 1}{e^\delta + 1} \leq \delta/2 \\ \Leftrightarrow &1 - \frac{2}{e^\delta + 1} \leq \frac{\delta}{2} \\ \Leftrightarrow &2 - \delta \leq \frac{4}{e^\delta + 1} \end{aligned}$$

which is true since function $\frac{4}{e^\delta + 1}$ is convex and is tangent to the function $2 - \delta$ at $\delta = 0$.

Proof of Theorem 1. Recall that

$$Pr_j(p_j;\Psi) = \int_{\boldsymbol{p}_{-j}} f_j(p_j,\boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Psi_k(p_k).$$

Define the ratio function

$$R_j(p_j;\Psi,\Phi) = \frac{Pr_j(p_j;\Psi)}{Pr_j(p_j;\Phi)}.$$

We show that for all $\Psi,\Phi \in \prod_j \Delta([\underline{p}_j,\overline{p}_j])$,

$$D(T\Psi,T\Phi) \leq \rho D(\Psi,\Phi).$$

Given Equation (7) and the definition of the metric d, we have

$$d((T\Psi)_j,(T\Phi)_j) \leq \sup_{p_j} \ln R_j(p_j;\Psi,\Phi) - \inf_{p_j} \ln R_j(p_j;\Psi,\Phi).$$

The equality holds when $\tilde{G}_j$ admits full support on $[\underline{p}_j,\overline{p}_j]$. Thus, it suffices to show that for all $j \in \mathcal{J}$

$$\sup_{p_j} \ln R_j(p_j;\Psi,\Phi) - \inf_{p_j} \ln R_j(p_j;\Psi,\Phi) \leq \rho D(\Psi,\Phi) \tag{15}$$

We evaluate how the log ratio changes with p_j,

$$\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j} = \frac{\int_{\boldsymbol{p}_{-j}} \frac{\partial f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j} \prod_{k,k\neq j} d\Psi_k(p_k)}{\int_{\boldsymbol{p}_{-j}} f_j(p_j,\boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Psi_k(p_k)} - \frac{\int_{\boldsymbol{p}_{-j}} \frac{\partial f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j} \prod_{k,k\neq j} d\Phi_k(p_k)}{\int_{\boldsymbol{p}_{-j}} f_j(p_j,\boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Phi_k(p_k)} \tag{16}$$

$$= \frac{\int_{\boldsymbol{p}_{-j}} \frac{\partial \ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j} f_j \prod_{k,k\neq j} d\Psi_k(p_k)}{\int_{\boldsymbol{p}_{-j}} f_j(p_j,\boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Psi_k(p_k)} - \frac{\int_{\boldsymbol{p}_{-j}} \frac{\partial \ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j} f_j \prod_{k,k\neq j} d\Phi_k(p_k)}{\int_{\boldsymbol{p}_{-j}} f_j(p_j,\boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Phi_k(p_k)} \tag{17}$$

Next, we define a new measure $f_j\Psi_{-j} \in \Delta(\prod_{k\neq j}[\underline{p}_k, \overline{p}_k])$

$$f_j\Psi_{-j}(y) = \frac{\int_{\underline{\boldsymbol{p}}_{-j}}^{y} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Psi_k(p_k)}{\int_{\boldsymbol{p}_{-j}} f_j(p_j, \boldsymbol{p}_{-j}) \prod_{k,k\neq j} d\Psi_k(p_k)}.$$

Similarly, we define measure $f_j\Phi_{-j} \in \Delta(\prod_{k\neq j}[\underline{p}_k, \overline{p}_k])$. (Both measures depend on p_j.) Given these measures, we can rewrite Equation (17)

$$\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j} = \mathop{\mathbb{E}}_{\boldsymbol{p}_{-j}\sim f_j\Psi_{-j}} \frac{\partial \ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j} - \mathop{\mathbb{E}}_{\boldsymbol{p}_{-j}\sim f_j\Phi_{-j}} \frac{\partial \ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j} \tag{18}$$

$$= \int_{\boldsymbol{p}_{-j}} \frac{\partial \ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j} [df_j\Psi_{-j}(\boldsymbol{p}_{-j}) - df_j\Phi_{-j}(\boldsymbol{p}_{-j})]. \tag{19}$$

We shall upper bound this integral under the constraint $D(\Psi,\Phi) \leq \delta$ for some arbitrary $\delta > 0$.

$$\begin{aligned}\sup_{D(\Psi,\Phi)\leq\delta} \left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\right| &= \sup_{D(\Psi,\Phi)\leq\delta} \left| \int_{\boldsymbol{p}_{-j}} \frac{\partial \ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j} [df_j\Psi_{-j}(\boldsymbol{p}_{-j}) - df_j\Phi_{-j}(\boldsymbol{p}_{-j})]\right| \\ &\leq M_j \sup_{D(\Psi,\Phi)\leq\delta} \frac{1}{2}||f_j\Psi_{-j} - f_j\Phi_{-j}||_{TV}\end{aligned}$$

The inequality follows by interpreting the integral as a transportation problem. We transport the mass from distribution $f_j\Phi_{-j}$ to $f_j\Psi_{-j}$. The function $\frac{\partial \ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j}$ is the height. Then the integral is the change in the gravitational potential, which is bounded by the product of the total transportation mass $\frac{1}{2}||f_j\Psi_{-j} - f_j\Phi_{-j}||_{TV}$ and the largest height difference, M_j. Note that given $D(\Psi,\Phi) \leq \delta$,

$$d(f_j\Psi_{-j}, f_j\Phi_{-j}) = d(\Psi_{-j},\Phi_{-j}) \leq (J-1)\delta,$$

as for all j, $d(\Psi_j,\Phi_j) \leq D(\Psi,\Phi) \leq \delta$. Thus, for all $\delta > 0$,

$$\begin{aligned}\sup_{D(\Psi,\Phi)\leq\delta} \left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\right| &\leq M_j \sup_{D(\Psi,\Phi)\leq\delta} \frac{1}{2}||f_j\Psi_{-j} - f_j\Phi_{-j}||_{TV} \\ &\leq M_j \sup_{d(f_j\Psi_{-j}, f_j\Phi_{-j})\leq(J-1)\delta} \frac{1}{2}||f_j\Psi_{-j} - f_j\Phi_{-j}||_{TV} \\ &\leq M_j \frac{1}{4}(J-1)\delta \end{aligned} \tag{20}$$

where the last inequality follows by Lemma 1. By Lemma 2,

$$\sup_{\Psi,\Phi}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| = \sup_{D(\Psi,\Phi)\leq\delta}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| \leq \frac{J-1}{4}M_j.$$

To see why the inequality holds, towards a contradiction, suppose it does not hold. Then there exists $\tilde{\Psi}$, $\tilde{\Phi}$ with $D(\tilde{\Psi},\tilde{\Phi}) = \delta_1$ and

$$\left|\frac{d\ln R_j(p_j;\tilde{\Psi},\tilde{\Phi})}{dp_j}\frac{1}{D(\tilde{\Psi},\tilde{\Phi})}\right| > \frac{J-1}{4}M_j$$

$$\left|\frac{d\ln R_j(p_j;\tilde{\Psi},\tilde{\Phi})}{dp_j}\right| > \frac{J-1}{4}M_j D(\tilde{\Psi},\tilde{\Phi})$$

which implies that

$$\sup_{D(\Psi,\Phi)\leq\delta_1}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| \geq \left|\frac{d\ln R_j(p_j;\tilde{\Psi},\tilde{\Phi})}{dp_j}\frac{1}{D(\tilde{\Psi},\tilde{\Phi})}\right| > \frac{J-1}{4}M_j$$

contradicting Equation (20) which holds for all $\delta > 0$.

