Identification and Estimation of Demand Models with Endogenous Product Entry and Exit*

Victor Aguirregabiria †
University of Toronto
CEPR

Alessandro Iaria ‡
University of Bristol
CEPR

Senay Sokullu §
University of Bristol

August 28, 2023

Abstract

This paper deals with the endogeneity of firms' entry and exit decisions in demand estimation. Product entry decisions lack a single crossing property in terms of demand unobservables, which causes the inconsistency of conventional methods dealing with selection. We present a novel and straightforward two-step approach to estimate demand while addressing endogenous product entry. In the first step, our method estimates a finite mixture model of product entry accommodating latent market types. In the second step, it estimates demand controlling for the propensity scores of all latent market types. We apply this approach to data from the airline industry.

Keywords: Demand for differentiated product; Endogenous product availability; Selection bias; Market entry and exit; Multiple equilibria; Identification; Estimation; Demand for airlines.

JEL codes: C14, C34, C35, C57, D22, L13, L93.

*We are grateful for helpful comments from Andreea Enache, Mathieu Marcoux, David Pacini, Yuanyuan Wan, Ao Wang, and from seminar participants at the Universities of Bolzano, Glasgow, Mannheim, Penn State, Rochester, Sciences Po, and at the conference of the International Association of Applied Econometrics in Oslo.

†Department of Economics, University of Toronto. 150 St. George Street, Toronto, ON, M5S 3G7, Canada, victor.aguirregabiria@utoronto.ca.

‡Department of Economics, University of Bristol. The Priory Road Complex, Priory Road, BS8 1TU, Bristol, UK. alessandro.iaria@bristol.ac.uk.

§Department of Economics, University of Bristol. The Priory Road Complex, Priory Road, BS8 1TU, Bristol, UK. senay.sokullu@bristol.ac.uk.

1 Introduction

Since the influential work by Tobin (1958), Amemiya (1973), and Heckman (1976), addressing endogenous selection has been a fundamental topic in microeconometrics. The need to handle censored observations, specifically zeros, in consumer demand estimation has made this economic application a significant driver in the development of methods to account for sample selection. While much of the early literature focused on analyzing demand for a single product, there have also been early applications to demand systems (Amemiya, 1974, Yen, 2005, Yen and Lin, 2006).

More recently, the selection problem in estimating demand systems has garnered attention within the context of structural models of demand and supply for differentiated products. A key aspect distinguishing the recently proposed methods is the source of the zeros in market shares. The first group of studies considers that the choice probabilities in the model are strictly positive, but observed market shares may be zero because they are sample means based on a small number of consumers (Gandhi, Lu, and Shi, 2023). A second approach assumes that some market shares are zero because consumers in certain markets exclude these products from their choice set or *consideration set* (Dubé, Hortaçsu, and Joo, 2021).[1] Finally, a third group of studies examines sample selection bias in demand estimation when zeros arise from firms' market entry decisions (Conlon and Mortimer, 2013, Ciliberto, Murry, and Tamer, 2021; Li, Mazur, Park, Roberts, Sweeting, and Zhang, 2022).

This paper focuses on estimating demand for differentiated products using market-level data while accounting for censoring or selection resulting from firms not offering certain products in specific markets or periods. Demand estimation typically relies on data collected from multiple geographic markets and periods, where it is common for certain products to be unavailable in specific markets or periods. When firms make their market entry decisions, they possess information about the demand for their products, particularly regarding demand components that are not observable to the researcher. Firms are more inclined to enter markets with higher expected demand. Failure to consider this selection process can introduce significant biases in the estimation of demand parameters. This issue arises across various demand applications and industries, such as the demand for airlines (Berry, Carnall, and Spiller, 2006; Berry and Jia, 2010; Aguirregabiria and Ho, 2012), supermarket chains (Smith, 2004), radio stations (Sweeting, 2013),

[1] There is a growing empirical literature on consideration sets, where typically all products are assumed to be available in the market, but consumers consider subsets of these due to inattention or costly search. These heterogeneous consideration sets are usually unobserved by the researcher. The estimators proposed to overcome this complication do not consider the endogenous selection problem addressed in this paper (Goeree, 2008; Abaluck and Adams-Prassl, 2021; Barseghyan, Coughlin, Molinari, and Teitelbaum, 2021; Lu, 2022; and Moraga-González, Sándor, and Wildenbeest, 2023). For more details on this literature, see Crawford, Griffith, and Iaria (2021).

personal computers (Eizenberg, 2014), or ice cream (Draganska, Mazzeo, and Seim, 2009).[2]

The selection problem in this demand-entry model exhibits two key characteristics that distinguish it from more standard cases. Firstly, the non-additive nature of demand unobservables in firms' entry decisions makes the selection function dependent not only on the entry probabilities (i.e., propensity scores) but also on all the observable variables influencing market entry. Consequently, the conventional identification results and two-step estimation methods found in the literature are not applicable (Newey, Powell, and Walker, 1990; Ahn and Powell, 1993; Powell, 2001; Das, Newey, and Vella, 2003; Aradillas-Lopez, Honoré, and Powell, 2007; Newey, 2009). Secondly, the model exhibits multiple equilibria, both in the entry game and the pricing game. Even with identical observable characteristics, different equilibria can be selected across markets. This introduces an additional source of unobserved heterogeneity that may impact sample selection.

This paper examines the identification of demand parameters within a structural model encompassing demand, price competition, and market entry, where the distribution of demand unobservables is nonparametrically specified. As in other models, the characterization of the selection problem depends crucially on the specification of firms' information at the time of their product entry decisions. We assume that firms observe two signals on demand unobservables: a *market type signal*, which is common knowledge among all firms in the market and has finite support; and a *private information signal*, which is independent across firms. This structural framework retains the standard assumption of complete information in the Bertrand pricing game but incorporates firms' uncertainty and private information regarding future demand, such that the product entry game is an incomplete information game. The distribution of all the unobservables in the model is nonparametric.

The paper presents three main contributions. First, it establishes the sequential (two-step) identification of demand parameters in this model. We demonstrate that the selection term in the demand equation is a convolution of the probabilities of product entry for each discrete 'market type' and the densities associated with these market types. Leveraging results from

[2]Recent research on nonparametric identification of demand systems accounts for product entry and exit as a source of variation in consumer choice sets which helps in the identification of demand. However, these studies focus on the exogenous component of such variation and do not examine the endogenous selection problem tied to product entry and exit. See the work of Berry and Haile, (2014, 2022), as well as the recent survey papers by Berry and Haile, (2021) and Gandhi and Nevo, (2021). One common approach to tackle this selection problem is to incorporate fixed effects, such as product, market, and time fixed effects while assuming that the remaining portion of the error term in the demand equation is unknown to firms when they make their market/product entry decisions. This approach is employed in studies by Aguirregabiria and Ho (2012), Sweeting (2013), and Eizenberg (2014). Although this fixed effects approach is convenient in practice, it relies on assumptions regarding firms' information that may not be realistic in certain empirical applications. Furthermore, these assumptions can be subject to testing and potential rejection by the data. Notably, our model in this paper encompasses the fixed effects model as a specific case.

the literature on nonparametric finite mixture models, we show that data on firms' product entry decisions nonparametrically identify the probabilities of product entry conditional on the market type and the density of unobserved market types. Lastly, we demonstrate that, under mild conditions on the observable variables, demand parameters are identified after effectively controlling for the nonparametric entry probabilities and densities for each market type. Our proof of sequential identification addresses two important issues in nonparametric finite mixture models: the identification up to label swapping, and the fact that sometimes we can only establish a non-binding lower bound for the number of unobserved market types.

Second, building on our constructive proof of identification, the paper proposes a simple two-step estimator to address endogenous selection. In the first step, a semiparametric mixture model is employed to estimate both the probability distribution of unobserved market types and the conditional choice probabilities of product entry in the case of a static entry game, or product entry and exit in the event of a dynamic game. In the second step, demand parameters are estimated using a Generalized Method of Moments (GMM) approach that considers both endogenous product availability and price endogeneity.

Third, the paper illustrates the proposed method by applying it to data from the airline industry. Our findings highlight the importance of accounting for endogenous product entry and for a finite mixture of latent market types, as failure to do so can lead to significant biases. Specifically, neglecting latent market types while accounting for endogenous selection results in significant attenuation biases in the estimates of demand own-price elasticities.

Our paper is motivated by and closely aligned with the work of Draganska, Mazzeo, and Seim (2009), Ciliberto, Murry, and Tamer (2021), and Li, Mazur, Park, Roberts, Sweeting, and Zhang (2022). These authors have developed methods for estimating structural models that combine the demand for differentiated products in Berry, Levinsohn, and Pakes (1995) with games of market/product entry as in Bresnahan and Reiss (1990, 1991) and Berry (1992). Their focus is on the joint estimation of all the structural parameters in the model, including demand, marginal costs, entry costs, and the probability distribution of unobservable factors. To jointly estimate the full model, these authors employ nested fixed point algorithms, which require solving multiple times for the equilibria of a two-step game. Consequently, they rely on strong parametric assumptions for all the structural functions and the distribution of unobservables. In contrast, we adopt a sequential approach to identify and estimate the structural parameters and functions within the model. Our approach yields identification outcomes that guarantee alignment of the supply side structure with an equilibrium model, all the while maintaining a nonparametric specification. This means we do not impose strong assumptions on marginal costs, entry costs, or the distribution of unobservables. Furthermore, our estimation method offers computational

simplicity since it does not necessitate the computation of equilibria. Finally, the method and its computational advantages apply both to static games of market entry and dynamic games of market entry and exit.[3]

Our model of product entry and exit relates to the literature of structural models of market entry/exit with incomplete information such as the static games in Seim (2006) and Draganska, Mazzeo, and Seim (2009) and the dynamic games in Aguirregabiria and Mira (2007), Pakes, Ostrovsky, and Berry (2007), or Sweeting (2009). The specification of the unobservables in the entry-decision part of our model — which combines common-knowledge unobservables with finite support and private information unobservables with continuous support — is closely related to the models in Xiao (2018) and Aguirregabiria and Mira (2019).

Our estimation method is rooted in and expands upon the literature on semiparametric estimation of sample selection models, with seminal contributions by Newey, Powell, and Walker (1990), Ahn and Powell (1993), Powell (2001), and Newey (2009). A distinctive aspect of our method lies in the multi-dimensionality of the selection term in the second step of estimation. Specifically, the selection term involves the convolution of functions representing market entry probabilities for each latent market type. Our method builds upon the semiparametric sieve method in Das, Newey, and Vella (2003) and Newey (2009). We enhance the applicability of this method by broadening its scope in the context of our model.

Our method also relates to the literature on estimating Conditional Average Treatment Effects (CATE) using finite mixture models for the propensity score (Haviland and Nagin, 2005, Haviland, Nagin, Rosenbaum, and Tremblay, 2008, Lanza, Coffman, and Xu, 2013). In the first step of our method, we follow a similar approach, with the only significant distinction being the crucial role played by the conditional independence between firms' entry decisions to obtain nonparametric identification. However, the second step of our method differs substantially. Previous studies in the latent propensity score literature assign each individual in the sample to a latent class using techniques such as the highest posterior probability (modal assignment) and treating the assigned class as an observable variable. In contrast, our method recognizes that an individual's (firm's) latent class remains unobservable, even after estimating the finite mixture model using an infinite sample. Consequently, controlling for selection bias in the second step requires the inclusion of propensity scores for all potential latent types.

[3]It is worth noting that, given estimates of demand parameters and unobservables from our method, one can obtain estimates of marginal costs and entry costs with less stringent parametric assumptions than those required for joint estimation of the full structural model. Similar to Ciliberto, Murry, and Tamer (2021) and Li, Mazur, Park, Roberts, Sweeting, and Zhang (2022), we can rely on the estimates of our model to implement a wide range of counterfactual experiments accounting for the endogeneity of product entry and exit. This is particularly valuable when simulating the effects of a merger, as demonstrated by Li, Mazur, Park, Roberts, Sweeting, and Zhang (2022). In section 6.4, we discuss the implementation of these counterfactuals.

The rest of the paper is organized as follows. Section 2 presents our model and assumptions. Section 3 describes the selection problem in this model. Section 4 presents our identification results. We describe our estimation method in section 5 and illustrate it using an empirical application on the US airline industry in section 6. We summarize and conclude in section 7.

2 Model

The demand system follows the BLP framework (Berry, Levinsohn, and Pakes, 1995). For the sake of notational simplicity, we focus on single-product firms. In section 2.4, we discuss how to adapt our model and methodology to the case of multi-product firms, and more specifically the application of Lemma 1 and Assumptions 1 and 2 to this scenario. There are J firms indexed by $j \in \mathcal{J} = \{1, 2, ..., J\}$ and T markets indexed by $t \in \{1, 2, ..., T\}$, where a market can be a geographic location, a time period, or a combination of both. Consumers living in a market t can buy only the products available in that market. Firms' market entry decisions, prices, and quantities are determined as an equilibrium of a two-stage game. In the first stage, firms maximize their expected profit by choosing whether to be active or not in the market. In the second stage, prices and quantities of the active firms are determined as a Nash-Bertrand equilibrium of a pricing game. This two-stage game is played separately across markets. Demand and price competition are static. Our model accommodates both static and dynamic games for firms' product entry (and exit) decisions.

2.1 Demand

The indirect utility of household h in market t from buying product j is:

$$U_{hjt} \equiv \delta(p_{jt}, \boldsymbol{x}_{jt}) + v(p_{jt}, \boldsymbol{x}_{jt}, \upsilon_{ht}) + \varepsilon_{hjt}, \tag{1}$$

where p_{jt} and $\boldsymbol{x}_{jt}$ are the price and other characteristics, respectively, of product j in market t; $\delta_{jt} \equiv \delta(p_{jt}, \boldsymbol{x}_{jt})$ is the average (indirect) utility of product j in market t; and $v(p_{jt}, \boldsymbol{x}_{jt}, \upsilon_{ht}) + \varepsilon_{hjt}$ represents a household-specific deviation from the average utility. The term $v(p_{jt}, \boldsymbol{x}_{jt}, \upsilon_{ht})$ depends on the vector of random coefficients υ_{ht} that is unobserved to the researcher with distribution $F_v(\cdot|\boldsymbol{\sigma})$, where $\boldsymbol{\sigma}$ is a vector of parameters. The term ε_{hjt} is unobserved to the researcher and is i.i.d. over (h, j, t) with type I extreme value distribution. Following the standard specification, the average utility of product j is:

$$\delta_{jt} \equiv \alpha\ p_{jt} + \boldsymbol{x}'_{jt}\ \boldsymbol{\beta} + \xi_{jt}, \tag{2}$$

where α and $\boldsymbol{\beta}$ are parameters. Variable ξ_{jt} captures the characteristics of product j in market t which are unobserved to the researcher. The outside option is represented by $j = 0$ and its indirect utility is normalized to $U_{h0t} = \varepsilon_{h0t}$.

