WARWICK
ECONOMICS

Sieve BLP: A Semi-Nonparametric Model of Demand for Differentiated Products

(Updated & previously circulated as:
Identifying the Distribution of Random Coefficients in BLP Demand Models Using One Single Variation in Product Characteristics)

Ao Wang

September 2020(updated June 2021) No: 1304

Warwick Economics Research Papers

ISSN 2059-4283 (online)
ISSN 0083-7350 (print)

Sieve BLP: A Semi-Nonparametric Model of Demand for Differentiated Products*

Ao Wang†

June 28, 2021

Abstract

We develop a semi-nonparametric approach to identify and estimate the demand for differentiated products. The proposed method adopts a random coefficients discrete choice logit model (i.e., mixed logit model) in which the distribution of random coefficients is nonparametrically specified. Our method minimizes misspecification error in this distribution to which routinely used parametric approach is subject. In addition, it overcomes the practical challenge of dimensionality in the number of products that remains the main hurdle in the nonparametric estimation of demand functions. We propose a sieve estimation procedure (referred to as *sieve BLP*) that remains simple to implement. Extensive Monte Carlo simulations show its robust finite-sample performance under various data generating processes. We use our method to investigate the welfare implications of a sugar tax in the ready-to-eat cereal industry in the US. This illustrative application underscores the advantages of sieve BLP in predicting the distributional effects of a policy change such as introducing a sugar tax.

*This paper was previously circulated as *Identifying the Distribution of Random Coefficients in BLP Demand Models Using One Single Variation in Product Characteristics*. I am grateful to Mingli Chen, Giovanni Compiani, Xavier d'Haultfœuille, Roland Rathelot and Eric Renault for useful discussions. I would like to thank seminar and workshop participants at the University of Warwick, SJT. University, and the 5th Dongbei Econometrics Workshop. I would also like to thank IRI for making the data used in this paper available. Empirical estimates and analyses in this paper, based on data provided by IRI, are by mine and not by IRI. All errors are my own.

†University of Warwick (ao.wang@warwick.ac.uk).

1 Introduction

Since the seminal work of Berry (1994) and Berry et al. (1995) (hereafter BLP), BLP-type models are widely used in the empirical literature of demand. This method of aggregate-level data allows for (unobserved) individual preference heterogeneity and enables to capture realistic substitution patterns among products.[1] Oftentimes, applied researchers adopt a BLP-type model with random coefficients in the utility structure to proxy individual heterogeneity and estimate a parametric distribution of the random coefficients. Despite computational convenience, the parametric assumptions may lack an economic foundation and be subject to severe misspecification errors. These assumptions can be particularly problematic when the researcher wishes to learn distributional effects of a policy change. Recent papers propose a fully nonparametric approach that remains agnostic about individual heterogeneity and directly estimates structural demand functions.[2] However, its application is limited by the dimensionality in the number of products: the number of parameters to be estimated in the finite sample may explode even when the number of products is moderate.

We propose a semi-nonparametric approach that is theoretically founded, simple to implement, and applicable in large applications. Our approach adopts a random coefficients discrete choice logit model (i.e., mixed logit model) and nonparametrically identifies and estimates the distribution of random coefficients. This method minimizes misspecification error in the distribution of random coefficients. Also, it avoids the dimensionality problem in the number of products by nonparametrically estimating the distribution of random coefficients (rather than structural demand functions) whose dimensionality does not increase with the number of products in most empirical settings. We prove that the distribution of random coefficients is nonparametrically identified under fairly weak conditions. In particular, our identification strategy only requires at most one single variation in product characteristics that interact with the random coefficients, largely relaxing the support requirements in the existing literature. We propose a sieve procedure (referred to as *sieve BLP*) to estimate the distribution of random coefficients that is implemented literally as a parametric GMM in the finite sample. Extensive Monte Carlo simulations show the robust performance of the sieve

[1] In particular, the property of Independence of Irrelevant Alternatives is not assumed.

[2] See Compiani (2019).

BLP estimator under various data generating processes and problem sizes (i.e., the number of products). We further demonstrate the economic relevance of the proposed method and investigate the welfare implications of a sugar tax in the ready-to-eat (RTE) cereal industry in the US. The demand estimates using sieve BLP suggest complex but realistic individual heterogeneity in preference and predict largely heterogeneous effects of the sugar tax on households of different sizes. In contrast, the parametric approach understates the magnitude of individual heterogeneity, predicting almost homogeneous welfare effects of the tax on these households. This illustrative exercise suggests the use of sieve BLP (over the routinely used parametric approach) when the researcher aims to quantify distributional effects of a policy change such as introducing a sugar tax.

The mixed-logit specification with flexibly distributed (i.e., nonparametric) random coefficients is central to our main results. This specification has various appealing economic and econometric properties. First, the mixed-logit model has clear micro-foundations. Saito (2017) provides necessary and sufficient conditions for a mixed-logit model to represent a random choice function. Second, the mixed-logit model has desirable approximation properties. McFadden and Train (2000) show that a mixed-logit model with flexibly distributed random coefficients can approximate any discrete choice model derived from random utility maximization. Third, and importantly, we prove that the distribution of random coefficients in the mixed-logit model can be nonparametrically identified using at most one single variation in product characteristics, satisfied in most theoretical and empirical settings. This new identification result is the key to ensure desirable asymptotic properties and robust finite-sample performance of the sieve BLP estimator.

While our proposed method circumvents the practical challenge of dimensionality in the number of products to which the fully nonparametric approach is subject, the two approaches complement each other. On one hand, the mixed-logit specification in this paper restricts products to be substitutes and rules out complementaries. In contrast, the fully nonparametric approach directly targets structural demand functions (e.g., Compiani (2019)) and allows for a broader range of consumer behaviors.[3] On the other, because the distribution of individual heterogeneity is identified, our approach enables researchers to quantify individual consumer welfare relevant to many policy

[3] It is possible to extend this paper's methodologies to the setting of Wang (2021) in which products are not restricted to be substitutes. We leave the extension as future research.

questions (e.g., mergers, welfare implication of excise tax), while the fully nonparametric approach cannot.[4] Depending on research questions, problem sizes, and availability of computational resources, applied researchers can choose one approach over the other.

Related Literature. This paper's identification strategy relies on recent progresses on identification of structural demand functions using aggregate data and provide conditions under which the distribution of random coefficients is further identified in mixed-logit model.[5] Importantly, our identification strategy employs a real analytic property of mixed-logit models and relies on a rank condition of product characteristics matrix that is substantially weaker than those in the existing literature.[6] Fox et al. (2012) identify the distribution of random coefficients in mixed logit models with continuous product characteristics and exclude interactive terms of product characteristics (e.g., polynomial terms of product characteristics) in utilities, while our identification strategy applies to both continuous and discrete product characteristics, as well as their interactions. Fox (2017) does not assume the mixed-logit specification and identifies the joint distribution of random coefficients and idiosyncratic errors. Similarly, his strategy relies on large support condition and does not allow for interactive terms of product characteristics. Both Dunker et al. (2017) and Allen and Rehbeck (2020) assume the identification of structural demand functions and study the identification of distribution of random coefficients. Dunker et al. (2017) employ Radon transformation to achieve the identification.[7] Their approach requires regressors to be continuous and with large support (or limited support together with additional restrictions on moments of the random coefficients).[8] In contrast, our identification strategy requires neither continuity nor

[4]See page 4 of Compiani (2019).

[5]These progresses include identification arguments using completeness conditions (Berry and Haile, 2014), in a simultaneous system of demand and supply (Matzkin (2008), Berry and Haile (2014, 2018)), in a triangular system (Chesher (2003), Imbens and Newey (2009), D'Haultfœuille and Février (2015), Torgovitsky (2015)), in perturbed utility models of demand (Allen and Rehbeck, 2019), and in models of demand for bundles (Fox and Lazzati (2017), Wang (2021)).

[6]Specifically, we employ Iaria and Wang (2019)'s result of real analyticity of mixed logit models that allows for any distribution of random coefficients. Fox et al. (2012) first prove this real analytic property in a mixed-logit model in which the support of random coefficients is compact. See il Kim et al. (2014) and Fox and Gandhi (2016) for discussions and applications of real analytic properties in the identification of random coefficients in multinomial choice models. See also Wang (2021) for an application in the identification of demand for bundles. Conlon and Gortmaker (2020) discuss numerical implications of this property in the estimation of BLP-type models.

[7]See also Hoderlein et al. (2010) and Gautier and Hoderlein (2013) for applications of Radon transformation in identification and estimation of random coefficient models.

[8]See their Assumptions 3.1-3.3. See also Masten (2018) for similar support conditions that achieve the identification of random coefficients.

large support of the regressors. In a perturbed utility model, Allen and Rehbeck (2020) identify moments of the random coefficients from the (higher-order) derivatives of structural demand functions at the origin.[9] Differently, we directly target the distribution of random coefficients.[10]

We also contribute to the literature of estimation of BLP-type models by proposing a practical sieve minimal-distance estimator built on our identification result.[11] Similarly to the nonparametric approach in this literature, the sieve BLP estimator relaxes parametric assumptions on the distribution of random coefficients routinely used in BLP-type models, minimizing misspecification errors. Differently, and importantly, our approach does not introduce the practical challenge of dimensionality in the number of products often faced by the fully nonparametric approach. For example, Compiani (2019) proposes to nonparametrically estimate structural demand functions. His approach remains agnostic about individual heterogeneity. However, the number of parameters to be estimated in the finite sample can increase quickly in the number of products. Lu et al. (2019) propose an alternative semi-nonparametric estimation procedure and transform the demand system into a partial linear model. Differently, our sieve BLP estimation follows the two-step procedure in BLP-type models to nonparametrically estimate the distribution of random coefficients.