By the fundamental theorem of calculus, for all $p_j, p_j' \in [\underline{p}_j, \overline{p}_j]$,

$$\sup_{\Psi,\Phi}\left|\frac{\ln R_j(p_j;\Psi,\Phi) - \ln R_j(p_j';\Psi,\Phi)}{D(\Psi,\Phi)}\right| \leq \frac{J-1}{4}M_j(\overline{p}_j - \underline{p}_j)$$

Finally, for all $j \in \mathcal{J}$, all Ψ, Φ,

$$\sup_{p_j}\ln R_j(p_j;\Psi,\Phi) - \inf_{p_j}\ln R_j(p_j;\Psi,\Phi) \leq \rho D(\Psi,\Phi).$$

□

Lemma 2. *For all $\delta > 0$,*

$$\sup_{\Psi,\Phi}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| = \sup_{\Psi,\Phi,D(\Psi,\Phi)\leq\delta}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right|. \tag{21}$$

Proof of Lemma 2. We prove this lemma through a continuous interpolation. Fixing any $\Psi, \Phi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$, we define a continuous interpolation $\Upsilon(\cdot;\lambda) \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$

parametrized by $\lambda \in [0,1]$:

$$\Upsilon_j(p_j;\lambda) = \frac{\int_{\underline{p}_j}^{p_j} d\Phi_j(y) \cdot \left(\frac{d\Psi_j}{d\Phi_j}(y)\right)^\lambda}{\int_{\underline{p}_j}^{\overline{p}_j} d\Phi_j(y) \cdot \left(\frac{d\Psi_j}{d\Phi_j}(y)\right)^\lambda}$$

Notice that $\Upsilon(\cdot;0) = \Phi$, $\Upsilon(\cdot;1) = \Psi$. Moreover,

$$d(\Upsilon_j(\cdot;\lambda_1), \Upsilon_j(\cdot;\lambda_2)) = |\lambda_1 - \lambda_2| d(\Psi_j, \Phi_j).$$

Thus, in our metric space, $\Upsilon(\cdot;\lambda)$ is an interpolation that is linear in the metric.[16] That is, for all $\lambda_1, \lambda_2 \in [0,1]$,

$$D(\Upsilon(\cdot;\lambda_1), \Upsilon(\cdot;\lambda_2)) = |\lambda_1 - \lambda_2| D(\Psi, \Phi).$$

We define a new function by adapting Equation (18).

$$k(\lambda) = \mathop{\mathbb{E}}_{\boldsymbol{p}_{-j} \sim f_j \Upsilon_{-j}(\cdot;\lambda)} \frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j} - \mathop{\mathbb{E}}_{\boldsymbol{p}_{-j} \sim f_j \Phi_{-j}} \frac{\partial \ln f_j(p_j, \boldsymbol{p}_{-j})}{\partial p_j}.$$

Notice that when $\lambda = 1$, this reduces to Equation (18). As k is continuously differentiable, there exists $0 \leq \underline{\lambda} < \underline{\lambda} + d\lambda \leq 1$ and $d\lambda \leq \frac{\delta}{D(\Psi,\Phi)}$ such that

$$|k(1)| \leq \left| \frac{k(\underline{\lambda} + d\lambda) - k(\underline{\lambda})}{d\lambda} \right|$$

[16] Note that $\Upsilon(\cdot;\lambda)$ is also a linear interpolation in the Kullback-Leibler divergence, since

$$D_{KL}(\Phi || \Upsilon(\cdot;\lambda)) = \lambda D_{KL}(\Phi || \Psi)$$

and

$$D_{KL}(\Psi || \Upsilon(\cdot;\lambda)) = (1-\lambda) D_{KL}(\Psi || \Phi).$$

This is equivalent to

$$\begin{aligned}
&\left|\frac{k(1)}{D(\Psi,\Phi)}\right| \leq \left|\frac{k(\underline{\lambda}+d\lambda)-k(\underline{\lambda})}{d\lambda D(\Psi,\Phi)}\right| \\
\Leftrightarrow &\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| \leq \left|\frac{d\ln R_j(p_j;\Upsilon(\cdot;\underline{\lambda}+d\lambda),\Upsilon(\cdot;\underline{\lambda}))}{dp_j}\frac{1}{d\lambda D(\Psi,\Phi)}\right| \\
\Leftrightarrow &\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| \leq \left|\frac{d\ln R_j(p_j;\Upsilon(\cdot;\underline{\lambda}+d\lambda),\Upsilon(\cdot;\underline{\lambda}))}{dp_j}\frac{1}{D(\Upsilon(\cdot;\underline{\lambda}+d\lambda),\Upsilon(\cdot;\underline{\lambda}))}\right|
\end{aligned}$$

As $D(\Upsilon(\cdot;\underline{\lambda}+d\lambda),\Upsilon(\cdot;\underline{\lambda}))=d\lambda D(\Psi,\Phi)\leq\delta$, we have established Equation (21). □

B.2 Proof of Theorem 2

Proof of Theorem 2. With Assumption 1, we can provide tighter bound on the right-hand side of Equation (19).

$$\begin{aligned}
&\sup_{D(\Psi,\Phi)\leq\delta}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\right| \\
=&\sup_{D(\Psi,\Phi)\leq\delta}\left|\int_{\boldsymbol{p}_{-j}}\frac{\partial\ln f_j(p_j,\boldsymbol{p}_{-j})}{\partial p_j}[df_j\Psi_{-j}(\boldsymbol{p}_{-j})-df_j\Phi_{-j}(\boldsymbol{p}_{-j})]\right| \\
\leq&\left[\frac{\partial\ln f_j(p_j,\overline{\boldsymbol{p}}_{-j})}{\partial p_j}-\frac{\partial\ln f_j(p_j,\underline{\boldsymbol{p}}_{-j})}{\partial p_j}\right]\sup_{D(\Psi,\Phi)\leq\delta}\frac{1}{2}||f_j\Psi_{-j}-f_j\Phi_{-j}||_{TV} \\
\leq&\left[\frac{\partial\ln f_j(p_j,\overline{\boldsymbol{p}}_{-j})}{\partial p_j}-\frac{\partial\ln f_j(p_j,\underline{\boldsymbol{p}}_{-j})}{\partial p_j}\right]\frac{J-1}{4}\delta
\end{aligned}$$

By Lemma 2,

$$\sup_{\Psi,\Phi}\left|\frac{d\ln R_j(p_j;\Psi,\Phi)}{dp_j}\frac{1}{D(\Psi,\Phi)}\right| \leq \frac{J-1}{4}\left[\frac{\partial\ln f_j(p_j,\overline{\boldsymbol{p}}_{-j})}{\partial p_j}-\frac{\partial\ln f_j(p_j,\underline{\boldsymbol{p}}_{-j})}{\partial p_j}\right]$$

By the fundamental theorem of calculus, for all $p_j,p_j'\in[\underline{p}_j,\overline{p}_j]$,

$$\sup_{\Psi,\Phi}\left|\frac{\ln R_j(p_j;\Psi,\Phi)-\ln R_j(p_j';\Psi,\Phi)}{D(\Psi,\Phi)}\right|\leq\rho^*$$

Finally, for all $j\in\mathcal{J}$, all Ψ, Φ,

$$\sup_{p_j}\ln R_j(p_j;\Psi,\Phi)-\inf_{p_j}\ln R_j(p_j;\Psi,\Phi)\leq\rho^*D(\Psi,\Phi).$$

□

B.3 Proof of Theorem 3

Proof of Proposition 1. Suppose $\rho < 1$. By Theorem 1, the operator T is a contraction. This implies that F is surjective, since for any $\tilde{G}$, we can take a $\Psi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$F(\lim_{n\to\infty} T^n \Psi) = \tilde{G}.$$

Moreover, F is injective. Towards a contradiction, suppose F maps both $G_1 \neq G_2 \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$ to the same $\tilde{G}$. Then both G_1 and G_2 are fixed points for operator T, contradicting contraction.

