

Estimating Nonseparable Selection Models: A Functional Contraction Approach*

Fan Wu Yi Xin

December 19, 2025

Abstract

We propose a novel method for estimating nonseparable selection models. We show that, for a given selection function, the potential outcome distributions are nonparametrically identified from the selected outcome distributions and can be recovered using a simple iterative algorithm based on a contraction mapping. This result enables a full-information approach to estimating selection models without imposing parametric or separability assumptions on the outcome equation. We propose a three-step estimation strategy for the potential outcome distributions and the parameters of the selection function and establish the consistency and asymptotic normality of the resulting estimators. Monte Carlo simulations demonstrate that our approach performs well in finite samples. The method is applicable to a wide range of empirical settings, including consumer demand models with only transaction prices, auctions with incomplete bid data, and Roy models with data on accepted wages.

Keywords: Sample Selection, Nonseparable Models, Functional Contraction, Potential Outcome Distribution, Semiparametric Estimation, Demand Estimation, Auction, Roy Models.

JEL Codes: C14, C24, C51, L11, D44, J31.

*Wu: Peking University HSBC Business School, University Town of Shenzhen, China. Email: fanwu@phbs.pku.edu.cn. 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 Jeremy Fox, Michael Keane, Philip Haile, Yingyao Hu, Yao Luo, Luciano Pomatto, Robert Sherman, Xun Tang, Omer Tamuz, and Ao Wang. We thank seminar and conference participants at Caltech, CUFE, Peking University, Tsinghua University, UC Irvine, University of Michigan, USC, 2025 Asian Summer School in Econometrics and Statistics, Econometric Society World Congress 2025, DSE Conference 2025.

1 Introduction

Sample selection issues arise in many empirical settings. In consumer demand studies, researchers often observe only the transaction prices of chosen products (Goldberg, 1996; Cicala, 2015; Crawford et al., 2018; Allen et al., 2019; D’Haultfœuille et al., 2019; Sagl, 2023; Cosconati et al., 2025). In auctions, available data may consist solely of the winning bids (Athey and Haile, 2002; Komarova, 2013; Guerre and Luo, 2019; Allen et al., 2024). In labor economics, wage data are typically observed only for individuals who choose to work (Gronau, 1974; Heckman, 1974), and in the Roy model (Roy, 1951), earnings within an occupation are observed only for those who self-select into that sector.

Observing only a selected sample of outcomes—such as prices, bids, or wages—poses significant challenges for estimating two key elements. First is the selection function, which specifies how agents choose among alternatives, for example, through a consumer demand system, an auction’s winning rule, or a labor force participation decision. Second is the distribution of outcomes *prior to selection*, often referred to as “potential outcomes” in the literature. 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.

Our paper proposes a new approach to estimating nonseparable selection models. A key novelty of our contribution is a constructive identification result showing that, for a given selection function, the potential outcome distributions are nonparametrically identified from the selected outcome distributions and can be recovered using a simple iterative algorithm. This result reveals a fundamental one-to-one mapping between the outcome distributions before and after selection, implying that estimation of the selection model essentially reduces to recovering the selection function from observed choice patterns. Our method demonstrates that a full-information approach to estimating selection models can be implemented without imposing parametric or separability assumptions on the outcome equation.

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 alternatives are independent conditional on observable and unobservable characteristics, which allows for correlation across outcomes when conditioning only on observables. In addition, we allow the unobservable component in the outcome equation to enter directly into agents' preferences over alternatives, thereby capturing selection on unobservables.

We 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 selected outcome distributions using Bayes' rule. This two-way relationship characterizes a fixed-point problem.

Building on this intuition, we construct an operator whose fixed point is the potential outcome distributions and establish sufficient conditions for it to be a functional contraction (Theorems 1 and 2). Our results imply that, given the selection function and the distributions of selected outcomes, we can nonparametrically identify 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. This process converges to the potential outcome distributions associated with the selection function.

We then embed this identification result into a three-step estimation strategy for the unobserved potential outcome distributions and the parameters of the selection function. In the first step, we estimate the selected outcome distribution conditional on both observable and unobservable covariates using instrumental variables. In the second step, we propose a nested fixed-point algorithm to estimate the parameters of the selection function: in the inner loop, for any candidate selection function, we recover the potential outcome distributions by iterating the operator, while in the outer loop we search for the parameter values that maximize the likelihood of the observed choice patterns. Finally, in the third step, we obtain the potential outcome distributions by reapplying the fixed-point algorithm using the estimated parameters of the selection function.

We establish the consistency and asymptotic normality of the proposed estimators

in Theorems 3 and 4. To examine their finite sample properties, we conduct Monte Carlo simulations across various designs of the outcome equation. Our results show that the biases in our estimators 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 classic Heckman parametric two-step approach and the quantile selection model of [Arellano and Bonhomme \(2017\)](#) with linear conditional quantiles and a Gaussian copula, particularly when the outcome equation contains nonseparable error terms. In addition, we show that our approach does not require an excluded variable in the selection equation and remains robust even when the selection function is misspecified by the econometrician.

In a companion paper [Cosconati et al. \(2025\)](#), we apply our method to estimate consumer demand for auto insurance products when only transaction prices are observed. We nonparametrically estimate the offered price distribution for each insurance company and allow these distributions to vary fully flexibly across firms. The substantial heterogeneity in the recovered price distributions reflects differences in firms' information technologies and cost structures, which are key primitives we estimate through a supply-side competition model. We omit the details of this application here and refer readers to [Cosconati et al. \(2025\)](#) for the full empirical setup and results.

Related Literature Our paper contributes to the extensive theoretical literature on sample selection models. An early solution to sample selection bias is full information maximum likelihood (FIML) estimation based on parametric assumptions, as in [Heckman \(1974\)](#) and [Lee \(1982, 1983\)](#). More commonly employed methods for sample selection models are the two-step control function approach pioneered by [Heckman \(1976, 1979\)](#). A substantial body of theoretical work has been developed to relax the distributional assumptions in the two stages of the estimation procedure (e.g., [Ahn and Powell, 1993](#); [Andrews and Schafgans, 1998](#); [Chen and Khan, 2003](#); [Das et al., 2003](#); [Newey, 2007, 2009](#); [Chernozhukov et al., 2023](#); [Fernández-Val et al., 2024](#)). For a comprehensive survey of semiparametric two-step estimation methods for selection models, see [Vella \(1998\)](#).

Compared with the existing methods, our approach offers several key advantages. First, we allow the outcome equation to be nonparametric and nonseparable in the error terms, and we exploit the full information in the selected outcome distribution

to recover the entire distribution of potential outcomes. [Newey \(2007\)](#) and [Fernández-Val et al. \(2024\)](#) use control function approaches to correct for sample selection in nonseparable models with binary and censored selection rules, respectively, and they focus on identifying certain global and local parameters of the outcome distribution.

Second, our method accommodates fully heterogeneous effects of covariates on outcomes, whereas most existing approaches that estimate conditional mean models restrict covariates to affecting only the location of the outcome distribution.¹ More recently, [Arellano and Bonhomme \(2017\)](#) propose a method to correct for sample selection in quantile regression models by modeling the copula of the error terms in the outcome and selection equations. Although identification under more general settings is discussed, their estimation strategy primarily considers linear conditional quantile models and copulas characterized by a low-dimensional set of parameters.

Third, our approach does not require an instrument to exogenously shift the choice probability, a key to achieving identification in the two-step method, nor does it rely on identification-at-infinity arguments. In practice, finding a suitable instrument can be quite challenging (see [Vella \(1998\)](#) for further discussion).² When the conditioning set of variables includes an unobservable component, our method requires instruments for estimating the distribution of selected outcomes conditional on the unobservable in the first step. We adopt the measurement error framework in [Hu \(2008\)](#) and [Hu and Schennach \(2008\)](#), where the key requirement is to find instruments such that, conditional on the latent variable, the outcome and the instruments are independent.

At a broader conceptual level, our reliance on the structural restrictions implied by the selection model resonates with the nonparametric identification literature on auction models with missing bids. [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

¹An exception is the recent paper by [Chernozhukov et al. \(2023\)](#), which proposes a semiparametric generalization of the Heckman selection model that allows for rich forms of heterogeneity in the effects of covariates on both outcomes and selection.

²[d'Haultfoeuille and Maurel \(2013\)](#) and [D'Haultfoeuille 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.

nonparametric identification of symmetric IPV first-price auctions with only winning bids, accounting for unobserved competition. In these auction models, the selection rule is deterministic conditional on bids (the highest bidder wins), which allows order-statistic arguments to be applied. In contrast, our selection model assigns a probability distribution over alternatives and is therefore closer in spirit to multi-attribute auction environments (see e.g., [Krasnokutskaya et al. \(2020\)](#)). Moreover, our framework can flexibly accommodate asymmetries across alternatives, whereas bidder asymmetries are known to pose significant challenges in auction models (see the discussion in the handbook chapter by [Athey and Haile \(2007\)](#)).

Outline The rest of the paper is organized as follows. Section 2 formally introduces our model and provides an illustrative example. Section 3 presents the main theoretical results. In Section 4, we describe our estimation strategy and establish the asymptotic properties of the estimators. Section 5 reports results from our Monte Carlo simulations, and Section 6 discusses the empirical application in [Cosconati et al. \(2025\)](#). Section 7 concludes. All proofs are collected in the appendix.

2 Model

In Sections 2–3, all analyses are conditional on a vector of characteristics (x, x^*) , where x denotes observables and x^* denotes unobservables. The structure of the model and the main theoretical results *do not* depend on whether the conditioning set includes unobserved components, although the presence of unobservables introduces additional challenges for estimation, which we address in Section 4. Because all results in these two sections are stated conditional on (x, x^*) , we omit these variables from the notation to simplify exposition. 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 $\mathcal{J} = \{1, \dots, J\}$. Each alternative is associated with a price distribution. Let $G_j \in \Delta([\underline{p}_j, \bar{p}_j])$ 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 condition on sufficient x and x^* such that $p_j \sim G_j$ are independently distributed across alternatives. 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 $f = (f_1, f_2, \dots, f_J)$ where f_j maps the prices of alternatives $\mathbf{p} = (p_1, \dots, p_J)$ to a strictly positive probability of selecting alternative $j \in \mathcal{J}$.³ We assume that the selection function is continuously differentiable,

$$f_j \in \mathscr{C}^1: \prod_j [\underline{p}_j, \bar{p}_j] \rightarrow (0, 1],$$

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

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

where $Pr_j(\cdot; G)$ is a function defined on $[\underline{p}_j, \bar{p}_j]$. Independent prices across different alternatives (conditional on x and x^*) allow us to express the joint distribution of \mathbf{p}_{-j} as the product of their individual marginal distribution functions.

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

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

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 $F: \prod_j \Delta([\underline{p}_j, \bar{p}_j]) \rightarrow \prod_j \Delta([\underline{p}_j, \bar{p}_j])$ denote this mapping, i.e., $\tilde{G} = F(G)$.

³The 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 $[\underline{p}_j, \bar{p}_j]$. 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.

In many empirical settings, researchers have access only to the selected price distribution, such as the distribution of transaction prices, accepted wages, or winning bids. However, the key primitives of interest are often the offered price distribution, such as the distributions of posted prices, wage offers, or submitted bids. Our research question is how to recover the offered price distribution G from the 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 utilities from products 1 and 2 are:

$$u_1 = \gamma p_1 + x^* \kappa + \varepsilon_1, \quad (3)$$

$$u_2 = \gamma p_2 + \varepsilon_2, \quad (4)$$

where p_j represents the price of product j for this consumer, and ε_j is an idiosyncratic utility shock. We also allow an unobservable characteristic x^* to enter directly into the utility from product 1. This captures settings in which unobserved traits affect preferences across alternatives. For example, high-risk consumers may prefer a product offered by Firm 1 in an insurance market; in a labor setting, workers with higher productivity may prefer certain types of jobs. Our model can flexibly allow utility to depend on observable consumer attributes as well, but we omit these terms here for simplicity of presentation.

In this model, the price sensitivity parameter γ , coefficient κ , and the distribution of ε_j determine the selection function f , for any fixed x^* . If $\varepsilon_1 - \varepsilon_2 \sim \mathcal{N}(0, 1)$, the

selection function for product 1 takes the standard binary probit form:

$$f_1(p_1, p_2; x^*) = 1 - \Phi_N(\gamma(p_2 - p_1) - x^* \kappa),$$

where Φ_N denotes the CDF for standard normal distribution. For simplicity, we denote the difference in unobservables across the two utilities as $\tilde{\varepsilon} = x^* \kappa + (\varepsilon_1 - \varepsilon_2)$.