Let $a_{jt} \in \{0, 1\}$ be the indicator for product j being available in market t, and let $\boldsymbol{a}_t \equiv (a_{jt} : j \in \mathcal{J})$ denote the vector with the indicators for the availability of every product in market t. The outside option $j = 0$ is always available in every market. Every household chooses the product that maximizes their utility. Let s_{jt} be the market share of product j in market t, i.e., the proportion of households choosing product j:

$$s_{jt} = d_{jt}(\boldsymbol{\delta}_t, \boldsymbol{a}_t) \equiv \int \frac{a_{jt} \ \exp\left(\delta_{jt} + v(p_{jt}, \boldsymbol{x}_{jt}, v)\right)}{1 + \sum_{i=1}^{J} a_{it} \ \exp\left(\delta_{it} + v(p_{it}, \boldsymbol{x}_{it}, v)\right)} \ dF_v(v). \tag{3}$$

This system of J equations represents the demand system in market t. We can represent this system in a vector form as: $\boldsymbol{s}_t = \boldsymbol{d}_t(\boldsymbol{\delta}_t, \boldsymbol{a}_t)$.

For our analysis, it is convenient to define the sub-system of demand equations that includes market shares, average utilities, and product characteristics of only those products available in the market. We represent this system as:

$$\boldsymbol{s}_t^{(\boldsymbol{a})} = \boldsymbol{d}_t^{(\boldsymbol{a})}(\boldsymbol{\delta}_t^{(\boldsymbol{a})}), \tag{4}$$

where $\boldsymbol{s}_t^{(\boldsymbol{a})}$ is the subvector of $\boldsymbol{s}_t$ containing the market shares for only those products available in the market and a similar definition applies to the subvector $\boldsymbol{\delta}_t^{(\boldsymbol{a})}$. Lemma 1 establishes that the invertibility property in Berry (1994) applies to demand system (4) for any value of $\boldsymbol{a}$.

LEMMA 1. *Suppose that the outside option $j = 0$ is always available. Then, for any value of the vector $\boldsymbol{a} \in \{0, 1\}^J$, the system $\boldsymbol{s}_t^{(\boldsymbol{a})} = \boldsymbol{d}_t^{(\boldsymbol{a})}(\boldsymbol{\delta}_t^{(\boldsymbol{a})})$ is invertible with respect to $\boldsymbol{\delta}_t^{(\boldsymbol{a})}$ such that for every product in this subsystem (i.e., for every product with $a_{jt} = 1$) the inverse function $\delta_{jt}^{(\boldsymbol{a})} = d_{jt}^{-1}(\boldsymbol{s}_t^{(\boldsymbol{a})})$ exists.* ■

Proof of Lemma 1. If the outside option $j = 0$ is available, then, for any value of the vector $\boldsymbol{a}$, the system of equations (4) satisfies the conditions for invertibility in Berry (1994). ■

For a product available in market t, we have:

$$d_{jt}^{-1}(\boldsymbol{s}_t^{(\boldsymbol{a}_t)}) \ = \ \delta_{jt} \ = \ \alpha \ p_{jt} + \boldsymbol{x}_{jt}' \ \boldsymbol{\beta} + \xi_{jt} \quad \text{if and only if} \quad a_{jt} = 1. \tag{5}$$

Importantly, this regression equation for product j only depends on the availability of product j and not on the availability of the other products. Therefore, the selection problem in the

estimation of the demand of product j can be described in terms of the conditional expectation

$$\mathbb{E}\left(\xi_{jt} \mid a_{jt} = 1\right). \tag{6}$$

This is an important implication of working directly with the inverse demand system, as represented by equation (5).

To appreciate the value of this property, consider instead the case of the *Almost Ideal Demand System* (AIDS) (Deaton and Muellbauer, 1980). In the AIDS, each value of the vector $\boldsymbol{a}_t$ implies a different set of regressors and slope parameters in the regression equation that relates the demand of product j to the log-prices of the available products. Therefore, in the AIDS model, the selection problem in the estimation of the demand of product j is not only related to the availability of that product but to the availability of all products in the system. In other words, the selection term cannot be represented in terms of $\mathbb{E}\left(\xi_{jt} \mid a_{jt} = 1\right)$ but must instead be expressed in terms of $\mathbb{E}\left(\xi_{jt} \mid a_{jt} = 1,\ \boldsymbol{a}_{-jt} = \boldsymbol{a}_{-j}\right)$. That is, in the AIDS model we have a different selection term for each value of the vector $\boldsymbol{a}_{-j}$ representing the availability of products other than j. This structure makes the selection problem multi-dimensional and significantly complicates identification and estimation when the number of products J is large.[4]

The next Example illustrates Lemma 1 in the case of a nested logit model.

EXAMPLE 1 (Nested logit model). The J products are partitioned into R mutually exclusive groups indexed by r. We denote by r_j the group to which product j belongs. The indirect utility function is $U_{htj} \equiv \delta_{jt} + (1-\sigma)\ v_{ht,r_j} + \varepsilon_{htj}$, where variables v and ε are independently distributed, ε and $(1-\sigma)\ v + \varepsilon$ are i.i.d. type I extreme value, and $\sigma \in [0,1]$ is a parameter (Cardell, 1997). This model implies $s_{jt} = d_j^{(\boldsymbol{a}_t)}(\boldsymbol{\delta}_t) = d_{r_j}^{(\boldsymbol{a}_t)}\ d_{j|r_j}^{(\boldsymbol{a}_t)}$ with

$$d_{j|r_j}^{(\boldsymbol{a}_t)} = \frac{a_{jt}\ e^{\delta_{jt}}}{\sum_{i \in r_j} a_{it}\ e^{\delta_{it}}} \quad \text{and} \quad d_{r_j}^{(\boldsymbol{a}_t)} = \frac{\left[\sum_{i \in r_j} a_{it}\ e^{\delta_{it}}\right]^{\frac{1}{1-\sigma}}}{1 + \sum_{r=1}^{R}\left[\sum_{i \in r} a_{it}\ e^{\delta_{it}}\right]^{\frac{1}{1-\sigma}}}. \tag{7}$$

If $a_{jt} = 1$ and $s_{0t} > 0$, the inverse function $d_j^{(\boldsymbol{a}_t)-1}(\cdot)$ exists — regardless of the value of a_{it} for any product i different from j. It is straightforward to show that this inverse function has the following form:

$$\delta_{jt} = \ln\left(\frac{s_{jt}}{s_{0t}}\right) - \sigma\ \ln\left(\frac{\sum_{i \in r_j} s_{it}}{s_{0t}}\right) \tag{8}$$

[4] As we explain in section 2.4, a similar dimensionality problem appears in our model in the case of multi-product firms.

and it implies the regression equation:

$$\ln\left(\frac{s_{jt}}{s_{0t}}\right) = \sigma \ \ln\left(\frac{\sum_{i\in r_j} s_{it}}{s_{0t}}\right) + \alpha \ p_{jt} + \boldsymbol{x}'_{jt} \ \boldsymbol{\beta} + \xi_{jt}. \tag{9}$$

Given $s_{0t} > 0$, this regression equation holds whenever $a_{jt} = 1$. ■

2.2 Price competition

Let Π_{jt} be the profit of firm j if active in market t. This equals revenues minus costs:

$$\Pi_{jt} = p_{jt} \ q_{jt} - c(q_{jt}; \boldsymbol{x}_{jt}, \omega_{jt}) - f(\boldsymbol{x}_{jt}, \eta_{jt}), \tag{10}$$

where q_{jt} is the quantity sold (i.e., market share s_{jt} times market size H_t), $c(q_{jt}; \boldsymbol{x}_{jt}, \omega_{jt})$ is the variable cost function, and $f(\boldsymbol{x}_{jt}, \eta_{jt})$ is the fixed entry cost. Variables ω_{jt} and η_{jt} are unobserved to the researcher.

Given firms' entry decisions, the best response function in the Bertrand pricing game implies the following system of pricing equations:

$$p_{jt} = mc_{jt} - d_{jt}^{(\boldsymbol{a}_t)} \ \left[\frac{\partial d_{jt}^{(\boldsymbol{a}_t)}}{\partial p_{jt}}\right]^{-1} \ \text{for every } j \in \mathcal{J}, \tag{11}$$

where mc_{jt} is the marginal cost $\partial c_{jt}/\partial q_{jt}$. A solution to this system of equations is a Nash-Bertrand equilibrium. The pricing game may have multiple equilibria. We do not impose restrictions on equilibrium selection and allow each market to potentially select a different equilibrium. We use scalar variable τ_t^2 to index the equilibrium type selected in the Bertrand game, i.e., in step 2 of the two-stage game.

Let $\boldsymbol{x}_t \equiv (\boldsymbol{x}_{jt} : j \in \mathcal{J})$ be the vector with all the exogenous variables that are observed to the researcher, affecting demand or costs. Vectors $\boldsymbol{\xi}_t$ and $\boldsymbol{\omega}_t$ have similar definitions. Let $\boldsymbol{a}_{-jt}$ be the vector with the entry decisions of every firm other than j. We use $VP_j(\boldsymbol{a}_{-jt}, \boldsymbol{x}_t, \boldsymbol{\xi}_t, \boldsymbol{\omega}_t, \tau_t^2)$ to denote the indirect variable profit function for firm j that results from plugging into the expression $p_{jt} \ q_{jt} - c(q_{jt}; \boldsymbol{x}_{jt}, \omega_{jt})$ the value of (p_{jt}, q_{jt}) from the Nash-Bertrand equilibrium given $(a_{jt} = 1, \boldsymbol{a}_{-jt}, \boldsymbol{x}_t, \boldsymbol{\xi}_t, \boldsymbol{\omega}_t, \tau_t^2)$.

2.3 Market entry game and information structure

Firms' entry decisions are determined as an equilibrium of a game of market entry. The profit of being inactive is normalized to zero for all firms. Firms have uncertainty about their profits if

active in the market. Their information about demand and costs plays a key role in their entry decisions and, therefore, on the selection problem in the estimation of demand. Assumptions 1 and 2 summarize our conditions on the information structure and on the entry cost function, respectively.

ASSUMPTION 1. *Firm j's information at the moment of its entry decision in market t consists of $(\boldsymbol{x}_t, \kappa_t, \tau_t^1, \eta_{jt})$.*

A. κ_t is a signal for the demand-cost variables $(\boldsymbol{\xi}_t, \boldsymbol{\omega}_t, \tau_t^2)$. It is common knowledge for the firms, it has discrete and finite support that we denote as $\mathcal{K}(\boldsymbol{x}_t)$, and its probability distribution conditional on $\boldsymbol{x}_t$ is $f_\kappa(\kappa_t|\boldsymbol{x}_t)$, which is nonparametrically specified.

B. Variable τ_t^1 represents the type of equilibrium selected in the entry game.

C. Variable η_{jt} is a component of the entry cost that is private information of firm j, independently distributed over firms, and independent of $(\kappa_t, \boldsymbol{x}_t)$ with CDF F_η, which is strictly increasing over the real line.

D. Vector $(\boldsymbol{\xi}_t, \boldsymbol{\omega}_t, \tau_t^2, \kappa_t, \tau_t^1, \eta_{jt})$ is unobserved to the researcher. Conditional on κ_t, variables $\boldsymbol{\xi}_t$, $\boldsymbol{\omega}_t$, and η_{jt} are independent of $\boldsymbol{x}_t$. ■

For our analysis, the payoff-relevant information in discrete variable κ_t plays the same role as the equilibrium-selection discrete variable τ_t^1. Therefore, for notational simplicity, we omit τ_t^1 and interpret κ_t as representing both equilibrium selection and payoff relevant variables. Let $|\mathcal{K}(\boldsymbol{x}_t)|$ be the number of points in the support of κ_t. We then represent κ_t as an index with support $\{1, 2, ..., |\mathcal{K}(\boldsymbol{x}_t)|\}$. Similarly, with some abuse of notation, for the rest of the paper we represent the vector of unobservables $(\boldsymbol{\xi}_t, \boldsymbol{\omega}_t, \tau_t^2)$ using the more compact notation $\boldsymbol{\xi}_t$.[5]

Let $\pi_j(\boldsymbol{a}_{-j}, \boldsymbol{x}_t, \kappa_t, \eta_{jt})$ be firm j's expected profit given its information about demand and costs and conditional on the hypothetical entry profile $\boldsymbol{a}_{-j} \in \{0,1\}^{J-1}$. Under Assumption 1:

$$\pi_j(\boldsymbol{a}_{-j}, \boldsymbol{x}_t, \kappa_t, \eta_{jt}) \;=\; \int VP_j(\boldsymbol{a}_{-j}, \boldsymbol{x}_t, \boldsymbol{\xi}_t)\; dF_{j,\xi}\left(\boldsymbol{\xi}_t \mid \kappa_t, \eta_{jt}\right) - f(\boldsymbol{x}_{jt}, \eta_{jt}), \tag{12}$$

where $F_{j,\xi}\left(\boldsymbol{\xi}_t \mid \kappa_t, \eta_{jt}\right)$ is the CDF of $\boldsymbol{\xi}_t$ conditional on (κ_t, η_{jt}).

ASSUMPTION 2. *For any value $(\boldsymbol{a}_{-j}, \boldsymbol{x}_t, \kappa_t)$, the function $\pi_j(\boldsymbol{a}_{-j}, \boldsymbol{x}_t, \kappa_t, \eta_{jt})$ is strictly monotonic in η_{jt}. Without loss of generality, we consider that this function is decreasing*

[5]Note that in the entry game of this model, the set of regular Bayesian Nash equilibria is finite. See Lemma 1 in Aguirregabiria and Mira (2019). This is consistent with our assumption of finite support for κ_t.

in η_{jt}. *Therefore, for any scalar value* π^0 *and any value* $(\boldsymbol{a}_{-j}, \boldsymbol{x}_t, \kappa_t)$, *the equation* $\pi^0 = \pi_j(\boldsymbol{a}_{-j}, \boldsymbol{x}_t, \kappa_t, \eta_{jt})$ *is invertible with respect to* η_{jt}. *That is, there is an inverse function* π_j^{-1} *such that* $\eta_{jt} = \pi_j^{-1}(\boldsymbol{a}_{-j}, \boldsymbol{x}_t, \kappa_t, \pi^0)$, *and for any other scalar* π^1 *with* $\pi^1 \leq \pi^0$, *we have that* $\pi_j^{-1}(\boldsymbol{a}_{-j}, \boldsymbol{x}_t, \kappa_t, \pi^1) \geq \eta_{jt}$. ■

The presence of firms' private information implies that the entry game is of incomplete information. Given $(\boldsymbol{x}_t, \kappa_t)$, a *Bayesian Nash Equilibrium (BNE)* of this game can be represented as a J-tuple of entry probabilities, one for each firm, $(P_{jt} : j \in \mathcal{J})$. To describe this BNE, we first define a firm's expected profit function that accounts for its uncertainty about other firms' entry decisions.

$$\pi_j^P(\boldsymbol{x}_t, \kappa_t, \eta_{jt}) = \sum_{\boldsymbol{a}_{-j} \in \{0,1\}^{J-1}} \left(\prod_{i \neq j} [P_{it}]^{a_i} [1 - P_{it}]^{1-a_i} \right) \pi_j(\boldsymbol{a}_{-j}, \boldsymbol{x}_t, \kappa_t, \eta_{jt}). \tag{13}$$

Under Assumption 2, the expected profit function $\pi_j^P(\boldsymbol{x}_t, \kappa_t, \eta_{jt})$ is strictly monotonic and invertible in η_{jt}.[6] Let $\pi_j^{P(-1)}$ be this inverse function. Then, given the entry probabilities of its competitors, firm j's best response is to enter in the market if and only if:

$$\eta_{jt} \leq \pi_j^{P(-1)}(\boldsymbol{x}_t, \kappa_t, 0) \tag{14}$$

Taking this into account, we can define a BNE in this game as follows.