The mapping F is continuous by Equation (1) and (2). Take two offered distributions G and G'. By Equation (2) and the definition of our metric,

$$\begin{aligned}
d(F(G)_j, F(G')_j) &= \ln \operatorname*{ess\,sup}_{p\in[\underline{p}_j,\overline{p}_j]} \left(\frac{dG_j}{dG'_j}(p)\frac{Pr_j(p;G)}{Pr_j(p;G')}\right) + \ln \operatorname*{ess\,sup}_{p\in[\underline{p}_j,\overline{p}_j]} \left(\frac{dG'_j}{dG_j}(p)\frac{Pr_j(p;G')}{Pr_j(p;G)}\right) \\
&\leq \ln \operatorname*{ess\,sup}_{p\in[\underline{p}_j,\overline{p}_j]} \frac{dG_j}{dG'_j}(p) + \ln \operatorname*{ess\,sup}_{p\in[\underline{p}_j,\overline{p}_j]} \frac{dG'_j}{dG_j}(p) \\
&+ \ln \sup_{p\in[\underline{p}_j,\overline{p}_j]} \left(\frac{Pr_j(p;G)}{Pr_j(p;G')}\right) + \ln \sup_{p\in[\underline{p}_j,\overline{p}_j]} \left(\frac{Pr_j(p;G')}{Pr_j(p;G)}\right) \\
&\leq D(G,G') + \rho D(G,G')
\end{aligned}$$

where the last inequality is by Equation (15). Consequently,

$$D(F(G), F(G')) \leq (1+\rho) D(G,G')$$

F is Lipschitz continuous with Lipschitz constant $1+\rho$.

Next, we show F^{-1} is Lipschitz continuous. Take two selected distributions $\tilde{G} \neq \tilde{G}' \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$ where $\tilde{G} = F(G)$. Let $T_{\tilde{G}}$ and $T_{\tilde{G}'}$ denote the corresponding operator T. Here we express dependence on the selected distribution. Note that

$$D(\tilde{G}, \tilde{G}') = D(T_{\tilde{G}}G, T_{\tilde{G}'}G) = D(G, T_{\tilde{G}'}G)$$

where the first equality is by the definition of the operator T and the metric D, while

the second equality is by G being a fixed point of $T_{\tilde{G}}$. Observe that

$$D(T^k_{\tilde{G}'}G, T^{k+1}_{\tilde{G}'}G) \leq \rho^k D(G, T_{\tilde{G}'}G) = \rho^k D(\tilde{G}, \tilde{G}')$$

$$\begin{aligned} D(F^{-1}(\tilde{G}), F^{-1}(\tilde{G}')) = D(G, F^{-1}(\tilde{G}')) =& D(G, T^{\infty}_{\tilde{G}'}G) \\ \leq& \sum_{k=0}^{\infty} D(T^k_{\tilde{G}'}G, T^{k+1}_{\tilde{G}'}G) \\ \leq& \sum_{k=0}^{\infty} \rho^k D(\tilde{G}, \tilde{G}') \\ =& \frac{1}{1-\rho} D(\tilde{G}, \tilde{G}') \end{aligned}$$

where the first inequality is by triangular inequality. This proves that F^{-1} is Lipschitz continuous with Lipschitz constant $\frac{1}{1-\rho}$. □

For proofs below, it suffices to prove the case without variable x. So we shall drop it. We next prove the consistency result (Proposition 3). The proof requires a combination of Lemma 3-6 below. We first collect useful notations below. Let

$$Q_0(\theta) = \sum_j \ln\big(Prob^*_j(\theta, \tilde{G})\big) \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; \theta_0) dG(\boldsymbol{p}).$$

$$\hat{Q}^*_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln\big(Prob^*_j(\theta, \hat{G})\big),$$

$$Prob^*_j(\theta, \hat{G}) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; \theta) dF^{-1}(\hat{G}, \theta)(\boldsymbol{p}).$$

$$\hat{Q}_{n,m}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln\big(Prob_j(\theta, \hat{G}, m)\big),$$

$$Prob_j(\theta, \hat{G}, m) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p}; \theta) d\big(T^m_{\hat{G},\theta}\Psi\big)(\boldsymbol{p}).$$

$$\hat{Q}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln\big(Prob_j(\theta, \hat{G}, m(n))\big),$$

$$\mathfrak{g}^*(z_i, \theta, \hat{G}) = \nabla_\theta \bigg(\sum_{j=1}^{J} y_{ij} \ln Prob^*_j(\theta, \hat{G}) \bigg),$$

$$\mathfrak{g}(z_i, \theta, \hat{G}, n) = \nabla_\theta \left(\sum_{j=1}^{J} y_{ij} \ln Prob_j(\theta, \hat{G}, m(n)) \right),$$

Lemma 3. *$F^{-1}(\hat{G}, \theta)$ is continuous in θ.*

Proof of Lemma 3. Let $\theta, \theta' \in \Theta$. Let

$$\begin{aligned} \tilde{G} &= F(G; \theta) \\ G' &= F^{-1}(\tilde{G}; \theta') \\ \tilde{G}' &= F(G'; \theta). \end{aligned}$$

As $\theta' \to \theta$, by $F(G'; \theta)$ being continuous in θ, $\tilde{G} \to \tilde{G}'$. By $F^{-1}(\tilde{G}; \theta)$ being continuous in $\tilde{G}$ (Proposition 1), $F^{-1}(\tilde{G}; \theta) \to F^{-1}(\tilde{G}'; \theta)$. This is equivalent to $G' \to G$, which is $F^{-1}(\tilde{G}; \theta') \to F^{-1}(\tilde{G}; \theta)$. This implies that F^{-1} is continuous in θ.

□

For the next lemma, we view $F^{-1}(\theta; \hat{G})$ as a function of θ parametrized by $\hat{G}$.