Organization. Section 2 introduces the model and necessary notations. Section 3 presents the main identification and estimation results. Section 4 summarizes results of Monte Carlo simulations. Section 5 shows our empirical illustration. Section 6 concludes. All examples, additional tables, and proofs can be found in Appendices A-F.

[9] This strategy is also used in the constructive identification arguments of Fox et al. (2012). Note that without further conditions, the distribution is not identified even if all its moments are. For a set of such conditions, see Assumption 7 of Fox et al. (2012).

[10] Allen and Rehbeck (2020) allows for independently distributed random slopes and (product-specific) random intercepts in utilities (see their Assumption 1). Oftentimes in empirical research, only random slopes are present and random intercepts are degenerated, i.e., constants (see Nevo (2000, 2001), Gentzkow (2007)). Our identification and estimation results in the main text cover these empirical settings. In Appendix B, we extend our identification strategy to allow for the presence of product-specific random intercepts.

[11] See Chen (2007) and Chen and Qiu (2016) for recent progresses on sieve estimation. See Fox et al. (2016) for a sieve estimation of the distribution of random coefficients in structural models that do not cover BLP-type ones.

2 Model

Denote market by $t \in \mathbf{T}$.[12] Let $\mathbf{J}_t$ be the set of J_t market-specific products in market t. Individuals in market t can either choose a product $j \in \mathbf{J}_t$ or the outside option denoted by 0. Let $x_{jt} \in \mathbb{R}^{1\times K}$ denote the vector of K observed characteristics of product j in market t. Since the main results of the paper do not necessitate a notational distinction between price and non-price characteristics, we use x_{jt} to refer to all observed characteristics of j that also include its price in market t. Following theoretical and empirical literature of BLP models, we assume a linear index restriction in individuals' indirect utilities.[13] For individual i, her indirect utility from purchasing product j in market t is:

$$\begin{aligned} U_{ijt} &= x_{jt}\beta_i + \xi_{jt} + \varepsilon_{ijt} \\ &= x_{jt}^{(1)}\beta^{(1)} + x_{jt}^{(2)}\beta_i^{(2)} + \xi_{jt} + \varepsilon_{ijt} \\ &= [x_{jt}^{(1)}\beta^{(1)} + \xi_{jt}] + x_{jt}^{(2)}\beta_i^{(2)} + \varepsilon_{ijt} \\ &= \delta_{jt} + x_{jt}^{(2)}\beta_i^{(2)} + \varepsilon_{ijt}, \end{aligned}$$

where $\delta_{jt} = x_{jt}^{(1)}\beta^{(1)} + \xi_{jt}$ is product j- and market t-specific mean utility, $x_{jt}^{(1)} \in \mathbb{R}^{1\times K_1}$ is a vector of product characteristics for which individuals have homogeneous taste $\beta^{(1)} \in \mathbb{R}^{K_1\times 1}$, $x_{jt}^{(2)} \in \mathbb{R}^{1\times K_2}$ is a vector that includes product characteristics for which individuals have heterogeneous tastes $\beta_i^{(2)} \in \mathbb{R}^{K_2\times 1}$ (e.g., price, sugar content) that can be unobserved to the econometrician (i.e., random coefficients)[14], and ε_{ijt} is an idiosyncratic error term. The term ξ_{jt} is product-market specific demand shock, observed to both firms and individuals but not to the econometrician. For individual i, the indirect utility from purchasing the outside option 0 in market t is normalized to

$$U_{i0t} = \varepsilon_{i0t}.$$

[12] The definition of the market depends on the empirical application. In the case of cross-sectional data, one could define markets as different geographic areas; in the case of panel data, they can be defined as different periods. In many cases, the market can be defined as a combination of both.

[13] See Assumption 1 and the discussions in Example 1 of Berry and Haile (2014).

[14] In applied work, the mean part of $\beta_i^{(2)}$, $\beta^{(2)}$, is often separated from $\beta_i^{(2)}$ and enters the index δ_{jt} via $x_{jt}^{(2)}\beta^{(2)}$. Then, the deviation $\beta_i^{(2)} - \beta^{(2)}$ is left outside of the index. Both writings are equivalent.

Assume that ε_{i0t} and ε_{ijt}, $j \in \mathbf{J}_t$, are i.i.d. Gumbel, and $\beta_i^{(2)}$ is distributed according to F and independent of $(x_{jt}, \xi_{jt})_{j \in \mathbf{J}_t}$ and $(\varepsilon_{ijt})_{j \in \mathbf{J}_t \cup \{0\}}$. Then, for any $j \in \mathbf{J}_t$,

$$s_{jt} = \int \frac{\exp\{\delta_{jt} + x_{jt}^{(2)} \beta_i^{(2)}\}}{1 + \sum_{j' \in \mathbf{J}_t} \exp\{\delta_{jt'} + x_{jt'}^{(2)} \beta_i^{(2)}\}} dF(\beta_i^{(2)}), \tag{1}$$

where s_{jt} is the observed market share of j in market t. Without loss of generality, suppose that $\mathbf{J}_t = \mathbf{J}$, i.e., there is no variation in choice set across markets.[15] Then, we can re-write (1) as:

$$\begin{aligned} s_{jt} &= \sigma_j(\delta_t; X_t^{(2)}, F) \\ &= \int \frac{\exp\{\delta_{jt} + x_{jt}^{(2)} \beta_i^{(2)}\}}{1 + \sum_{j' \in \mathbf{J}} \exp\{\delta_{jt'} + x_{jt'}^{(2)} \beta_i^{(2)}\}} dF(\beta_i^{(2)}), \end{aligned} \tag{2}$$

where $\delta_t = (\delta_{jt})_{j \in \mathbf{J}}$ and $X_t^{(2)} = (x_{jt}^{(2)})_{j \in \mathbf{J}} \in \mathbb{R}^{J \times K_2}$. Similarly, let $X_t^{(1)}$ denote $(x_{jt}^{(1)})_{j \in \mathbf{J}} \in \mathbb{R}^{J \times K_1}$ and $s_t = (s_{jt})_{j \in \mathbf{J}}$. The econometrician observes $(X_t^{(1)}, X_t^{(2)}, s_t)$ for all $t \in \mathbf{T}$ and wishes to identify and estimate the distribution F.[16]

3 Identification and Estimation of F

3.1 Identification

We identify F using instrumental variable approach in two steps. In the first step, given F and $X_t^{(2)}$, we invert the observed market shares s_t to δ_t:[17]

$$s_{jt} = \sigma_j(\delta_t; X_t^{(2)}, F) \iff \delta_{jt} = \sigma_j^{-1}(s_t; X_t^{(2)}, F) = X_{jt}^{(1)} \beta^{(1)} + \xi_{jt}, \tag{3}$$

where $\sigma^{-1} = (\sigma_j^{-1})_{j \in \mathbf{J}}$ is the inverse function of $\sigma = (\sigma_j)_{j \in \mathbf{J}}$. In the second step, we assume that the econometrician observes a vector of variables $Z_{jt} \in \mathbb{R}^P$ for $j = 1, ..., J$ and $t \in \mathbf{T}$.[18] Moreover, the following exogeneity conditions hold: for $j \in \mathbf{J}$ and $Z_j \in$

[15] The distribution of random coefficients, F, is allowed to change as $\mathbf{J}_t$ changes. Then, the results in this paper hold conditional on a given choice set $\mathbf{J}_t = \mathbf{J}$.

[16] Once F is identified, one can invert s_t to δ_t for each $t \in \mathbf{T}$ (see Berry et al. (2013)) and apply standard arguments in linear models using instrument variables to identify $\beta^{(1)}$.

[17] As noted by Berry and Haile (2014), mixed-logit model (2) is invertible.

[18] Z_{jt} includes both exogenous characteristics X_{jt} and additional (excluded) instruments, e.g., cost shifters, exogenous characteristics of other products.

$\mathscr{D}_{Z_j}$,

$$\mathbb{E}\left[\xi_{jt}|Z_{jt}=Z_j\right]=0, \tag{4}$$

Then, combining (3) and (4), one obtains the following moment conditions:

$$m_j(Z_j;\beta^{(1)},F)=\mathbb{E}\left[\sigma_j^{-1}(s_t;X_t^{(2)},F)-X_{jt}^{(1)}\beta^{(1)}|Z_{jt}=Z_j\right]=0. \tag{5}$$

Using moment conditions (5), Theorem 1 of Berry and Haile (2014) proves that $\sigma=(\sigma_j)_{j\in\mathbf{J}}$ are identified as functions of δ_t for any given $X_t^{(2)}\in\mathscr{X}$.[19] However, without further assumptions, this does not automatically imply that the distribution F is identified, i.e., given two distributions F and F',

$$\sigma(\delta_t;X_t^{(2)},F)=\sigma(\delta_t;X_t^{(2)},F')\not\Rightarrow F=F'.$$

The aim of identification in this paper is to establish conditions under which F is identified in model (1), i.e., $F=F'$. To ease the exposition, we drop the notation t. Denote the support of $X^{(2)}=(x_j^{(2)})_{j\in\mathbf{J}}$ by $\mathscr{X}\subset\mathbb{R}^{J\times K_2}$.

Theorem 1. *Suppose that the following two conditions hold:*

1. *For any $X^{(2)}\in\mathscr{X}$ and any $j\in\mathbf{J}$, $\sigma_j(\delta;X^{(2)},F)$ is identified in $\mathscr{D}\ni\delta$, where $\mathscr{D}$ is an open set in $\mathbb{R}^J$.*

2. *There exists $X^{(2)}\in\mathscr{X}$, such that $X^{(2)}$ is of full column rank.*

Then, F is identified.