DEFINITION 1 (BNE). *Under Assumptions 1-2 and given* $(\boldsymbol{x}_t, \kappa_t)$, *a Bayesian Nash Equilibrium (BNE) can be represented as a* J*-tuple of probabilities* $\{P_{jt} \equiv P_j(\boldsymbol{x}_t, \kappa_t) : j \in \mathcal{J}\}$ *that solve the following system of* J *best response equations in the space of probabilities:*

$$P_{jt} = F_\eta \left(\pi_j^{P(-1)}(\boldsymbol{x}_t, \kappa_t, 0) \right), \tag{15}$$

where $\pi_j^{P(-1)}$ *is the inverse function with respect to* η_{jt} *of the expected profit in (13).* ■

This framework encompasses various information structures concerning the unobservable factors, all corresponding to distinct scenarios considered within structural models of market entry and product introduction. When $\text{Var}(\kappa_t) = 0$, the scenario aligns with an entry game featuring solely private information unobservables, as examined in studies such as Seim (2006), Sweeting (2009), and Bajari, Hong, Krainer, and Nekipelov (2010). As $\text{Var}(\eta_{jt})$ approaches zero, we converge to a scenario characterized by an entry game involving exclusively complete

[6] If π_j is strictly monotonic in η_{jt} for any possible entry profile $\boldsymbol{a}$, then a convex linear combination of π_j for different entry profiles is also strictly monotonic in η_{jt}.

information unobservables, as in the work by Ciliberto and Tamer (2009) and Ciliberto, Murry, and Tamer (2021). In instances where $\text{Var}(\kappa_t) > 0$ and $\text{Var}(\eta_{jt}) > 0$, the model describes an entry game encompassing both categories of unobservable factors, as in the work by Grieco (2014) and Aguirregabiria and Mira (2019).

2.4 Multi-product firms

We briefly discuss here how the model, the results above, and the characterization of the selection problem in section 3 below can be extended to the case of multi-product firms. We still use $j \in \mathcal{J}$ to index products but now we introduce the firm sub-index f and define $\mathcal{J}_f \subseteq \mathcal{J}$ as the set of products owned by firm f. The product entry decisions of firm f are described by vector $\boldsymbol{a}_{ft} \equiv (a_{jt} : j \in \mathcal{J}_f) \in \{0,1\}^{|\mathcal{J}_f|}$.

First, it is important to note that Lemma 1's applicability remains unaffected by the product ownership structure. This Lemma only rests on the structure of the demand system. Therefore, regardless of the product ownership structure, the selection problem in the estimation of the demand of product j is still described in terms of the conditional expectation $\mathbb{E}(\xi_{jt} \mid a_{jt} = 1)$.

Second, Assumption 1, which describes a firm's information at the moment of its entry decisions in a market t, remains basically the same when we consider multi-product firms. The only difference is that we need to represent a firm's private information entry cost using a vector with as many elements as possible products for this firm; that is, $\boldsymbol{\eta}_{ft} \equiv (\eta_{ft}(\boldsymbol{a}_f) : \boldsymbol{a}_f \in \{0,1\}^{|\mathcal{J}_f|})$. For instance, in the case of a two-product firm, $\eta_{ft}(1,0)$ is the latent component of entry cost when the firm offers product 1 while excluding product 2. Under Assumption 1, equation (12), describing the expected profit of a firm, readily extends to multi-product firms as follows:

$$\pi_f\left(\boldsymbol{a}_f, \boldsymbol{a}_{-f}, \boldsymbol{x}_t, \kappa_t, \boldsymbol{\eta}_{ft}\right) = \int VP_f\left(\boldsymbol{a}_f, \boldsymbol{a}_{-f}, \boldsymbol{x}_t, \boldsymbol{\xi}_t\right) \, dF_{f,\xi}\left(\boldsymbol{\xi}_t \mid \kappa_t, \boldsymbol{\eta}_{ft}\right) - f(\boldsymbol{x}_{ft}, \boldsymbol{\eta}_{ft}), \tag{16}$$

where $F_{f,\xi}\left(\boldsymbol{\xi}_t \mid \kappa_t, \boldsymbol{\eta}_{ft}\right)$ is the CDF of $\boldsymbol{\xi}_t$ conditional on $(\kappa_t, \boldsymbol{\eta}_{ft})$.

The following Assumption 2-Multi is the extension of Assumption 2 to the case of multi-product firms.

ASSUMPTION 2-Multi. *The expected profit function π_f has a structure with respect to $\boldsymbol{\eta}_{ft}$ implying that a choice of product portfolio maximizes expected profit if and only if the elements in vector $\boldsymbol{\eta}_{ft}$ satisfy $|\mathcal{J}_f| - 1$ thresholds conditions, where the threshold values are functions only of $(\boldsymbol{x}_t, \kappa_t)$, say, $\boldsymbol{g}_f(\boldsymbol{x}_t, \kappa_t)$. For instance $\boldsymbol{a}_{ft} = \boldsymbol{a}_f$ if and only if $\boldsymbol{\eta}_{ft} \leq \boldsymbol{g}_f(\boldsymbol{x}_t, \kappa_t)$.* ■

Under Assumption 2-Multi, the probability of choosing a product portfolio is a function of the vector of thresholds $\boldsymbol{g}_f(\boldsymbol{x}_t, \kappa_t)$. There is a one-to-one relationship between the $|\mathcal{J}_f| - 1$ dimension

vector of portfolio-choice probabilities and the vector of threshold values. This structure implies that the characterization of the selection term in the demand regression model is similar with multi-product or single-product firms. In section 3 below, for single-product firms, we show that this selection term is the average over all possible unobserved market types κ_t of a function of the product entry probability $P_j(\boldsymbol{x}_t, \kappa_t)$. Similarly, in the multi-product case, the selection term is the average over all possible unobserved market types of a function of the vector of probabilities of every possible product portfolio.

While the preceding discussion illustrates the similar structure shared by the selection problem with single- and multi-product firms, it also underscores a noteworthy practical difference. The dimension of the vector of choice probabilities we need to control for to deal with sample selection bias grows exponentially with the number of products per firm. In some applications, this can pose a substantial challenge in the practical implementation of our method.

2.5 Dynamic game of product entry and exit

Our framework and identification results can accommodate cases in which firms' decisions about product availability come from a Markov Perfect Equilibrium (MPE) of a dynamic game of product entry and exit, where firms are forward-looking. In this dynamic game, a firm's fixed cost is denoted as $f(a_{it}, a_{i,t-1}, \boldsymbol{x}_{jt}, \eta_{jt})$, where $f(1, 0, \boldsymbol{x}_{jt}, \eta_{jt})$ represents the cost of entry, $f(0, 1, \boldsymbol{x}_{jt}, \eta_{jt})$ is the cost of exit, $f(1, 1, \boldsymbol{x}_{jt}, \eta_{jt})$ is the fixed cost when a product stays in the market, and $f(0, 0, \boldsymbol{x}_{jt}, \eta_{jt})$ can be normalized to zero.

ASSUMPTION 3. *Suppose that t represents time. (A) The vector of state variables at period t, $\boldsymbol{x}_t$, includes the entry decisions of all the firms at the previous period, $(a_{j,t-1} : j = 1, 2, ..., J)$. (B) The exogenous product characteristics in vector $\boldsymbol{x}_t$ and the latent market type κ_t are either time-invariant or follow a first-order Markov process. (C) The private information shock η_{jt} is independently and identically distributed over time and independent across firms.* ■

Assumptions 3(B) and 3(C) are standard in the literature of empirical dynamic discrete choice games (see Aguirregabiria, Collard-Wexler, and Ryan, 2021). Under Assumption 3, the value of being or not in the market depends on the state variables $(\boldsymbol{x}_t, \kappa_t)$ and on the private information shock η_{jt}. Let $v_j^P(\boldsymbol{x}_t, \kappa_t, \eta_{jt})$ be the difference between the value functions of being in the market and not being in the market at period t. This function can be represented as the sum of two functions: the difference between current profits and the difference between expected continuation values. Assumption 3(C) implies that η_{jt} enters in the current profit but not in the expected continuation value. Therefore, by Assumption 2 on the strict monotonicity of the profit function with respect to η_{jt}, we have that the value function $v_j^P(\boldsymbol{x}_t, \kappa_t, \eta_{jt})$ is also strictly

monotonic in η_{jt}. This implies that the best response of a firm in the dynamic game can be described by a threshold condition as in equation (14). Consequently, and similar to a BNE in a static entry game, a MPE in a dynamic game can be characterized in terms of J conditional choice probabilities.

DEFINITION 2 (MPE). *Suppose that Assumptions 1-3 hold. Then, a Markov Perfect Equilibrium (MPE) can be represented as a J-tuple of probability functions $\{P_j(\boldsymbol{x}_t, \kappa_t) : j \in \mathcal{J})\}$ that solve the following system of best response equations in the space of probability functions:*

$$P_j\left(\boldsymbol{x}_t, \kappa_t\right) \;=\; F_\eta\left(v_j^{P(-1)}\left(\boldsymbol{x}_t, \kappa_t, 0\right)\right), \tag{17}$$

where $v_j^{P(-1)}$ is the inverse function with respect to η_{jt} of the difference between the value of being in the market and the value of not being in the market at period t. ■

For the rest of the paper, we will not distinguish whether the choice probabilities $P_j\left(\boldsymbol{x}_t, \kappa_t\right)$ come from a BNE of a static entry game or from a MPE of a dynamic entry/exit game. All our identification results apply to both cases.

3 Selection problem

For simplicity and concreteness, we describe our sample selection problem using the nested logit demand model from Example 1. We use the starred variables s_{jt}^* and p_{jt}^* to represent *latent variables*. That is, s_{jt}^* and p_{jt}^* represent the latent market share and price, respectively, that we would observe if product j were offered in market t. Using these latent variables, we can write the following demand system:

$$\ln\left(\frac{s_{jt}^*}{s_{0t}}\right) \;=\; \sigma\;\ln\left(\frac{s_{jt}^* + S_{-jt}}{s_{0t}}\right) + \alpha\; p_{jt}^* + \boldsymbol{x}_{jt}'\,\boldsymbol{\beta} + \xi_{jt}, \tag{18}$$

where $S_{-jt} \equiv \sum_{i \neq j, i \in r_j} s_{it}$ is the aggregate market share of all products in group r_j other than product j. Latent variables (s_{jt}^*, p_{jt}^*) are equal to the observed variables (s_{jt}, p_{jt}) if and only if product j is offered in market t:

$$\{s_{jt}^* = s_{jt} \text{ and } p_{jt}^* = p_{jt}\} \text{ if and only if } a_{jt} = 1. \tag{19}$$

Firm j's best response entry decision completes the econometric model:[7]

$$a_{jt} \;=\; 1\left\{\eta_{jt} \;\leq\; \pi_j^{P(-1)}(\boldsymbol{x}_t, \kappa_t)\right\}. \tag{20}$$

Equations (18) to (20) imply the following regression equation for any product with $a_{jt} = 1$:

$$\ln\left(\frac{s_{jt}}{s_{0t}}\right) \;=\; \sigma\;\ln\left(\frac{s_{jt}+S_{-jt}}{s_{0t}}\right) + \alpha\; p_{jt} + \boldsymbol{x}'_{jt}\;\boldsymbol{\beta} + \lambda_j(\boldsymbol{x}_t) + \widetilde{\xi}_{jt}, \tag{21}$$

where $\lambda_j(\boldsymbol{x}_t)$ is the *selection bias function*, $\mathbb{E}\left(\xi_{jt} \mid \boldsymbol{x}_t, a_{jt} = 1\right)$. That is,

$$\lambda_j(\boldsymbol{x}_t) = \int \xi_{jt}\; 1\left\{\eta_{jt} \;\leq\; \pi_j^{P(-1)}(\boldsymbol{x}_t, \kappa_t)\right\}\; \frac{f_{\xi,\eta,\kappa}\left(\xi_{jt}, \eta_{jt}, \kappa_t | \boldsymbol{x}_t\right)}{\Pr\left(a_{jt} = 1 | \boldsymbol{x}_t\right)}\; d\left(\xi_{jt}, \eta_{jt}, \kappa_t\right), \tag{22}$$

and:

$$\Pr\left(a_{jt} = 1 | \boldsymbol{x}_t\right) \equiv \mathbb{E}\left(a_{jt} \mid \boldsymbol{x}_t\right) = \int 1\left\{\eta_{jt} \;\leq\; \pi_j^{P(-1)}(\boldsymbol{x}_t, \kappa_t)\right\}\; f_{\eta,\kappa}\left(\eta_{jt}, \kappa_t | \boldsymbol{x}_t\right)\; d\left(\eta_{jt}, \kappa_t\right), \tag{23}$$

where $f_{\eta,\kappa}$ and $f_{\xi,\eta,\kappa}$ are the joint density functions of (η_{jt}, κ_t) and $(\xi_{jt}, \eta_{jt}, \kappa_t)$, conditional of $\boldsymbol{x}_t$, respectively.

Note that the selection bias function $\lambda_j(\boldsymbol{x}_t)$ is a nonparametric function of all its arguments $\boldsymbol{x}_t$. Therefore, based on equation (21), and without further restrictions, the demand parameters σ, α, and $\boldsymbol{\beta}$ are not identified. That is, we cannot disentangle the direct effect of $\boldsymbol{x}_{jt}$ on demand (as represented by the vector of parameters $\boldsymbol{\beta}$) from the indirect effect that comes from the selection bias function $\lambda_j(\boldsymbol{x}_t)$.

Examples 2 and 3 below present restrictions that imply the identification of the demand parameters. Example 2 is simple but it imposes strong restrictions on the unobservables. Our main identification results in section 4 are closely related to Example 3.

EXAMPLE 2 (No signals κ_t). Suppose that: (1^*) $\kappa_t = 0$ such that, at the moment of the entry decision, the only information that a firm has about the demand/cost variables $\boldsymbol{\xi}_t$ is its private information variable η_{jt}; and (2^*) a unique equilibrium is played across all entry games with the same observables $\boldsymbol{x}_t$. Under conditions (1^*) and (2^*), the selection term $\lambda_j(\boldsymbol{x}_t)$ only depends on the CCP $P_j\left(\boldsymbol{x}_t\right)$: that is, $\lambda_j(\boldsymbol{x}_t) = \rho_j(P_j\left(\boldsymbol{x}_t\right))$ for some function $\rho_j(\cdot)$.