Lemma 4. *The function $F^{-1}(\theta; \hat{G})$ is equicontinuous in θ, i.e., for all $\theta \in \Theta$, $\epsilon > 0$, there exists a $\delta > 0$ such that for all $|\theta' - \theta| < \delta$, $\hat{G} \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,*

$$D(F^{-1}(\theta; \hat{G}), F^{-1}(\theta'; \hat{G})) \leq \epsilon.$$

Proof of Lemma 4. Since the function f is continuous on a compact set $\prod_j [\underline{p}_j, \overline{p}_j] \times \Theta$ and the image of f is in the interior of the simplex, there exists $\underline{f}$ and $\overline{f}$, $0 < \underline{f} \leq \overline{f} < 1$ such that for all $j \in \mathcal{J}$, $\theta \in \Theta$, $\boldsymbol{p} \in \prod_j [\underline{p}_j, \overline{p}_j]$,

$$\underline{f} < f_j(\boldsymbol{p}; \theta) < \overline{f}.$$

Consequently, for all $j \in \mathcal{J}$, $\theta \in \Theta$, $p_j \in [\underline{p}_j, \overline{p}_j]$, $G \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$\underline{f} < Pr_j(p_j; G, \theta) < \overline{f}. \tag{22}$$

Moreover, since the function f is continuous on a compact set $\prod_j [\underline{p}_j, \overline{p}_j] \times \Theta$, f is uniformly continuous. Thus, for any $\epsilon' > 0$, there exists a $\delta' > 0$ such that for all $j \in \mathcal{J}$, $\boldsymbol{p} \in \prod_j [\underline{p}_j, \overline{p}_j]$, $\theta, \theta' \in \Theta$ with $|\theta - \theta'| < \delta'$,

$$|f_j(\boldsymbol{p}, \theta) - f_j(\boldsymbol{p}, \theta')| < \epsilon'.$$

Therefore, for all $j \in \mathcal{J}$, $p_j \in [\underline{p}_j, \overline{p}_j]$, $G \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$, $\theta, \theta' \in \Theta$ with $|\theta - \theta'| < \delta'$,

$$\begin{aligned}
&|Pr_j(p_j; G, \theta) - Pr_j(p_j; G, \theta')| \\
=&\left| \int_{\boldsymbol{p}_{-j}} [f_j(p_j, \boldsymbol{p}_{-j}; \theta) - f_j(p_j, \boldsymbol{p}_{-j}; \theta')] \prod_{k, k \neq j} dG_k(p_k) \right| < \epsilon'. \qquad (23)
\end{aligned}$$

Take an arbitrary $\hat{G} \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$. Let $G_\theta = F^{-1}(\theta; \hat{G})$, $G_{\theta'} = F^{-1}(\theta'; \hat{G})$. Let T_θ and $T_{\theta'}$ be the operator T associated with selected distribution $\hat{G}$, when the parameter is θ and θ', respectively: for any $\Psi \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$(T_\theta \Psi)_j(p) = \frac{\int_{\underline{p}_j}^{p} d\hat{G}_j(y) / Pr_j(y; \Psi, \theta)}{\int_{\underline{p}_j}^{\overline{p}_j} d\hat{G}_j(y) / Pr_j(y; \Psi, \theta)}.$$

By the definition of metric D,

$$D(T_\theta G_\theta, T_{\theta'} G_\theta) \leq \max_j \left[\sup_p \ln \frac{Pr_j(p; G_\theta, \theta)}{Pr_j(p; G_\theta, \theta')} + \sup_p \ln \frac{Pr_j(p; G_\theta, \theta')}{Pr_j(p; G_\theta, \theta)} \right].$$

By Equation (22) and (23), for all $\hat{G} \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$, $\theta, \theta' \in \Theta$ with $|\theta - \theta'| < \delta'$,

$$D(T_\theta G_\theta, T_{\theta'} G_\theta) \leq 2 \ln \frac{\overline{f} + \epsilon'}{\underline{f}},$$

$$\begin{aligned}
D(F^{-1}(\theta; \hat{G}), F^{-1}(\theta'; \hat{G})) =& D(G_\theta, T_{\theta'}^{\infty} G_\theta) \\
\leq& \sum_{k=0}^{\infty} D(T_{\theta'}^{k} G_\theta, T_{\theta'}^{k+1} G_\theta) \\
\leq& \sum_{k=0}^{\infty} \bar{\rho}^k D(G_\theta, T_{\theta'} G_\theta) \\
=& \frac{1}{1 - \bar{\rho}} D(T_\theta G_\theta, T_{\theta'} G_\theta) \\
\leq& \frac{2}{1 - \bar{\rho}} \ln \frac{\overline{f} + \epsilon'}{\underline{f}}.
\end{aligned}$$

Finally, for any $\epsilon > 0$, let ϵ' be such that $\frac{2}{1-\bar{\rho}} \ln \frac{\overline{f}+\epsilon'}{\underline{f}} = \epsilon$. The δ' corresponding to this ϵ' is the desired δ in the statement of the Lemma.

□

Lemma 5. *$\hat{Q}_n^*(\theta)$ converges uniformly in probability to $Q_0(\theta)$.*

Proof of Lemma 5. By Lemma 4 and the uniform continuity of f, for all j, $Prob_j^*(\theta; \hat{G})$ is equicontinuous in θ, parametrized by $\hat{G}$. That is, for all $\theta \in \Theta$, $\epsilon > 0$, there exists a $\delta > 0$ such that for all $|\theta' - \theta| < \delta$, $\hat{G} \in \prod_j \Delta([\underline{p}_j, \overline{p}_j])$,

$$|Prob_j^*(\theta; \hat{G}) - Prob_j^*(\theta'; \hat{G})| \leq \epsilon.$$

Consequently, by Equation (22), for all $\theta \in \Theta$, $\epsilon > 0$, there exists a $\delta > 0$ such that for all $|\theta' - \theta| < \delta$, $\{z_i\}_{i=1}^n$,

$$|\hat{Q}_n^*(\theta) - \hat{Q}_n^*(\theta')| \leq \ln \frac{\underline{f} + \epsilon}{\underline{f}}.$$

Thus, $\hat{Q}_n^*(\theta)$ is equicontinuous in θ.

For all $\theta \in \Theta$, $\hat{Q}_n^*(\theta)$ converges in probability to $Q_0(\theta)$, by the weakly law of large numbers, $\hat{G} \xrightarrow{p} \tilde{G}$, and F^{-1} being continuous (Proposition 1). Lastly, $\hat{Q}_n^*(\theta)$ converges uniformly in probability to $Q_0(\theta)$, as $\hat{Q}_n^*$ is equicontinuous in θ (Lemma 2.8 in Newey and McFadden (1994)). □

Lemma 6. *$\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$.*

Proof of Lemma 6. Pick a $\Psi \sim \hat{G}$.[17] Fix some $\epsilon' > 0$. As $\hat{G} \xrightarrow{p} \tilde{G}$, there exists some $\delta(n) \to 0$ as $n \to \infty$ such that

$$D(\hat{G}, \tilde{G}) < \epsilon' \quad \text{with probability above } 1 - \delta(n).$$

Moreover, with probability approaching 1, we have $\hat{G} \sim \tilde{G}$ and thus $\Psi \sim G$. Given $\tilde{G}$, let

$$\overline{D} = \max_{\theta \in \Theta, D(\hat{G}, \tilde{G}) < \epsilon'} D(\Psi, F^{-1}(\hat{G}; \theta)) \leq \max_{\theta \in \Theta} D(\Psi, F^{-1}(\tilde{G}; \theta)) + \frac{\epsilon'}{1 - \bar{\rho}}$$

where the second inequality follows by for all θ, $F^{-1}(\tilde{G}; \theta)$ being Lipschitz continuous in $\tilde{G}$ with Lipschitz constant $\frac{1}{1-\bar{\rho}}$ and the triangle inequality. Note that with probability approaching 1, $\max_{\theta \in \Theta} D(\Psi, F^{-1}(\tilde{G}; \theta))$ is well-defined, since (1). $\Psi \sim \tilde{G}$ with

[17] Even if Ψ is not equivalent to $\hat{G}$, $T_{\hat{G}}\Psi$ is equivalent to $\hat{G}$.

probability approaching 1, (2). $F^{-1}(\tilde{G};\theta)$ is continuous in θ by Lemma 3, (3). metric D is continuous and Θ is compact.