Proof. See Appendix A. □

Condition 1 of Theorem 1 is built on Theorem 1 of Berry and Haile (2014) and assumes the identification of $\sigma=(\sigma_j)_{j\in\mathbf{J}}$ as functions of δ in an open set $\mathscr{D}$, for any given $X^{(2)}\in\mathscr{X}$. We only require $\mathscr{D}$ to be local (i.e., an open set) rather than to have large support, i.e., $\mathbb{R}^J$. This requirement is sufficient in the setting of model (2) due to its real analytic property. As we show in Lemma 1 in Appendix A.1, this property implies that Condition 1 is sufficient for $\sigma=(\sigma_j)_{j\in\mathbf{J}}$ being identified as functions of δ in the

[19] The key argument is the completeness of the joint distribution of $(Z_{jt},X_{jt}^{(1)},X_{jt}^{(2)},s_t,p_t)$ with respect (s_t,p_t). See also Newey and Powell (2003) for the application of completeness conditions in nonlinear IV models.

entire $\mathbb{R}^J$.[20] This local condition is convenient in practice. It can be achieved by a price-setting game with J supply-side variables varying in a local open set. Take cost shifters $c = (c_j)_{j\in\mathbf{J}}$ for example. Because the price vector p (a column vector of $X^{(2)}$) is outcome of the price-setting game, it is typically a function of (δ, c). Then, for a given p, the pricing equations define a relationship between c and δ. Under suitable regularity conditions, this relationship allows generating variation of δ in an open set as long as c varies in an open set.[21]

As shown in Appendix A.1, Condition 1 implies the identification of $\mu_i = X^{(2)}\beta_i^{(2)}$ given $X^{(2)}$. We propose Condition 2 to further achieve the identification of F. This condition relies on the rank of $X^{(2)}$ and has several advantages. First, it is weaker than and implied by most support conditions in the literature, e.g., $X^{(2)}$ is continuous and $\mathscr{X}$ contains an open set (Fox et al., 2012). Second, it complements existing support arguments and covers important situations that they rule out, e.g., $X^{(2)}$ includes same regressors across products ($x_{jt}^{(2)} = x_{j't}^{(2)}$), or polynomial terms of regressors.[22] To the extreme, Condition 2 can still apply even when $X^{(2)}$ does not vary. Intuitively, once the distribution of $\mu_i = X^{(2)}\beta_i^{(2)}$ given $X^{(2)}$ is identified, we can identify the distribution of $X^{(2)\mathrm{T}}\mu_i = X^{(2)\mathrm{T}}X^{(2)}\beta_i^{(2)}$. Then, because of the full column rank of $X^{(2)}$, $X^{(2)\mathrm{T}}X^{(2)}$ is invertible and, as a consequence, we identify the distribution of $\beta_i^{(2)}$ from that of $X^{(2)\mathrm{T}}X^{(2)}\beta_i^{(2)}$. Note that this identification argument using Condition 2 relies on the feature of model (2) that the random coefficients are individual specific rather than individual-product specific. This feature is widely adopted in empirical research.[23] However, it rules out the specification in which $X^{(2)}$ contains J product-specific intercepts, i.e., $\beta_i^{(2)}$ contains J random intercepts. In Appendix B, we provide further conditions that achieve the identification of F in the presence of such random intercepts

[20] In the absence of mixed-logit assumption and when one wishes to identify the joint distribution of random coefficients and idiosyncratic errors, this local support condition may not be sufficient for the identification, and the large support assumption on δ may be needed. See Fox (2017) for an example.

[21] Using cost shifters as instruments, Theorem 2 of Wang (2021) provides a set of sufficient conditions under which this requirement is satisfied and Condition 1 holds.

[22] See Fox (2017); Fox et al. (2012) and Masten (2018) among others for such support arguments.

[23] See McFadden and Train (2000), Nevo (2000, 2001), Petrin (2002), Berto Villas-Boas (2007), Fan (2013) among others.

using only one single variation in $X^{(2)}$.[24]

Similar to Condition 1, Condition 2 easily holds in many empirical settings. If product characteristics $X^{(2)}$ vary continuously across markets (e.g., prices, advertisement amount) and the support $\mathscr{X}$ has an open subset, then Condition 2 is satisfied.[25] When $X^{(2)}$ contain characteristics that do not vary across markets (e.g., car color) and the corresponding sub-matrix in $X^{(2)}$ remains constant, Condition 2 holds if products are differentiated enough along the dimensions of these product characteristics, i.e., the column vectors of product characteristics in $X^{(2)}$ are independent. This independence condition holds generically. To see this, suppose that the column vectors in $X^{(2)}$ are co-linear. Then, one can re-define the vectors of product characteristics as $\tilde{X}^{(2)} \in \mathbb{R}^{J\times\tilde{K}_2}$ with $\tilde{K}_2 = \text{rank}(X^{(2)}) < K_2$ by linearly combining column vectors in $X^{(2)}$, such that $\tilde{X}^{(2)}$ is of full column rank $\tilde{K}_2$. Accordingly, the corresponding random coefficients $\tilde{\beta}_i^{(2)}$ are linear combinations of $\beta_i^{(2)}$.

3.2 Estimation

Theorem 1 provides a theoretical foundation for nonparametrically estimating the distribution F. This brings several advantages. First, allowing for nonparametric F minimizes misspecification errors due to parametric assumptions imposed on F that applied researchers are often concerned about. Second, nonparametrically estimating F avoids the dimensionality problem in the number of products that remains the main hurdle in the direct nonparametric estimation of structural demand functions. As a result, even when the number of products is large, one can still flexibly estimate F and therefore the demand functions. In this section, we propose a sieve procedure to estimate F (*sieve BLP*) in the setting of model (2). The estimator has desirable asymptotic properties and is convenient to implement in practice. We refer to Chen (2007) for a general discussion of sieve estimation of semi-nonparametric models.

Let $\theta = (\beta^{(1)}, F) \in \Theta_{\beta^{(1)}} \times \Theta_F$ be the true parameters, where $\Theta_{\beta^{(1)}}$ is a compact subset of $\mathbb{R}^{K_1}$ and Θ_F is a set of distribution functions defined in $\mathbb{R}^{K_2}$. Suppose that

[24] In the existing literature, identification of the distribution of individual-product specific random coefficients often requires strong support conditions or restrictions on the distribution of F (see Masten (2018) for a recent example). Our identification strategy in Appendix B minimizes the support requirement and only relies on a single variation in $X^{(2)}$. Instead, it imposes an independence condition between random intercepts and random slopes, which is also assumed in some recent papers such as Allen and Rehbeck (2020).

[25] This is because the subset of full-column rank matrices in $\mathbb{R}^{J\times K_2}$ is dense.

the conditions of Theorem 1 hold and F and $\beta^{(1)}$ are identified. Then, the vector of the true parameters $(\beta^{(1)}, F)$ is the unique solution to (5):

$$(\beta^{(1)}, F) = \underset{(\beta^{(1)\prime}, F') \in \Theta_{\beta^{(1)}} \times \Theta_F}{\operatorname{argmin}} \mathbb{E}\left[\left| (m_j(Z_j; \beta^{(1)}, F))_{j \in \mathbf{J}} \right|_{[\Sigma(Z)]^{-1}} \right], \tag{6}$$

where $|v|_W = v^{\mathrm{T}} W v$ and $\Sigma(Z)$ is a weighting matrix that depends on $Z = (Z_j)_{j \in \mathbf{J}}$.

Two challenges arise when we directly estimate $(\beta^{(1)}, F)$ using the finite-sample counterpart of (6). First, the functional form of the distribution $s_t | Z_t = Z$ is unknown to the econometrician and therefore $m_j(Z_j; \beta^{(1)}, F)$ is unknown. Consequently, one needs to (nonparametrically) estimate $m_j(Z_j; \beta^{(1)}, F)$. Second, Θ_F is typically of infinite dimension; the estimator of F obtained by directly minimizing over Θ_F may be ill-posed.

The sieve BLP estimation procedure overcomes these two challenges. First, the sieve procedure nonparametrically estimates $m_j(Z_j; \beta^{(1)}, F)$ using a linear sieve space generated by a set of basis functions $\{\phi_k(\cdot)\}_{k=1}^{k_{1T}}$, where k_{1T} is the dimension of the sieve space and increases with T. In practice, it suffices to run a linear regression of $\sigma_j^{-1}(s_t; X_t^{(2)}, F') - X_{jt}^{(1)} \beta^{(1)\prime}$ over $(\phi_1(Z_{jt}), ..., \phi_{k_{1T}}(Z_{jt}))$, for each $j = 1 \in \mathbf{J}$, and obtain $\hat{m}_j(Z_j; \beta^{(1)\prime}, F')$ as the linear prediction evaluated at $(\phi_1(Z_j), ..., \phi_{k_{1T}}(Z_j))$. Second, the proposed procedure estimates F by minimizing over a k_{2T} dimensional sieve space $\Theta_{k_{2T}, F}$ that is a "good" approximation of Θ_F. Concretely, we require k_{2T} to increase with T and $\Theta_{k_{2T}, F}$ to be asymptotically dense in Θ_F. This minimization in the finite sample is essentially parametric and simple to implement.

Denote by $\hat{\Sigma}(.)$ a consistent estimate of $\Sigma(.)$. The sieve BLP estimation procedure is implemented as follows:

Implementation *(Sieve BLP).*

1. Given $(\beta^{(1)\prime}, F') \in \Theta_{\beta^{(1)}} \times \Theta_{k_{2T}, F}$, for $t = 1, ..., T$, implement the demand inverse and obtain $\sigma^{-1}(s_t; X_t^{(2)}, F') - X_t^{(1)} \beta^{(1)\prime}$.

2. For each $j = 1, ..., J$, run a linear regression of $\sigma_j^{-1}(s_t; X_{jt}^{(2)}, F') - X_{jt}^{(1)} \beta^{(1)\prime}$ over $(\phi_k(Z_{jt}))_{k=1}^{k_{1T}}$; denote by $\pi_{j,k_{1T}}(\beta^{(1)\prime}, F')$ column vector of the coefficients of the linear regression. Obtain $\hat{m}_j(Z_j; \beta^{(1)\prime}, F') = (\phi_1(Z_j), ..., \phi_{k_{1T}}(Z_j)) \pi_{j,k_{1T}}(\beta^{(1)\prime}, F')$.