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:

$$p_j = x\beta_j + x^*\delta_j + \eta_j \equiv x\beta_j + \eta_j^*,$$

where x denotes observable characteristics, x^* denotes unobservable characteristics, and η_j is an idiosyncratic shock. We define the composite error in the pricing equation as $\eta_j^* \equiv x^*\delta_j + \eta_j$, which, for simplicity, is assumed to be independent of x . We assume that the true underlying price shocks η_1 and η_2 are independent. However, when x^* is unobserved by the econometrician, the composite errors η_1^* and η_2^* may be correlated through 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:

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

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 error in the selection model (ε^*) and the error in the outcome equation (η_1^*). Specifically,

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

Equation (6) shows that the error term η_1^* directly enters the composite error ε^* , generating the first term $\gamma \text{var}(\eta_1^*) \neq 0$ unless $\gamma = 0$. This correlation is *by construction* in selection models, as agents make decisions after observing the potential outcomes. The second term in Equation (6) reflects the correlation between the composite errors in the two outcome equations. It is important to emphasize that our independence assumption—that p_j are independently distributed across alternatives conditional on (x, x^*) —requires independence of the underlying shocks η_1 and η_2 , not of the composite errors η_1^* and η_2^* . Thus, our framework accommodates settings in which prices across alternatives are correlated conditional on observables. For instance, if x^* captures a worker’s unobserved productivity, then wage offers from two firms may appear correlated when x^* is not conditioned on. 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 by allowing unobservable characteristics x^* to enter both the selection equation and the outcome equation.

This simple two-product example illustrates how our notation for the selection function and the offered price distribution maps to the conventions commonly used in the existing literature, and highlights how our setting accommodates all key sources of correlation that give rise to selection bias. The general framework introduced in Section 2 is substantially more flexible than this illustrative case. In particular, our model allows the outcome equation to be fully flexible and nonparametrically specified with an nonseparable error term. 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 Equations (3) and (4); 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

We now present our main theoretical results on how to recover the offered price distribution from the selected 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):

$$G_j(p_j) = \frac{\int_{\underline{p}_j}^{p_j} d\tilde{G}_j(p)/Pr_j(p; G)}{\int_{\underline{p}_j}^{\bar{p}_j} d\tilde{G}_j(p)/Pr_j(p; G)}. \quad (7)$$

Note that if the selection probability $Pr_j(\cdot; G)$ were known—that is, the probability of selecting product j conditional on its offered price—then recovering the offered price distribution from Equation (7) would be straightforward.

We illustrate this inversion process using the simulated example in Figure 1. The red solid line plots the selected price density for an alternative. Dividing this density by the probability that the alternative is chosen at each price, $Pr_j(p; G)$, yields the unnormalized offered price density shown by the blue dashed line. The gap between these two densities captures the selection mechanism: when a lower price is offered, agents are more likely to accept it, whereas higher prices make them more likely to choose other alternatives. The offered price distribution shown by the blue solid line is then obtained after normalization, which corresponds to the denominator in Equation (7).

However, $Pr_j(\cdot; G)$ is *not* known, because it depends on the offered price distribution G , which we seek to recover. A tentative solution is to start with a conjecture Ψ for the offered price distribution and use it to compute the implied selection probability $Pr_j(\cdot; \Psi)$. Equation (7) then delivers an *updated* conjecture of the offered price distribution. This procedure, which maps a conjectured offered price distribution into its updated version, defines an operator T : $\prod_j \Delta([\underline{p}_j, \bar{p}_j]) \rightarrow \prod_j \Delta([\underline{p}_j, \bar{p}_j])$ as follows.

$$(T\Psi)_j(p_j) = \frac{\int_{\underline{p}_j}^{p_j} d\tilde{G}_j(p)/Pr_j(p; \Psi)}{\int_{\underline{p}_j}^{\bar{p}_j} d\tilde{G}_j(p)/Pr_j(p; \Psi)} \quad (8)$$

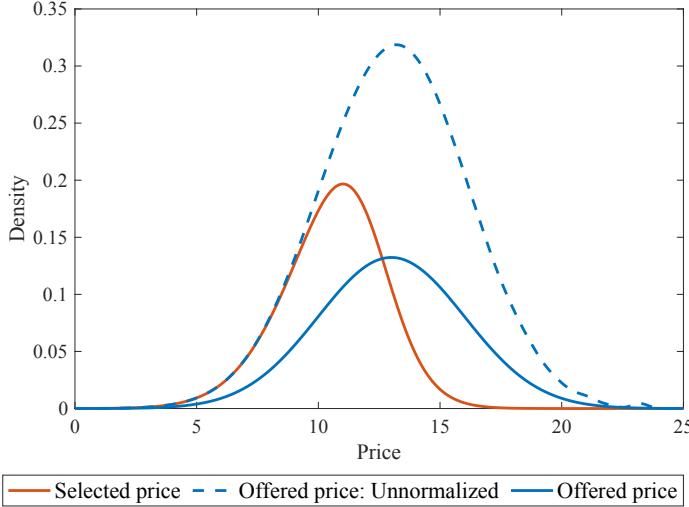


Figure 1: Densities of offered and selected prices. We draw offered prices from $\mathcal{N}(13, 3)$, and the probability that the agent given price p chooses this alternative is given by $\exp(10 - p)/(0.1 + \exp(10 - p))$.

where $\Psi = (\Psi_1, \Psi_2, \dots, \Psi_J) \in \prod_j \Delta([\underline{p}_j, \bar{p}_j])$. Importantly, if the conjecture Ψ is correct, i.e., $\Psi = G$, then the selection probability $Pr_j(\cdot; \Psi)$ is correctly specified, ensuring that the updated conjecture $T\Psi$ also equals G . Thus, the offered price distribution G is 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 $\Psi, \Phi \in \prod_j \Delta([\underline{p}_j, \bar{p}_j])$,

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

given some metric D .⁴ In the remainder 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 $\Delta([\underline{p}_j, \bar{p}_j])$. 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. Since $f_j > 0$, $G_j \sim \tilde{G}_j$. When

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

$\Psi_j \sim \Phi_j$, the Radon-Nikodym derivative,

$$\frac{d\Psi_j}{d\Phi_j} : [\underline{p}_j, \bar{p}_j] \rightarrow (0, \infty),$$

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

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

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

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

Given our operator T in Equation (8), for all $\Psi_j, \Phi_j \in \Delta([\underline{p}_j, \bar{p}_j])$,

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

Thus,

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

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 $(T\Phi)_j$ relies only on the

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

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

ratio between selection probabilities:

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

(The equality holds when \tilde{G}_j admits full support on $[\underline{p}_j, \bar{p}_j]$.) Since $f_j > 0$ is continuous with compact support, Pr_j is bounded away from 0. Thus, $d((T\Psi)_j, G_j)$, $d((T\Psi)_j, \tilde{G}_j)$ and $d((T\Psi)_j, (T\Phi)_j)$ are all finite.

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

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

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

3.2 Functional Contraction

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

$$M_j = \sup_{p_j, \mathbf{p}_{-j}, \mathbf{p}'_{-j}} \left| \frac{\partial \ln f_j(p_j, \mathbf{p}_{-j})}{\partial p_j} - \frac{\partial \ln f_j(p_j, \mathbf{p}'_{-j})}{\partial p_j} \right|. \quad (9)$$

The quantity $\frac{\partial \ln f_j}{\partial p_j}$ measures the sensitivity of the log choice probability to price and is therefore referred to as the semi-elasticity. Let

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

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

Proof. See Appendix B.1. □

Theorem 1 establishes a key identification result for selection models. 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 . Notably, the theorem requires no assumptions on the functional form of the potential outcome distributions, allowing

the outcome equation to be fully nonparametric and nonseparable in the error terms, and it applies to any form of the selection function f .

Moreover, the result in Theorem 1 provides a *constructive* method for solving for G . Take any $\Psi \in \prod_j \Delta([\underline{p}_j, \bar{p}_j])$, 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

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

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

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.

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

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

This is difficult for two reasons. First, the domain of the supreme, $\prod_j \Delta([\underline{p}_j, \bar{p}_j])$, is a large space. For instance, if $J = 10$, the supreme is over 20 functions. Second, the selection function f is very general as we have not impose much structure. In the proof of this theorem in Appendix B.1, we employ a technique called a change of measure, also known as the tilted measure, and combine it with insights from transportation problem.⁶

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 $\bar{p}_j - \underline{p}_j$, and the maximum semi-elasticity difference.⁷ Our condition requires this product to be small. We emphasize that this is a sufficient condition; the mapping may remain a contraction even when the product is above 1.

⁶In Appendix A, we connect our contraction result with quantal response equilibria (McKelvey and Palfrey, 1995).

⁷Note that by definition ρ is unitless. Changing the unit of price does not affect ρ .

To build intuition for the role of each component in the bound on the modulus, we now examine these factors in turn. First, if we expand the support $[\underline{p}_j, \bar{p}_j]$ to $[\underline{p}'_j, \bar{p}'_j]$ where

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

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 $\prod_{k \neq j} \Delta([\underline{p}_k, \bar{p}_k]) \times \Delta([\underline{p}'_j, \bar{p}'_j])$ 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.

Second, the maximum semi-elasticity difference M_j is small when the own-price responsiveness of product j stays stable even as competitors' prices vary. This arises, for example, when product j is highly differentiated and exhibits weak substitutability with other products. In the extreme case where the demand for product j is completely independent of competitors' prices, we have $M_j = 0$. In this situation, the selection probability for product j does not depend on the distribution of competing prices \mathbf{p}_{-j} , so the integral over \mathbf{p}_{-j} in Equation (1) reduces to a constant. Consequently, the offered price distribution for product j can be obtained directly from its observed selected price distribution given the selection function f .

Extending this intuition to more general settings, when M_j is small, the influence of competing offer distributions on the demand for product j is limited. As a result, if we plug a conjecture Ψ into the selection probability, even if this conjecture is not perfectly accurate, the resulting selection probability will remain close to the truth. Using this approximate selection probability to recover the offered price distribution therefore yields an estimate that is also close to the true distribution. Thus, smaller values of M_j make it easier for the operator T to be a contraction.

Finally, the effect of J on ρ is more subtle because J not only directly affects the dimensionality of the unknown offered price distributions but also influences the maximum semi-elasticity difference M_j . For example, consider the multinomial logit model, arguably the most popular model for discrete choices due to its analytical form and ease of estimation:

$$f_j(p_1, \dots, p_J) = \frac{\exp(\gamma p_j + \xi_j)}{\sum_{k=1}^J \exp(\gamma p_k + \xi_k)}, \quad (10)$$

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

the logit model,

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

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. Hence, the modulus can remain small even when the number of alternatives J is large.

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 $\mathbf{p}_{-j}, \mathbf{p}'_{-j}$ 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 $p_j \in [\underline{p}_j, \bar{p}_j]$, $\frac{\partial \ln f_j(p_j, \mathbf{p}_{-j})}{\partial p_j}$ is weakly increasing in each p_k with $k \neq j$.

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

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

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

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

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.⁸ 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.

For illustration, we compute ρ^* for the simple multinomial logit model in Equation (10) with two alternatives. We fix $\gamma = 1$, normalize $\xi_1 = 0$, and set the lower bounds of both price distributions to zero. In Figure 2, the x- and y-axes represent the upper bounds of the price distributions for alternatives 1 and 2, respectively. For various values of ξ_2 , we plot the contour of the region where $\rho^* > 1$, and we highlight this region with shaded areas. Figure 2 shows patterns that are consistent with our theoretical results. As the price range widens, the bound on the modulus is more likely to exceed 1, so the shaded areas are concentrated in the upper-right corner of the figure. We again emphasize that $\rho^* < 1$ is only 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 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. With this powerful identification result, which exhausts all the information contained in the selected outcome distributions, estimation of the selection model essentially reduces to recovering the selection function from observed choice patterns. We discuss the estimation strategy in the next section.

⁸To see this, we compute the log derivative for the binary probit model:

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

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.

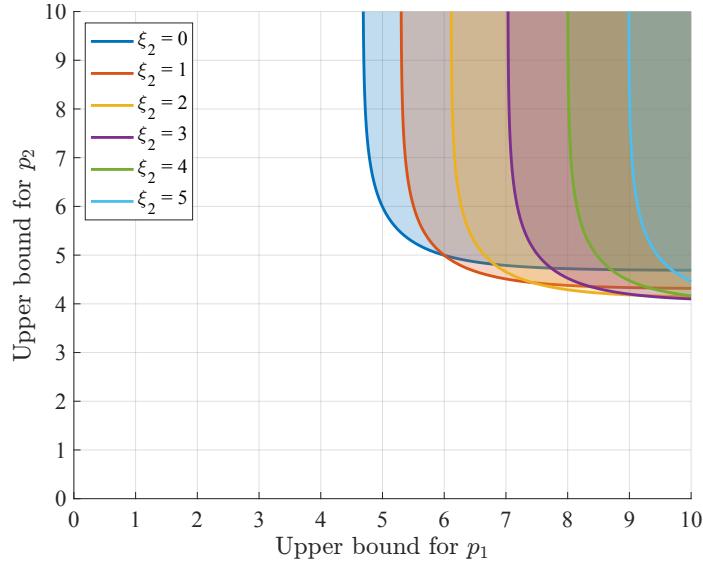


Figure 2: Region of price ranges where the modulus exceeds 1. The x- and y-axes show the upper bounds of the price distributions for alternatives 1 and 2. The shaded areas indicate the regions where $\rho^* > 1$ for various values of ξ_2 , based on the multinomial logit model in Equation (10) with two alternatives. We fix $\gamma = 1$, normalize $\xi_1 = 0$, and set the lower bounds of both price distributions to zero.