Next, I show that with probability above $1-\delta(n)$, $\hat{Q}_{n,m} \to \hat{Q}^*_n$ uniformly in probability as $m \to +\infty$, and the convergence speed does not depend on n. Fix some n. Note

$$Prob_j(\theta, \hat{G}, m) - Prob^*_j(\theta, \hat{G}) = \int_{\boldsymbol{p}} f_j(\boldsymbol{p};\theta) d\big(T^m_{\hat{G},\theta}\Psi - F^{-1}(\hat{G},\theta)\big)(\boldsymbol{p}).$$

With probability above $1-\delta(n)$, we have $D(\hat{G},\tilde{G}) < \epsilon'$,

$$D(T^m_{\hat{G},\theta}\Psi, F^{-1}(\hat{G},\theta)) \leq \bar{\rho}^m D(\Psi, F^{-1}(\hat{G},\theta)) \leq \bar{\rho}^m \overline{D}$$

$$\begin{aligned}
\big|Prob_j(\theta,\hat{G},m) - Prob^*_j(\theta,\hat{G})\big| &\leq \left| \sup_{D(\Phi,\Upsilon)\leq \bar{\rho}^m\overline{D}} \int_{\boldsymbol{p}} f_j(\boldsymbol{p};\theta) d\big(\Phi - \Upsilon\big)(\boldsymbol{p}) \right| \\
&\leq (\overline{f} - \underline{f}) \frac{1}{2} \sup_{D(\Phi,\Upsilon)\leq \bar{\rho}^m\overline{D}} ||\Phi - \Upsilon||_{TV} \\
&\leq (\overline{f} - \underline{f}) \frac{1}{2}\frac{1}{2} J \bar{\rho}^m \overline{D}
\end{aligned}$$

where the last inequality is by applying Lemma 1 to the product measure. (Here we have an additional factor of J.[18]) Consequently,

$$\big|\hat{Q}_{n,m}(\theta) - \hat{Q}^*_n(\theta)\big| \leq \ln \frac{\underline{f} + \frac{1}{4}(\overline{f} - \underline{f}) J \bar{\rho}^m \overline{D}}{\underline{f}} \quad \text{with probability above } 1-\delta(n).$$

Note this bound does not depend on θ or n.

Lastly,

$$\sup_{\theta\in\Theta} \big|\hat{Q}_n(\theta) - Q_0(\theta)\big| \leq \sup_{\theta\in\Theta} \big|\hat{Q}_n(\theta) - \hat{Q}^*_n(\theta)\big| + \sup_{\theta\in\Theta} \big|\hat{Q}^*_n(\theta) - Q_0(\theta)\big|$$

where $\sup_{\theta\in\Theta} \big|\hat{Q}^*_n(\theta) - Q_0(\theta)\big| \xrightarrow{p} 0$ by Lemma 5. By $\delta(n) \to 0$, $m(n) \to +\infty$, and

$$\lim_{m\to+\infty} \ln \frac{\underline{f} + \frac{1}{4}(\overline{f} - \underline{f}) J \bar{\rho}^m \overline{D}}{\underline{f}} = 0,$$

we have $\sup_{\theta\in\Theta} \big|\hat{Q}_n(\theta) - \hat{Q}^*_n(\theta)\big| \xrightarrow{p} 0$. Thus, $\sup_{\theta\in\Theta} \big|\hat{Q}_n(\theta) - Q_0(\theta)\big| \xrightarrow{p} 0$ □

[18] This bound is not tight.

Proof of Theorem 3. We are ready to apply Theorem 2.1 in Newey and McFadden (1994). (1). By the identification assumption 3, $Q_0(\theta)$ is uniquely maximized at θ_0. (2). Θ is compact. (3). As $Prob_j^*(\theta; \tilde{G})$ is also bounded below by $\underline{f}$ and continuous in θ by Lemma 3, $Q_0(\theta)$ is continuous. (4). $\hat{Q}_n(\theta)$ converges in probability to $Q_0(\theta)$, by Lemma 6. Thus, $\hat{\theta}$ is consistent.

To see $T^{m(n)}_{\hat{G},\hat{\theta}}\Psi \xrightarrow{p} G$, note that

$$D(T^{m(n)}_{\hat{G},\hat{\theta}}\Psi, G) \leq D(T^{m(n)}_{\hat{G},\hat{\theta}}\Psi, F^{-1}(\hat{G},\hat{\theta}))+D(F^{-1}(\hat{G},\hat{\theta}), F^{-1}(\hat{G},\theta_0))+D(F^{-1}(\hat{G},\theta_0), G).$$

The first term

$$D(T^{m(n)}_{\hat{G},\hat{\theta}}\Psi, F^{-1}(\hat{G},\hat{\theta})) \to 0 \quad \text{as} \quad m(n) \to \infty.$$

The second term

$$D(F^{-1}(\hat{G},\hat{\theta}), F^{-1}(\hat{G},\theta_0)) \xrightarrow{p} 0, \quad \text{as} \quad \hat{\theta} \xrightarrow{p} \theta_0$$

and F^{-1} is continuous in θ by Lemma 3. The third term

$$D(F^{-1}(\hat{G},\theta_0), G) \xrightarrow{p} 0, \quad \text{as} \quad \hat{G} \xrightarrow{p} \tilde{G}$$

and F is a homeomorphism by Proposition 1.

□

B.4 Proof of Theorem 4

Lemma 7. *If Assumption 2, 3, and 4 hold, then $\hat{\theta}^*$ is asymptotically normal and $\sqrt{n}(\hat{\theta}^* - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$.*

Proof of Lemma 7. We shall first rewrite the estimator as a generalized method of moment estimator. We let $\hat{Prob} = (\hat{Prob}_1, \hat{Prob}_2, \cdots, \hat{Prob}_J)'$ denote the observed frequency of alternatives. Let $\mathbf{1}_p$ denote the cumulative indicator vector that assigns 0 for entries $p_j < p$ and 1 for entries $p_j \geq p$. Estimator $\hat{\theta}^*$ solves the first-order condition of Equation (12)

$$\frac{1}{n}\sum_{i=1}^{n} \mathfrak{g}^*(z_i, \theta, \hat{G}) = 0$$

where $\hat{G}$ satisfies the moment condition

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{Prob} - y_i) = 0 \tag{24}$$

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{G}_j - y_{ij}\mathbf{1}_{p_i}/\hat{Prob}_j) = 0 \quad \text{for all } j \in \mathcal{J} \tag{25}$$

where p_i is the observed selected price for individual i.