3. Compute the objective function $\frac{1}{T} \sum_{t=1}^{T} \left| (\hat{m}_j(Z_{jt}; \beta^{(1)\prime}, F'))_{j=1}^{J} \right|_{[\hat{\Sigma}(Z_t)]^{-1}}$.

4. Obtain the sieve estimator $(\hat{\beta}^{(1)}, \hat{F})$ by minimizing this objective with respect to $(\beta^{(1)'}, F') \in \Theta_{\beta^{(1)}} \times \Theta_{k_{2T},F}$.

Break the challenge of dimensionality in J. When directly nonparametrically estimating σ_j^{-1} in (5), one will encounter the challenge of dimensionality in the number of products: σ_j^{-1} is a function of at least $2J$ arguments, i.e., J market shares $s_t = (s_{j't})_{j' \in \mathbf{J}}$ and J prices $p_t = (p_{j't})_{j' \in \mathbf{J}}$, if p_t enter $X_{jt}^{(2)}$. In the finite sample, the number of necessary terms to approximate σ_j^{-1} (i.e., parameters to be estimated) typically grows exponentially in J.[26] As a result, even when J is moderate, the number of parameters can already too large to be handled.[27]

The proposed sieve BLP overcomes this challenge of dimensionality. The key insight is that the dependence of σ_j^{-1} on (s_t, p_t) in model (2) is expressed via the distribution of K_2 random coefficients (F) rather than directly via $(s_t, p_t) \in \mathbb{R}^{2J}$. Then, the number of parameters in the finite-sample estimation is determined by the extent to which F is approximated (which is a function of K_2) and does not depend on J. To see this, suppose $\Sigma(.) = \mathbf{I}$. The proposed procedure can be interpreted as a parametric GMM using the following unconditional moment conditions: for $j \in \mathbf{J}$ and $k = 1, ..., k_{1T}$,

$$\mathbb{E}[(\sigma_j^{-1}(s_t; X_t^{(2)}, F) - X_t^{(1)} \beta^{(1)}) \phi_k(Z_{jt})] = 0, \tag{7}$$

where F belongs to a k_{2T}-dimensional parametric sieve space $\Theta_{k_{2T},F}$.[28] As usual, we require that the number of parameters be at most equal to that of moment conditions:

$$k_{2T} + K_1 \leq J \times k_{1T}. \tag{8}$$

Suppose that we choose $\Theta_{k_{2T},F}$ as a Hermite-form sieve space of order m.[29] Then, given

[26] For instance, Compiani (2019) approximates σ_j^{-1} using Bernstein polynomials up to order m for each argument. The number of parameters to be estimated in the finite sample is at least $(1+m)^{2J}$.

[27] When $J = 10$ and the order of Bernstein polynomials is 2, without further restriction, the number of parameters is of order 10^9 (Table 1 of Compiani (2019)). Imposing restrictions on σ_j^{-1} reduces the number of parameters in the finite sample. However, it can still be large after applying these restrictions. For example, imposing both exchangeability and index restrictions reduces the number of parameters to the order of J^m, where m is the order of Bernstein polynomials (see section 3.2 and Appendix 1 of Compiani (2019)).

[28] When $\Sigma(.) = \mathbf{I}$, the minimal-distance criterion function in Step 3 of the sieve BLP implementation becomes a GMM criterion function with the weighting matrix $\text{Diag}\left(\left(P_1^{\mathrm{T}} P_1\right)^{-1}, ..., \left(P_J^{\mathrm{T}} P_J\right)^{-1}\right)$, where $P_j = (\phi_k(Z_{jt}))_{k=1,...,k_{1T}; t=1,...,T} \in \mathbb{R}^{T \times k_{1T}}$. See page 1799 of Ai and Chen (2003) for more details.

[29] See equation (9).

T, the number of parameters, $k_{2T} + K_1 = K_2^m + K_1$, is fixed and does increase in J. Especially, in many empirical settings, K_2 is much smaller than J. Consequently, even when J is large, the number of parameters remains manageable in the finite sample.[30] In contrast, in the fully nonparametric approach, k_{2T} will be at least of order J^m. Even when J is moderate, this will introduce a huge amount of parameters to be estimated.

Asymptotic properties of $(\hat{\beta}^{(1)}, \hat{F})$. Several recent papers characterize the asymptotic properties of sieve estimators in a general framework defined by conditional moment restrictions.[31] In Appendix C, we leverage the mixed-logit setting of model (2) and provide a set of low-level sufficient conditions for the consistency of $(\hat{\beta}^{(1)}, \hat{F})$ that verify the conditions proposed by Ai and Chen (2003) and Newey and Powell (2003). Applied researchers often wish to conduct inferences on functionals of $(\beta^{(1)}, F)$ (e.g., price elasticities, equilibrium prices, consumer surplus changes) on the basis of $(\hat{\beta}^{(1)}, \hat{F})$, i.e., plug-in estimators of such functionals. When one is only interested in the inference of the parametric part $\beta^{(1)}$ (or its functions), the inference procedure is quite standard: under appropriate regularity conditions, the sieve estimator $\hat{\beta}^{(1)}$ is asymptotically normal and can attain the semiparametric efficiency bound (see section 4 and Theorem 4.1 of Ai and Chen (2003) for details). When the object of interest is also a function of F (e.g., price elasticities), the inference becomes less standard. One can adopt the penalized version of (6) proposed by Chen and Pouzo (2015) and conduct such inferences using sieve Wald statistic (or its bootstrap version).[32] In section 4, we conduct a Monte Carlo study of bootstrap confidence intervals for these functions using their plug-in sieve BLP estimators and provide some practical suggestions to improve their finite-sample coverages.

[30] If we choose $\{\phi_k\}_{k=1}^{k_{1T}}$ as polynomial basis up to order m, then the number of available moment conditions is of at least order of P^m. Moreover, the valid instruments usually include $X_t = (X_t^{(1)}, X_t^{(2)})$ and hence $P \geq K_1 + K_2$. As a consequence, The requirement (8) is generically satisfied.

[31] See Ai and Chen (2003), Newey and Powell (2003), Chen (2007); Chen and Pouzo (2015); Chen and Qiu (2016) among others.

[32] See Theorem 5.2 of their paper. A related paper is Chen and Christensen (2018) that provides bootstrap-based uniform confidence bands for a collection of nonlinear functionals in a nonparametric instrumental variables model. However, the paper does not cover model (2). In a fully nonparametric setting of demand estimation, Compiani (2019) shows that the plug-in estimators of functionals are asymptotically normal.

3.3 Choices of $\{\phi_k\}_{k=1}^{k_{1T}}$ and $\Theta_{k_{2T},F}$

In this section, we provide some practical guidances on the choice of $\{\phi_k\}_{k=1}^{k_{1T}}$ and $\Theta_{k_{2T},F}$ in the setting of model (2).

Denote by Θ_Z the functional space which $m_j(Z;\beta^{(1)},F)$, $j \in \mathbf{J}$, belong to and by $\Theta_{k_{1T},Z}$ the sieve space generated by $\{\phi_k\}_{k=1}^{k_{1T}}$. A natural choice of $\{\phi_k\}_{k=1}^{k_{1T}}$ is polynomial basis of Z. The first-order polynomials of Z include exogenous product characteristics $X = (X^{(1)}, X^{(2)})$ and/or excluded instruments (e.g., cost shifters). One can leverage certain symmetries among products imposed in model (2) to construct polynomial basis that improve efficiency. For example, Gandhi and Houde (2019) assume the exchangeability of $\{\xi_{jt}\}_{j\in\mathbf{J}}$ and propose differentiation IVs as basis functions.[33]

Depending on the support and tail behavior of F, applied researchers can use different sieve spaces $\Theta_{k_{2T},F}$. For instance, for F with unbounded support and thin tail, one can use sieve space of Hermite form:[34]

$$\Theta_{k,F} = \{P_k^2(v)\varphi^2\left(v;\Sigma_F\right) : P_k \in \mathscr{P}_k\}, \tag{9}$$

where $\mathscr{P}_k$ is the set of polynomials of $v \in \mathbb{R}^{K_2}$ up to order k and $\varphi\left(v;\Sigma_F\right)$ is the density function of centered Gaussian random vector v with $\Sigma_F = \operatorname{diag}\left(\sigma_1^2,...,\sigma_{K_2}^2\right)$ being the variance-covariance matrix. For F with bounded support, one can use a sieve space of similar form:

$$\Theta_{k,F} = \{P_k^2(v)\varphi_\Omega^2\left(v\right) : P_k \in \mathscr{P}_k\}, \tag{10}$$

where $\varphi_\Omega(v)$ is a (known and smooth enough) positive density function defined on $\Omega \subset \mathbb{R}^{K_2}$. In particular, one can choose $\varphi_\Omega(v)$ as the density of uniform distribution on Ω.[35]

Because F is a probabilist distribution function, we should ensure for any $F \in \Theta_{k_{2T},F}$, $F \geq 0$ and $\int_\Omega dF = 1$. Empirical researchers often specify the mean of $\beta_i^{(2)}$, $\beta^{(2)}$, in

[33]For product j, differentiation IVs are defined as functions of characteristics differences relative to product j. Under the exchangeability assumption, their Proposition 2 proves that the σ_j^{-1} is a symmetric function of characteristics differences relative to product j. Then, one can use the sum of characteristics differences relative to product j (first-order term) and their higher-order terms as basis functions to approximate σ_j^{-1}. These terms are all polynomial of Z.

[34]See Gallant and Nychka (1987) and Chen (2007) for more details of the approximation properties of the Hermite form sieve space.

[35]In the setting of nonparametric likelihood estimation of the distribution of random coefficients, Fox et al. (2016) propose a estimator of F with compact support using a growing grid of points in Ω. Their estimator is computationally attractive in the setting of likelihood estimation. Adopting this approach in model (2) is beyond the scope of this paper. We leave this as future research.