4 Estimation

We now turn to the estimation of the model’s primitives, which include (1) the unobserved offered price distributions G and (2) the parameters in the selection function f . We propose a three-step estimation procedure. In the first step, we estimate the selected outcome distribution conditional on both observable and unobservable covariates using instruments. Once the selected outcome distribution has been recovered, for any given selection function f , the potential outcome distribution can be recovered iteratively using the contraction mapping results in Section 3. The second step nests this fixed-point problem within an estimation routine that recovers the parameters of the selection function f from agents’ observed choice patterns. In the third step, we re-run the fixed-point algorithm using the estimated parameters of the selection function in order to recover the offered price distribution.

In the data, for each individual i , we observe their choice $y_i \in \mathcal{J}$ and the price of the selected product p_i . Let x_{ij} denote a vector of observable characteristics, and define $x_i = (x'_{i1}, \dots, x'_{iJ})' \in X$. We let $x_i^* \in X^*$ denote an unobservable characteristic which may affect both the selection decision and the distribution of potential

outcomes.

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

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

where v_j is a known function indexed by a finite-dimensional parameter vector θ . Here, p_{ij} is the offered price of alternative j for individual i , and the vector of unobserved shocks $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{iJ})$ follows a known joint distribution, such as Type 1 extreme value. Each individual chooses the alternative that maximizes utility, and the selection function f is captured by the parameter θ .

A widely used specification takes the following form:

$$u_{ij} = \gamma p_{ij} + x'_{ij}\beta + \xi_j + x_i^*\kappa_j + \varepsilon_{ij}, \quad j = 1, 2, \dots, J, \quad (11)$$

where ξ_j represents a scalar-valued unobserved characteristic of alternative j , such as product quality or brand loyalty. The term $x_i^*\kappa_j$ allows preferences for product j to vary with the unobservable characteristic x_i^* . In this example, $\theta = (\gamma, \beta, \boldsymbol{\xi}, \boldsymbol{\kappa})$, where $\boldsymbol{\xi} = (\xi_1, \dots, \xi_J)$ and $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_J)$. Throughout the paper, we use θ_0 to denote the true parameter.

4.1 Three-Step Estimation Strategy

Step 1: Estimating Selected Outcome Distribution The key inputs for our contraction-mapping results are the selected outcome distributions \tilde{G} conditional on (x, x^*) . When all relevant covariates are observed, so that no unobserved component x^* is present, \tilde{G} conditional on x can be easily estimated nonparametrically from the data, for example using kernel methods. We therefore do not elaborate on this case. The more challenging setting arises when an unobserved covariate x^* is present. In this case, \tilde{G} conditional on (x, x^*) cannot be directly estimated from the observed data, and additional information about the unobserved covariates is required in order to recover this distribution.

In this paper, we follow the instrumental variable approach of [Hu \(2008\)](#) to estimate the selected outcome distribution conditional on the unobservable x^* in the first step. We assume that the variables $\omega_i = \{x_i, y_i, p_i, z_{1i}, z_{2i}\}$ are observed in an

i.i.d. sample and take values in a finite support.⁹ The variables z_1 and z_2 serve as instrumental variables and are required to satisfy the following condition:

$$\begin{aligned} & h_{p,z_1|z_2,x,y}(p, z_1 | z_2, x, y) \\ &= \sum_{x^*} h_{p|x^*,x,y}(p | x^*, x, y) h_{z_1|x^*,x,y}(z_1 | x^*, x, y) h_{x^*|z_2,x,y}(x^* | z_2, x, y), \end{aligned} \quad (12)$$

where $h(\cdot)$ represents probability mass functions. Equation (12) shows that the joint distribution of (p, z_1) conditional on (z_2, x, y) can be expressed as a mixture over the latent variable x^* . This condition requires, first, that the two instrumental variables are informative about the latent variable x^* , and second, that once we condition on x^* , the price and the instruments are independent. In practice, finding such instruments is often feasible. In insurance pricing, for example, the latent variable x^* may represent a consumer's unobserved risk type, which influences premiums. Realized claims can serve as proxy variables for this latent type. In labor applications, the latent variable might correspond to a worker's unobserved productivity, which affects wages. Measures such as work-performance evaluations or test scores can provide useful proxies in these settings.

Theorem 1 in [Hu \(2008\)](#) shows that, under additional rank and ordering assumptions, the unknown probability mass functions on the right-hand side of Equation (12), $h = (h_{p|x^*,x,y}, h_{z_1|x^*,x,y}, h_{x^*|z_2,x,y}) \in H$, are nonparametrically identified. We do not restate these additional assumptions here and instead refer readers to [Hu \(2008\)](#) for the technical details.

Given Equation (12), a maximum-likelihood estimator of h can be obtained in a straightforward manner. We denote this estimator by $\hat{h} = (\hat{h}_{p|x^*,x,y}, \hat{h}_{z_1|x^*,x,y}, \hat{h}_{x^*|z_2,x,y})$. The term $\hat{h}_{p|x^*,x,y}$ represents the estimate of the selected price distribution conditional on (x, x^*) , which corresponds to $\tilde{G}(x, x^*)$. The only distinction is that $h_{p|x^*,x,y}$ is a probability mass function, whereas $\tilde{G}(x, x^*)$ is its associated cumulative mass function. We do not distinguish between these two objects in what follows. Finally, by taking the expectation of $\hat{h}_{x^*|z_2,x,y}$ with respect to the distribution of z_2 , we obtain

⁹The finite support assumption is not essential for identifying and estimating the selected outcome distributions in the first step. [Hu and Schennach \(2008\)](#) extends the results in [Hu \(2008\)](#) to settings with continuously distributed variables, so similar identification argument and estimation procedure remain valid without discreteness. This assumption is adopted here primarily to simplify the asymptotic normality results, which we discuss further in Section 4.2.

an estimate of the distribution of the latent variable x^* conditional on (x, y) , which we denote by $\hat{h}_{x^*|x,y}$. From this object, we can easily derive the choice probability for each alternative $j \in \mathcal{J}$ conditional on (x, x^*) , which are essential for identifying parameters of the selection function, as discussed in Section 4.2.

Step 2: Estimating Parameters in the Selection Function Given the first-step estimates $\hat{h}_{p|x^*,x,y}$ and $\hat{h}_{x^*|x,y}$, we propose a semiparametric maximum likelihood estimator for parameter θ in the selection function:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \hat{Q}_n(\theta), \quad (13)$$

where

$$\hat{Q}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \sum_{x^*} \hat{h}_{x^*|x,y}(x^*|x_i, y_i) \ln Prob_{y_i}(x_i, x^*, \theta, \hat{h}_{p|x^*,x,y}), \quad (14)$$

$$Prob_j(x, x^*; \theta, \tilde{G}) = \int_{\mathbf{p}} f_j(\mathbf{p}; x, x^*, \theta) d(F^{-1}(\tilde{G}(x, x^*); \theta, x, x^*))(\mathbf{p}). \quad (15)$$

Equation (15) derives the probability that alternative j is chosen conditional on (x, x^*) for any utility parameters θ and any selected outcome distributions \tilde{G} . This probability is obtained by integrating the selection function $f_j(\mathbf{p}; x, x^*, \theta)$ with respect to the distribution of offered prices for all competing alternatives. To recover the offered price distribution, we rely on the contraction-mapping results in Theorem 1. Recall that F denotes the mapping from the offered price distribution G to the selected outcome distribution \tilde{G} as defined in Equations (1) and (2). The inverse mapping F^{-1} in Equation (15) maps \tilde{G} back to G . Theorem 1 guarantees that we can replicate F^{-1} by iterating the operator T .

Our second-step estimation follows a nested fixed-point algorithm. In the inner loop, for any candidate value of the parameter θ in the selection function, we solve the fixed point of the operator T to obtain the offered price distribution given θ and the first-step estimate of the selected outcome distribution \tilde{G} , namely $\hat{h}_{p|x^*,x,y}$. With the offered price distribution in hand, we compute agents' choice probabilities using Equation (15) and then construct the sample analogue of the likelihood function in Equation (14). In the outer loop, we then search over θ to maximize this likelihood.

Step 3: Recovering the Offered Price Distribution Once $\hat{\theta}$ is obtained, a plug-in estimator of the offered price distribution G can be constructed by iterating the operator T until convergence. The operator T depends on two objects estimated in the previous steps: (1) the parameters of the selection function, $\hat{\theta}$, and (2) the selected outcome distributions, $\hat{h}_{p|x^*,x,y}$. We use \hat{T} to denote the operator formed using these estimates. The plug-in estimator of the offered price distribution is

$$\hat{G} = \hat{T}^\infty \Psi,$$

where $\hat{T}^\infty \Psi$ denotes the limit of the iterates of \hat{T} starting from an initial distribution Ψ . This step essentially repeats the inner-loop procedure, except that we replace θ with its estimate $\hat{\theta}$. In practice, the algorithm used to solve the fixed point is terminated after a finite number of iterations. We show that the resulting approximation error is asymptotically negligible, provided that the number of iterations grows fast enough compared to the logarithm of the sample size. Further details are provided at the end of Section 4.2.

4.2 Consistency and Asymptotic Normality

We now discuss the asymptotic properties of our proposed estimators $\hat{\theta}$ and \hat{G} . When constructing the model-implied choice probabilities in Equation (15), the inverse mapping F^{-1} appears, which maps the selected price distribution \tilde{G} back to the offered price distribution G . We therefore begin by analyzing the properties of this inverse mapping F^{-1} .

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. □

Proposition 1 has three important implications. First, because F is a homeomorphism, its inverse F^{-1} is well-defined, and we have $G = F^{-1}(\tilde{G})$. Second, the continuity of F^{-1} implies that if a consistent estimator \tilde{G}_n of the selected outcome distribution is used in place of \tilde{G} , then

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

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

We now turn to the consistency and asymptotic normality of our estimators. To establish consistency, we rely on the fundamental consistency theorem for extremum estimators (Theorem 2.1 in [Newey and McFadden \(1994\)](#)). We construct the true population objective function as follows:

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

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

We maintain the previous assumptions on the selection function, namely that $f_j \in \mathcal{C}^1: \prod_j [\underline{p}_j, \bar{p}_j] \rightarrow (0, 1]$. The following additional technical conditions are required to establish the consistency of $\hat{\theta}$.

Assumption 2. (i) The space Θ of parameter θ is compact; (ii) for each x, x^* , the selection function $f(\mathbf{p}; x, x^*, \theta)$ is jointly continuous in θ and \mathbf{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 $\bar{\rho}$.

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

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

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

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 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.

The identification condition merits additional discussion. The unknown objects in our model are the parameter vector θ in the selection function f and the offered price distribution G . A key insight from our contraction mapping result (Theorem 1) is that, for any given selection function f , the operator T admits a *unique* fixed point, and this fixed point corresponds to the offered price distribution associated with f .

In other words, given f , the offered price distribution G is fully nonparametrically identified from the accepted price distribution. This is *not* an assumption; it is an implication of the model's structure. As a result, identification of the full model reduces to identification of the parameters θ in the selection function f .

Assumption 3 essentially requires that variation in the observed choice probabilities conditional on (x, x^*) —which we have already identified in the first-step estimation—is sufficient to uniquely pin down the selection-function parameters. For example, under the commonly used demand specification in Equation (11), the unknown parameters include the price sensitivity parameter γ , the coefficients (β, κ_j) on the covariates (x, x^*) , and the unobserved product characteristics ξ_j (with one of them normalized to zero without loss). The number of unknowns is $\dim(x_{ij}) + 2J$, whereas the number of moments (i.e., conditional choice probabilities) available for identification is $|X||X^*|(J - 1)$. As the dimensions of x and x^* increase, the variation in choice probabilities expands, generating an overidentified system for the utility parameters.

Moreover, if additional instrumental variables are available, such as exogenous cost shocks that shift the offered price distribution, these provide extra moment conditions for identifying the price sensitivity parameter as in the classical demand estimation literature. In the paper, we provide a high-level version of the identification condition for simplicity, but our framework can readily incorporate any additional instrumental variables when available. These extra moments can be included in the outer loop of the Step 2 estimation procedure described in Section 4.1. We summarize the consistency result in the following theorem.

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

Proof. See Appendix B.3. □

Next, we show that the estimator defined in Equation (13) is asymptotically normal. Let

$$\mathbf{g}(\omega; \theta, h) = \nabla_\theta \left(\sum_{x^*} h_{x^*|x,y}(x^*|x, y) \ln Prob_y(x, x^*, \theta, h_{p|x^*,x,y}) \right),$$

where ∇_θ denote the gradient operator with respect to θ . The estimator $\hat{\theta}$ solves the

first-order condition

$$\frac{1}{n} \sum_{i=1}^n \mathbf{g}(\omega_i; \theta, \hat{h}) = 0.$$

Moreover, we define

$$\mathbf{m}(\omega_i, h) = \nabla_h \ln \left(\sum_{x^*} h_{p|x^*, x, y}(p_i|x^*, x_i, y_i) h_{z_1|x^*, x, y}(z_{1i}|x^*, x_i, y_i) h_{x^*|z_2, x, y}(x^*|z_{2i}, x_i, y_i) \right).$$

We stack \mathbf{g} and \mathbf{m} to form

$$\tilde{\mathbf{g}}(\omega, \theta, h) = [\mathbf{g}(\omega, \theta, h)', \mathbf{m}(\omega, h)']',$$

then the estimators in the first two steps can be viewed as a GMM estimator. We impose the following standard regularity conditions.