The proof is straightforward. Under conditions (1^*)-(2^*), the inverse profit function $\pi_j^{P(-1)}$ and the equilibrium entry probability P_j only depend on $\boldsymbol{x}_t$ but not on κ_t. The equilibrium entry probability $P_j\left(\boldsymbol{x}_t\right)$ is equal to the conditional expectation $\mathbb{E}\left(a_{jt} | \boldsymbol{x}_t\right)$, which is nonpara-

[7] With some abuse of notation, we use function $\pi_j^{P(-1)}(\boldsymbol{x}_t, \kappa_t)$ to represent $\pi_j^{P(-1)}(\boldsymbol{x}_t, \kappa_t, 0)$.

metrically identified. Furthermore, this probability satisfies the equilibrium condition $P_j(\boldsymbol{x}_t) = F_\eta(\pi_j^{P(-1)}(\boldsymbol{x}_t))$, such that $\pi_j^{P(-1)}(\boldsymbol{x}_t) = F_\eta^{-1}(P_j(\boldsymbol{x}_t))$, and the entry condition can be represented as $a_{jt} = 1\{\eta_{jt} \leq F_\eta^{-1}(P_j(\boldsymbol{x}_t))\}$. Independence between (ξ_{jt}, η_{jt}) and $\boldsymbol{x}_t$ then implies that:

$$\lambda_j(\boldsymbol{x}_t) \;=\; \int \xi_{jt}\; 1\left\{\eta_{jt} \leq F_\eta^{-1}(P_j\left(\boldsymbol{x}_t\right))\right\}\; \frac{f_{\xi,\eta}\left(\xi_{jt}, \eta_{jt}\right)}{P_j\left(\boldsymbol{x}_t\right)}\; d\xi_{jt}\; d\eta_{jt} \;=\; \rho_j(P_j\left(\boldsymbol{x}_t\right)) \tag{24}$$

Therefore, the demand equation can be represented as:

$$\ln\left(\frac{s_{jt}}{s_{0t}}\right) = \sigma\;\ln\left(\frac{s_{jt}+S_{-jt}}{s_{0t}}\right) + \alpha\; p_{jt} + \boldsymbol{x}_{jt}'\;\boldsymbol{\beta} + \rho_j(P_j\left(\boldsymbol{x}_t\right)) + \widetilde{\xi}_{jt}. \tag{25}$$

The result in equation (25) has important implications for identification and estimation. Regression equation (25) is a standard semiparametric partially linear model with two endogenous regressors, $\ln[(s_{jt}+S_{-jt})/s_{0t}]$ and p_{jt}, and with the nonparametric component $\rho_j(P_j\left(\boldsymbol{x}_t\right))$ only depending on the CCP of firm j. As such, identification and estimation follows a standard two-step procedure. In a first step, we nonparametrically estimate $P_j\left(\boldsymbol{x}_t\right)$ from data on $(a_{jt}, \boldsymbol{x}_t)$. Then, in a second step, one can apply the semiparametric series estimator in Das, Newey, and Vella (2003) and Newey (2009), or the pairwise differencing method in Powell (2001) and Aradillas-Lopez (2012). Valid instruments in this regression are observed $\boldsymbol{x}$ characteristics of products other than j, i.e., the so-called BLP instruments. ■

Under the assumptions of Example 2, the market entry condition has only one unobservable variable, η_{jt}, which, after inverting the profit function, enters additively in the inequality that defines the selection/entry decision. The model becomes a standard semiparametric sample selection model. However, though practically convenient, the conditions in Example 2 are likely to be rejected in many empirical applications.[8] In particular, restriction (1*) — i.e., η_{jt} is the only relevant information that firm j has about demand/cost variables $\boldsymbol{\xi}_t$ when making its entry decision — seems unrealistic. If this restriction does not hold, the estimation approach described above will be inconsistent as it will not control for the correct selection bias function.

EXAMPLE 3 (κ_t has finite support). Consider the model under Assumptions 1-2, including Assumption 1(A) on the finite support of κ_t. Let $P_j(\boldsymbol{x}_t, \kappa_t)$ be the equilibrium probabilities in the entry game in market t. By definition, we have that $P_j(\boldsymbol{x}_t, \kappa_t) \equiv \mathbb{E}(a_{jt}|\boldsymbol{x}_t, \kappa_t)$, and:

$$P_j(\boldsymbol{x}_t, \kappa_t) \;=\; F_\eta\left(\pi_j^{P(-1)}(\boldsymbol{x}_t, \kappa_t)\right). \tag{26}$$

[8]The model of Example 2 is over-identified, allowing for the testability of its over-identifying restrictions.

Similar to Example 2, the model implies a one-to-one relationship between $P_j(\boldsymbol{x}_t, \kappa_t)$ and the inverse expected profit function: i.e., $\pi_j^{P(-1)}(\boldsymbol{x}_t, \kappa_t) = F_\eta^{-1}\left(P_j(\boldsymbol{x}_t, \kappa_t)\right)$. Define $\widetilde{\lambda}_j(\boldsymbol{x}_t, \boldsymbol{\kappa}_t) \equiv \mathbb{E}(\xi_{jt}|\boldsymbol{x}_t, \kappa_t, a_{jt} = 1)$. Applying the one-to-one relationship between CCP and inverse profit function, we have that — conditional on κ_t — the selection function is:

$$\begin{aligned} \widetilde{\lambda}_j(\boldsymbol{x}_t, \kappa_t) &= \int \xi_{jt}\ 1\left\{\eta_{jt} \leq F_\eta^{-1}(P_j(\boldsymbol{x}_t, \kappa_t))\right\}\ \frac{f_{\xi,\eta|\kappa}\left(\xi_{jt}, \eta_{jt}|\kappa_t\right)}{P_j(\boldsymbol{x}_t, \kappa_t)}\ d\left(\xi_{jt},\eta_{jt}\right) \\ &\equiv \psi_j\left(P_j\left(\boldsymbol{x}_t, \kappa_t\right), \kappa_t\right). \end{aligned} \tag{27}$$

However, because κ_t is unobserved to the researcher, we must integrate $\widetilde{\lambda}_j$ over its distribution. By definition, we then have the following relationship between $\widetilde{\lambda}_j$ and the selection bias function $\lambda_j(\boldsymbol{x}_t) \equiv \mathbb{E}(\xi_{jt}|\boldsymbol{x}_t, a_{jt} = 1)$ that appears in the demand equation:

$$\lambda_j(\boldsymbol{x}_t) = \sum_{\kappa_t \in \mathcal{K}(\boldsymbol{x}_t)} \widetilde{\lambda}_j(\boldsymbol{x}_t, \kappa_t)\ f_\kappa\left(\kappa_t|\boldsymbol{x}_t\right). \tag{28}$$

Combining the demand equation with equations (27) and (28), we obtain the regression equation:

$$\ln\left(\frac{s_{jt}}{s_{0t}}\right) = \sigma\ \ln\left(\frac{s_{jt} + S_{-jt}}{s_{0t}}\right) + \alpha\ p_{jt} + \boldsymbol{x}_{jt}'\ \boldsymbol{\beta} + \sum_{\kappa_t \in \mathcal{K}(\boldsymbol{x}_t)} \psi_j\left(P_j\left(\boldsymbol{x}_t, \kappa_t\right), \kappa_t\right) f_\kappa\left(\kappa_t|\boldsymbol{x}_t\right) + \widetilde{\xi}_{jt}. \tag{29}$$

Given the structure of this selection bias function, we obtain identification on the basis of two key properties. First, $P_j\left(\boldsymbol{x}_t, \kappa_t\right)$ and the distribution $f_\kappa\left(\kappa_t|\boldsymbol{x}_t\right)$ are identifiable using data on firms' entry decisions. Second, conditional on $\{P_j\left(\boldsymbol{x}_t, \kappa_t\right), f_\kappa\left(\kappa_t|\boldsymbol{x}_t\right) : \kappa_t \in \mathcal{K}(\boldsymbol{x}_t)\}$, the vector of product characteristics $\boldsymbol{x}_{jt}$ has full-rank. We establish these results in section 4. ■

4 Identification

4.1 Data and sequential identification

Suppose that each of the J firms is a potential entrant in every local market. The researcher observes these firms in a random sample of T markets. For every market t, the researcher observes the vector of exogenous variables $\boldsymbol{x}_t \in \mathcal{X}$ and the vectors of firms' entry decisions $\boldsymbol{a}_t \in \{0, 1\}^J$. The space $\mathcal{X}$ can be discrete or continuous. For those firms active in market t, the researcher observes prices $\boldsymbol{p}_t$ and market shares $\boldsymbol{s}_t$.

Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ be the vector of all the parameters in the model, where $\boldsymbol{\Theta}$ is the parameter space. This vector has infinite dimension because some of the structural parameters are real-valued

functions. The vector $\boldsymbol{\theta}$ has the following components: demand parameters $\boldsymbol{\theta}_\delta \equiv (\alpha, \boldsymbol{\beta}, \boldsymbol{\sigma})$; probability distribution of demand/cost signals, $\boldsymbol{f}_\kappa \equiv (f_\kappa(\kappa|\boldsymbol{x}) : \text{for every } \kappa, \boldsymbol{x})$; equilibrium entry probabilities, $\boldsymbol{P}_\kappa \equiv (P_j(\boldsymbol{x}, \kappa) : \text{for every } j, \boldsymbol{x}, \kappa)$; the probability distribution of private information F_η, and the distribution of unobserved demand conditional on signals, $f_{\xi|\eta,\kappa}$.

$$\boldsymbol{\theta} \equiv \left(\boldsymbol{\theta}_\delta,\ \boldsymbol{P}_\kappa,\ \boldsymbol{f}_\kappa,\ f_{\xi|\eta,\kappa},\ F_\eta \right). \tag{30}$$

In this paper, we are interested in the identification of demand parameters $\boldsymbol{\theta}_\delta$ when the distributions $\boldsymbol{f}_\kappa$ and $f_{\xi|\eta,\kappa}$ and the equilibrium choice probabilities $\boldsymbol{P}_\kappa$ are nonparametrically specified.

We consider a two-step sequential procedure for the identification of $\boldsymbol{\theta}_\delta$. First, given the empirical distribution of firms' entry decisions, we establish the identification of the equilibrium probabilities $\boldsymbol{P}_\kappa$ and the distribution f_κ. Then, given the structure of the selection bias function in (29), we show the identification of $\boldsymbol{\theta}_\delta$.

4.2 First step: Game of market entry

The identification of the equilibrium probabilities $\boldsymbol{P}_\kappa$ and the distribution $\boldsymbol{f}_\kappa$ is based on the structure of the joint probability distribution of the entry decisions of the J firms. For any value $(\boldsymbol{a}, \boldsymbol{x}) \in \{0,1\}^J \times \mathcal{X}$:

$$\Pr\left(\boldsymbol{a}_t = \boldsymbol{a} | \boldsymbol{x}_t = \boldsymbol{x}\right) = \sum_{\kappa=1}^{|\mathcal{K}(\boldsymbol{x})|} f_\kappa(\kappa|\boldsymbol{x}) \left[\prod_{j=1}^{J} \left[P_j(\boldsymbol{x}, \kappa)\right]^{a_j} \left[1 - P_j(\boldsymbol{x}, \kappa)\right]^{1-a_{jt}} \right] \tag{31}$$

This system of equations describes a nonparametric finite mixture model. The identification of this class of models has been studied by Hall and Zhou (2003), Hall, Neeman, Pakyari, and Elmore (2005), Allman, Matias, and Rhodes (2009), and Kasahara and Shimotsu (2014), among others. Identification is based on the assumption of independence between firms' entry decisions once we condition on $\boldsymbol{x}_t$ and κ_t.

In this first step, the proof of identification is pointwise for each value of $\boldsymbol{x}$. To simplify notation, for the rest of this subsection we omit any further reference to $\boldsymbol{x}$ and to the market subscript t .

4.2.1 Identification of the number of latent market types

The number of components $|\mathcal{K}|$ in finite mixture (31) is typically unknown to the researcher. Following ideas similar to Bonhomme, Jochmans, and Robin (2016), Xiao (2018), and Aguirregabiria and Mira (2019), we start our first step identification argument by providing sufficient

conditions for the unique determination of $|\mathcal{K}|$ from observables. In particular, we adapt to our context Proposition 2 in Aguirregabiria and Mira (2019) and Lemma 1 in Xiao (2018).

Suppose that $J \geq 3$ and let (Y_1, Y_2, Y_3) be three random variables that represent a partition of the vector of firms' entry decisions $(a_1, a_2, ..., a_J)$ such that Y_1 is equal to the entry decision of one firm (if J is odd) or two firms (if J is even), and variables Y_2 and Y_3 evenly divide the entry decisions of the rest of the firms. Denote by $\tilde{J}$ the number of firms collected in Y_i, $i = 2, 3$, such that $\tilde{J} = (J-1)/2$ if J is odd, and $\tilde{J} = (J-2)/2$ if J is even. For $i = 1, 2, 3$, let $\boldsymbol{P}_{Y_i}(\kappa)$ be the matrix of probabilities for each possible value of Y_i — in the rows of the matrix — conditional on every possible value of κ — in the columns of the matrix. The main idea is then to identify the number of components $|\mathcal{K}|$ from the observed joint distribution of Y_2 and Y_3:

$$\Pr(Y_2 = y_2, Y_3 = y_3) \;=\; \sum_{\kappa=1}^{|\mathcal{K}|} \Pr(Y_2 = y_2|\kappa)\;\Pr(Y_3 = y_3|\kappa)\;f_\kappa(\kappa) \tag{32}$$

or, in matrix notation,

$$\boldsymbol{P}_{Y_2,Y_3} \;=\; \boldsymbol{P}_{Y_2|\kappa}\; diag(\boldsymbol{f}_\kappa)\; \boldsymbol{P}'_{Y_3|\kappa}, \tag{33}$$

where: $\boldsymbol{P}_{Y_2,Y_3}$ is the $2^{\tilde{J}} \times 2^{\tilde{J}}$ matrix with elements $P(y_2, y_3)$; $\boldsymbol{P}_{Y_i|\kappa}$ is the $2^{\tilde{J}} \times |\mathcal{K}|$ matrix with elements $\Pr(Y_i = y|\kappa)$; and $diag(\boldsymbol{f}_\kappa)$ is the $|\mathcal{K}| \times |\mathcal{K}|$ diagonal matrix with the probabilities $f_\kappa(\kappa)$.

LEMMA 2. *Without further restrictions, Rank*$(\boldsymbol{P}_{Y_2,Y_3})$ *is a lower bound for the true value of parameter* $|\mathcal{K}|$. *Furthermore, if (i)* $|\mathcal{K}| < 2^{\tilde{J}}$ *and (ii) for* $i = 2, 3$ *the* $|\mathcal{K}|$ *vectors* $\boldsymbol{P}_{Y_i}(\kappa = 1)$, $\boldsymbol{P}_{Y_i}(\kappa = 2)$, ..., $\boldsymbol{P}_{Y_i}(\kappa = |\mathcal{K}|)$ *are linearly independent, then* $|\mathcal{K}| = Rank(\boldsymbol{P}_{Y_2,Y_3})$. ■

The point identification of the number of components $|\mathcal{K}|$ from the observed matrix $\boldsymbol{P}_{Y_2,Y_3}$ hinges on a "large enough" number of firms $\tilde{J}$ and on the matrices $\boldsymbol{P}_{Y_2|\kappa}$ and $\boldsymbol{P}_{Y_3|\kappa}$ being of full column rank, so that the entry probabilities associated to each component κ cannot be obtained as linear combinations of the others.