$(\delta_j)_{j\in\mathbf{J}}$ and, consequently, the first moments of F are restricted to be zero, which adds K_2 restrictions. When using sieve spaces (9) or (10), such moment restrictions on F become simple quadratic functions of parameters to be estimated (i.e., the polynomial coefficients in P_k). See Appendix D for more details.

4 Monte Carlo Simulations

In this section, we conduct Monte Carlo simulations to investigate finite-sample performance of the sieve BLP estimator. These simulation exercises serve two purposes. First, we compare the performance of the BLP sieve estimator to that of parametric approach in different scenarios (e.g., bounded/unbounded random coefficients, the parametric approach is correctly or wrongly specified, different numbers of products). Specifically, we want to compare the precision of estimates and of the corresponding confidential intervals (coverage probability and length). Second, we compare sive BLP estimators with different configurations (e.g., different choices of Σ_F, different degrees of polynomials in sieve spaces (9)) and provide practical suggestions for applied researchers.

Settings. We simulate 100 independent samples by randomly perturbing product-market specific demand shocks $((\xi_{tj})_{j\in\mathbf{J},t\in\mathbf{T}})$. Each sample consists of $T = 500$ markets and J products. For each $t = 1, ..., T$, we simulate J observed market shares using model (2) in which price coefficient α_i is random. We generate J equilibrium prices from a Bertrand pricing competition in a duopoly and each product is produced by one firm with constant marginal cost. For each sample, we implement a parametric BLP estimator and sieve BLP estimators using the same set of differentiation instrumental variables along the lines of Gandhi and Houde (2019). For details of the data generating process, see Appendix E.

Tables 1 reports the Root of Mean Square Errors (RMSE) of estimated demand parameters (α, β, δ), density function of α_i for $J = 25$ across the 100 samples. In the top and middle panels, the parametric model is correctly specified. In the top panel, α_i is assumed to follow a Gaussian distribution. The sieve estimators are obtained using polynomial sieve spaces of form (9). In the middle panel, α_i is uniformly distributed in $[-1, 1]$. The parametric estimation assumes uniform distribution of α_i in $[-a, a]$, $a > 0$, and is correctly specified. The sieve estimators are obtained using polynomial sieve spaces of form (10). In the bottom panel, α_i is assumed to be distributed according to a mixture of two Gaussian distributions. The parametric estimation still assumes Gaussian specification and is misspecified. The sieve estimators are obtained using the same sieve spaces as those in the first panel. For the sieve estimates in each specification, we report results obtained by using 4 sieve BLP estimators. Those in the second and third rows of each panel are obtained by using scaled Hermite sieve space (9) with $f(v;\hat{\sigma}) = P_d^2(v/\hat{\sigma})\varphi_0\left(v/\hat{\sigma}\right)/\hat{\sigma}$, where φ_0 is the standard normal distribution, $\hat{\sigma}$ is the parametric estimate of the standard deviation of α_i, and $d = 2, 3$. Differently, those on the fourth and fifth rows are obtained using $f(v;\sigma_0)$ where σ_0 is the true standard deviation of α_i. Theoretically, as long as the random variable is continuous, the choice of σ in (9) does not affect the approximation properties of the sieve space. In finite sample, however, different choices of σ may affect approximation quality of the sieve spaces and lead to different precisions. We aim to investigate this issue and provide practical suggestions regarding the choice of Σ_F in (9) and (10).

Precision of sieve BLP estimators. When the parametric model is correctly specified (Specifications I-II), sieve BLP achieves similar (or better) precision to the parametric approach. In particular, in the middle panel, when we use scaled sieve estimators with the true scale of α_i (i.e., the true support a_0 in Specification II), the sieve BLP estimates of the average price coefficient α (first column) and the density function of α_i (forth column) have much smaller RMSEs than their parametric estimates. When the parametric model is misspecified (Specification III), sieve BLP can achieve better performance than the parametric approach when the dimension of the sieve space, d, is well chosen. When we use Hermite sieve space with $d = 3$, no matter which scale we use (the parametric estimate $\hat{\sigma}$, or its true value σ_0), the sieve estimates of demand parameters, density function unanimously achieve smaller RMSEs than their parametric estimates. On one hand, sieve space with larger d approximates better the true distribution of α_i, attenuating misspecification error in finite sample. On the other, enlarging sieve space introduces also more parameters to be estimated in finite sample, which may decrease the precision of the inference (i.e., larger confidence interval. See Table 2). In practice, we suggest to implement sieve estimators with different d's, and choose $d = d^*$ such that the sieve estimators with $d > d^*$ generate similar estimates and counterfactual

Table 1: RMSE, Parametric and Sieve Estimations, $J = 25$, $T = 500$

	Demand parameters			Density function
	$\alpha = -3$	$\beta = 1$	$\{\delta_j\}_{j \in \mathbf{J}}$	f
Specification I	$\alpha_i \sim \mathscr{N}(0, 0.64)$			
Parametric	0.0991	0.0091	0.1262	0.1067
Hermite Sieve with $\hat{\sigma}$				
$d = 2$	0.1312	0.0091	0.1693	0.1494
$d = 3$	0.1233	0.0091	0.1585	0.1547
Hermite Sieve with $\sigma_0 = 0.8$				
$d = 2$	0.0900	0.0087	0.1183	0.1436
$d = 3$	0.0947	0.0089	0.1226	0.1488
Specification II	$\alpha_i \sim \mathscr{U}([-1, 1])$			
Parametric	0.0886	0.0091	0.1129	0.2042
Polynomial Sieve with $\hat{a}$				
$d = 2$	0.0894	0.0091	0.1251	0.1635
$d = 3$	0.0900	0.0091	0.1418	0.1679
Polynomial Sieve with $a_0 = 1$				
$d = 2$	0.0238	0.0090	0.0924	0.0480
$d = 3$	0.0245	0.0090	0.1146	0.0642
Specification III	$\alpha_i \sim 0.5\mathscr{N}(-1, 0.2) + 0.5\mathscr{N}(1, 0.2)$			
Parametric	0.2280	0.0098	0.2711	0.4308
Hermite Sieve with $\hat{\sigma}$				
$d = 2$	0.2660	0.0094	0.3144	0.4275
$d = 3$	0.1565	0.0092	0.1993	0.4241
Hermite Sieve with $\sigma_0 = \sqrt{1.04}$				
$d = 2$	0.2005	0.0088	0.2358	0.4183
$d = 3$	0.1317	0.0091	0.1681	0.4172

Notes: The RMSEs are computed on the basis of 100 independently simulated samples, each with $J = 25$ products and $T = 500$ markets. For product-specific intercepts $\{\delta_j\}_{j \in \mathbf{J}}$, we report the median of the RMSEs of the estimates $\{\hat{\delta}_j\}_{j \in \mathbf{J}}$. For the density function, we report the root of median of $\mathbb{E}\left[(\hat{f} - f)^2\right]$, where f is the true density function and $\mathbb{E}$ is with respect to the true distribution f. All the sieve estimates are obtained by using identity weighting matrix, i.e., $\Sigma(Z) = \mathbf{I}$.

predictions.

Choice of Σ_F. In all the three specifications, sieve estimators obtained by using the true scale σ_0 (fourth and fifth rows in each panel) achieve smaller RMSEs than those obtained by using the parametric estimate $\hat{\sigma}$. Intuitively, using sieve space with the true σ correctly captures the scale of the random coefficient. Unfortunately, the true scale is unknown to the researcher in practice. In contrast, the parametric estimate $\hat{\sigma}$ provides an approximation of σ. Interestingly, when the parametric estimation is misspecified (bottom panel), which is probably the most relevant scenario to empirical research, the sieve BLP estimator obtained by using $\hat{\sigma}$ (third row) achieves comparable RMSE to that obtained by using the true scale σ (fifth row), and the RMSE is much smaller than that of the parametric estimator. This suggests the use of parametrically estimated scale $\Sigma_F = \text{Diag}\left(\hat{\sigma}_1^2, ..., \hat{\sigma}_{K_2}^2\right)$ in (9) and (10).

We replicate the same exercise for $J = 50$ (see Table 5 in Appendix E). These findings remain valid.

Bootstrap confidence intervals. In this section, we investigate the precision of the bootstrap confidence intervals for demand parameters in Table 1 and self-price elasticities evaluated at observed prices by using their plug-in estimators. In particular, we compare the coverages obtained by using parametric and sieve BLP estimators in the case of misspecification. To this purpose, we focus on Hermite sieve with $\hat{\sigma}$ in the setting of specification III in Table 1. Each bootstrap confidence interval is constructed using 200 bootstrap estimators. To compute the coverage probability of a confidence interval, we independently repeat the construction of a confidence interval 500 times and report the average coverage. s

Table 2 reports the results. For product-specific intercepts $\{\delta_j\}_{j\in\mathbf{J}}$ and self-price elasticities $\{\epsilon_{jj}\}_{j\in\mathbf{J}}$, we report average coverage of bootstrap confidence intervals across $J = 25$ products. The first two columns correspond to quantile intervals that achieve asymptotically 95% coverage, with the first using symmetric quantiles and the second asymmetric quantiles. The third column corresponds to symmetric quantile interval that achieve asymptotically 99% coverage. The panel "parametric" reports the results obtained by using the parametric estimation in the bottom panel of Table 1. As most empirical researchers, we implement a parametric bootstrap procedure for this panel. The panels $d = 2$ and $d = 3$ report results obtained by using respectively sieve BLP estimators with $d = 2$ and $d = 3$. For these panels, we implement an empirical bootstrap procedure.