Assumption 4. (i) (θ_0, h_0) is in the interior of $\Theta \times H$. (ii) f is twice continuously differentiable in θ . (iii) $\mathbb{E}\nabla_{\theta, h}\tilde{\mathbf{g}}(\omega; \theta_0, h_0)$ is nonsingular.

Theorem 4 (Asymptotic Normality). *Suppose that Assumption 2, 3, and 4 hold. Then $\hat{\theta}$ is asymptotically normal and $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V)$.¹⁰ $\hat{T}^\infty \Psi$ converges to G in probability at rate \sqrt{n} .*

Proof. See Appendix B.4. □

So far, the asymptotic results have been stated under the assumption that the operator is iterated infinitely many times. In practice, however, the iteration used to obtain the offered price distribution is stopped after a finite number of steps. The resulting approximation error is asymptotically negligible as long as the number of iterations grows fast enough relative to the logarithm of the sample size. Formally, let $m(n)$ denote the number of iterations given the sample size n . Consistency of our estimator (Theorem 3) can be achieved as long as $\lim_{n \rightarrow +\infty} m(n) \rightarrow \infty$. Asymptotic normality (Theorem 4) continues to hold if in addition, $\liminf_{n \rightarrow +\infty} \frac{m(n)}{\ln n} > \frac{1}{2}(\ln(1/\bar{\rho}))^{-1}$.

Finally, we discuss the finite support assumption imposed on the outcome p_i . This assumption is not essential for the consistency result in Theorem 3. As long as the estimator of the selected price distribution is consistent, our proposed estimator remains consistent even when p_i is continuous. Assuming that p_i has finite support

¹⁰See the analytical form of V in the proof of Theorem 4.

mainly keeps the proof of asymptotic normality in Theorem 4 tractable. If p_i is instead continuous, establishing asymptotic normality for 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 setting, this would require expanding the function \mathbf{g} around $\hat{h}_{p|x^*,x,y}$. A standard argument would apply if $\hat{h}_{p|x^*,x,y}$ entered Equation (15) directly. However, in our case it enters only through F^{-1} , for which no analytic expression is available. As a result, working with the infinite-dimensional distribution \tilde{G} is extremely challenging.

In practice, when the selected outcome distribution is estimated nonparametrically, even if p_i is conceptually continuous, the estimator necessarily evaluates its CDF on a finite grid of points. For this reason, this assumption does not impose a substantive restriction and aligns with how nonparametric estimators are implemented in applied work.

5 Monte Carlo Simulations

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

$$\begin{aligned} u_{i1} &= -\gamma \log(p_{i1}) + \xi_1 + \beta x_{i1} + \kappa x_i^* + \varepsilon_i, \\ u_{i2} &= -\gamma \log(p_{i2}) + \xi_2, \end{aligned}$$

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; $x_i^* \in \{-1, 1\}$ is a binary unobservable with $Pr(x_i^* = 1) = 0.5$; and $\varepsilon_i \sim N(0, 1)$ is the error term. Throughout the simulation exercises, we set the utility parameters as follows: $\gamma = 1$, $\xi_1 = 0$, $\xi_2 = 0.5$, $\beta = 0.5$ and $\kappa = 0.1$. Let $y_i \in \{1, 2\}$ denote the choice of individual i .

We consider four data generating processes for the offered prices. Let x_{i2} denote the observable characteristic of individual i that enters the pricing equation. We assume that x_{i2} takes values in $\{0, 0.25, 0.5, 0.75, 1\}$ with equal probability.

DGP 1: $\log(p_{ij}) = \delta_{0j} + \delta_{1j}x_{i2} + \delta_{2j}x_i^* + \eta_{ij}$, where $\eta_{ij} \sim N(0, \sigma_j^2)$. For alternative 1, we set $\delta_{01} = 0.2$, $\delta_{11} = 0.5$, $\delta_{21} = 0.1$, $\sigma_1 = 0.1$. For alternative 2, we set

$$\delta_{02} = 0.1, \delta_{12} = 1, \delta_{22} = 0.1, \sigma_2 = 0.2.$$

DGP 2: $\log(p_{ij}) = \delta_{0j} + \delta_{1j}x_{i2}^2 + \delta_{2j}x_i^* + \eta_{ij}$, where $\eta_{ij} \sim N(0, \sigma_j^2)$. For alternative 1, we set $\delta_{01} = 0.2, \delta_{11} = 0.5, \delta_{21} = 0.1, \sigma_1 = 0.1$. For alternative 2, we set $\delta_{02} = 0.1, \delta_{12} = 1, \delta_{22} = 0.1, \sigma_2 = 0.2$.

DGP 3: $\log(p_{ij}) = \exp((\delta_{0j} + \delta_{1j}x_{i2} + \delta_{2j}x_i^*)(1 + \eta_{ij}))$, where $\eta_{ij} \sim N(0, \sigma_j^2)$. For alternative 1, we set $\delta_{01} = 0.2, \delta_{11} = 0.5, \delta_{21} = 0.1, \sigma_1 = 0.1$. For alternative 2, we set $\delta_{02} = 0.1, \delta_{12} = 1, \delta_{22} = 0.1, \sigma_2 = 0.2$.

DGP 4: $\log(p_{ij}) = (\delta_{0j} + \delta_{1j}x_{i2}^2)(\delta_{2j}x_i^* + \eta_{ij})^{-1}$, where $\eta_{ij} \sim N(-2, \sigma_j^2)$. For alternative 1, we set $\delta_{01} = 0.2, \delta_{11} = 0.1, \delta_{21} = 0.1, \sigma_1 = 0.1$. For alternative 2, we set $\delta_{02} = 0.1, \delta_{12} = 0.3, \delta_{22} = 0.1, \sigma_2 = 0.2$.

Across all data generating processes, the unobserved characteristic x_i^* enters the pricing equations for both alternatives, which induces correlation in prices conditional on observables. In addition, x_i^* also enters the utility specification, allowing the unobserved type to jointly affect the prices individuals face and their preferences over alternatives. DGP 1 specifies an additively separable linear pricing equation, which is commonly assumed in empirical applications. DGP 2 introduces a nonlinear term. DGPs 3 and 4 consider scenarios where the pricing function takes a nonseparable form.¹¹

For each DGP, we simulate offered prices and individual choices. To implement our estimator, we require an instrument z_i to recover the selected price distribution conditional on (x_{i1}, x_{i2}, x_i^*) in the first step, since x_i^* is unobserved. We construct such an instrument by assuming $z_i \sim Poisson(x_i^*)$ when $x_i^* = 1$, and $z_i = 0$ otherwise. This choice is motivated by settings where x_i^* can be interpreted as an individual's unobserved risk type, and such risk types may be reflected in the ex post realization of accidents, which are often modeled using a Poisson distribution. Because we impose a parametric relationship between the instrument and the unobservable, only one instrument is needed.

We assume that the econometrician observes $(y_i, x_{i1}, x_{i2}, p_i, z_i)$, where p_i denotes the price of the chosen alternative. Using these data, we apply the procedure described

¹¹Although 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.

in Section 4 to estimate the parameters of the selection function, $\theta = (\gamma, \xi_2, \beta, \kappa)$ with ξ_1 normalized to 0, along with the offered price distribution for each alternative.¹² For comparison, we first implement the classic Heckman two-step method, assuming that the pricing equations are linearly separable and that the error terms in the selection and pricing equations follow a bivariate normal distribution. We also compare our estimator with the quantile selection model of [Arellano and Bonhomme \(2017\)](#). To implement their approach, we follow the standard practice of assuming that the quantile functions are linear in x_{i2} and that the dependence structure is governed by a Gaussian copula. Although [Arellano and Bonhomme \(2017\)](#) discuss identification under more general settings, their empirical implementation focuses on cases in which the copula depends on a low-dimensional vector of parameters, which is the specification we adopt here. For each design, we run 500 simulations with sample sizes of 2,000 and 5,000 observations.

Table 1 reports the Monte Carlo biases, standard deviations, and root mean squared errors for the estimates of θ obtained using our method. Overall, the estimator performs well in finite samples across all DGPs, including those with nonseparable pricing equations. The biases are small, and the root mean squared errors remain modest for all parameters in the selection function. The standard deviation decreases as the sample size increases in all simulation designs.

For the cumulative distribution functions of $\log(price)$, Tables 2 and 3 report the integrated squared biases and integrated mean squared errors for our proposed estimator, the Heckman two-step estimator, and the copula-based sample-selection correction estimator for quantile regression, separately for the two alternatives. Each row of the tables corresponds to the price distribution conditional on a specific value of x_{i2} . We also plot the true CDFs for alternatives 1 and 2 alongside the estimates produced by these models conditional on $x_{i2} = 0.25$ and $x_{i2} = 0.75$ in Figures 3 and 4, respectively. To save space, we report the CDF results only for the sample size of 2,000 observations, and we omit the figures for other values of the observable covariates.

Our method allows for nonparametric estimation of the offered price distributions, whereas the alternative approaches impose parametric restrictions on either the conditional mean or the conditional quantiles of the pricing distributions, or on the dependence structure through the copula. Tables 2 and 3 show that our estimator

¹²We estimate the cumulative distribution function of prices at 300 grid points.

Table 1: Simulation Results for Utility Parameters

DGP 1						
	$N = 2000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	-0.1017	0.1762	0.2033	-0.0697	0.1087	0.1290
β	-0.0058	0.0619	0.0621	-0.0013	0.0372	0.0371
κ	-0.0224	0.0491	0.0540	-0.0095	0.0350	0.0362
ξ_2	-0.0267	0.0570	0.0629	-0.0156	0.0349	0.0382
DGP 2						
	$N = 2000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	-0.1003	0.1672	0.1949	-0.0705	0.1067	0.1278
β	-0.0049	0.0626	0.0627	-0.0013	0.0379	0.0379
κ	-0.0185	0.0487	0.0521	-0.0072	0.0344	0.0351
ξ_2	-0.0197	0.0508	0.0544	-0.0114	0.0319	0.0338
DGP 3						
	$N = 2000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	-0.0866	0.0883	0.1236	-0.0665	0.0611	0.0903
β	-0.0041	0.0724	0.0724	-0.0015	0.0470	0.0470
κ	0.0199	0.0584	0.0616	0.0324	0.0370	0.0491
ξ_2	-0.0400	0.0615	0.0733	-0.0243	0.0384	0.0454
DGP 4						
	$N = 2000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	0.0934	0.8041	0.8087	0.0666	0.4806	0.4847
β	0.0005	0.0613	0.0613	0.0038	0.0361	0.0363
κ	-0.0194	0.0474	0.0512	-0.0103	0.0335	0.0350
ξ_2	-0.0003	0.0451	0.0451	0.0016	0.0280	0.0280

achieves very low integrated squared bias and integrated mean squared error for the CDFs of $\log(price)$ across all simulation designs and for all values of x_{i2} . In contrast, while the classic Heckman two-step method and the quantile selection model perform well in DGP 1, their biases and mean squared errors increase substantially as the pricing equation becomes more complex in DGPs 2–4. These results are expected, since the parametric assumptions underlying these methods, such as linear conditional mean or quantile functions and a Gaussian copula, are severely violated in these designs.

Figures 3 and 4 provide a visual illustration of these results. We can see that across all simulation designs, the estimated CDFs of $\log(price)$ for both alternatives produced by our functional contraction approach closely track the true CDFs, as indicated by the black curves with “+” markers and the red solid curve in Figures 3 and 4. By comparison, the biases of the Heckman two-step method (blue dashed curves) and the quantile selection model (purple dash-dotted curves) can be substantial, particularly in DGPs 3 and 4. The direction and magnitude of these biases also vary with the values of the observable covariates.

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 practice, it is also difficult to find variables that affect selection but can be excluded from the outcome equation.