4.2.2 Identification of equilibrium CCPs and distribution of latent types

Allman, Matias, and Rhodes (2009) study the identification of nonparametric multinomial finite mixtures that include our binary choice model as a particular case. They establish that a mixture with $|\mathcal{K}|$ components is identified if $J \geq 3$ and $|\mathcal{K}| \leq 2^J/(J+1)$. The following Lemma 3 is an application to our model of Theorem 4 and Corollary 5 by Allman, Matias, and Rhodes (2009).

LEMMA 3. *Suppose that: (i)* $J \geq 3$; (ii) $|\mathcal{K}| \leq 2^J/(J+1)$; *and (iii) for* $i = 1, 2, 3$, *the* $|\mathcal{K}|$ *vectors* $\boldsymbol{P}_{Y_i}(\kappa = 1)$, $\boldsymbol{P}_{Y_i}(\kappa = 2)$, ..., $\boldsymbol{P}_{Y_i}(\kappa = |\mathcal{K}|)$ *are linearly independent. Then, the probability*

distribution of κ *— i.e.,* $f_\kappa(\kappa)$ *for* $\kappa = 1, 2, ..., |\mathcal{K}|$ *— and the equilibrium CCPs — i.e.,* $P_j(\kappa)$ *for* $j = 1, 2, ..., J$ *and* $\kappa = 1, 2, ..., |\mathcal{K}|$ *— are uniquely identified up to label swapping.* ■

Note that the order condition (i) in Lemma 2 is in general more stringent than the order condition (ii) of Lemma 3: that is, for $J \geq 3$, we have that $2^{\tilde{J}} \leq 2^J/(J+1)$. In this sense, for any $J \geq 3$, when the conditions in Lemma 2 hold and the $|\mathcal{K}|$ vectors $\boldsymbol{P}_{Y_1}(\kappa = 1)$, $\boldsymbol{P}_{Y_1}(\kappa = 2)$, ..., $\boldsymbol{P}_{Y_1}(\kappa = |\mathcal{K}|)$ are linearly independent, then $|\mathcal{K}| = \text{Rank}(\boldsymbol{P}_{Y_2,Y_3})$ and the distribution of κ and the equilibrium CCPs are uniquely identified.

The identification of the distribution of κ_t and the equilibrium CCPs is up to label swapping, and "pointwise" or separately for each value of the observable $\boldsymbol{x}_t$. In the absence of additional assumptions, the combination of these two features leads to an identification problem in the estimation of demand parameters in the second step of our method. In the second step, we need to include $f_\kappa(\kappa|\boldsymbol{x}_t)$ and $P_j(\kappa, \boldsymbol{x}_t)$ for every value of κ as additional regressors, or more precisely as control variables, in the estimation of the demand equation. To construct these regressors, we need to be able to "match" the same latent type κ across the different observed values of $\boldsymbol{x}_t$ in the sample. However, this task is not feasible without further assumptions.

In the work by Aguirregabiria and Mira (2019), the authors discuss alternative assumptions that can solve this matching-latent-types problem. In our empirical application, we opt for the independence between κ_t and $\boldsymbol{x}_t$. This assumption addresses the challenge associated with matching latent types. It does so by altering the nature of identification in the first step: rather than being "pointwise" over $\boldsymbol{x}_t$, identification now holds uniformly across all values of the variable $\boldsymbol{x}$. Therefore, though identification is still up to label swapping, we have the same label κ over all values of $\boldsymbol{x}_t$ such that there is not a problem about matching latent types.

4.3 Second Step: Identification of Demand Parameters

Following the discussion in section 2.1, we represent the demand system using the inverse $d_j^{(\boldsymbol{a})-1}(\boldsymbol{s}_t^{(\boldsymbol{a})}, \boldsymbol{p}_t^{(\boldsymbol{a})}, \boldsymbol{x}_t^{(\boldsymbol{a})})$ from Lemma 1. For those markets with $a_{jt} = 1$, the demand equation can be expressed as:

$$\delta_j(\boldsymbol{s}_t, \boldsymbol{\sigma}) \;=\; \alpha\, p_{jt} + \boldsymbol{x}'_{jt}\, \boldsymbol{\beta} + \xi_{jt}, \qquad \text{for } a_{jt} = 1 \tag{34}$$

where we use the notation $\delta_j(\boldsymbol{s}_t, \boldsymbol{\sigma})$ to emphasize that δ_{jt} is a function of the parameters $\boldsymbol{\sigma}$ characterizing the distribution of the random coefficients υ_h. The selection problem arises because the unobservable ξ_{jt} is not mean independent of the market entry (or product availability) condition $a_{jt} = 1$. Therefore, moment conditions that are valid under exogenous product selection are no longer valid when ξ_{jt} and a_{jt} are not independent.

Suppose for a moment that the market type κ_t were observable to the researcher after identification in the first step. In this case, the selection term would be $\psi_j\left(P_j(\boldsymbol{x}_t, \kappa_t), \kappa_t\right)$ from equation (27) and we would have — as in Example 2 — a relatively standard selection problem represented by the semiparametric partially linear model:

$$\delta_j(\boldsymbol{s}_t, \boldsymbol{\sigma}) = \alpha\ p_{jt} + \boldsymbol{x}'_{jt}\ \boldsymbol{\beta} + \psi_j\left(P_j(\boldsymbol{x}_t, \kappa_t), \kappa_t\right) + \widetilde{\xi}_{jt}. \tag{35}$$

A key complication of the selection problem in our model is that the market type κ_t is unobserved to the researcher. After the first step of the identification procedure, we do not know the unobserved type of a market but only its probability distribution conditional on $\boldsymbol{x}_t$. Therefore, in the second step, we cannot condition on κ_t as in equation (35). We instead need to deal with the more complex selection bias function:

$$\lambda_j(\boldsymbol{x}_t) \equiv \mathbb{E}\left(\xi_{jt} | \boldsymbol{x}_t, a_{jt} = 1\right) = \sum_{\kappa=1}^{|\mathcal{K}(\boldsymbol{x}_t)|} f_\kappa(\kappa|\boldsymbol{x}_t)\ \psi_j\left(P_j(\boldsymbol{x}_t, \kappa), \kappa\right) \ = \ \boldsymbol{f}'_{\kappa,t}\ \boldsymbol{\psi}_j(\boldsymbol{P}_{j,t}), \tag{36}$$

where $\boldsymbol{f}_{\kappa,t}$, $\boldsymbol{P}_{j,t}$, and $\boldsymbol{\psi}_j(\boldsymbol{P}_{j,t})$ are all vectors of dimension $|\mathcal{K}(\boldsymbol{x}_t)| \times 1$ such that $\boldsymbol{f}_{\kappa,t} \equiv (f_\kappa(\kappa|\boldsymbol{x}_t) : \kappa = 1, 2, ..., |\mathcal{K}(\boldsymbol{x}_t)|)$, $\boldsymbol{P}_{j,t} \equiv (P_j(\boldsymbol{x}_t, \kappa) : \kappa = 1, 2, ..., |\mathcal{K}(\boldsymbol{x}_t)|)$, and $\boldsymbol{\psi}_j(\boldsymbol{P}_{j,t}) \equiv (\psi_j\left(P_j(\boldsymbol{x}_t, \kappa), \kappa\right) : \kappa = 1, 2, ..., |\mathcal{K}(\boldsymbol{x}_t)|)$. Therefore, the regression equation of our model is:

$$\delta_j(\boldsymbol{s}_t, \boldsymbol{\sigma}) = \alpha\ p_{jt} + \boldsymbol{x}'_{jt}\ \boldsymbol{\beta} + \boldsymbol{f}'_{\kappa,t}\ \boldsymbol{\psi}_j(\boldsymbol{P}_{j,t}) + \widetilde{\xi}_{jt}. \tag{37}$$

where $\mathbb{E}\left(\widetilde{\xi}_{jt} | \boldsymbol{x}_t\right) = \mathbb{E}\left(\widetilde{\xi}_{jt} | \boldsymbol{P}_{j,t}, \boldsymbol{f}_{\kappa,t}\right) = 0$. This equation establishes that $(\boldsymbol{P}_{j,t}, \boldsymbol{f}_{\kappa,t})$ is a sufficient statistic for the selection bias function.

Proposition 1 establishes a necessary and sufficient condition for the identification of $\boldsymbol{\theta}_\delta \equiv (\alpha, \boldsymbol{\beta}, \boldsymbol{\sigma})$ from equation (37). It is an application of Theorem 6 in Rothenberg (1971).

PROPOSITION 1. *Define the vector* $\boldsymbol{Z}_{jt} \equiv \left(\mathbb{E}\left(\frac{\partial \delta_j(\boldsymbol{s}_t, \boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} \mid \boldsymbol{x}_t\right),\ \mathbb{E}\left(p_{jt} \mid \boldsymbol{x}_t\right),\ \boldsymbol{x}'_{jt}\right)'$, *and let* $\widetilde{\boldsymbol{Z}}_{jt}$ *be the deviation (or residual)* $\boldsymbol{Z}_{jt} - \mathbb{E}(\boldsymbol{Z}_{jt} \mid \boldsymbol{P}_{j,t}, \boldsymbol{f}_{\kappa,t})$. *Then, given that* $\mathbb{E}\left(\widetilde{\xi}_{jt} | \boldsymbol{x}_t\right) = \mathbb{E}\left(\widetilde{\xi}_{jt} | \boldsymbol{P}_{j,t}, \boldsymbol{f}_{\kappa,t}\right) = 0$, *a necessary and sufficient condition for the identification of* $\boldsymbol{\theta}_\delta \equiv (\alpha, \boldsymbol{\beta}, \boldsymbol{\sigma})$ *in equation (37) is that matrix* $\mathbb{E}\left(\widetilde{\boldsymbol{Z}}_{jt}\ \widetilde{\boldsymbol{Z}}'_{jt}\right)$ *is full-rank.* ■

Intuitively, Proposition 1 says that the identification of $\boldsymbol{\theta}_\delta$ requires that, after differencing out any dependence with respect to $(\boldsymbol{P}_{j,t}, \boldsymbol{f}_{\kappa,t})$, there should be no perfect collinearity in the vector of explanatory variables $\boldsymbol{Z}_{jt} \equiv (\mathbb{E}(\partial \delta_{jt}/\partial \boldsymbol{\sigma} \mid \boldsymbol{x}_t),\ \mathbb{E}(p_{jt} \mid \boldsymbol{x}_t),\ \boldsymbol{x}'_{jt})'$.

Proposition 1 does not provide identification conditions that apply directly to the primitives

of the model. However, on the basis of this Proposition, it is straightforward to establish necessary identification conditions that apply to primitives of the model, or to objects which are more closely related to primitives. First, we need $J \geq 2$, otherwise there would not be exclusion restrictions to deal with price endogeneity, i.e., $\mathbb{E}(p_{jt} \mid \boldsymbol{x}_t)$ would be a linear combination of $\boldsymbol{x}_{jt}$. Second, the vector of entry probabilities $\boldsymbol{P}_{j,t}$ should depend on $\boldsymbol{x}_{it}$ for $i \neq j$. Otherwise, keeping $\boldsymbol{P}_{j,t}$ fixed would also imply fixing $\boldsymbol{x}_{jt}$ and the vector of parameters $\boldsymbol{\beta}$ would not be identified. Hence, there should be effective competition in firms' market entry decisions. For instance, in the absence of observable variables affecting entry but not demand, the model would not be identified under monopolistic competition. Third, the number of points in the support of κ should be smaller than the number of variables in vector $\boldsymbol{x}_t$: i.e., $|\mathcal{K}(\boldsymbol{x}_t)| < \dim(\boldsymbol{x}_t)$. Otherwise, controlling for $\boldsymbol{P}_{j,t}$ would be equivalent to controlling for the whole vector $\boldsymbol{x}_t$, and no parameter in $\boldsymbol{\theta}_\delta$ would be identified.

5 Estimation and inference

In this section, we present a two-step estimation method that mimics our sequential identification result. In the first step, we use a nonparametric sieve maximum likelihood method to estimate the distribution of unobserved market types, the vector of entry probabilities for each unobserved type, and the number of market types. In the second step, we use sieves to approximate the selection bias term as a function of the densities and entry probabilities estimated in the first step.[9] Then, we apply GMM to jointly estimate the coefficients in the sieve approximation and the structural demand parameters. We calculate asymptotic standard errors of the estimates in the second step using the method and formulas in Newey (2009).

5.1 First step: Estimation of CCPs and distribution of latent types

We use sieves to approximate the nonparametric functions $f_\kappa(\kappa_t|\boldsymbol{x}_t)$ and $P_j(\boldsymbol{x}_t, \kappa_t)$ (Hirano, Imbens, and Ridder, 2003, Chen, 2007). Let $\boldsymbol{r}_t^f \equiv \left(r_1^f(\boldsymbol{x_t}), r_2^f(\boldsymbol{x_t}), ..., r_{L_f}^f(\boldsymbol{x_t})\right)'$ be a vector with a finite number L_f of basis functions. The density function $f_\kappa(\kappa_t|\boldsymbol{x}_t)$ has the following sieves multinomial logit structure:

$$f_\kappa(\kappa|\boldsymbol{x}_t) \;=\; \frac{\exp\{\boldsymbol{r}_t^{f\prime}\ \boldsymbol{\gamma}_\kappa^f\}}{\sum_{\kappa'=1}^{K}\exp\{\boldsymbol{r}_t^{f\prime}\ \boldsymbol{\gamma}_{\kappa'}^f\}} \tag{38}$$

[9] The second step could alternatively be based on differencing out the selection bias term using a matching estimator as in Ahn and Powell (1993), Powell (2001), and Aradillas-Lopez, Honoré, and Powell (2007).

where, for $\kappa = 1, 2, ..., |\mathcal{K}|$, $\boldsymbol{\gamma}_{\kappa}^{f}$ is a vector of parameters with dimension $L_f \times 1$ and normalization $\boldsymbol{\gamma}^{f}(1) = 0$. Similarly, let $\boldsymbol{r}_t^P \equiv \left(r_1^P(\boldsymbol{x_t}), r_2^P(\boldsymbol{x_t}), ..., r_{L_P}^P(\boldsymbol{x_t})\right)'$ be a vector with a finite number L_P of basis functions. For any product j and any unobserved type κ, the entry probability function $P_j(\boldsymbol{x}_t, \kappa)$ has the following sieves binary Logit structure:

$$P_j(\boldsymbol{x}_t, \kappa) \;=\; \Lambda\left(\boldsymbol{r}_t^{P\prime}\, \boldsymbol{\gamma}_{j\kappa}^{P}\right) \tag{39}$$

where $\Lambda(\cdot)$ is the Logistic function. For $j = 1, 2, ..., J$ and $\kappa = 1, 2, ..., |\mathcal{K}|$, we have that $\boldsymbol{\gamma}_{j\kappa}^{P}$ is a vector of parameters of dimension $L_P \times 1$. The log-likelihood function of this *nonparametric finite mixture model* is:

$$\ell(\boldsymbol{\gamma}^{f,P}) \;=\; \sum_{t=1}^{T} \ln\left(\sum_{\kappa=1}^{|\mathcal{K}|} f_{\kappa}(\kappa|\boldsymbol{x}_t, \boldsymbol{\gamma}^{f})\; \prod_{j=1}^{J} \Lambda\left(\boldsymbol{r}_t^{P\prime}\, \boldsymbol{\gamma}_{j\kappa}^{P}\right)^{a_{jt}} \left[1 - \Lambda\left(\boldsymbol{r}_t^{P\prime}\, \boldsymbol{\gamma}_{j\kappa}^{P}\right)\right]^{1-a_{jt}}\right) \tag{40}$$

where $\boldsymbol{\gamma}^{f,P}$ is a vector with the parameters $\{\boldsymbol{\gamma}_{\kappa}^{f}, \boldsymbol{\gamma}_{j\kappa}^{P} : \kappa = 1, 2, ..., K; j = 1, 2, ..., J\}$, with a total of $L_f(|\mathcal{K}| - 1) + L_P|\mathcal{K}|J$ parameters.