For this standard GMM estimator, we can directly invoke Theorem 6.1 in Newey and McFadden (1994). Note that our $\mathfrak{g}^*$ is their g and our $(\hat{Prob}, \hat{G})$ is their $\hat{\gamma}$ in Newey and McFadden (1994). Let

$$\mathfrak{m}_1(z_i, \hat{Prob}) = \hat{Prob} - y_i,$$

$$\mathfrak{m}_2(z_i, \hat{Prob}, \hat{G}) = [[\hat{G}_1 - y_{i1}\mathbf{1}_{p_i}/\hat{Prob}_1]', [\hat{G}_2 - y_{i2}\mathbf{1}_{p_i}/\hat{Prob}_2]', \cdots, [\hat{G}_J - y_{J2}\mathbf{1}_{p_i}/\hat{Prob}_J]']'$$

We stack $\mathfrak{g}^*$, $\mathfrak{m}_1$, $\mathfrak{m}_2$ to form $\tilde{\mathfrak{g}}^*$

$$\tilde{\mathfrak{g}}^*(z, \theta, \hat{Prob}, \hat{G}) = [\mathfrak{g}^*(z, \theta, \hat{G})', \mathfrak{m}_1(z, \hat{Prob})', \mathfrak{m}_2(z, \hat{Prob}, \hat{G})']'$$

By the proof of Theorem 3 and Lemma 5, $\hat{\theta}^* \xrightarrow{p} \theta_0$. By the weak law of large numbers, $\hat{G} \xrightarrow{p} \tilde{G}$ and $\hat{Prob} \xrightarrow{p} Prob_0 = Prob^*(\theta_0, \tilde{G})$. By Assumption 4, $\theta_0 \in \Theta^o$. Next, we verify that $\tilde{\mathfrak{g}}^*(z, \theta, \hat{Prob}, \hat{G})$ is continuously differentiable in $\theta, \hat{Prob}, \hat{G}$.

First, we verify that $\mathfrak{g}^*(z, \theta, \hat{G})$ is continuously differentiable in θ. It suffices to show that $Prob^*(\theta, \hat{G})$ is twice continuously differentiable in θ. As f is twice continuously differentiable in θ, we only need to show that $F^{-1}(\hat{G}, \theta)$ is twice continuously differentiable in θ. By Equation (1), (2) and f being twice continuously differentiable in θ, $F(G, \theta)$ is twice continuously differentiable in θ and infinitely continuously differentiable in G. Thus, by the implicit function theorem,

$$\nabla_\theta F^{-1}(\tilde{G}, \theta) = -\left[\nabla_G F(G, \theta)\right]^{-1} \nabla_\theta F(G, \theta)$$

where the matrix $\nabla_G F(G, \theta)$ is non-singular by F^{-1} being Lipschitz continuous. Consequently, F^{-1} is twice continuously differentiable in θ.

Next, we verify that $\mathfrak{g}^*(z, \theta, \hat{G})$ is continuously differentiable in $\hat{G}$. It suffices to show that $F^{-1}(\hat{G}, \theta)$ is continuously differentiable in $\hat{G}$. As $F(G, \theta)$ is infinitely

continuously differentiable in G and $F^{-1}(\hat{G},\theta)$ is Lipschitz continuous in $\hat{G}$, we have

$$\nabla_{\hat{G}} F^{-1}(\hat{G},\theta) = [\nabla_G F(G,\theta)]^{-1}$$

which is continuous in $\hat{G}$. Additionally, $\mathfrak{m}_1$ and $\mathfrak{m}_2$ are infinitely continuously differentiable in all parameters $\theta, \hat{G}, \hat{Prob}$. Consequently, we have show that $\tilde{\mathfrak{g}}^*(z,\theta,\hat{Prob},\hat{G})$ is continuously differentiable in $\theta, \hat{Prob}, \hat{G}$.

In addition,

$$\mathbb{E}[\mathfrak{g}^*(z,\theta_0,\tilde{G})] = 0$$

by the first-order condition of Q_0. Since f_j is bounded from 0, $||\mathfrak{g}^*(z,\theta_0,\tilde{G})||$ is finite for each z. Furthermore, as supp(G) is finite, there is only a finite possible values of z. Thus, $\mathbb{E}[||\mathfrak{g}^*(z,\theta_0,\tilde{G})||^2]$ is finite. By $\tilde{\mathfrak{g}}^*(z,\theta,\hat{Prob},\hat{G})$ being continuously differentiable in $(\theta,\hat{Prob},\hat{G})$ and a finite possible values of z,

$$\mathbb{E}[\sup_{\theta,\hat{Prob},\hat{G}} ||\nabla_{\theta,\hat{Prob},\hat{G}} \tilde{\mathfrak{g}}^*(z,\theta,\hat{Prob},\hat{G})||] < \infty.$$

The last condition we need is that

$$\mathbb{E}[\nabla_{\theta,\hat{Prob},\hat{G}} \tilde{\mathfrak{g}}^*(z,\theta_0,Prob_0,\tilde{G})]$$

being nonsingular. The matrix $\nabla_{\theta,\hat{Prob},\hat{G}} \tilde{\mathfrak{g}}^*(z,\theta_0,Prob_0,\tilde{G})$ is

$$\begin{pmatrix} \nabla_\theta \mathfrak{g}^*(z,\theta_0,\tilde{G}) & \mathbf{0} & \nabla_{\hat{G}} \mathfrak{g}^*(z,\theta_0,\tilde{G}) \\ \mathbf{0} & \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \nabla_{\hat{Prob}} \mathfrak{m}_2(z,Prob_0,\tilde{G}) & \boldsymbol{I} \end{pmatrix}$$

Its expectation being nonsingular is equivalent to $\mathbb{E}\nabla_\theta \mathfrak{g}^*(z,\theta_0,\tilde{G})$ being nonsingular, which is in Assumption 4.

We can write down the variance matrix V by Theorem 6.1 in Newey and McFadden (1994).

$$\begin{aligned} A(z) =& \mathfrak{g}^*(z,\theta_0,\tilde{G}) + \big(\mathbb{E}\nabla_{\hat{G}} \mathfrak{g}^*(z,\theta_0,\tilde{G})\big) \times \\ & \big[\big(\mathbb{E}\nabla_{\hat{Prob}} m_2(z,Prob_0,\tilde{G})\big) \times \mathfrak{m}_1(z,Prob_0) - \mathfrak{m}_2(z,Prob_0,\tilde{G})\big]. \end{aligned}$$

$$V = \left(\mathbb{E}\nabla_\theta \mathfrak{g}^*(z, \theta_0, \tilde{G})\right)^{-1} \times \mathbb{E}(A(z)A(z)') \times \left(\left(\mathbb{E}\nabla_\theta \mathfrak{g}^*(z, \theta_0, \tilde{G})\right)^{-1}\right)'$$

□

Proof of Theorem 4. Recall $\hat{\theta}$ solves the first-order condition

$$\frac{1}{n}\sum_{i=1}^{n} \mathfrak{g}(z_i; \hat{\theta}, \hat{G}, n) = 0.$$

We expand this equation around θ_0 and solve for $\sqrt{n}(\hat{\theta} - \theta_0)$

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left[\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \mathfrak{g}(z_i, \bar{\theta}, \hat{G}, n)\right]^{-1} \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \mathfrak{g}(z_i, \theta_0, \hat{G}, n)$$

where the second summation is

$$\sum_{i=1}^{n} \frac{1}{\sqrt{n}} \mathfrak{g}(z_i, \theta_0, \hat{G}, n) = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \left(\mathfrak{g}^*(z_i, \theta_0, \hat{G}) + \mathcal{o}_p(\frac{1}{\sqrt{n}})\right) = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \mathfrak{g}^*(z_i, \theta_0, \hat{G}) + \mathcal{o}_p(1)$$

by Assumption 4 (v). Similarly,

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \mathfrak{g}(z_i, \bar{\theta}, \hat{G}, n) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \mathfrak{g}^*(z_i, \bar{\theta}, \hat{G}) + \mathcal{o}_p(1).$$