In contrast, our approach does not require such an excluded variable. 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 four DGPs for $\log(price)$. The results for this specification are reported in Tables 4–5 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 the econometrician to correctly specify the functional form of the selection function. To evaluate how the estimator performs under misspecification, we conduct a series of Monte Carlo simulations in which the econometrician assumes

Table 2: Simulation Results for CDF of $\log(p_1)$

DGP 1						
	Functional Contraction		Heckman Two-Step		Quantile Selection	
	IBias ²	IMSE	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0005	0.0017	0.0001	0.0039	0.0001	0.0046
$x_{i2} = 0.25$	0.0004	0.0015	0.0001	0.0033	0.0001	0.0038
$x_{i2} = 0.5$	0.0002	0.0012	0.0001	0.0027	0.0001	0.0031
$x_{i2} = 0.75$	0.0002	0.0010	0.0001	0.0023	0.0001	0.0026
$x_{i2} = 1$	0.0001	0.0010	0.0001	0.0021	0.0001	0.0023
DGP 2						
	Functional Contraction		Heckman Two-Step		Quantile Selection	
	IBias ²	IMSE	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0005	0.0017	0.0235	0.0275	0.0185	0.0231
$x_{i2} = 0.25$	0.0005	0.0017	0.0017	0.0053	0.0031	0.0072
$x_{i2} = 0.5$	0.0003	0.0014	0.0124	0.0153	0.0145	0.0177
$x_{i2} = 0.75$	0.0002	0.0011	0.0037	0.0062	0.0052	0.0080
$x_{i2} = 1$	0.0001	0.0010	0.0133	0.0154	0.0106	0.0129
DGP 3						
	Functional Contraction		Heckman Two-Step		Quantile Selection	
	IBias ²	IMSE	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0095	0.0106	0.0598	0.0648	0.0218	0.0298
$x_{i2} = 0.25$	0.0026	0.0036	0.0163	0.0201	0.0035	0.0070
$x_{i2} = 0.5$	0.0008	0.0016	0.0040	0.0063	0.0040	0.0062
$x_{i2} = 0.75$	0.0003	0.0010	0.0011	0.0023	0.0004	0.0020
$x_{i2} = 1$	0.0002	0.0009	0.0061	0.0068	0.0028	0.0042
DGP 4						
	Functional Contraction		Heckman Two-Step		Quantile Selection	
	IBias ²	IMSE	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0014	0.0023	0.0432	0.0463	0.0390	0.0425
$x_{i2} = 0.25$	0.0014	0.0024	0.0147	0.0182	0.0153	0.0190
$x_{i2} = 0.5$	0.0011	0.0021	0.0412	0.0443	0.0358	0.0392
$x_{i2} = 0.75$	0.0012	0.0021	0.0106	0.0141	0.0059	0.0100
$x_{i2} = 1$	0.0005	0.0018	0.0270	0.0304	0.0346	0.0384

Note: The IBias² of a function h is calculated as follows. Let \hat{h}_r be the estimate of h from the r -th simulated dataset, and $\bar{h}(p) = \frac{1}{R} \sum_{r=1}^R \hat{h}_r(p)$ be the point-wise average over R simulations. The integrated squared bias is calculated by numerically integrating the point-wise squared bias $(\bar{h}(p) - h(p))^2$ over the distribution of p . The integrated MSE is computed in a similar way. The values reported in each row correspond to the price distributions conditional on a given value of x_{i2} . The results shown in this table are based on a 500 Monte Carlo replications with a sample size of 2,000.

Table 3: Simulation Results for CDF of $\log(p_2)$

DGP 1						
	Functional Contraction		Heckman Two-Step		Quantile Selection	
	IBias ²	IMSE	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0002	0.0008	0.0000	0.0016	0.0001	0.0018
$x_{i2} = 0.25$	0.0002	0.0009	0.0000	0.0019	0.0001	0.0020
$x_{i2} = 0.5$	0.0002	0.0010	0.0000	0.0023	0.0001	0.0024
$x_{i2} = 0.75$	0.0002	0.0011	0.0000	0.0028	0.0001	0.0030
$x_{i2} = 1$	0.0002	0.0012	0.0000	0.0034	0.0001	0.0038
DGP 2						
	Functional Contraction		Heckman Two-Step		Quantile Selection	
	IBias ²	IMSE	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0002	0.0008	0.0231	0.0246	0.0211	0.0228
$x_{i2} = 0.25$	0.0002	0.0008	0.0058	0.0075	0.0067	0.0086
$x_{i2} = 0.5$	0.0002	0.0009	0.0200	0.0220	0.0213	0.0234
$x_{i2} = 0.75$	0.0002	0.0010	0.0030	0.0058	0.0040	0.0069
$x_{i2} = 1$	0.0003	0.0011	0.0368	0.0401	0.0329	0.0365
DGP 3						
	Functional Contraction		Heckman Two-Step		Quantile Selection	
	IBias ²	IMSE	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0320	0.0326	0.0646	0.0670	0.0035	0.0050
$x_{i2} = 0.25$	0.0015	0.0022	0.0308	0.0339	0.0162	0.0204
$x_{i2} = 0.5$	0.0004	0.0013	0.0040	0.0078	0.0009	0.0058
$x_{i2} = 0.75$	0.0002	0.0020	0.0099	0.0139	0.0082	0.0129
$x_{i2} = 1$	0.0004	0.0037	0.0563	0.0598	0.0388	0.0428
DGP 4						
	Functional Contraction		Heckman Two-Step		Quantile Selection	
	IBias ²	IMSE	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0011	0.0016	0.1443	0.1459	0.1123	0.1141
$x_{i2} = 0.25$	0.0009	0.0015	0.0568	0.0592	0.0455	0.0483
$x_{i2} = 0.5$	0.0005	0.0011	0.1095	0.1109	0.0952	0.0966
$x_{i2} = 0.75$	0.0003	0.0009	0.0290	0.0308	0.0133	0.0155
$x_{i2} = 1$	0.0002	0.0007	0.0530	0.0546	0.0722	0.0741

Note: The IBias² of a function h is calculated as follows. Let \hat{h}_r be the estimate of h from the r -th simulated dataset, and $\bar{h}(p) = \frac{1}{R} \sum_{r=1}^R \hat{h}_r(p)$ be the point-wise average over R simulations. The integrated squared bias is calculated by numerically integrating the point-wise squared bias $(\bar{h}(p) - h(p))^2$ over the distribution of p . The integrated MSE is computed in a similar way. The values reported in each row correspond to the price distributions conditional on a given value of x_{i2} . The results shown in this table are based on a 500 Monte Carlo replications with a sample size of 2,000.

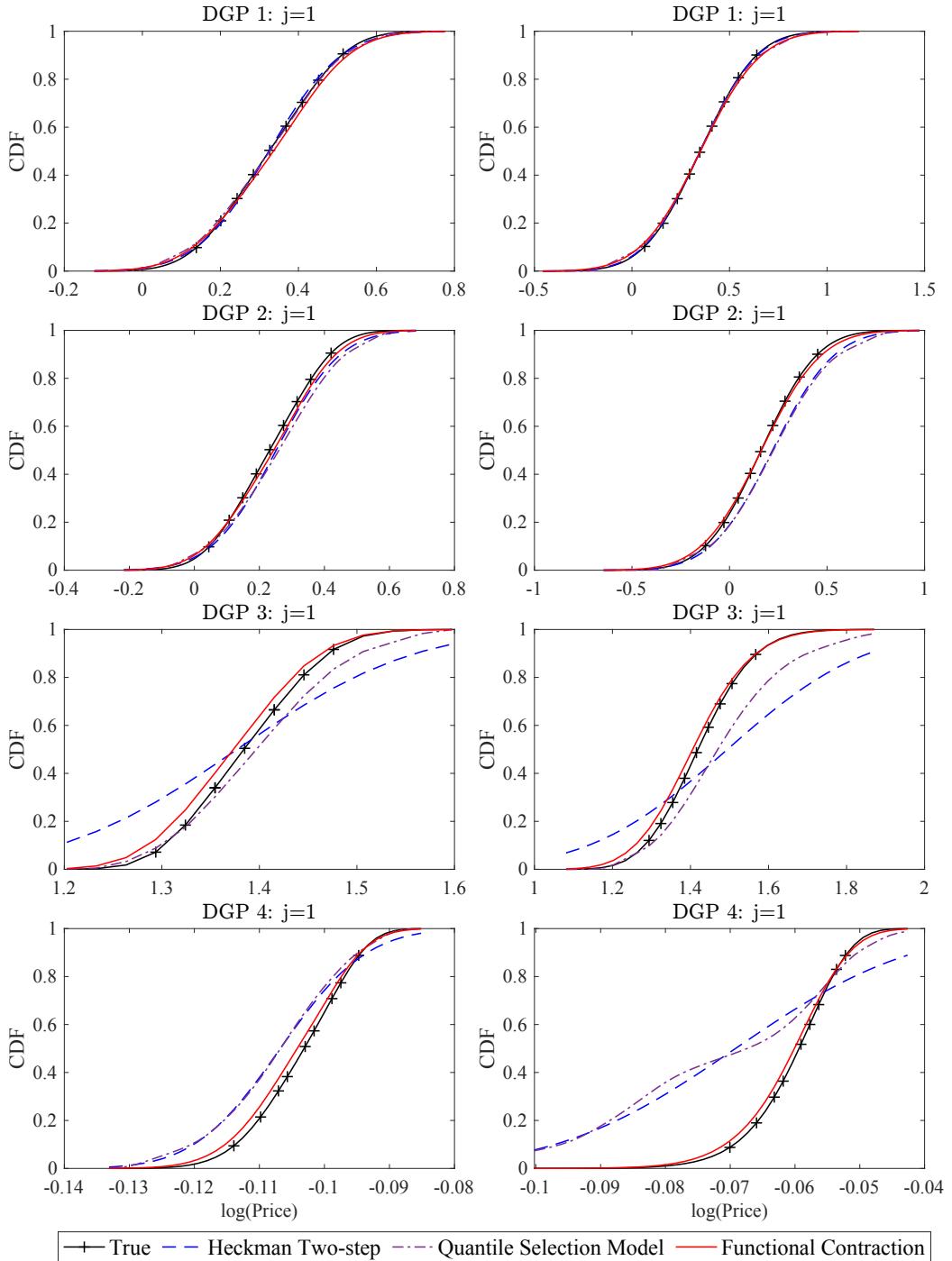


Figure 3: CDFs of $\log(\text{price})$ for alternatives 1 and 2, conditional on $x_{i2} = 0.25$. The black curve with “+” markers, the blue dashed curve, the purple dash-dotted curve, and the red solid curve correspond to the true CDF, the Heckman two-step estimate, the quantile selection estimate, and the functional contraction estimate, respectively, based on 500 Monte Carlo replications with a sample size of 2,000.

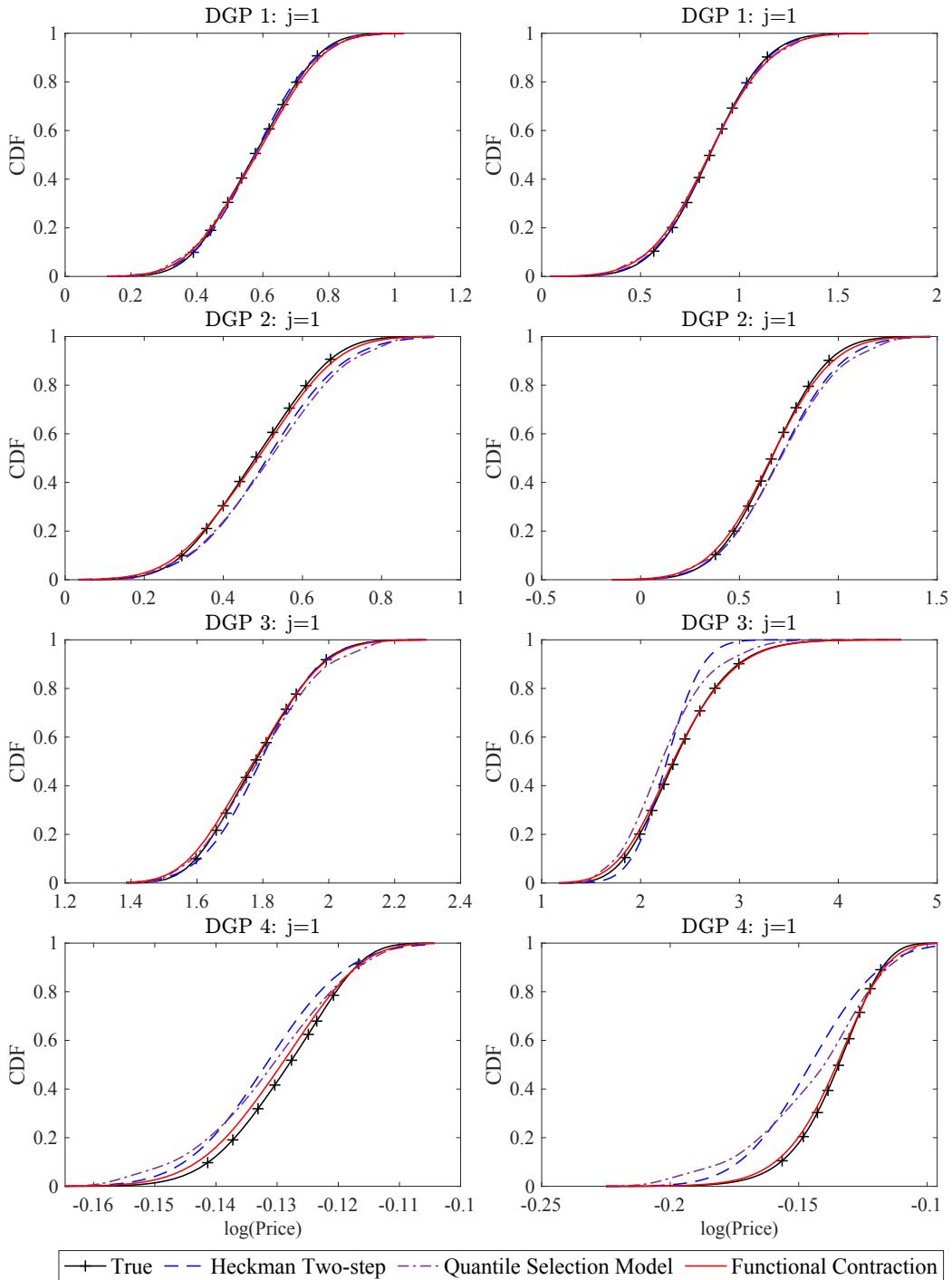


Figure 4: CDFs of $\log(\text{price})$ for alternatives 1 and 2, conditional on $x_{i2} = 0.75$. The black curve with “+” markers, the blue dashed curve, the purple dash-dotted curve, and the red solid curve correspond to the true CDF, the Heckman two-step estimate, the quantile selection estimate, and the functional contraction estimate, respectively, based on 500 Monte Carlo replications with a sample size of 2,000.

that ε follows a logistic distribution, while in truth it is generated from a normal distribution. In Tables 6–7 in Appendix C, we report the estimation results for the utility parameters and CDFs of $\log(\text{price})$ under this misspecification. For the utility parameters, we rescale the estimates by the scale parameter of the logit model to make them comparable to those in the original probit specification. After this adjustment, the biases are small. The estimator for the offered price distributions also performs well: the integrated squared biases and mean squared errors of the CDFs remain close to those in Tables 2 and 3. These results suggest that the estimator of the offered price distributions is robust to misspecification of the selection function, an appealing feature in practice, particularly when the econometrician has limited prior information about the correct functional form.