We estimate the vector of parameters $\boldsymbol{\gamma}$ by Maximum Likelihood (MLE) using the EM algorithm (Pilla and Lindsay, 2001). Recent papers considering MLE and the EM algorithm to estimate nonparametric mixtures in discrete choice models include Bunting (2022), Bunting, Diegert, and Maurel (2022), Hu and Xin (2022), and Williams (2020). We use Bayesian Information Criterion (BIC) to determine the number of mixtures $|\mathcal{K}|$.

When $\boldsymbol{x}_t$ is discrete, the nonparametric MLE is $\sqrt{T}$-consistent and asymptotically normal. With continuous variables in $\boldsymbol{x}_t$, the nonparametric sieve MLE cannot achieve a $\sqrt{T}$ rate. However, under standard regularity conditions, this does not affect the $\sqrt{T}$-consistency and asymptotic normality of the estimator of the demand parameters in the second step. The proof of this result follows from Hirano, Imbens, and Ridder (2003) and Das, Newey, and Vella (2003).

5.2 Second step: Estimation of demand parameters

Following Das, Newey, and Vella (2003), we use the method of sieves and approximate each function $\psi_j\left(P_j(\boldsymbol{x}_t, \kappa), \kappa\right)$ using a polynomial of order L_ψ in the logarithm of the entry probability $P_j(\boldsymbol{x}_t, \kappa)$

$$\begin{aligned} \psi_j\left(P_j(\boldsymbol{x}_t, \kappa), \kappa\right) \;&\approx\; \boldsymbol{r}^{\psi}\left(P_j(\boldsymbol{x}_t, \kappa)\right)' \boldsymbol{\gamma}_{j\kappa}^{\psi} \\ &=\; \left[1,\;\; \ln P_j(\boldsymbol{x}_t, \kappa),\;\; \ln P_j(\boldsymbol{x}_t, \kappa)^2,\;\; ...,\;\; \ln P_j(\boldsymbol{x}_t, \kappa)^{L_\psi}\right] \boldsymbol{\gamma}_{j\kappa}^{\psi} \end{aligned} \tag{41}$$

where $\boldsymbol{\gamma}^{\psi}_{j\kappa} \equiv (\gamma^{\psi}_{0,j\kappa}, \gamma^{\psi}_{1,j\kappa}, ..., \gamma^{\psi}_{L_\psi,j\kappa})'$ is a vector of parameters. Given this approximation, the selection function is linear in $\boldsymbol{\gamma}^{\psi}_{j\kappa}$ and has the following expression:

$$\boldsymbol{f}'_{\kappa,t}\boldsymbol{\psi}_j(\boldsymbol{P}_{j,t}) \;\approx\; \boldsymbol{h}'_{j,t}\,\boldsymbol{\gamma}^{\psi}_j \;=\; \sum_{\kappa=1}^{|\mathcal{K}|}\sum_{\ell=0}^{L_\psi} \gamma^{\psi}_{j,\ell}(\kappa)\; f_\kappa(\kappa|\boldsymbol{x}_t)\;\left[\ln P_j(\boldsymbol{x}_t,\kappa)\right]^\ell \tag{42}$$

where $\boldsymbol{h}'_{j,t}$ is a vector with dimension $1 \times (L_\psi + 1)|\mathcal{K}|$ and elements $\{f_\kappa(\kappa|\boldsymbol{x}_t)\;\left[\ln P_j(\boldsymbol{x}_t,\kappa)\right]^\ell : \ell = 0, 1, ..., L_\psi; \kappa = 1, ..., |\mathcal{K}|\}$, and $\boldsymbol{\gamma}^{\psi}_j$ is a vector of parameters with the same dimension and with elements $\{\gamma^{\psi}_{j,\ell}(\kappa) : \ell = 0, 1, ..., L_\psi; \kappa = 1, ..., |\mathcal{K}|\}$.

Plugging equation (42) into the demand equation (37), we have the regression equation:

$$\delta_j(\boldsymbol{s}_t, \boldsymbol{\sigma}) = \alpha\; p_{jt} + \boldsymbol{x}'_{jt}\;\boldsymbol{\beta} + \boldsymbol{h}'_{j,t}\;\boldsymbol{\gamma}^{\psi}_j + \widetilde{\xi}_{jt}. \tag{43}$$

Equation (43) can be estimated by GMM. Following Das, Newey, and Vella (2003), one can show that this two-step estimator of the vector of demand parameters $\boldsymbol{\theta}_\delta$ is $\sqrt{T}$-consistent and asymptotically normal.

6 Empirical application

6.1 Data and descriptive statistics

We apply our method to estimate demand in the US airline industry. The challenge of accounting for endogenous product entry in demand estimation in this industry has recently been explored by Ciliberto, Murry, and Tamer (2021) and Li, Mazur, Park, Roberts, Sweeting, and Zhang (2022).

We use publicly available data from the US Department of Transportation for our analysis. Our working sample consists of data from the DB1B and T100 databases. Specifically, we use quarterly data spanning from 2012-Q1 to 2013-Q4 for routes between the airports at the 100 largest Metropolitan Statistical Areas (MSA) in the United States. This accounts for 108 airports, as there are a few MSAs with more than one airport.

In terms of airlines' entry decisions, we define a market as a non-directional airport-pair, where, for example, Chicago O'Hare (ORD) to New York La Guardia (LGA) is considered equivalent to LGA to ORD. There are potentially $5,778$ non-directional markets between the 108 airports, i.e., $108 * 107/2$. However, many of these markets have never had an incumbent airline with non-stop flights for several decades. These are typically airport pairs that are geographically too close or in smaller MSAs. In our sample, we consider only non-directional

markets with at least one incumbent airline in one year over the four decades of data from the USDOT database. This accounts for 2,230 non-directional markets in our sample, and 17,541 market-quarter observations. We say that an airline is an *entrant* in a non-directional airport pair if it operates non-stop flights between the two airports.

A product is defined as the combination of directional airport-pair, airline, and the dummy for non-stop flight. For example, an American Airlines non-stop flight from LGA to ORD is a product. The airlines included in our analysis are American (AA), Delta (DL), United (UA), US Airways (US), Southwest (WN), a combined group of Low-Cost Carriers (LCC), and a combined group of the remaining carriers (*Others*).[10]

Following the empirical literature on the airline industry, we define the *hub-size* of an airline in an airport as the number of nonstop routes that the airline operates from that airport. For the construction of market shares in the estimation of demand, we define *market size* as the geometric mean of the populations in the metropolitan regions of the two airports that define the route.

Table 1 presents the distribution of the number of entrants and the average value of market characteristics. Notably, in a significant portion of these markets (32%), there are no airlines providing non-stop flights, and they are exclusively served with stop flights. Among the markets with non-stop flights, more than 90% are monopolies or duopolies. Furthermore, there is a strong positive correlation between the number of incumbents and market size and distance.

Table 1: Distribution of Markets by Number of Entrants

Number of airlines	Frequency # markets-quarters (%)	Avg. market size in million people	Avg. market distance in miles
0 airlines	5,583 (31.83%)	2.88	737
1 airline	8,204 (46.77%)	3.42	916
2 airlines	2,614 (14.90%)	4.44	955
3 airlines	844 (4.81%)	5.36	1,112
4 airlines	221 (1.26%)	5.44	1,136
5 airlines	68 (0.39%)	8.61	1,185
$\geq$ 6 airlines	7 (0.04%)	6.95	314
Total	17,541 (100.00%)	3.54	881

Table 2 presents entry frequencies for each airline and the average market size and distance

[10]Following Ciliberto, Murry, and Tamer (2021), the list of airlines included in the group LCC is: Alaska, JetBlue, Frontier, Allegiant, Spirit, Sun Country, and Virgin. The carriers in the group *Others* are small regional carriers, charters, and private jets.

associated with their entry. We observe significant variation in airlines' entry probabilities, with WN and AA having the highest (26.2%) and the lowest (10.3%) entry propensities, respectively. Furthermore, substantial heterogeneity exists across airlines concerning the correlations between entry and market size and distance. For example, while WN enters markets that are not significantly different in size from the markets it does not enter (3.63 million people versus 3.54 million people), AA tends to enter markets with much larger average size (5.32 million people versus 3.54 million people). Moreover, different entry strategies are evident based on market distance. DL and US typically enter markets with an average distance of around 876 miles, whereas the markets served by LCC have an average distance of 1,171 miles.

Table 2: Entry Frequency by Airline

Airline	Frequency # markets-quarters (%)	Avg. market size in million people	Avg. market distance in miles
WN	4,602 (26.23%)	3.63	981
DL	3,257 (18.56%)	4.07	876
UA	3,221 (18.36%)	4.50	968
LCC	2,382 (13.57%)	4.61	1,171
US	1,933 (11.02%)	3.98	879
AA	1,815 (10.34%)	5.32	962

6.2 Estimation of the model for market entry

For the entry decisions, we consider the nonparametric sieve finite mixture Logit described in equations (38) and (39). The vector of explanatory variables in $\boldsymbol{x}_t$ includes: market size (`msize`), as measured by the sum of populations in the MSAs of the two airports; market distance (`mktdistance`), as the geodesic distance between the two airports; the airline's own hub-size in the market (`ownhub-size`), as measured by the sum of the airline's hub-size in the two airports; the average hub-size of the other airlines (`comphub-size`); and time dummies for each of the eight quarters in the sample.

We have estimated various versions of the Logit entry model based on two key factors: the polynomial order in $\boldsymbol{x}_t$ used to construct the basis $\boldsymbol{r}_t^P$ and the number of points in the support of κ_t. Notably, the parameters of the entry model are specific to each airline and are unrestricted across airlines. Our estimation results for demand parameters are robust in relation to the selection of the basis $\boldsymbol{r}_t^P$ in the entry model. For brevity, we then present results only for the specification with $\boldsymbol{r}_t^P = \boldsymbol{x}_t$. Regarding the selection of the number of unobserved

market types $|\mathcal{K}|$ and its influence on the estimation of demand parameters, as illustrated in Table 4 below, the primary and most substantial effect arises from allowing for *some* unobserved market heterogeneity κ_t. Specifically, transitioning from a model without κ_t (i.e., one unobserved market type) to a model with two unobserved market types induces a noteworthy impact. Once we account for this type of unobserved market heterogeneity, demand estimates are however very similar among specifications that allow for two, three, or four unobserved market types.[11]

In this section, we present the estimation results for mixture Logit models, focusing on cases where the distribution of κ_t is independent of $\boldsymbol{x}_t$. Our choice stems from the robustness of our demand estimates to incorporating dependence between κ_t and $\boldsymbol{x}_t$, and because this independence assumption effectively addresses the identification challenge of matching latent market types due to label swapping (see discussion in Section 4.2.2). The robustness of our results to relaxing independence between κ_t and $\boldsymbol{x}_t$ is intuitive, as this assumption does not impose any exclusion restriction necessary to control for selection bias in the estimation of the demand parameters.

Furthermore, there is a practical computational rationale behind this decision. While estimating the mixture Logit model under the assumption of independence and with two or three unobserved market types requires only a few hours, and the Expectation-Maximization (EM) algorithm consistently converges, introducing dependence significantly complicates computations. The estimation process, even with just two unobserved market types, extends over several days of EM iterations, and the EM algorithm often fails to converge. While the model with dependence is theoretically identified, the practical implementation of the estimator considerably complicates in our empirical application.

Table 3 presents the goodness-of-fit statistics obtained from estimating four nested specifications of the market entry model. The selection of the preferred model is guided by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), alongside other crucial considerations such as the convergence performance of the EM algorithm, the accuracy of parameter estimates of the entry model, and the robustness of the demand parameter estimates.

The introduction of unobserved market heterogeneity κ_t results in a significant enhancement in the model's goodness-of-fit. This is demonstrated by a substantial increase in the log-likelihood and a decrease in both AIC and BIC when comparing the model without κ_t and the model with two unobserved market types. This form of unobserved market heterogeneity captures a strong correlation among airlines' entry decisions, a correlation beyond the explanatory power of the observable market and airline characteristics in the vector $\boldsymbol{x}_t$.

[11] To simplify exposition, the estimation results for the model with four unobserved market types are not reported in Table 4 but are qualitatively similar to those for models with two or three unobserved market types.

The inclusion of additional unobserved market types continues to positively impact the goodness of fit. However, this impact shows diminishing returns, with the improvement being negligible when we move from three to four unobserved market types. While the EM algorithm converges rapidly to the MLE in the specifications with two and three unobserved market types, we experience convergence problems in the model with four unobserved market types. In this case, we obtain imprecise estimates for some of the parameters in the entry model. These factors, combined with the marginal improvement observed in the AIC and BIC criteria as we transition from three to four unobserved market types, lead us to favor the model featuring $|\mathcal{K}| = 3$ unobserved market types.

Table 3: Estimation of Market Entry Model — Goodness-of-Fit Statistics

Statistics	**Logit** **# types = 1**	**Mixture Logit** **# types = 2**	**Mixture Logit** **# types = 3**	**Mixture Logit** **# types = 4**
Observations	17,155	17,155	17,155	17,155
Parameters	72	145	218	287
Log-likelihood	−20,378	−18,985	−18,022	−17,621
AIC	40,900	38,261	36,481	35,817
BIC	41,458	39,385	38,170	38,041

6.3 Estimation of demand parameters

For the demand system, we follow Ciliberto, Murry, and Tamer (2021) and estimate a nested logit demand with two nests: a nest for all the airlines and another nest for the outside alternative.

$$\ln\left(\frac{s_{jt}}{s_{0t}}\right) = \alpha\, p_{jt} + \boldsymbol{x}'_{jt}\boldsymbol{\beta} + \sigma\, \ln\left(\frac{s_{jt}}{1-s_{0t}}\right) + \boldsymbol{h}'_{jt}\,\boldsymbol{\gamma}^{\psi}_{j} + \widetilde{\xi}_{jt} \tag{44}$$

In this equation, the vector of product characteristics $\boldsymbol{x}_{jt}$ includes *mktdistance*, square-*mktdistance*, airline j's *hub-size* in the origin airport, airline j's *hub-size* in the destination airport, and airline × quarter fixed effects (dummies). The expression for the selection bias term, $\boldsymbol{h}'_{jt}\,\boldsymbol{\gamma}^{\psi}_{j}$, varies across the specifications of the market entry model, from the more restrictive parametric Logit model to the more general semiparametric finite mixture Logit model.