Thus, $\sqrt{n}(\hat{\theta} - \theta_0)$ converges to

$$\left(\mathbb{E}\nabla_\theta \mathfrak{g}^*(z; \theta_0, \tilde{G})\right)^{-1} \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \mathfrak{g}^*(z_i, \theta_0, \hat{G}) + \mathcal{o}_p(1)$$

which has the same limiting distribution as $\sqrt{n}(\hat{\theta}^* - \theta_0)$. Thus, $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$ by Lemma 7. □

To see the convergence rate of $T^{m(n)}_{\hat{G},\hat{\theta}}\Psi$, note that

$$D(T^{m(n)}_{\hat{G},\hat{\theta}}\Psi, G) \leq D(T^{m(n)}_{\hat{G},\hat{\theta}}\Psi, F^{-1}(\hat{G}, \hat{\theta})) + D(F^{-1}(\hat{G}, \hat{\theta}), F^{-1}(\hat{G}, \theta_0)) + D(F^{-1}(\hat{G}, \theta_0), G).$$

The first term goes to 0 at rate faster than $\sqrt{n}$ by Assumption 4 (v). By the proof of Lemma 7, F^{-1} is continuously differentiable in θ; as Θ is compact, F^{-1} is Lipschitz

continuous in θ. As $\hat{\theta} \xrightarrow{p} \theta_0$ at rate $\sqrt{n}$,

$$D(F^{-1}(\hat{G}, \hat{\theta}), F^{-1}(\hat{G}, \theta_0)) \xrightarrow{p} 0 \quad \text{at rate } \sqrt{n}.$$

The last term converges in probability to 0 at rate $\sqrt{n}$, as $\hat{G} \xrightarrow{p} \tilde{G}$ at rate $\sqrt{n}$ and by Proposition 1.

C Tables

Table 3: Simulation Results for Utility Parameters: $N = 5000$

	Functional Contraction			Two-Step Method		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
			DGP 1			
γ	-0.0022	0.0864	0.0864	0.0017	0.0917	0.0916
ξ_2	0.0028	0.0345	0.0345	0.0045	0.0427	0.0429
β	-0.0007	0.0412	0.0411	0.0003	0.0417	0.0416
			DGP 2			
γ	0.0011	0.0858	0.0857	0.0157	0.0982	0.0994
ξ_2	0.0001	0.0344	0.0344	0.0051	0.0510	0.0512
β	-0.0009	0.0427	0.0427	0.0052	0.0434	0.0436
			DGP 3			
γ	-0.0133	0.0707	0.0719	0.1611	0.0984	0.1887
ξ_2	0.0019	0.0317	0.0317	0.0137	0.0363	0.0388
β	-0.0012	0.0414	0.0414	-0.0002	0.0417	0.0417
			DGP 4			
γ	0.0026	0.1544	0.1543	0.0433	0.1666	0.1720
ξ_2	0.0009	0.0306	0.0306	-0.0001	0.0314	0.0314
β	-0.0003	0.0404	0.0404	-0.0002	0.0406	0.0405
			DGP 5			
γ	-0.0054	0.4395	0.4391	-0.0043	0.4274	0.4270
ξ_2	0.0009	0.0304	0.0304	0.0010	0.0305	0.0305
β	0.0002	0.0399	0.0398	0.0003	0.0398	0.0398

Table 4: Simulation Results for CDF of log(Price): $N = 5000$

	Func. Contraction		Two-Step Method	
	IBias^2	IMSE	IBias^2	IMSE
		DGP 1		
$F_1(\cdot\|x_{i2}=0)$	0.0002	0.0007	0.0000	0.0040
$F_2(\cdot\|x_{i2}=0)$	0.0001	0.0002	0.0000	0.0003
$F_1(\cdot\|x_{i2}=1)$	0.0002	0.0009	0.0000	0.0024
$F_2(\cdot\|x_{i2}=1)$	0.0001	0.0003	0.0000	0.0008
		DGP 2		
$F_1(\cdot\|x_{i2}=0)$	0.0003	0.0009	0.0024	0.0078
$F_2(\cdot\|x_{i2}=0)$	0.0001	0.0002	0.0020	0.0023
$F_1(\cdot\|x_{i2}=1)$	0.0004	0.0010	0.0024	0.0056
$F_2(\cdot\|x_{i2}=1)$	0.0001	0.0003	0.0020	0.0029
		DGP 3		
$F_1(\cdot\|x_{i2}=0)$	0.0059	0.0064	0.0209	0.0269
$F_2(\cdot\|x_{i2}=0)$	0.0028	0.0029	0.0493	0.0499
$F_1(\cdot\|x_{i2}=1)$	0.0005	0.0011	0.0037	0.0057
$F_2(\cdot\|x_{i2}=1)$	0.0001	0.0003	0.0143	0.0152
		DGP 4		
$F_1(\cdot\|x_{i2}=0)$	0.0006	0.0011	0.0024	0.0077
$F_2(\cdot\|x_{i2}=0)$	0.0007	0.0008	0.0272	0.0277
$F_1(\cdot\|x_{i2}=1)$	0.0003	0.0012	0.0016	0.0047
$F_2(\cdot\|x_{i2}=1)$	0.0001	0.0003	0.0098	0.0104
		DGP 5		
$F_1(\cdot\|x_{i2}=0)$	0.0014	0.0018	0.0011	0.0053
$F_2(\cdot\|x_{i2}=0)$	0.0014	0.0015	0.0202	0.0206
$F_1(\cdot\|x_{i2}=1)$	0.0007	0.0018	0.0016	0.0055
$F_2(\cdot\|x_{i2}=1)$	0.0001	0.0003	0.0081	0.0084

Note: The IBias^2 of a function h is calculated as follows. Let $\hat{h}_r$ be the estimate of h from the r-th simulated dataset, and $\bar{h}(x) = \frac{1}{R}\sum_{r=1}^{R}\hat{h}_r(x)$ be the point-wise average over R simulations. The integrated squared bias is calculated by numerically integrating the point-wise squared bias $(\bar{h}(x) - h(x))^2$ over the distribution of x. The integrated MSE is computed in a similar way.

Table 5: Simulation Results for Utility Parameters: Removing the Excluded Variable

	$N = 1000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
			DGP 1			
γ	-0.0011	0.2082	0.2080	0.0018	0.0866	0.0865
ξ_2	0.0074	0.0570	0.0574	0.0000	0.0254	0.0253
			DGP 2			
γ	-0.0018	0.2066	0.2064	0.0024	0.1000	0.0999
ξ_2	0.0033	0.0535	0.0535	0.0028	0.0264	0.0265
			DGP 3			
γ	-0.0163	0.1581	0.1587	-0.0043	0.0728	0.0729
ξ_2	0.0061	0.0542	0.0544	0.0007	0.0238	0.0238
			DGP 4			
γ	0.0019	0.3660	0.3656	0.0059	0.1563	0.1563
ξ_2	0.0050	0.0498	0.0500	-0.0005	0.0225	0.0225
			DGP 5			
γ	0.0000	1.0797	1.0786	-0.0146	0.4409	0.4407
ξ_2	0.0014	0.0531	0.0530	-0.0007	0.0233	0.0233

Note: In these specifications, we remove the excluded variable from the selection function, so the parameter β in u_{i1} is not estimated.