Finally, we briefly discuss how the functional contraction performs computationally in practice. We compute ρ^* for all four simulation designs and find that it is below 1 in every case except DGP 3. This result is unsurprising: as illustrated in Figure 4, the price range in DGP 3 is substantially wider than in the other designs, making a violation of the sufficient condition $\rho^* < 1$ more likely. Importantly, even in DGP 3 where the bound on the modulus exceeds 1, the fixed-point iteration still converges quickly. The average numbers of iterations needed to reach convergence (with a tolerance of 10^{-5}) are 3.8, 3.8, 5.1, and 2.1 for DGPs 1 through 4, respectively (averaged over 500 replications). These findings demonstrate that our estimator remains computationally efficient and stable, even in settings where the sufficient contraction condition does not strictly hold.

6 Discussion of Empirical Applications

Our estimator introduced in Section 4 is broadly applicable to a wide range of empirical settings. It effectively addresses the challenge of selection bias that arises when only the outcomes of chosen alternatives are observed. The method has three features that are particularly important for empirical applications. First, it imposes no parametric restrictions on the potential outcome distributions and allows them to vary flexibly across alternatives. Second, the framework accommodates unobservable characteristics in both the outcome distributions and the selection model. Third, the selection function can incorporate alternative-specific unobserved heterogeneity and does not require an excluded variable, which is desirable in many empirical settings.

An important empirical application that illustrates these advantages is consumer demand estimation in markets where only transaction prices are observed. In classic differentiated product demand estimation pioneered by [Berry \(1994\)](#) and [Berry et al. \(1995\)](#), 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 ([D'Haultfœuille et al., 2019](#); [Sagl, 2023](#); [Buchholz et al., 2020](#); [Dubé and Misra, 2023](#)), discount negotiation ([Goldberg, 1996](#); [Allen et al., 2014](#)), or risk-based pricing ([Crawford et al., 2018](#); [Cosconati et al., 2025](#)). 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., 2025](#)), 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.

In this application, we assume that the offered prices across different firms are independent conditional on observable characteristics and the consumer's true unobserved risk type. At the same time, the consumer's risk type may also influence their preferences over insurance products. For example, higher-risk consumers may prefer insurers with higher service quality. We therefore allow the true risk type to enter both the pricing distributions and the utility parameters. Our data include realized claim records for each consumer over multiple years, and we use these records as instruments for the latent risk type in the first-step estimation.

We nonparametrically estimate each insurance company's offered price distribution using our functional contraction approach. In Figure 5, we plot the CDFs of offered price for several firms based on estimates in [Cosconati et al. \(2025\)](#). The distributions differ substantially across firms, indicating significant heterogeneity in their pricing strategies. Building on this result, we estimate each firm's information precision parameter using supply-side model restrictions. These estimates provide

important insights for analyzing competition under heterogeneous information structures in this market.

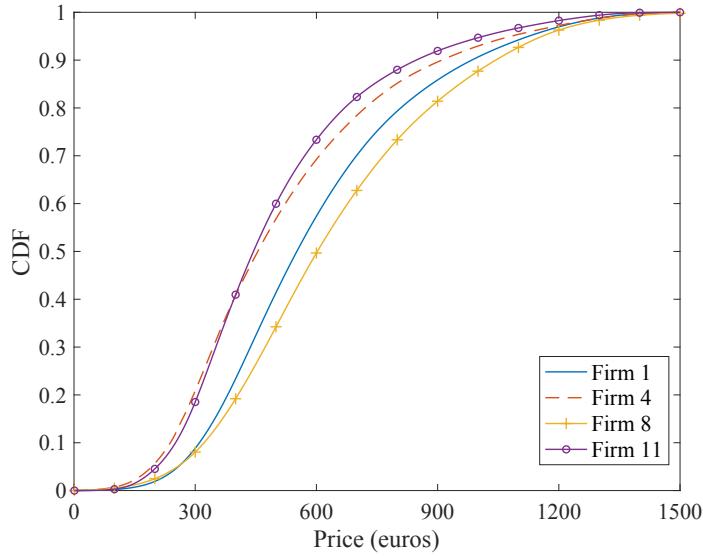


Figure 5: CDFs of the offered price distributions for firms 1, 4, 8, and 11 based on estimates from [Cosconati et al. \(2025\)](#). The CDFs are averaged across different characteristic groups.

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.

The usefulness of our method is not limited to consumer demand. It can also be applied to auction models and Roy models, where similar selection issues arise. For example, in multi-attribute auctions, our approach can be used to nonparametrically recover the full bid distribution and the auctioneer’s scoring weights when only the winning bids and the winner’s identity are observed, even in the presence of bidder asymmetry.¹³ Auctions in many settings have used the scoring rule that departs from the lowest bid criterion by accounting for quality differences,¹⁴ and our framework can

¹³Flexibly 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 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](#)).

¹⁴See for examples [Asker and Cantillon \(2008\); Lewis and Bajari \(2011\); Nakabayashi \(2013\); Yoganarasimhan \(2016\); Takahashi \(2018\); Krasnokutskaya et al. \(2020\); Allen et al. \(2024\)](#).

flexibly accommodate these multi-attribute scoring mechanisms with both observable and unobservable components. A similar application arises in Roy models, where our method can recover the distribution of potential wages when only realized wages in the chosen sector are observed. Our framework enables researchers to recover these distributions flexibly and without relying on excluded variables in the selection equation, which are often difficult to find in empirical settings. This capability provides a valuable tool for studying key questions in labor economics, such as occupational choice and wage inequality.

7 Conclusion

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

Our method has several important features. First, we allow the outcome equation to be fully nonparametric and nonseparable in error terms, and we recover the entire distribution of potential outcomes rather than focusing on specific moments or quantiles. In essence, we correct for sample selection bias by examining how the bias is *systematically* generated by the selection model. Second, our 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. Finally, 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. \(2025\)](#)). 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 J. L. POWELL (1993): “Semiparametric estimation of censored selection models with a nonparametric selection mechanism,” *Journal of Econometrics*, 58, 3–29.
- ALLEN, J., R. CLARK, B. HICKMAN, AND E. RICHERT (2024): “Resolving failed banks: Uncertainty, multiple bidding and auction design,” *Review of Economic Studies*, 91, 1201–1242.
- ALLEN, J., R. CLARK, AND J.-F. HOUDE (2014): “Price dispersion in mortgage markets,” *The Journal of Industrial Economics*, 62, 377–416.
- (2019): “Search frictions and market power in negotiated-price markets,” *Journal of Political Economy*, 127, 1550–1598.
- ANDREWS, D. W. AND M. M. SCHAFGANS (1998): “Semiparametric estimation of the intercept of a sample selection model,” *The Review of Economic Studies*, 65, 497–517.
- ARELLANO, M. AND S. BONHOMME (2017): “Quantile selection models with an application to understanding changes in wage inequality,” *Econometrica*, 85, 1–28.
- ASKER, J. AND E. CANTILLON (2008): “Properties of scoring auctions,” *The RAND Journal of Economics*, 39, 69–85.
- ATHEY, S. AND P. A. HAILE (2002): “Identification of standard auction models,” *Econometrica*, 70, 2107–2140.
- (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, 1362–1370.
- BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): “Automobile prices in market equilibrium,” *Econometrica*, 63, 841–890.
- BERRY, S. T. (1994): “Estimating discrete-choice models of product differentiation,” *The RAND Journal of Economics*, 242–262.
- BUCHHOLZ, N., L. DOVAL, J. KASTL, F. MATĚJKA, AND T. SALZ (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, 96–107.
- CHEN, S. AND S. KHAN (2003): “Semiparametric estimation of a heteroskedastic sample selection model,” *Econometric Theory*, 19, 1040–1064.
- CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND S. LUO (2023): “Distribution regression with sample selection and UK wage decomposition,” Tech. rep., cemmap working paper.
- CICALA, S. (2015): “When does regulation distort costs? lessons from fuel procurement in us electricity generation,” *American Economic Review*, 105, 411–444.
- COSCONATI, M., Y. XIN, F. WU, AND Y. JIN (2025): “Competing under Information Heterogeneity: Evidence from auto insurance,” *The Review of Economic Studies*, forthcoming.
- CRAWFORD, G. S., N. PAVANINI, AND F. SCHIVARDI (2018): “Asymmetric information and imperfect competition in lending markets,” *American Economic Review*, 108, 1659–1701.
- DAS, M., W. K. NEWEY, AND F. VELLA (2003): “Nonparametric estimation of sample selection models,” *The Review of Economic Studies*, 70, 33–58.
- DE SILVA, D. G., G. KOSMOPOULOU, AND C. LAMARCHE (2009): “The effect of information on the bidding and survival of entrants in procurement auctions,” *Journal of Public Economics*, 93, 56–72.
- DUBÉ, J.-P. AND S. MISRA (2023): “Personalized pricing and consumer welfare,” *Journal of Political Economy*, 131, 131–189.
- D’HAULTFOUILLE, X., I. DURRMEYER, AND P. FÉVRIER (2019): “Automobile prices in market equilibrium with unobserved price discrimination,” *The Review of Economic Studies*, 86, 1973–1998.
- D’HAULTFOUILLE, X. AND A. MAUREL (2013): “Another look at the identification at infinity of sample selection models,” *Econometric Theory*, 29, 213–224.
- D’HAULTFOUILLE, X., A. MAUREL, AND Y. ZHANG (2018): “Extremal quantile regressions for selection models and the black–white wage gap,” *Journal of Econometrics*, 203, 129–142.
- FERNÁNDEZ-VAL, I., A. VAN VUUREN, AND F. VELLA (2024): “Nonseparable sample selection models with censored selection rules,” *Journal of Econometrics*, 240, 105088.

- FLAMBARD, V. AND I. PERRIGNE (2006): “Asymmetry in procurement auctions: Evidence from snow removal contracts,” *The Economic Journal*, 116, 1014–1036.
- GOLDBERG, P. K. (1996): “Dealer price discrimination in new car purchases: Evidence from the consumer expenditure survey,” *Journal of Political Economy*, 104, 622–654.
- GRONAU, R. (1974): “Wage comparisons—A selectivity bias,” *Journal of political Economy*, 82, 1119–1143.
- GUERRE, E. AND Y. LUO (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*, 679–694.
- (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*, NBER, 475–492.
- (1979): “Sample selection bias as a specification error,” *Econometrica*, 47, 153–161.
- HENDRICKS, K. AND R. H. PORTER (1988): “An empirical study of an auction with asymmetric information,” *The American Economic Review*, 865–883.
- HORTAÇSU, A., F. LUCO, S. L. PULLER, AND D. ZHU (2019): “Does strategic ability affect efficiency? Evidence from electricity markets,” *American Economic Review*, 109, 4302–4342.
- HU, Y. (2008): “Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution,” *Journal of Econometrics*, 144, 27–61.
- HU, Y. AND S. M. SCHENNACH (2008): “Instrumental variable treatment of non-classical measurement error models,” *Econometrica*, 76, 195–216.
- KOMAROVA, T. (2013): “A new approach to identifying generalized competing risks models with application to second-price auctions,” *Quantitative Economics*, 4, 269–328.
- KRASNOKUTSKAYA, E., K. SONG, AND X. TANG (2020): “The role of quality in internet service markets,” *Journal of Political Economy*, 128, 75–117.