The first finding is that all the coverages obtained by using the parametric approach are zero. Intuitively, and without surprise, when the distribution of the random coefficients is misspecified, the parametric estimates are inconsistent. In contrast, sieve BLP achieves reasonable finite-sample coverage. For α, β, and $\{\delta_j\}_{j\in\mathbf{J}}$, the confidence intervals of level 95% all achieves coverages near or above 95%. For self-price elasticities, the coverage is lower than the level of the corresponding confidence interval, but it

Table 2: Bootstrap Confidence Intervals, Coverage

Quantile interval (%)	$[2.5, 97.5]$	$[4, 99]$	$[0.5, 99.5]$
Parametric			
α	0	0	0
β	0	0	0
$\{\delta_j\}_{j\in\mathbf{J}}$	0	0	0
$\{\epsilon_{jj}\}_{j\in\mathbf{J}}$	0	0	0
Hermite Sieve with $\hat{\sigma}$, $d = 2$			
α	98.20%	100%	100%
β	100%	100%	100%
$\{\delta_j\}_{j\in\mathbf{J}}$	99.33%	94.60%	100%
$\{\epsilon_{jj}\}_{j\in\mathbf{J}}$	6.57%	38.25%	57.26%
Hermite Sieve with $\hat{\sigma}$, $d = 3$			
α	99.80%	98.80%	100%
β	100%	100%	100%
$\{\delta_j\}_{j\in\mathbf{J}}$	100%	100%	100%
$\{\epsilon_{jj}\}_{j\in\mathbf{J}}$	15.00%	43.17%	61.91%

improves substantially as d increases. Intuitively, for any finite d, the sieve space in use is a parametric space and does not cover the true DGP used in Table 2. As d increases, the sieve space becomes larger, and we should expect that the bias diminishes and the coverage improves. Moreover, the existence of this finite-sample bias also implies that confidence intervals using asymmetric quantiles may achieve better finite-sample coverage. The results in the second column confirm this point: we use quantiles 4% and 99% to construct the confidence interval, and the coverage is largely improved. Finally, for self-price elasticities, confidence intervals of level 99% achieve much better coverage than those of level 95% (third column). Interestingly, their lengths (see Table 6 in Appendix F) do not increase substantially relative to those of intervals of level 95%. In particular, the lengths are still much smaller than the magnitude of self-price elasticities (≈ -3.5). These findings have two implications. First, they further confirm the severity of estimating a parametric BLP when the parametric distribution is misspecified and suggest the use of sieve BLP to alleviate this misspecification error in inference. Second, these results suggest that the researcher can implement sieve BLP with potentially large d (as long as the number of moment conditions is greater than that of unknowns), compare bootstrap confidence intervals of the same level using symmetric and asymmetric quantiles, and vary the confidence level, to check the robustness of empirical findings.

5 Empirical Application

In this section, we apply sieve BLP to investigate the welfare implications of sugar tax in the RTE cereal industry in the US.

Motivation. Excessive sugar consumption is known to be linked with many health problems and particularly detrimental to children (WHO, 2015). RTE cereals marketed to children usually contain more sugar than average, leading to potentially excessive sugar intake for children who regularly consume RTE cereals.[36] Sugar tax provides a policy tool to reduce excessive sugar consumption of targeted demographic groups.[37]

In general, introducing a sugar tax has two opposite effects: direct consumer welfare loss (due to increased prices by the tax) and potential gain from reduced sugar intake.[38] Researchers often aim to quantify these two effects for the different demographic groups (e.g., households of different sizes). To this end, allowing for flexibly distributed (unobserved) individual heterogeneity in preference is crucial. Routinely used parametric approach may impose too restrictive assumptions on the distribution of individual heterogeneity, failing to meet this requirement and potentially leading to erroneous predictions. In contrast, the proposed sieve BLP allows for nonparametrically distributed random coefficients in preference and fits well this empirical requirement.[39]

This illustrative exercise with real data accents this economic relevance of our proposed method. First, we estimate the demand for RTE cereal using the proposed sieve BLP. The estimation results suggest complex (but realistic) preference heterogeneity across households, especially sugar preference. Second, we simulate a sugar tax on high-sugar RTE cereals. We find large heterogeneous effects of the tax across households of different sizes, in both consumer surplus loss and reduction of sugar intake from RTE

[36] American Heart Association recommends 6 teaspoons of added sugar per day for children between 2 and 18 years old (see Vos et al. (2017)). While, a report published in 2014 by Environmental Working Group, an American activist group, estimated that eating one serving per day of a kids' cereal containing the average amount of sugar would already consume nearly 1000 teaspoons of sugar in a year.

[37] Many papers have studied the impact of sugar tax in popular product categories such as soft drinks (also known as soda tax). See Allcott et al. (2019a,b), Dubois et al. (2020) for some recent examples.

[38] The former is quantified by the change of consumer surplus, and the latter is proxied by the reduction of sugar intake from consuming the products in consideration, before and after the introduction of sugar tax.

[39] Another approach to deal with this issue is to use micro data that contains individual-level transactions and estimate heterogeneity in preference by using individual fixed effects (see Dubois et al. (2020) for example). Differently, our method (as standard BLP models) relies on market-level data that is largely available to researchers in most industries. In practice, the two approaches are complementary; when micro data is not (yet) available, one can use our method as a first-step investigation.

cereals. Finally, we replicate the excise using a parametric approach (i.e., Gaussian). The demand estimates understate heterogeneity in preference across households, especially those of different sizes. Consequently, in contrast to the welfare predictions by BLP sieve, the parametric approach predicts almost the same magnitude of consumer surplus loss and sugar intake reduction for small and large-size households.

5.1 Data Description

We use the store-week level datasets of the RTE cereal category from the IRI data.[40] We focus on the period 2008-2011 and the city of Pittsfield in the US. The dataset contains $t = 1, ..., 1381$ markets, each being defined as a combination of store and week. In each market, the sales (in lbs and dollars) of RTE cereals are observed at Universal Product Code (UPC) level. We define a RTE cereal product j as a combination of brand and flavour (e.g., Kellogg's Special K Fruit & Yoghurt) The sales of product j in market t is the sum of the sales in lbs of all the UPC's that j collects. The price of j in market t, p_{jt}, is defined as the ratio between its sales in dollars and in lbs. We define the choice set as the union of the 50 largest RTE cereal products (in terms of sales in lbs), denoted by $\mathbf{J}$, and the outside option, denoted by 0.[41] The outside option groups all the other smaller products and other forms of breakfast cereal consumption (e.g., cereal biscuit).[42] For each market, we consider the weekly consumption of breakfast cereals as the market size. The market size for RTE cereal category is the product of weekly per capita consumption of breakfast cereals and population size in this market. Finally, the market share of $j \in \mathbf{J}$ in market t is then the ratio between its total sales in lbs and the market size. Table 7 in Appendix F provides detailed descriptive statistics of the products.

5.2 Empirical Specification

For household i in market t, the indirect utility from purchasing product j is:

$$U_{ijt} = -\alpha_i \text{Price}_{jt} + \beta_i \text{Sugar}_j + \gamma_i \text{Flavor}_j + \delta_j + \xi_{jt} + \varepsilon_{ijt}$$

[40] The IRI data has been used in the empirical literature of demand (see Nevo (2000, 2001)). We refer to these papers and also Bronnenberg et al. (2008) for a thorough discussion.

[41] Not all the 50 products are available in each market. We ignore the notation t in $\mathbf{J}$ for simplicity.

[42] The total sales quantity of the top 50 RTE cereal products represents 67% of that of the RTE cereal category in the IRI data in Pittsfield between 2008 and 2011.

where Sugar_j is the sugar amount per 100 grams of product j, δ_j is product j specific intercept, $\text{Flavor}_j = 1$ if product j is flavored and 0 otherwise, and ξ_{jt} is unobserved demand shock of product j in market t.[43] Random coefficients α_i, β_i, and γ_i capture household i's price sensitivity, sugar preference, and flavor preference, respectively. We specify $\alpha_i = \alpha_g + \Delta\alpha_i$ with $g = 1, 2, 3$ being income group of i, $\beta_i = \beta_f + \Delta\beta_i$, $\gamma_i = \gamma_f$ with $f = 1, 2, 3+$ being household size of i where $\Delta\alpha_i$ and $\Delta\beta_i$ respectively represent unobserved heterogeneity in price sensitivity and sugar preference and are jointly distributed according to F.[44] The market share of product j in market t is then

$$s_{jt} = \int \sum_{f,g=1}^{3} w_{gf} \frac{\exp\{-\alpha_i \text{Price}_{jt} + \beta_i \text{Sugar}_j + \gamma_f \text{Flavor}_j + \delta_j + \xi_{jt}\}}{1 + \sum_{j'=1}^{50} \exp\{-\alpha_i \text{Price}_{j't} + \beta_i \text{Sugar}_{j'} + \gamma_f \text{Flavor}_{j'} + \delta_{j'} + \xi_{j't}\}} dF(\Delta\alpha_i, \Delta\beta_i), \tag{11}$$

where w_{gf} is the weight of demographic group with income g and family size f.[45]

5.3 Demand Estimates

Table 3 reports demand estimates obtained by using parametric (columns 1-2) and sieve BLP estimators (columns 3-4). In the parametric approach, we assume that $(\Delta\alpha_i, \Delta\beta_i)$ follows a joint normal distribution. In column 1, we estimate the full co-variance matrix of $(\Delta\alpha_i, \Delta\beta_i)$. We find that the estimated correlation between $\Delta\alpha_i$ and $\Delta\beta_i$ is almost -1. In column 2, we restrict the correlation to be -1. We find that the point estimates of other parameters remain the same and the difference in the value of GMM objective function is negligible. As a result, we use the estimates in column 2 as the parametric benchmark. In columns 3-4, we use Hermite sieve spaces (9) with scaling parameters $\Sigma_F = \text{Diag}(\hat{\sigma}_\alpha^2, \hat{\sigma}_\beta^2)$ and $d = 4, 5$, where $\hat{\sigma}_\alpha$ and $\hat{\sigma}_\beta$ is the parametric estimates. Column 5 reports the bootstrap confidence interval of level 95% using sieve BLP estimator $d = 4$ and symmetric quantiles.[46] For all the columns, we use the same instruments: the price of the same product in the same week-chain combination in other cities (e.g., Boston),

[43] The IRI data does not contain precise sugar content information. We collect such information by matching brand and flavor of product j with its nutrition label obtained from the website of j's producer.