Table 6: Simulation Results for CDF of log(Price): Removing the Excluded Variable

	$N = 1000$		$N = 5000$	
	IBias^2	IMSE	IBias^2	IMSE
	DGP 1			
$F_1(\cdot\|x_{i2} = 0)$	0.0002	0.0018	0.0002	0.0004
$F_2(\cdot\|x_{i2} = 0)$	0.0001	0.0003	0.0000	0.0001
$F_1(\cdot\|x_{i2} = 1)$	0.0003	0.0019	0.0002	0.0005
$F_2(\cdot\|x_{i2} = 1)$	0.0001	0.0007	0.0001	0.0002
	DGP 2			
$F_1(\cdot\|x_{i2} = 0)$	0.0004	0.0017	0.0003	0.0006
$F_2(\cdot\|x_{i2} = 0)$	0.0002	0.0004	0.0001	0.0001
$F_1(\cdot\|x_{i2} = 1)$	0.0005	0.0018	0.0003	0.0006
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0007	0.0001	0.0002
	DGP 3			
$F_1(\cdot\|x_{i2} = 0)$	0.0058	0.0073	0.0061	0.0064
$F_2(\cdot\|x_{i2} = 0)$	0.0029	0.0031	0.0028	0.0028
$F_1(\cdot\|x_{i2} = 1)$	0.0006	0.0021	0.0005	0.0008
$F_2(\cdot\|x_{i2} = 1)$	0.0001	0.0007	0.0000	0.0002
	DGP 4			
$F_1(\cdot\|x_{i2} = 0)$	0.0006	0.0021	0.0006	0.0008
$F_2(\cdot\|x_{i2} = 0)$	0.0008	0.0010	0.0007	0.0007
$F_1(\cdot\|x_{i2} = 1)$	0.0004	0.0024	0.0003	0.0007
$F_2(\cdot\|x_{i2} = 1)$	0.0001	0.0006	0.0000	0.0002
	DGP 5			
$F_1(\cdot\|x_{i2} = 0)$	0.0014	0.0025	0.0013	0.0016
$F_2(\cdot\|x_{i2} = 0)$	0.0013	0.0015	0.0014	0.0014
$F_1(\cdot\|x_{i2} = 1)$	0.0007	0.0033	0.0006	0.0012
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0006	0.0001	0.0002

Note: In these specifications, we remove the excluded variable from the selection function. The IBias^2 of a function h is calculated as follows. Let $\hat{h}_r$ be the estimate of h from the r-th simulated dataset, and $\bar{h}(x) = \frac{1}{R}\sum_{r=1}^{R} \hat{h}_r(x)$ be the point-wise average over R simulations. The integrated squared bias is calculated by numerically integrating the point-wise squared bias $(\bar{h}(x) - h(x))^2$ over the distribution of x. The integrated MSE is computed in a similar way.

Table 7: Simulation Results for Utility Parameters: Misspecifying the Selection Function

	$N = 1000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
			DGP 1			
γ	-0.0793	0.1826	0.1989	-0.0743	0.0806	0.1096
ξ_2	-0.0754	0.0714	0.1038	-0.0752	0.0342	0.0826
β	-0.0309	0.0856	0.0909	-0.0306	0.0392	0.0497
			DGP 2			
γ	-0.0748	0.1864	0.2007	-0.0667	0.0806	0.1045
ξ_2	-0.0714	0.0727	0.1018	-0.0742	0.0340	0.0816
β	-0.0315	0.0902	0.0954	-0.0282	0.0407	0.0495
			DGP 3			
γ	-0.1051	0.1475	0.1810	-0.0940	0.0650	0.1142
ξ_2	-0.0789	0.0698	0.1053	-0.0768	0.0317	0.0830
β	-0.0323	0.0887	0.0943	-0.0293	0.0398	0.0494
			DGP 4			
γ	-0.0746	0.3249	0.3330	-0.0584	0.1442	0.1554
ξ_2	-0.0776	0.0679	0.1030	-0.0755	0.0307	0.0814
β	-0.0263	0.0900	0.0937	-0.0222	0.0392	0.0450
			DGP 5			
γ	0.0029	0.9169	0.9160	-0.0606	0.4177	0.4217
ξ_2	-0.0827	0.0667	0.1063	-0.0801	0.0302	0.0856
β	-0.0315	0.0847	0.0902	-0.0266	0.0381	0.0465

Note: In these specifications, we misspecify the selection model, assuming that the error term ε_i is drawn from $Logistic(0, 1)$.

Table 8: Simulation Results for CDF of log(Price): Misspecifying the Selection Function

	$N = 1000$		$N = 5000$	
	IBias^2	IMSE	IBias^2	IMSE
	DGP 1			
$F_1(\cdot\|x_{i2} = 0)$	0.0004	0.0029	0.0002	0.0007
$F_2(\cdot\|x_{i2} = 0)$	0.0001	0.0006	0.0001	0.0002
$F_1(\cdot\|x_{i2} = 1)$	0.0004	0.0032	0.0002	0.0009
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0013	0.0001	0.0003
	DGP 2			
$F_1(\cdot\|x_{i2} = 0)$	0.0006	0.0033	0.0005	0.0010
$F_2(\cdot\|x_{i2} = 0)$	0.0002	0.0006	0.0001	0.0002
$F_1(\cdot\|x_{i2} = 1)$	0.0007	0.0037	0.0004	0.0010
$F_2(\cdot\|x_{i2} = 1)$	0.0003	0.0015	0.0001	0.0004
	DGP 3			
$F_1(\cdot\|x_{i2} = 0)$	0.0062	0.0087	0.0061	0.0066
$F_2(\cdot\|x_{i2} = 0)$	0.0028	0.0032	0.0028	0.0029
$F_1(\cdot\|x_{i2} = 1)$	0.0007	0.0033	0.0005	0.0011
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0013	0.0001	0.0003
	DGP 4			
$F_1(\cdot\|x_{i2} = 0)$	0.0008	0.0034	0.0006	0.0012
$F_2(\cdot\|x_{i2} = 0)$	0.0008	0.0012	0.0007	0.0008
$F_1(\cdot\|x_{i2} = 1)$	0.0006	0.0046	0.0003	0.0012
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0011	0.0001	0.0003
	DGP 5			
$F_1(\cdot\|x_{i2} = 0)$	0.0014	0.0034	0.0014	0.0019
$F_2(\cdot\|x_{i2} = 0)$	0.0014	0.0018	0.0014	0.0015
$F_1(\cdot\|x_{i2} = 1)$	0.0008	0.0058	0.0007	0.0018
$F_2(\cdot\|x_{i2} = 1)$	0.0002	0.0011	0.0001	0.0003

Note: In these specifications, we misspecify the selection model, assuming that the error term ε_i is drawn from $Logistic(0, 1)$. The IBias^2 of a function h is calculated as follows. Let $\hat{h}_r$ be the estimate of h from the r-th simulated dataset, and $\bar{h}(x) = \frac{1}{R}\sum_{r=1}^{R} \hat{h}_r(x)$ be the point-wise average over R simulations. The integrated squared bias is calculated by numerically integrating the point-wise squared bias $(\bar{h}(x) - h(x))^2$ over the distribution of x. The integrated MSE is computed in a similar way.