- LEE, L.-F. (1982): “Some approaches to the correction of selectivity bias,” *The Review of Economic Studies*, 49, 355–372.
- (1983): “Generalized econometric models with selectivity,” *Econometrica: Journal of the Econometric Society*, 507–512.
- LEWIS, G. AND P. BAJARI (2011): “Procurement contracting with time incentives: Theory and evidence,” *The Quarterly Journal of Economics*, 126, 1173–1211.
- MCKELVEY, R. D. AND T. R. PALFREY (1995): “Quantal response equilibria for normal form games,” *Games and Economic Behavior*, 10, 6–38.
- MEILIJSSEN, I. (1981): “Estimation of the lifetime distribution of the parts from the autopsy statistics of the machine,” *Journal of Applied Probability*, 18, 829–838.
- NAKABAYASHI, J. (2013): “Small business set-asides in procurement auctions: An empirical analysis,” *Journal of Public Economics*, 100, 28–44.
- NEWHEY, W. K. (2007): “Nonparametric continuous/discrete choice models,” *International Economic Review*, 48, 1429–1439.
- (2009): “Two-step series estimation of sample selection models,” *The Econometrics Journal*, 12, S217–S229.
- NEWHEY, W. K. AND D. MCFADDEN (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, 135–146.
- SAGL, S. (2023): “Dispersion, discrimination, and the price of your pickup,” *Working paper*.
- TAKAHASHI, H. (2018): “Strategic design under uncertain evaluations: Structural analysis of design-build auctions,” *The RAND Journal of Economics*, 49, 594–618.
- THOMPSON, A. C. (1963): “On certain contraction mappings in a partially ordered vector space.” *Proceedings of the American Mathematical Society*, 14, 438–443.
- VELLA, F. (1998): “Estimating models with sample selection bias: a survey,” *Journal of Human Resources*, 127–169.
- YOGANARASIMHAN, H. (2016): “Estimation of beauty contest auctions,” *Marketing Science*, 35, 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 $\mathcal{J} = \{1, 2, \dots, J\}$ of players. For each player $j \in \mathcal{J}$, there is a finite set $P_j = \{p_{j1}, p_{j2}, \dots, p_{jn_j}\} \subset [\underline{p}_j, \bar{p}_j]$ consisting of n_j pure strategies. A payoff function $f: \prod_{j \in \mathcal{J}} P_j \rightarrow \Delta(\mathcal{J})$ assigns payoff f_j to player j . Let $g_j \in \Delta P_j$ denote player j 's 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

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

We define the quantal response operator $\mathbb{T}: \prod_j \Delta(P_j) \rightarrow \prod_j \Delta(P_j)$ 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,

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

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 probability 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 $d(S, Q) = \delta$,

$$\ln \frac{p}{p - \epsilon} + \ln \frac{1 - p + \epsilon}{1 - p} = \delta. \quad (16)$$

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

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

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 $\delta \rightarrow 0$, $p = \frac{1}{2}$, $\epsilon = \frac{\delta}{4}$ 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

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

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

which implies

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

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 (16). Thus, the total variation norm

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

□

¹⁵To 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

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

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 $\Psi, \Phi \in \prod_j \Delta([\underline{p}_j, \bar{p}_j])$,

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

Given Equation (8) 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 $[\underline{p}_j, \bar{p}_j]$. Thus, it suffices to show that for all $j \in \mathcal{J}$

$$\sup_{p_j} \ln R_j(p_j; \Psi, \Phi) - \inf_{p_j} \ln R_j(p_j; \Psi, \Phi) \leq \rho D(\Psi, \Phi) \quad (17)$$

We evaluate how the log ratio changes with p_j ,

$$\begin{aligned} \frac{d \ln R_j(p_j; \Psi, \Phi)}{dp_j} &= \frac{\int_{\mathbf{p}_{-j}} \frac{\partial f_j(p_j, \mathbf{p}_{-j})}{\partial p_j} \prod_{k, k \neq j} d\Psi_k(p_k)}{\int_{\mathbf{p}_{-j}} f_j(p_j, \mathbf{p}_{-j}) \prod_{k, k \neq j} d\Psi_k(p_k)} - \frac{\int_{\mathbf{p}_{-j}} \frac{\partial f_j(p_j, \mathbf{p}_{-j})}{\partial p_j} \prod_{k, k \neq j} d\Phi_k(p_k)}{\int_{\mathbf{p}_{-j}} f_j(p_j, \mathbf{p}_{-j}) \prod_{k, k \neq j} d\Phi_k(p_k)} \\ &= \frac{\int_{\mathbf{p}_{-j}} \frac{\partial \ln f_j(p_j, \mathbf{p}_{-j})}{\partial p_j} f_j \prod_{k, k \neq j} d\Psi_k(p_k)}{\int_{\mathbf{p}_{-j}} f_j(p_j, \mathbf{p}_{-j}) \prod_{k, k \neq j} d\Psi_k(p_k)} - \frac{\int_{\mathbf{p}_{-j}} \frac{\partial \ln f_j(p_j, \mathbf{p}_{-j})}{\partial p_j} f_j \prod_{k, k \neq j} d\Phi_k(p_k)}{\int_{\mathbf{p}_{-j}} f_j(p_j, \mathbf{p}_{-j}) \prod_{k, k \neq j} d\Phi_k(p_k)} \end{aligned} \quad (18)$$

(19)

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

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

Similarly, we define measure $f_j \Phi_{-j} \in \Delta(\prod_{k \neq j} [\underline{p}_k, \bar{p}_k])$. (Both measures depend on p_j .) Given these measures, we can rewrite Equation (19)

$$\frac{d \ln R_j(p_j; \Psi, \Phi)}{dp_j} = \mathbb{E}_{\mathbf{p}_{-j} \sim f_j \Psi_{-j}} \frac{\partial \ln f_j(p_j, \mathbf{p}_{-j})}{\partial p_j} - \mathbb{E}_{\mathbf{p}_{-j} \sim f_j \Phi_{-j}} \frac{\partial \ln f_j(p_j, \mathbf{p}_{-j})}{\partial p_j} \quad (20)$$

$$= \int_{\mathbf{p}_{-j}} \frac{\partial \ln f_j(p_j, \mathbf{p}_{-j})}{\partial p_j} [df_j \Psi_{-j}(\mathbf{p}_{-j}) - df_j \Phi_{-j}(\mathbf{p}_{-j})]. \quad (21)$$

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

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

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 $\frac{\partial \ln f_j(p_j, \mathbf{p}_{-j})}{\partial p_j}$ 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 $\delta > 0$,

$$\begin{aligned} \sup_{D(\Psi, \Phi) \leq \delta} \left| \frac{d \ln R_j(p_j; \Psi, \Phi)}{dp_j} \right| &\leq M_j \sup_{D(\Psi, \Phi) \leq \delta} \frac{1}{2} \|f_j \Psi_{-j} - f_j \Phi_{-j}\|_{TV} \\ &\leq M_j \sup_{d(f_j \Psi_{-j}, f_j \Phi_{-j}) \leq (J-1)\delta} \frac{1}{2} \|f_j \Psi_{-j} - f_j \Phi_{-j}\|_{TV} \\ &\leq M_j \frac{1}{4} (J-1)\delta \end{aligned} \quad (22)$$

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 (22) which holds for all $\delta > 0$.

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

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

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 $\delta > 0$,*

$$\sup_{\Psi, \Phi} \left| \frac{d \ln R_j(p_j; \Psi, \Phi)}{dp_j} \frac{1}{D(\Psi, \Phi)} \right| = \sup_{\Psi, \Phi, D(\Psi, \Phi) \leq \delta} \left| \frac{d \ln R_j(p_j; \Psi, \Phi)}{dp_j} \frac{1}{D(\Psi, \Phi)} \right|. \quad (23)$$

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

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

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

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 (20).

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

Notice that when $\lambda = 1$, this reduces to Equation (20). As k is continuously differentiable, there exists $0 \leq \underline{\lambda} < \lambda + d\lambda \leq 1$ and $d\lambda \leq \frac{\delta}{D(\Psi, \Phi)}$ such that

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

¹⁶Note 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 (23). \square

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 (21).

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

By Lemma 2,

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

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

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

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 $\Psi \in \prod_j \Delta([\underline{p}_j, \bar{p}_j])$,

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

Moreover, F is injective. Towards a contradiction, suppose F maps both $G_1 \neq G_2 \in \prod_j \Delta([\underline{p}_j, \bar{p}_j])$ 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,

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

where the last inequality is by Equation (17). 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 $\tilde{G} \neq \tilde{G}' \in \prod_j \Delta([\underline{p}_j, \bar{p}_j])$ 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

$$\begin{aligned} D(T_{\tilde{G}'}^k G, T_{\tilde{G}'}^{k+1} G) &\leq \rho^k D(G, T_{\tilde{G}'} G) = \rho^k D(\tilde{G}, \tilde{G}') \\ D(F^{-1}(\tilde{G}), F^{-1}(\tilde{G}')) &= D(G, F^{-1}(\tilde{G}')) = D(G, T_{\tilde{G}'}^\infty G) \\ &\leq \sum_{k=0}^{\infty} D(T_{\tilde{G}'}^k G, T_{\tilde{G}'}^{k+1} G) \\ &\leq \sum_{k=0}^{\infty} \rho^k D(\tilde{G}, \tilde{G}') \\ &= \frac{1}{1-\rho} D(\tilde{G}, \tilde{G}') \end{aligned}$$

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

We next prove the consistency result (Theorem 3). For proofs below, we shall suppress the dependence on variable x and x^* . The proof requires a combination of Lemma 3-5 below.

Lemma 3. $F^{-1}(\tilde{G}, \theta)$ 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 $\theta' \rightarrow \theta$, by $F(G'; \theta)$ being continuous in θ , $\tilde{G}' \rightarrow \tilde{G}$. By $F^{-1}(\tilde{G}; \theta)$ being continuous in \tilde{G} (Proposition 1), $F^{-1}(\tilde{G}; \theta) \rightarrow F^{-1}(\tilde{G}'; \theta)$. This is equivalent to $G' \rightarrow G$, which is $F^{-1}(\tilde{G}; \theta') \rightarrow F^{-1}(\tilde{G}; \theta)$. This implies that F^{-1} is continuous in θ . \square

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

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

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

Proof of Lemma 4. Since the function f is continuous on a compact set $\prod_j [\underline{p}_j, \bar{p}_j] \times \Theta$ and $f_j > 0$, there exists $\underline{f} > 0$ such that for all $j \in \mathcal{J}$, $\theta \in \Theta$, $\mathbf{p} \in \prod_j [\underline{p}_j, \bar{p}_j]$,

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

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

$$\underline{f} < Pr_j(p_j; G, \theta). \quad (24)$$

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

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

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

$$\begin{aligned} & |Pr_j(p_j; G, \theta) - Pr_j(p_j; G, \theta')| \\ &= \left| \int_{\mathbf{p}_{-j}} [f_j(p_j, \mathbf{p}_{-j}; \theta) - f_j(p_j, \mathbf{p}_{-j}; \theta')] \prod_{k, k \neq j} dG_k(p_k) \right| < \epsilon'. \end{aligned} \quad (25)$$

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

$$(T_\theta \Psi)_j(p_j) = \frac{\int_{\underline{p}_j}^{p_j} d\tilde{G}_j(p) / Pr_j(p; \Psi, \theta)}{\int_{\underline{p}_j}^{\bar{p}_j} d\tilde{G}_j(p) / Pr_j(p; \Psi, \theta)}.$$

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 (24) and (25), for all $\tilde{G} \in \prod_j \Delta([\underline{p}_j, \bar{p}_j])$, $\theta, \theta' \in \Theta$ with $|\theta - \theta'| < \delta'$,

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

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

Finally, for any $\epsilon > 0$, let ϵ' be such that $\frac{2}{1 - \bar{\rho}} \ln \frac{\underline{f} + \epsilon'}{\underline{f}} = \epsilon$. 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 , $Prob_j(\theta; \hat{h}_{p|y})$ is equicontinuous in θ , parametrized by $\hat{h}_{p|y}$. That is, for all $\theta \in \Theta$, $\epsilon > 0$, there exists a $\delta > 0$ such that for all $|\theta' - \theta| < \delta$, $\hat{h}_{p|y} \in \prod_j \Delta([\underline{p}_j, \bar{p}_j])$,

$$|Prob_j(\theta; \hat{h}_{p|y}) - Prob_j(\theta'; \hat{h}_{p|y})| \leq \epsilon.$$

Consequently, by Equation (24), for all $|\theta' - \theta| < \delta$, $\{\omega_i\}_{i=1}^n$,

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

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

By Hu (2008)'s identification result, $\hat{h} \xrightarrow{p} h_0$. Then for all $\theta \in \Theta$, $\hat{Q}_n(\theta)$ pointwise converges in probability to $Q_0(\theta)$, by the weakly law of large numbers, $\hat{h}_{p|y} \xrightarrow{p} \tilde{G}$, 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)). \square

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 $\underline{f} > 0$ and continuous in θ by Lemma 3, $Q_0(\theta)$ is continuous. (4). $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$, by Lemma 5. Thus, $\hat{\theta}$ is consistent.