1. *Parametric Logit* specification. We consider the entry model $a_{jt} = 1\{\eta_{jt} \leq \boldsymbol{x}'_{jt}\boldsymbol{\gamma}^{P}_{j}\}$, with $\eta_{jt} \sim$ *Logistic*, and $\xi_{jt} = \gamma^{\psi}_{j,1}\,\eta_{jt} + v_{jt}$, with v_{jt} independent of η_{jt} and $\boldsymbol{x}_t$. Under this

parametric Logit selection model, the selection term has the following form:

$$\mathbb{E}\left(\xi_{jt} \mid a_{jt}=1, \boldsymbol{x}_t\right) \;=\; \gamma_{j,1}^{\psi}\, \mathbb{E}\left(\eta_{jt} \mid \eta_{jt} \leq \boldsymbol{x}_{jt}'\boldsymbol{\gamma}_j^P\right) \;=\; \gamma_{j,1}^{\psi}\, \left[Euler - \ln\Lambda\left(\boldsymbol{x}_t'\,\boldsymbol{\gamma}_j^P\right)\right] \quad (45)$$

where *Euler* represents *Euler's constant* ≈ 0.5772. For the parametric Logit model, the term $Euler - \ln\Lambda\left(\boldsymbol{x}_t'\boldsymbol{\gamma}_j^P\right)$ is analogous to the inverse Mills ratio in the context of the parametric Probit model.

2. *Semiparametric Logit without* κ_t. The entry model is still the Logit $a_{jt} = 1\{\eta_{jt} \leq \boldsymbol{x}_{jt}'\boldsymbol{\gamma}_j^P\}$, with $\eta_{jt} \sim Logistic$, but now $\mathbb{E}\left(\xi_{jt} \mid a_{jt}=1, \boldsymbol{x}_t\right)$ is a third order polynomial in $Euler - \ln\Lambda\left(\boldsymbol{x}_t'\boldsymbol{\gamma}_j^P\right)$. Therefore, the vector of regressors controlling for endogenous selection is:

$$\boldsymbol{h}_{jt}' = \left[\left(Euler - \ln\Lambda\left(\boldsymbol{x}_t'\boldsymbol{\gamma}_j^P\right)\right)^{\ell} : \ell = 1,2,3\right] \quad (46)$$

This semiparametric approach to control for selection follows Newey (2009).

3. *Semiparametric mixture Logit.* The entry model is the mixture Logit with entry decision $a_{jt} = 1\{\eta_{jt} \leq \boldsymbol{x}_{jt}'\boldsymbol{\gamma}_{j\kappa}^P\}$ for unobserved market type κ, and with mixture distribution $\Pr(\kappa_t = \kappa) = f_\kappa(\kappa)$. Conditional on $\kappa_t = \kappa$, the selection term $\mathbb{E}\left(\xi_{jt} \mid a_{jt}=1, \boldsymbol{x}_t, \kappa_t=\kappa\right)$ is a third order polynomial in $Euler - \ln\Lambda\left(\boldsymbol{x}_t'\boldsymbol{\gamma}_{j\kappa}^P\right)$. Accordingly, the vector of regressors controlling for endogenous selection is:

$$\boldsymbol{h}_{jt}' = \left[f_\kappa(\kappa)\left(Euler - \ln\Lambda\left(\boldsymbol{x}_t'\boldsymbol{\gamma}_{j\kappa}^P\right)\right)^{\ell} : \ell = 1,2,3, \text{ and } \kappa = 1,2,...,|\mathcal{K}|\right] \quad (47)$$

For all the 2SLS estimators, we use as instrumental variables the number of competitors in the market and the average *hub-size* of the rest of the airlines (separately for origin and destination airports).

Table 4 presents the estimates of demand parameters, while Table 5 provides the average demand elasticities and Lerner indexes derived from these estimates. Comparing the estimates obtained using OLS with those from various 2SLS methods — whether accounting for selection effects or not — we observe a significant adjustment in all parameter estimates when addressing the endogeneity of price and within-nest market share. This correction notably impacts the average own-price elasticity, shifting it from -1.59 to values below -5.54, and the corresponding Lerner index, which transitions from 69% to less than 20%.

In the context of this paper, the most significant findings arise from our investigation into the effects of controlling for the endogeneity of market entry. Remarkably, the most pronounced impacts materialize when we introduce finite mixture unobserved heterogeneity to address se-

lection bias. Upon incorporating a finite mixture, parameters α and σ experience absolute value increases of more than 18% and 34%, respectively. This change translates to an absolute value increase exceeding 50% in the average own-price elasticities. Consequently, the corresponding average Lerner index shifts from approximately 19% to 15%. These effects hold substantial importance, carrying meaningful economic implications.

Notably, the estimates of demand parameters and their corresponding elasticities exhibit considerable robustness concerning the selection of the number of mixtures. We observe a modest uptick in price sensitivity of demand as we transition from two to three unobserved market types. The main impact stems from the introduction of a finite mixture to address selection bias, with the number of mixtures contributing overall less significantly.

Figure 1 presents the empirical distributions of estimated own-price elasticities. Each row corresponds to an airline, while each column pertains to a different 2SLS estimator: the first column presents the estimator without controlling for selection, the second column illustrates the estimator that controls for selection using a sieve method but no mixture, and the third column presents the estimator with a three-type mixture.

The histograms in this figure are constructed based on estimates of elasticities at the airline-market-quarter level. The equation describing each elasticity solely depends on data on price p_{jmt}, market shares s_{jmt} and s_{0mt}, and parameter estimates $\widehat{\alpha}$ and $\widehat{\sigma}$. It is important to note that the data regarding prices and market shares remain constant across the various columns in the figure. Therefore, any shift in the distribution can be attributed solely to changes in the values of estimates $\widehat{\alpha}$ and $\widehat{\sigma}$.

Table 4: Estimation of Demand Parameters

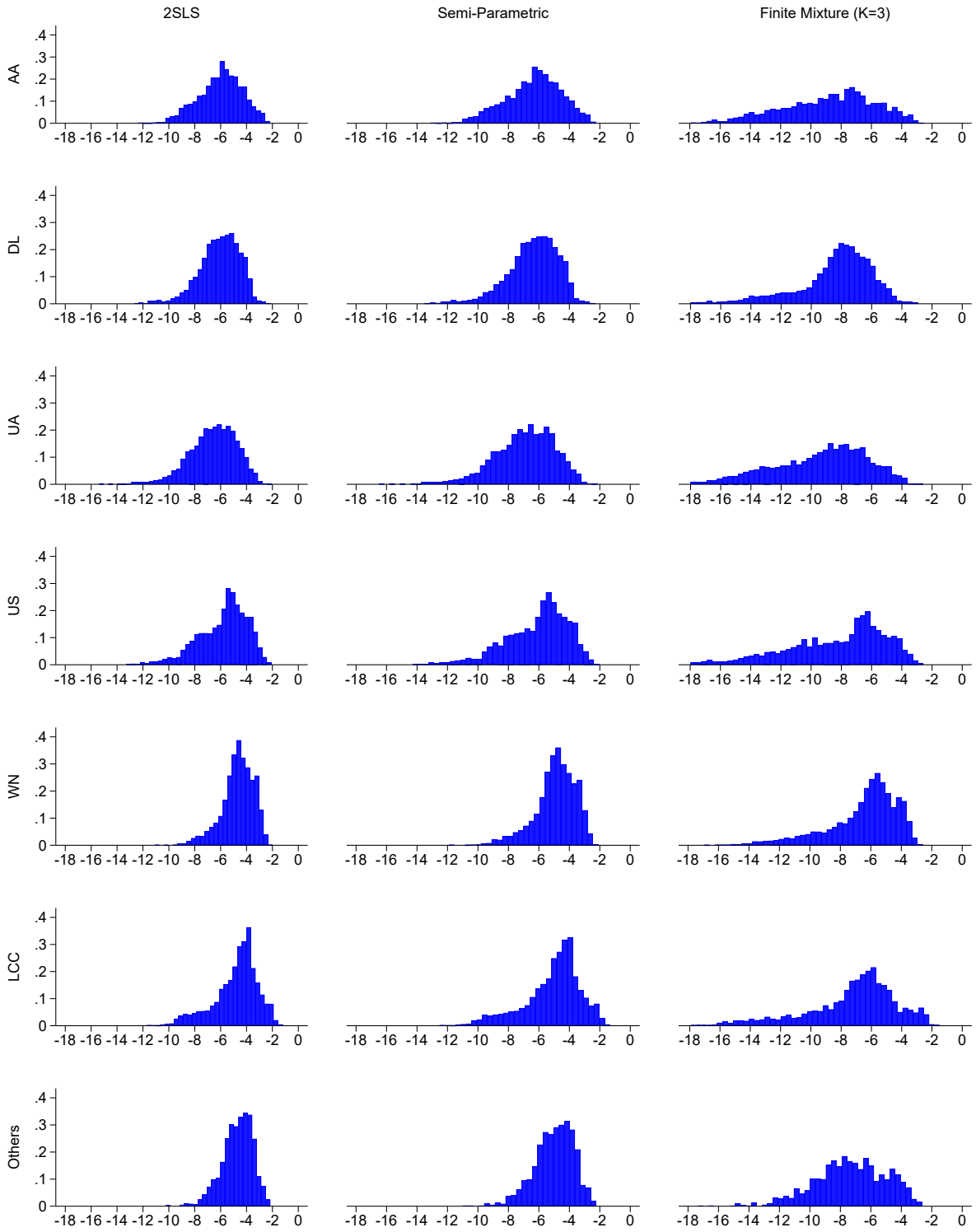
	Not control. for sel.		*Controlling for endogenous selection*			
	OLS	**2SLS**	**2SLS Heckman**	**2SLS Semi-P.**	**2SLS Fin-Mix $\lvert\mathcal{K}\rvert = 2$**	**2SLS Fin-Mix $\lvert\mathcal{K}\rvert = 3$**
Price (100\$) (α)	−0.643	−2.180	−2.193	−2.261	−2.574	−2.708
	(0.0105)	(0.1378)	(0.1348)	(0.1298)	(0.1549)	(0.1662)
Within Share (σ)	0.371	0.409	0.413	0.431	0.547	0.570
	(0.0058)	(0.0351)	(0.0389)	(0.0372)	(0.0509)	(0.0565)
Distance (1000mi)	0.729	2.130	2.196	2.264	2.497	2.472
	(0.0306)	(0.1372)	(0.1365)	(0.1310)	(0.1524)	(0.1572)
Distance^2	−0.216	−0.424	−0.453	−0.462	−0.496	−0.511
	(0.0112)	(0.0244)	(0.0252)	(0.0250)	(0.0276)	(0.0289)
hub-size orig. (100s)	1.637	2.272	1.999	1.320	1.593	1.383
	(0.0263)	(0.0382)	(0.0593)	(0.0625)	(0.0869)	(0.0989)
hub-size dest. (100s)	1.613	2.242	1.995	1.310	1.587	1.377
	(0.0267)	(0.0385)	(0.0595)	(0.0633)	(0.0872)	(0.0994)
Airline×Quarter FE	Y	Y	Y	Y	Y	Y
# control var. entry	0	0	6	18	36	54
Observations	35,763	35,763	35,763	35,763	35,763	35,763

Asymptotic standard errors account for estimation error in the first step using the method in Newey (2009).

Table 5: Average Own-Price Elasticities and Lerner Indexes

	Not control. for sel.		*Controllin for endogenous selection*			
	OLS	**2SLS**	**2SLS Heckman**	**2SLS Semi-P.**	**2SLS Fin-Mix $\lvert\mathcal{K}\rvert = 2$**	**2SLS Fin-Mix $\lvert\mathcal{K}\rvert = 3$**
Own-Price Elasticity	−1.596	−5.549	−5.601	−5.849	−7.406	−8.000
AA	−1.722	−6.013	−6.071	−6.363	−8.169	−8.857
DL	−1.761	−6.082	−6.133	−6.382	−7.871	−8.450
UA	−1.887	−6.573	−6.636	−6.936	−8.847	−9.573
US	−1.665	−5.801	−5.856	−6.122	−7.809	−8.450
WN	−1.354	−4.680	−4.719	−4.913	−6.068	−6.517
LCC	−1.370	−4.808	−4.857	−5.095	−6.674	−7.265
Others	−1.332	−4.705	−4.757	−5.006	−6.706	−7.337
Lerner Index	68.8%	19.9%	19.7%	18.9%	15.4%	14.4%
AA	62.7%	18.0%	17.9%	17.1%	13.8%	12.8%
DL	60.4%	17.5%	17.3%	16.7%	13.7%	12.8%
UA	56.9%	16.4%	16.2%	15.6%	12.6%	11.7%
US	65.9%	19.0%	18.9%	18.1%	14.8%	13.8%
WN	78.4%	22.8%	22.6%	21.8%	18.2%	17.1%
LCC	82.1%	23.5%	23.3%	22.2%	17.5%	16.3%
Others	79.2%	22.5%	22.3%	21.3%	16.4%	15.2%
Observations	35,763	35,763	35,763	35,763	35,763	35,763

Figure 1: Empirical Distribution of Estimated Elasticities (Airline-Market-Quarter level)



The histograms depicted in the first two columns are very similar. In contrast, the histograms based on the finite mixture estimates showcase significant alterations in both the location and the dispersion of the distribution of elasticities. Across all airlines, the incorporation of larger estimates for $\widehat{\alpha}$ and $\widehat{\sigma}$ using the mixture method leads to a noticeable leftward shift and an amplification in the spread of the histograms. These changes in the distributions' location and dispersion have important economic implications.

6.4 Estimation of costs and counterfactual experiments

In this paper, we focus on the consistent estimation of demand parameters in the presence of endogenous product entry. However, relying on the structure of our market equilibrium model, it is straightforward for researchers to estimate marginal costs, entry costs, and the joint distribution of unobservable variables. Armed with these estimated primitives, a wide array of counterfactual experiments can be undertaken. In this subsection, and within the framework of our empirical application, we offer a succinct overview of these supplementary estimation procedures.

6.4.1 Marginal costs

Based on an assumption about the nature of competition, such as Nash-Bertrand competition, we are able to derive marginal cost estimates at the airline-market-quarter level by treating the residuals in the pricing equation as estimates of these costs. It is important to note that these marginal costs can be computed only for those products that we observe being active in the market.

For some empirical questions, the researcher may need to estimate the marginal cost function: that is, the function that represents the causal effect of product characteristics and output on marginal costs. For this purpose, the researcher needs to estimate the parameters of a regression equation wherein the dependent variable is the marginal cost estimate, and exogenous product characteristics $\boldsymbol{x}_{jt}$ and output q_{jt} are the explanatory variables. A crucial consideration is that this regression equation is subject to selection bias due to endogenous product entry. Remarkably, the structure of the selection term in this equation mirrors that used in the estimation of the demand equation. We can then control for selection bias in the estimation of the marginal cost function using exactly the same control variables that we have used for the estimation of the demand parameters.

6.4.2 Demand and marginal cost unobservables

The consistent estimation of demand and marginal cost parameters inherently yields consistent estimates for the corresponding unobservable variables: ξ_{jmt} and ω_{jmt}. These unobservables are estimated as residuals within the estimated equations. While the estimation of these equations is subject to selection bias, the introduction of controls for selection enables us to achieve consistent estimation of the structural parameters and of the variables ξ_{jmt} and ω_{jmt} for the products observed being active in the market.[12]

Naturally, the more complex estimation of the probability distribution or stochastic process governing these unobservables for all products, both those observed being active and inactive in the market, requires one to address the issue of endogenous selection.