[44] Household size 3+ refers to household size equal to or larger than 3, i.e., household with child(ren). Also, note that product-specific intercept δ_j in U_{ijt} absorbs $\beta_1 \text{Sugar}_j$ and $\gamma_1 \text{Flavor}_j$. As a consequence, we will estimate the differences $\beta_f - \beta_1$ and $\gamma_f - \gamma_1$ for $f = 2, 3+$.

[45] When constructing w_{gf}, we adjust the population weight of each group by its size f. This is to control for the mechanical difference in breakfast cereal consumption due to having more (or fewer) persons in a household.

[46] Using asymmetric quantiles or higher confidence level to construct the confidence intervals does not change our empirical findings.

the absolute difference of sugar content from the average of other products available in the market, and the interactive term of the two.[47]

Table 3: Demand estimates, parametric approach and sieve BLP

Specification	Parametric, Gaussian	Constrained	Hermite sieve $d=4$	$d=5$	Confidence Interval $d=4$, 95%
Price coefficients					
α_1	-0.7100	-0.7100	-0.7481	-0.6967	$[-0.8562, -0.6332]$
	(0.1099)	(0.0234)	(0.0550)		
$\alpha_2 - \alpha_1$	-0.0609	-0.0609	-0.1113	-0.1081	$[-0.2603, -0.0118]$
	(0.0216)	(0.0214)	(0.0601)		
$\alpha_3 - \alpha_1$	0.2136	0.2136	0.2599	0.2660	$[0.0668, 0.3342]$
	(0.0128)	(0.0120)	(0.0716)		
Sugar preference					
$\beta_2 - \beta_1$	0.0268	0.0268	-0.0548	-0.0489	$[-0.2931, 0.0958]$
	(0.0272)	(0.0274)	(0.1073)		
$\beta_{3+} - \beta_1$	0.0212	0.0212	0.1628	0.1661	$[0.0140, 0.2944]$
	(0.0188)	(0.0189)	(0.0778)		
Flavor preference					
$\gamma_2 - \gamma_1$	0.0020	0.0020	2.6638	2.6580	$[-1.0109, 6.6412]$
	(0.1952)	(0.1952)	(2.3355)		
$\gamma_{3+} - \gamma_1$	-0.0007	-0.0007	-4.0207	-4.0203	$[-7.9577, -0.2806]$
	(0.1336)	(0.1336)	(2.2112)		
Unobs. preference					
σ_α	0.1326	0.1326	0.1518	0.1424	$[0.1051, 0.1970]$
	(0.0140)	(0.0132)	(0.0250)		
σ_β	0.0251	0.0251	0.0324	0.0305	$[0.0233, 0.0423]$
	(0.0533)	(0.0028)	(0.0049)		
$\mathrm{Corr}(\Delta\alpha_i, \Delta\beta_i)$	-1.000	-1	-0.4958	-0.4885	$[-0.7886, -0.2512]$
	(0.2570)		(0.1463)		
Val. of Obj. Fun. ($\times 10^{-3}$)	1.310	1.310	1.236	1.216	

Notes: The standard errors in columns 1 and 2 are computed using the estimated asymptotic variance-covariance matrix. The confidence intervals in column 5 and the standards errors in column 3 are computed using 200 bootstrap samples and symmetric quantiles (2.5% and 97.5%).

Points estimates of the price coefficients are quantitatively similar across all specifications (top panel), while those of sugar and flavor preferences differ much between the parametric and sieve BLP approaches, especially for large household, i.e., household with child(ren) represented by 3+. First, large household is estimated to slightly more sugar-seeking than single-person household (positive $\beta_{3+} - \beta_1$) by the parametric approach, but the difference is insignificant. This difference is estimated to be much larger and significantly different from zero by the sieve BLP approach. Second, the extent to which large household prefers flavored RTE cereals is estimated to be almost identical to that of smaller-size households by the parametric approach, whereas the sieve BLP estimates show that large household significantly dislikes flavored RTE cereals more than other

[47] The second instrument measures substitutions between products along the dimension of sugar content. Even though sugar content of a product does not vary across markets, the availability of the product may differ across markets. Consequently, the second instrument also varies across markets.

households. Finally, the estimated distribution of unobserved heterogeneity $(\Delta\alpha_i, \Delta\beta_i)$ differs significantly as well: the parametric approach suggests an almost perfect negative correlation between $\Delta\alpha_i$ and $\Delta\beta_i$, while this negative correlation is estimated to be mild (still significantly) by the sieve BLP estimators.

These findings suggest that restricting the distribution of the random coefficients as in the parametric approach has non-negligible impact on the estimates of other parameters in preference.[48] In particular, these restrictions seem to substantially understate heterogeneity in sugar preference across households of different sizes. This may lead to misleading policy predictions in which such individual heterogeneity plays an important role, e.g., distributional effects (sugar intake reduction, consumer surplus loss) of sugar tax. We investigate these issues more in detail in the next section.

5.4 Counterfactual: tax on excessive sugar

In this section, we simulate a tax on high-sugar RTE cereal products and compare welfare predictions of the parametric and sieve BLP approaches.[49] To do so, we assume that the observed prices are generated from a Bertrand price-setting game with complete information and that the marginal production cost of product j in market t, c_{jt}, is constant (not a function of production quantities). Then, given demand estimates of each approach, we first recover c_{jt} using the First-Order Conditions (FOCs) of the price-setting game for all j's and t's. Second, we define a tax as a function of the amount of excessive sugar that product j contains. We adopt the definition of excessive sugar by the USDA (Food and Nutrition Service) and consider RTE cereals with no less than 21.2g of sugar per 100 grams as high-sugar products.[50] The tax added on the ex-factory price of product j is the defined as:

$$t_j = \$\max\{0.05 \times (\text{Sugar}_j - 21.2), 0\}. \tag{12}$$

[48]We also estimate the model using Hermite sieve with $d = 2$ and 3. We find that using these more restrictive sieve spaces delivers similar results to the parametric approach. Moreover, the value of the objective function is larger than that of the parametric approach. This suggests that sieve spaces with small d's may not be flexible enough to capture the true distribution of the random coefficients, resulting in a worse fit of the data. In contrast, the sieve estimates using $d \geq 4$ remain stable and the value of the objective function is smaller than that of the parametric approach.

[49]We implicitly assume that RTE cereal products grouped in the outside option are not affected by the tax. As a consequence, in this counterfactual exercise, only high-sugar major RTE cereals are taxed.

[50]See the breakfast cereals section of https://www.fns.usda.gov/cacfp/grain-requirements-cacfp-qas.

Only products with more than 21.2g sugar per 100 grams will be taxed. For example, General Mills Lucky Charms contains 33.3g sugar per 100 grams and the ex-factory price set by General Mills is \$3 per lbs. Then, the tax is $(33.3 - 21.2) \times 0.05 = \$0.65$ and the final price of General Mills Lucky Charms faced by consumers is \$3.65 per lbs. When computing the new equilibrium prices, imposing such a tax t_j on product j is equivalent to assuming that marginal costs of production are increased to $c_{jt} + t_j$ and that producers directly set final prices based on new marginal costs $c_{jt} + t_j$, $j \in \mathbf{J}$.

Table 4: Counterfactuals, tax on excessive sugar

Specification	Parametric Gaussian	Hermite sieve $d = 4$	Hermite sieve $d = 5$	Confidence Interval $d = 4$, 95%
Average				
ΔConsumer surplus	−9.22%	−9.78%	−9.71%	[−10.87%, −8.61%]
		(0.58%)		
Sugar intake before tax	7.6175	7.6175	7.6175	
ΔSugar intake (g per person)	−1.0027	−1.0353	−1.0327	[−1.1533, −0.9000]
		(0.0662)		
Income groups				
ΔConsumer surplus				
Low income	−10.62%	−11.46%	−11.46%	[−13.27%, −8.89%]
		(1.14%)		
Medium income	−11.93%	−13.27%	−12.23%	[−14.94%, −10.84%]
		(2.08%)		
High income	−6.94%	−7.19%	−7.07%	[−8.70%, −6.03%]
		(0.67%)		
ΔSugar intake (g per person)				
Low income	−0.9782	−0.9949	−0.9914	[−1.1090, −0.8506]
		(0.0678)		
Medium income	−0.9879	−1.0003	−1.0081	[−1.0905, −0.8868]
		(0.0516)		
High income	−1.0405	−1.0990	−1.0880	[−1.2730, −0.9340]
		(0.0908)		
Family sizes				
ΔConsumer surplus				
Family size ≤ 2	−9.16%	−6.92%	−6.93%	[−8.50%, −3.70%]
	(1.26%)	(1.29%)		
Family size ≥ 3	−8.34%	−10.60%	−10.50%	[−12.99%, −9.39%]
	(1.49%)	(0.97%)		
Sugar intake before tax				
Family size ≤ 2	7.3532	5.4747	5.5588	[3.3664, 6.2035]
		(0.7895)		
Family size ≥ 3	7.8506	9.4976	9.4246	[8.8600, 11.3507]
		(0.6932)		
ΔSugar intake (g per person)				
Family size ≤ 2	−1.0065	−0.7348	−0.7461	[−0.9356, −0.3546]
		(0.1515)		
Family size ≥ 3	−1.0004	−1.3015	−1.2867	[−1.6193, −1.1466]
		(0.1236)		

Notes: The standard errors in column 1 are computed using parametric bootstrap method that resamples from the asymptotic distribution of the estimators. The confidence intervals in column 4 and the standard errors in column 2 are computed using 200 bootstrap samples and symmetric quantiles (2.5% and 97.5%). ΔSugar intake is measured by gram of sugar per person for 100g RTE cereal consumption.