To see $\hat{T}^\infty \Psi \xrightarrow{p} G$, note that $\hat{T}^\infty \Psi = F^{-1}(\hat{h}_{p|y}, \hat{\theta})$,

$$D(\hat{T}^\infty \Psi, G) \leq D(F^{-1}(\hat{h}_{p|y}, \hat{\theta}), F^{-1}(\hat{h}_{p|y}, \theta_0)) + D(F^{-1}(\hat{h}_{p|y}, \theta_0), G).$$

The first term

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

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

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

since F is a homeomorphism by Proposition 1. \square

B.4 Proof of Theorem 4

Proof of Theorem 4. Our GMM estimator is

$$\frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{g}}(\omega_i, \hat{\theta}, \hat{h}) = 0.$$

For this GMM estimator, we can directly invoke Theorem 6.1 in Newey and McFadden (1994). Note that our \mathbf{g} is their g and our h is their γ in Newey and McFadden (1994).

By the proof of Theorem 3, $\hat{\theta} \xrightarrow{p} \theta_0$. By standard argument of MLE and identification result in Hu (2008), $\hat{h} \xrightarrow{p} h_0$. By Assumption 4, (θ_0, h_0) is in the interior of $\Theta \times H$. Next, we verify that $\tilde{\mathbf{g}}(\omega, \theta, h)$ is continuously differentiable in a neighborhood \mathcal{N} of (θ_0, h_0) .

First, we verify that $\mathbf{g}(\omega, \theta, h)$ is continuously differentiable in θ . It suffices to show that $Prob(\theta, h_{p|y})$ is twice continuously differentiable in θ . As f is twice continuously

differentiable in θ , we only need to show that $F^{-1}(\tilde{G}, \theta)$ is twice continuously differentiable in θ . By Equation (1), (2), $F(G, \theta)$ is infinitely continuously differentiable in G . By f being twice continuously differentiable in θ , $F(G, \theta)$ is twice continuously differentiable in θ . Moreover, matrix $\nabla_G F(G, \theta)$ is non-singular by F^{-1} being Lipschitz continuous. Thus, by the implicit function theorem,

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

and F^{-1} is twice continuously differentiable in θ .

Next, we verify that $\mathbf{g}(\omega, \theta, h)$ is continuously differentiable in h . It suffices to show that $Prob(\theta, h_{p|y})$ is continuously differentiable in $h_{p|y}$, which is equivalent to show that $F^{-1}(\tilde{G}, \theta)$ is continuously differentiable in \tilde{G} . As $F(G, \theta)$ is infinitely continuously differentiable in G and $F^{-1}(\tilde{G}, \theta)$ is Lipschitz continuous in \tilde{G} , we have

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

which is continuous in \tilde{G} . Additionally, \mathbf{m} is infinitely continuously differentiable in all parameters θ, h . Consequently, we have show that $\tilde{\mathbf{g}}(\omega, \theta, h)$ is continuously differentiable in θ, h .

In addition,

$$\mathbb{E}[\tilde{\mathbf{g}}(\omega, \theta_0, h_0)] = 0$$

by the first-order condition of MLE and Q_0 . Since for each observation ω , the value of Equation (12) is strictly positive, $\|\mathbf{m}(\omega, h_0)\|$ is finite. Since $f_j \geq \underline{f} > 0$ is bounded from 0 and $Prob(\theta, h_{p|y})$ is continuously differentiable in θ , $\|\mathbf{g}(\omega, \theta_0, h_0)\|$ is finite for each ω . Since there are only finite possible realizations of ω ,

$$\mathbb{E}[\|\tilde{\mathbf{g}}(\omega, \theta_0, h_0)\|^2] < \infty$$

By $\tilde{\mathbf{g}}(\omega, \theta, h)$ being continuously differentiable in (θ, h) and a finite possible values of ω ,

$$\mathbb{E}[\sup_{(\theta, h) \in \mathcal{N}} \|\nabla_{\theta, h} \tilde{\mathbf{g}}(\omega, \theta, h)\|] < \infty.$$

The last condition we need is that $\mathbb{E}\nabla_{\theta, h} \tilde{\mathbf{g}}(\omega; \theta_0, h_0)$ is nonsingular, which is in Assumption 4.

We can write down the variance matrix V by Theorem 6.1 in Newey and McFadden

(1994).

$$V = \left(\mathbb{E} \nabla_{\theta} \mathbf{g}(\omega, \theta_0, h_0) \right)^{-1} \times \mathbb{E} (\mathcal{A}(\omega) \mathcal{A}(\omega)') \times \left(\left(\mathbb{E} \nabla_{\theta} \mathbf{g}(\omega, \theta_0, h_0) \right)^{-1} \right)'$$

where

$$\mathcal{A}(\omega) = \mathbf{g}(\omega, \theta_0, h_0) - \mathbb{E}[\nabla_h \mathbf{g}(\omega, \theta_0, h_0)] \left(\mathbb{E}[\nabla_h \mathbf{m}(\omega, h_0)] \right)^{-1} m(\omega, h_0).$$

To see the convergence rate of $\hat{T}^\infty \Psi$, note that

$$D(\hat{T}^\infty \Psi, G) \leq D(F^{-1}(\hat{h}_{p|y}, \hat{\theta}), F^{-1}(\hat{h}_{p|y}, \theta_0)) + D(F^{-1}(\hat{h}_{p|y}, \theta_0), G).$$

By the proof above, F^{-1} is continuously differentiable in θ . Moreover, as Θ is compact, $F^{-1}(\tilde{G}, \theta)$ is Lipschitz continuous in θ . As $\hat{\theta} \xrightarrow{p} \theta_0$ at rate \sqrt{n} , the first term

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

By [Hu and Schennach \(2008\)](#), $\hat{h}_{p|y} \xrightarrow{p} \tilde{G}$ at rate \sqrt{n} . By Proposition 1, $F^{-1}(\tilde{G}, \theta)$ is Lipschitz continuous in \tilde{G} . Thus, the second term converges in probability to 0 at rate \sqrt{n} .

□

C Additional Tables

Table 4: Simulation Results for Utility Parameters: Removing the Excluded Variable

DGP 1						
	N = 2000			N = 5000		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	-0.0658	0.1743	0.1861	-0.0358	0.1159	0.1212
κ	-0.0223	0.0527	0.0572	-0.0081	0.0351	0.0360
ξ_2	-0.0162	0.0458	0.0485	-0.0090	0.0309	0.0322
DGP 2						
	N = 2000			N = 5000		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	-0.0708	0.1636	0.1781	-0.0452	0.1091	0.1180
κ	-0.0192	0.0546	0.0579	-0.0070	0.0342	0.0349
ξ_2	-0.0130	0.0381	0.0402	-0.0083	0.0252	0.0265
DGP 3						
	N = 2000			N = 5000		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	-0.0562	0.0778	0.0959	-0.0424	0.0554	0.0697
κ	0.0224	0.0582	0.0623	0.0305	0.0375	0.0484
ξ_2	-0.0201	0.0432	0.0477	-0.0129	0.0268	0.0297
DGP 4						
	N = 2000			N = 5000		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	0.0453	0.8115	0.8119	0.0062	0.5276	0.5271
κ	-0.0138	0.0501	0.0519	-0.0068	0.0323	0.0330
ξ_2	-0.0001	0.0308	0.0308	-0.0003	0.0189	0.0189

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

Table 5: Simulation Results for CDF of log(Price): Removing the Excluded Variable

DGP 1				
	$j = 1$		$j = 2$	
	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0004	0.0019	0.0001	0.0006
$x_{i2} = 0.25$	0.0002	0.0016	0.0001	0.0007
$x_{i2} = 0.5$	0.0001	0.0014	0.0001	0.0008
$x_{i2} = 0.75$	0.0001	0.0013	0.0002	0.0009
$x_{i2} = 1$	0.0001	0.0011	0.0001	0.0009
DGP 2				
	$j = 1$		$j = 2$	
	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0003	0.0019	0.0001	0.0006
$x_{i2} = 0.25$	0.0003	0.0018	0.0001	0.0007
$x_{i2} = 0.5$	0.0002	0.0016	0.0001	0.0007
$x_{i2} = 0.75$	0.0001	0.0015	0.0002	0.0008
$x_{i2} = 1$	0.0001	0.0011	0.0001	0.0008
DGP 3				
	$j = 1$		$j = 2$	
	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0096	0.0108	0.0315	0.0320
$x_{i2} = 0.25$	0.0029	0.0042	0.0016	0.0022
$x_{i2} = 0.5$	0.0007	0.0018	0.0003	0.0010
$x_{i2} = 0.75$	0.0003	0.0012	0.0001	0.0014
$x_{i2} = 1$	0.0002	0.0010	0.0000	0.0021
DGP 4				
	$j = 1$		$j = 2$	
	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0014	0.0025	0.0011	0.0016
$x_{i2} = 0.25$	0.0013	0.0023	0.0009	0.0014
$x_{i2} = 0.5$	0.0012	0.0022	0.0005	0.0010
$x_{i2} = 0.75$	0.0011	0.0023	0.0002	0.0008
$x_{i2} = 1$	0.0005	0.0019	0.0001	0.0006

Note: In these specifications, we remove the excluded variable from the selection function. The IBias² 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. The results shown in this table are based on a 500 Monte Carlo replications with a sample size of 2,000.

Table 6: Simulation Results for Utility Parameters: Misspecifying the Selection Function

DGP 1						
	$N = 2000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	-0.0851	0.1829	0.2016	-0.0540	0.1123	0.1245
β	0.0009	0.0634	0.0633	0.0052	0.0378	0.0381
κ	-0.0210	0.0499	0.0541	-0.0080	0.0356	0.0365
ξ_2	-0.0193	0.0596	0.0626	-0.0083	0.0364	0.0373
DGP 2						
	$N = 2000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	-0.0834	0.1739	0.1927	-0.0540	0.1103	0.1228
β	0.0028	0.0642	0.0642	0.0062	0.0387	0.0391
κ	-0.0167	0.0497	0.0523	-0.0054	0.0351	0.0355
ξ_2	-0.0116	0.0532	0.0544	-0.0034	0.0333	0.0334
DGP 3						
	$N = 2000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	-0.0202	0.0936	0.0957	0.0018	0.0603	0.0603
β	0.0148	0.0736	0.0750	0.0192	0.0455	0.0493
κ	0.0210	0.0603	0.0638	0.0330	0.0381	0.0504
ξ_2	-0.0182	0.0628	0.0653	-0.0020	0.0371	0.0371
DGP 4						
	$N = 2000$			$N = 5000$		
	Bias	Std. Dev.	RMSE	Bias	Std. Dev.	RMSE
γ	0.1083	0.8175	0.8239	0.0813	0.4889	0.4951
β	0.0053	0.0624	0.0626	0.0086	0.0367	0.0376
κ	-0.0183	0.0480	0.0514	-0.0091	0.0340	0.0351
ξ_2	0.0047	0.0466	0.0468	0.0065	0.0289	0.0296

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

Table 7: Simulation Results for CDF of log(Price): Misspecifying the Selection Function

DGP 1				
	$j = 1$		$j = 2$	
	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0004	0.0016	0.0002	0.0009
$x_{i2} = 0.25$	0.0004	0.0015	0.0002	0.0009
$x_{i2} = 0.5$	0.0002	0.0012	0.0002	0.0010
$x_{i2} = 0.75$	0.0002	0.0010	0.0002	0.0011
$x_{i2} = 1$	0.0001	0.0010	0.0002	0.0012
DGP 2				
	$j = 1$		$j = 2$	
	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0005	0.0016	0.0002	0.0008
$x_{i2} = 0.25$	0.0005	0.0017	0.0002	0.0008
$x_{i2} = 0.5$	0.0003	0.0013	0.0002	0.0009
$x_{i2} = 0.75$	0.0002	0.0011	0.0002	0.0010
$x_{i2} = 1$	0.0001	0.0010	0.0003	0.0011
DGP 3				
	$j = 1$		$j = 2$	
	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0094	0.0105	0.0320	0.0325
$x_{i2} = 0.25$	0.0025	0.0035	0.0014	0.0021
$x_{i2} = 0.5$	0.0008	0.0016	0.0003	0.0012
$x_{i2} = 0.75$	0.0003	0.0010	0.0001	0.0018
$x_{i2} = 1$	0.0002	0.0009	0.0008	0.0038
DGP 4				
	$j = 1$		$j = 2$	
	IBias ²	IMSE	IBias ²	IMSE
$x_{i2} = 0$	0.0014	0.0023	0.0011	0.0016
$x_{i2} = 0.25$	0.0014	0.0024	0.0009	0.0015
$x_{i2} = 0.5$	0.0011	0.0021	0.0005	0.0011
$x_{i2} = 0.75$	0.0012	0.0021	0.0003	0.0009
$x_{i2} = 1$	0.0005	0.0018	0.0002	0.0007

Note: In these specifications, we misspecify the selection model, assuming that the error term ε_i is drawn from $\text{Logistic}(0, 1)$. The IBias² 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. The results shown in this table are based on a 500 Monte Carlo replications with a sample size of 2,000.