6.4.3 Counterfactuals at the intensive margin

Given estimates of demand and marginal cost parameters, researchers can perform counterfactual experiments involving changes to exogenous characteristics of demand and marginal costs but keeping the market structure and, more specifically, the set of active products, as constant. Such counterfactual scenarios are often encountered within applications of the widely employed BLP framework.

Importantly, once the challenge of endogenous selection has been addressed in the estimation of demand and marginal cost parameters, counterfactual experiments that uphold constant the set of products and market structure can be performed without further complications.

6.4.4 Counterfactuals at the extensive margin

Another class of counterfactual experiments involves changes to the ensemble of active firms and/or products within the market. In this category, the most straightforward experiment entails the exogenous removal of certain products from the market. Given the availability of data on the exogenous demand and marginal cost attributes of all products, performing this type of experiment does not significantly differ from the *counterfactuals at the intensive margin* previously discussed. This type of counterfactual includes as a particular case the evaluation of a counterfactual merger which ignores firms' endogenous responses at the extensive margin.[13]

Counterfactual experiments that involve the introduction of new products necessitate data

[12] Importantly, in calculating $\widehat{\xi}_{jmt}$ and $\widehat{\omega}_{jmt}$, one should not remove the estimated selection term from these residuals.

[13] In this class of models, the evaluation of the effects of a counterfactual merger requires making an assumption about the values of exogenous product characteristics for the new merging entity/firm. However, this complication is present regardless of the endogenous product selection issue that we address in this paper.

on the exogenous attributes of the new or hypothetical products. In our empirical analysis of the airline industry, we observe $\boldsymbol{x}_{jmt}$ for every airline-market-quarter product, irrespective of whether the product is active in the market. Specifically, data on the airline's hub size at both the origin and destination airports, as well as the airline-quarter fixed effects, are available for both existing products and potential entrants. However, the unobservable factors ξ_{jmt} and ω_{jmt} are unknown to the researcher for potential entrants. To perform this type of counterfactual, the researcher needs to determine the values of these unobservables also for the potential entrants.

In principle, the researcher could set values for ξ_{jmt} and ω_{jmt} for the new products at the unconditional mean of these variables, which is zero. However, this approach raises a significant concern: it contradicts the underlying reality that the airline opted not to enter in this particular market. To establish values of ξ_{jmt} and ω_{jmt} that align with observed endogenous entry decisions, one must consider $\mathbb{E}\left(\xi_{jmt}|\boldsymbol{x}_{mt}, a_{jmt}=0\right)$ and $\mathbb{E}\left(\omega_{jmt}|\boldsymbol{x}_{mt}, a_{jmt}=0\right)$ respectively.

While our estimation method yields consistent semiparametric estimates for the expected values $\mathbb{E}(\xi_{jmt}|\boldsymbol{x}_{mt}, a_{jmt}=1)$ and $\mathbb{E}(\omega_{jmt}|\boldsymbol{x}_{mt}, a_{jmt}=1)$, it is silent with respect to $\mathbb{E}(\xi_{jmt}|\boldsymbol{x}_{mt}, a_{jmt}=0)$ and $\mathbb{E}(\omega_{jmt}|\boldsymbol{x}_{mt}, a_{jmt}=0)$. Achieving point identification for the latter requires supplementary constraints, such as parametric assumptions or symmetry restrictions. An alternative approach instead involves estimating semiparametric bounds for these expected values. This information can then be harnessed to select appropriate values for ξ_{jmt} and ω_{jmt}.

7 Conclusions

In local geographic markets, we typically find only a subset of all the differentiated products available in an industry. Firms strategically select specific products that better match the preferences of local consumers. When making market entry decisions, firms possess information about the demand for their products, particularly regarding unobservable demand components. Firms tend to enter markets with higher expected demand. Neglecting this selection process can introduce significant biases in the estimation of demand parameters. This issue is common across various demand applications and industries. Existing methods to address this issue typically rely on strong parametric assumptions about demand unobservables and firms' information.

In this paper, we investigate the identification of demand parameters within a structural model that encompasses demand, price competition, and market entry (static or dynamic), while specifying the distribution of demand unobservables in a nonparametric finite mixture manner. The paper makes three main contributions. First, it establishes sequential identification of the demand parameters in this model. We demonstrate that the selection term in the demand equation results from a convolution of the probabilities of product entry for each discrete unobserved

market type and the densities associated with these market types. We show that data on firms' product entry decisions nonparametrically identify the probabilities of product entry conditional on the market type and the density of unobserved market types. Under mild conditions on the observable variables, demand parameters are identified after controlling for the nonparametric entry probabilities and densities for each market type.

Second, we propose a simple two-step estimator to address endogenous selection. In the first step, we estimate a nonparametric finite mixture model to determine the choice probabilities of product entry. In the second step, demand parameters are estimated using a Generalized Method of Moments (GMM) approach that accounts for both endogenous product availability and price endogeneity.

Third, we illustrate the proposed method by applying it to data from the airline industry. The findings highlight the importance of allowing for a finite mixture of unobserved market types when controlling for endogenous product entry, as failure to do so can lead to significant biases.

References

Abaluck, J., and A. Adams-Prassl (2021): "What do consumers consider before they choose? Identification from asymmetric demand responses," *The Quarterly Journal of Economics*, 136(3), 1611–1663.

Aguirregabiria, V., A. Collard-Wexler, and S. Ryan (2021): "Dynamic games in empirical industrial organization," in *Handbook of Industrial Organization, Volume 4*, ed. by K. Ho, A. Hortaçsu, and A. Lizzeri, pp. 225–343. Elsevier.

Aguirregabiria, V., and C.-Y. Ho (2012): "A dynamic oligopoly game of the US airline industry: Estimation and policy experiments," *Journal of Econometrics*, 168(1), 156–173.

Aguirregabiria, V., and P. Mira (2007): "Sequential estimation of dynamic discrete games," *Econometrica*, 75(1), 1–53.

——— (2019): "Identification of games of incomplete information with multiple equilibria and unobserved heterogeneity," *Quantitative Economics*, 10(4), 1659–1701.

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

Allman, E., C. Matias, and J. Rhodes (2009): "Identifiability of parameters in latent structure models with many observed variables," *The Annals of Statistics*, 37(6A), 3099–3132.

Amemiya, T. (1973): "Regression analysis when the dependent variable is truncated normal," *Econometrica*, 41(6), 997–1016.

——— (1974): "Multivariate regression and simultaneous equation models when the dependent variables are truncated normal," *Econometrica*, 42(6), 999–1012.

Aradillas-Lopez, A. (2012): "Pairwise-difference estimation of incomplete information games," *Journal of Econometrics*, 168(1), 120–140.

Aradillas-Lopez, A., B. Honoré, and J. Powell (2007): "Pairwise difference estimation with nonparametric control variables," *International Economic Review*, 48(4), 1119–1158.

Bajari, P., H. Hong, J. Krainer, and D. Nekipelov (2010): "Estimating static models of strategic interactions," *Journal of Business & Economic Statistics*, 28(4), 469–482.

Barseghyan, L., M. Coughlin, F. Molinari, and J. Teitelbaum (2021): "Heterogeneous choice sets and preferences," *Econometrica*, 89(5), 2015–2048.

Berry, S. (1992): "Estimation of a Model of Entry in the Airline Industry," *Econometrica*, 60(4), 889–917.

——— (1994): "Estimating discrete-choice models of product differentiation," *The RAND Journal of Economics*, 25(2), 242–262.

Berry, S., M. Carnall, and P. T. Spiller (2006): "Airline hubs: costs, markups and the implications of customer heterogeneity," *Competition Policy and Antitrust*, pp. 183–213.

Berry, S., and P. Haile (2014): "Identification in differentiated products markets using market level data," *Econometrica*, 82(5), 1749–1797.

——— (2021): "Foundations of demand estimation," in *Handbook of Industrial Organization, Volume 4*, ed. by K. Ho, A. Hortaçsu, and A. Lizzeri, pp. 1–62. Elsevier.

——— (2022): "Nonparametric Identification of Differentiated Products Demand Using Micro Data," *arXiv working paper*, 2204.06637.

Berry, S., and P. Jia (2010): "Tracing the woes: An empirical analysis of the airline industry," *American Economic Journal: Microeconomics*, 2(3), 1–43.

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

Bonhomme, S., K. Jochmans, and J.-M. Robin (2016): "Non-parametric estimation of finite mixtures from repeated measurements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1), 211–229.

Bresnahan, T., and P. Reiss (1991): "Entry and competition in concentrated markets," *Journal of Political Economy*, 99(5), 977–1009.

Bresnahan, T. F., and P. C. Reiss (1990): "Entry in monopoly market," *The Review of Economic Studies*, 57(4), 531–553.

Bunting, J. (2022): "Continuous permanent unobserved heterogeneity in dynamic discrete choice models," *arXiv preprint arXiv:2202.03960*.

Bunting, J., P. Diegert, and A. Maurel (2022): "Heterogeneity, Uncertainty and Learning: Semiparametric Identification and Estimation," *Working Paper*.

Cardell, S. (1997): "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity," *Econometric Theory*, 13(2), 185–213.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6, 5549–5632.

Ciliberto, F., C. Murry, and E. Tamer (2021): "Market structure and competition in airline markets," *Journal of Political Economy*, 129(11), 2995–3038.

Ciliberto, F., and E. Tamer (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77(6), 1791–1828.

Conlon, C., and J. Mortimer (2013): "Demand estimation under incomplete product availability," *American Economic Journal: Microeconomics*, 5(4), 1–30.

Crawford, G., R. Griffith, and A. Iaria (2021): "A survey of preference estimation with unobserved choice set heterogeneity," *Journal of Econometrics*, 222(1), 4–43.

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

Deaton, A., and J. Muellbauer (1980): "An almost ideal demand system," *The American Economic Review*, 70(3), 312–326.

Draganska, M., M. Mazzeo, and K. Seim (2009): "Beyond plain vanilla: Modeling joint product assortment and pricing decisions," *Quantitative Marketing and Economics*, 7, 105–146.

Dubé, J.-P., A. Hortaçsu, and J. Joo (2021): "Random-coefficients logit demand estimation with zero-valued market shares," *Marketing Science*, 40(4), 637–660.

Eizenberg, A. (2014): "Upstream innovation and product variety in the us home pc market," *The Review of Economic Studies*, 81(3), 1003–1045.

Gandhi, A., Z. Lu, and X. Shi (2023): "Estimating demand for differentiated products with zeroes in market share data," *Quantitative Economics*, 14(2), 381–418.

Gandhi, A., and A. Nevo (2021): "Empirical models of demand and supply in differentiated products industries," in *Handbook of Industrial Organization, Volume 4*, ed. by K. Ho, A. Hortaçsu, and A. Lizzeri, pp. 63–139. Elsevier.

Goeree, M. S. (2008): "Limited information and advertising in the US personal computer industry," *Econometrica*, 76(5), 1017–1074.

Grieco, P. (2014): "Discrete games with flexible information structures: An application to local grocery markets," *The RAND Journal of Economics*, 45(2), 303–340.

Hall, P., A. Neeman, R. Pakyari, and R. Elmore (2005): "Nonparametric inference in multivariate mixtures," *Biometrika*, 92(3), 667–678.

Hall, P., and X.-H. Zhou (2003): "Nonparametric estimation of component distributions in a multivariate mixture," *The Annals of Statistics*, 31(1), 201–224.

Haviland, A., and D. Nagin (2005): "Causal inferences with group based trajectory models," *Psychometrika*, 70(3), 557–578.

Haviland, A., D. Nagin, P. Rosenbaum, and R. Tremblay (2008): "Combining group-based trajectory modeling and propensity score matching for causal inferences in nonexperimental longitudinal data.," *Developmental Psychology*, 44(2), 422.

Heckman, J. (1976): "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement*, 5(4), 475–492.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71(4), 1161–1189.

Hu, Y., and Y. Xin (2022): "Identification and estimation of dynamic structural models with unobserved choices," *Available at SSRN 3634910*.

Kasahara, H., and K. Shimotsu (2014): "Non-parametric identification and estimation of the number of components in multivariate mixtures," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 97–111.

Lanza, S., D. Coffman, and S. Xu (2013): "Causal inference in latent class analysis," *Structural Equation Modeling: A Multidisciplinary Journal*, 20(3), 361–383.

Li, S., J. Mazur, Y. Park, J. Roberts, A. Sweeting, and J. Zhang (2022): "Repositioning and market power after airline mergers," *The RAND Journal of Economics*, 53(1), 166–199.

Lu, Z. (2022): "Estimating multinomial choice models with unobserved choice sets," *Journal of Econometrics*, 226(2), 368–398.

Moraga-González, J., Z. Sándor, and M. Wildenbeest (2023): "Consumer search and prices in the automobile market," *The Review of Economic Studies*, 90(3), 1394–1440.

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

Newey, W., J. Powell, and J. Walker (1990): "Semiparametric estimation of selection models: some empirical results," *The American Economic Review*, 80(2), 324–328.

Pakes, A., M. Ostrovsky, and S. Berry (2007): "Simple estimators for the parameters of discrete dynamic games (with entry/exit examples)," *The RAND Journal of Economics*, 38(2), 373–399.

Pilla, R. S., and B. G. Lindsay (2001): "Alternative EM methods for nonparametric finite mixture models," *Biometrika*, 88(2), 535–550.

Powell, J. (2001): "Semiparametric estimation of censored selection models," *Nonlinear Statistical Modeling*, pp. 165–96.

Rothenberg, T. (1971): "Identification in parametric models," *Econometrica*, 39(3), 577–591.

Seim, K. (2006): "An empirical model of firm entry with endogenous product-type choices," *The RAND Journal of Economics*, 37(3), 619–640.

Smith, H. (2004): "Supermarket choice and supermarket competition in market equilibrium," *The Review of Economic Studies*, 71(1), 235–263.

Sweeting, A. (2009): "The strategic timing incentives of commercial radio stations: An empirical analysis using multiple equilibria," *The RAND Journal of Economics*, 40(4), 710–742.

——— (2013): "Dynamic product positioning in differentiated product markets: The effect of fees for musical performance rights on the commercial radio industry," *Econometrica*, 81(5), 1763–1803.

Tobin, J. (1958): "Estimation of relationships for limited dependent variables," *Econometrica*, 26(1), 24–36.

Williams, B. (2020): "Nonparametric identification of discrete choice models with lagged dependent variables," *Journal of Econometrics*, 215(1), 286–304.

Xiao, R. (2018): "Identification and estimation of incomplete information games with multiple equilibria," *Journal of Econometrics*, 203(2), 328–343.

Yen, S. (2005): "A multivariate sample-selection model: Estimating cigarette and alcohol demands with zero observations," *American Journal of Agricultural Economics*, 87(2), 453–466.

Yen, S., and B.-H. Lin (2006): "A sample selection approach to censored demand systems," *American Journal of Agricultural Economics*, 88(3), 742–749.