Table 4 summarizes the results. First, in the top panel, both the parametric and sieve BLP approaches predict quantitatively similar average effects of the tax: it reduces consumer surplus by around 9.5% and per capita sugar intake by 1g (or, 13.12%) for 100g RTE cereal consumption. Second, in the middle panel, both the parametric and sieve BLP approaches predict that consumer surplus loss of lower-income households is significantly larger than high-income households, while both households' per capita sugar intake reductions are almost the same. This is aligned with the finding in the existing literature that excise type taxes (such as soda tax) are regressive: the poor often bear more the burden of the tax than the rich (Allcott et al., 2019a,b; Dubois et al., 2020; Gruber and Kőszegi, 2004). Third, the most divergent predictions lie in the effects of the tax on households of different sizes. In the bottom panel, the parametric approach predicts almost homogeneous per capita sugar intake before tax, consumer surplus loss, and per capita sugar intake reduction due to the tax for households of large (family size ≥ 3, i.e., family with child(ren)) and small (family size ≤ 2, i.e., family without child) sizes, whereas the predictions by the sieve BLP approach are much more heterogeneous. When using Hermite sieve with $d = 4$, we find that a person in large-size household has 73% more sugar intake than a person in small-size household before tax. Moreover, large-size household's surplus loss is almost 54% larger than that of small-size household, and a person in large-size household reduces her sugar intake 78% more than a person in small-size household due to the tax.

This counterfactual exercise has two important implications. First, when the objects of interest are population-average welfare outcomes, the parametric and sieve BLP approaches seem to agree on the predictions of these outcomes. The applied researcher can use sieve BLP as a robustness check for the simpler parametric approach. Second, when the researcher aims to quantify distributional effects of a policy change before its implementation, the parametric approach may be too restrictive to allow for flexible individual heterogeneity in preference and leads to misleading predictions. In our exercise, the parametric approach substantially understates the magnitude of individual heterogeneity and misallocate the effects of sugar tax across households of different sizes. In contrast, sieve BLP minimizes potential misspecification error in the parametric approach, and is able to produce more reliable predictions of the distributional effects of such a policy change. Therefore, we suggest to use sieve BLP when the applied researcher is also interested in predictions of distributional welfare outcomes.

6 Conclusion

In this paper, we propose a semi-nonparametric approach to identify and estimate the demand for differentiated products. This approach adopts a mixed logit model in which the distribution of random coefficients is nonparametrically specified. This method has two remarkable advantages compared to existing ones in the literature. First, it minimizes misspecification error in the distribution of random coefficients to which the usual parametric BLP is subject. Second, it overcomes the dimensionality problem in the number of products by nonparametrically estimating the distribution of random coefficients (rather than structural demand functions) whose dimensionality does not increase with the number of products in most empirical settings. We provide a new strategy to identify this distribution that only requires eat most one single variation in product characteristics interacting with the random coefficients, substantially relaxing the support requirements in the existing literature. We propose a sieve BLP estimator for the distribution of random coefficients. This estimator is simple to implement (literally as a parametric GMM) and has desirable asymptotic properties and robust finite-sample performance. We apply the proposed method to investigate the welfare implications of a sugar tax in the RTE cereal industry in the US. This illustrative application suggests the use of sieve BLP over the routinely used parametric approach when the researcher aims to quantify the distributional effects of a policy change such as introducing a sugar tax.

Appendix

A Proof of Theorem 1

The identification strategy consists of two steps. In the first step, we use Condition 1 and identify the distribution of random deviations $\mu_i = (\mu_{ij})_{j \in \mathbf{J}} = X^{(2)}\beta_i$ conditional on $X^{(2)}$. In the second step, we combine this result and Condition 2 of Theorem 1 to identify F.

A.1 Identification of the distribution of $\mu_i | X^{(2)}$

Denote the distribution function of μ_i conditional on $X^{(2)}$ by $G_{\mu|X^{(2)}}(\cdot)$. Then, according to Condition 1, we obtain that for any $j \in \mathbf{J}$,

$$\begin{aligned}\sigma_j(\delta; G_{\mu|X^{(2)}}) &= \sigma_j(\delta; X^{(2)}, F) \\ &= \int \frac{\exp\{\delta_j + \mu_{ij}\}}{1 + \sum_{j' \in \mathbf{J}} \exp\{\delta_{j'} + \mu_{ij'}\}} dG_{\mu|X^{(2)}}(\mu_i)\end{aligned}$$

is identified for all $\delta \in \mathscr{D}$. Suppose that there exist $G'_{\mu|X^{(2)}}(\cdot)$ such that $s_j(\delta; G_{\mu|X^{(2)}}) = s_j(\delta; G'_{\mu|X^{(2)}})$ for any $\delta \in \mathscr{D}$. In this section, we prove $G_{\mu|X^{(2)}} = G'_{\mu|X^{(2)}}$ in three steps.

Step 1.

Lemma 1. *Suppose that Condition 1 of Theorem 1 holds. Then, for any $X^{(2)} \in \mathscr{X}$ and $j \in \mathbf{J}$, $\sigma_j(\delta; G_{\mu|X^{(2)}}) = \sigma_j(\delta; G'_{\mu|X^{(2)}})$ for $\delta \in \mathbb{R}^J$.*

Remark 1. *When the price coefficient is homogeneous across individuals, the utility structure of model* (2) *satisfies Assumption 5 in section 4.2 of Berry and Haile (2014). Consequently, keeping other product characteristics fixed, any price change can be equivalently expressed via the change in δ. Then, the change in consumer welfare due to price change is already identified as long as the corresponding path of δ is included in $\mathscr{D}$. Lemma 1 enhances their result in mixed-logit models of demand and already allows to identify consumer welfare change due to* any *price change (and therefore any path of δ in $\mathbb{R}^J$), without identifying F. This is due to the real-analytic property of demand system* (2) *with respect to δ_t.*[51]

[51] Some papers in the literature have also employed this property in the identification and estimation of mixed-logit models of demand. See Fox et al. (2012), il Kim et al. (2014), Iaria and Wang (2019), Wang (2021).

Proof. According to Theorem 2 (Real Analytic Property) of Iaria and Wang (2019), $\sigma_j(\delta; G_{\mu|X^{(2)}})$ and $\sigma_j(\delta; G'_{\mu|X^{(2)}})$ are both real analytic with respect to δ in $\mathbb{R}^J$. Then, $\sigma_j(\delta; G_{\mu|X^{(2)}}) - \sigma_j(\delta; G'_{\mu|X^{(2)}})$ is also real analytic with respect to δ in $\mathbb{R}^J$. According to Assumption 1, $\sigma_j(\delta; G_{\mu|X^{(2)}}) - \sigma_j(\delta; G'_{\mu|X^{(2)}}) = 0$ in open set $\mathscr{D}$. Then, $\sigma_j(\delta; G_{\mu|X^{(2)}}) - \sigma_j(\delta; G'_{\mu|X^{(2)}}) = 0$ for any $\delta \in \mathbb{R}^J$. $\square$

Because of Lemma 1, we obtain that for any $\delta \in \mathbb{R}^J$, $\frac{\partial^J \sigma_0(\delta; G_{\mu|X^{(2)}})}{\prod_{j=1}^J \partial\delta_j} = \frac{\partial^J \sigma_0(\delta; G'_{\mu|X^{(2)}})}{\prod_{j=1}^J \partial\delta_j}$, where $\sigma_0 = 1 - \sum_{j=1}^J \sigma_j$. Equivalently,

$$\begin{aligned}
&\frac{\partial^J \sigma_0(\delta; G_{\mu|X^{(2)}})}{\prod_{j=1}^J \partial\delta_j} - \frac{\partial^J \sigma_0(\delta; G'_{\mu|X^{(2)}})}{\prod_{j=1}^J \partial\delta_j} \\
&= (-1)^J J! \int \frac{1}{1+\sum_{j'\in\mathbf{J}} \exp\{\delta_{j'} + \mu_{ij'}\}} \prod_{j=1}^J \frac{\exp\{\delta_j + \mu_{ij}\}}{1+\sum_{j'\in\mathbf{J}} \exp\{\delta_{j'} + \mu_{ij'}\}} d(G_{\mu|X^{(2)}} - G'_{\mu|X^{(2)}})(\mu_i) \\
&= (-1)^J J! \int \frac{1}{1+\sum_{j'\in\mathbf{J}} \exp\{\lambda_{j'}\}} \prod_{j=1}^J \frac{\exp\{\lambda_j\}}{1+\sum_{j'\in\mathbf{J}} \exp\{\lambda_{j'}\}} d(G_{\mu|X^{(2)}} - G'_{\mu|X^{(2)}})(\lambda_i - \delta) \\
&= (-1)^J J! \int \phi(\lambda_i) d(G_{\mu|X^{(2)}} - G'_{\mu|X^{(2)}})(\lambda_i - \delta) \\
&= 0,
\end{aligned} \tag{A.1}$$

where $\phi(\lambda) = \frac{1}{1+\sum_{j'\in\mathbf{J}} e^{\lambda_{j'}}} \prod_{j=1}^J \frac{e^{\lambda_j}}{1+\sum_{j'\in\mathbf{J}} e^{\lambda_{j'}}}$.

Step 2.

Lemma 2. *$\phi \in L^1(\mathbb{R}^J)$.*

Proof. First, by transforming λ to $\exp\{\lambda\}$, we obtain:

$$\begin{aligned}
\int \phi(\lambda) d\lambda &= \int_{\mathbb{R}_+^J} \left(\frac{1}{1+\sum_{j'\in\mathbf{J}} y_{j'}}\right)^{J+1} dy \\
&= \sum_{I=\otimes_{j=1}^J I_j, I_j \in \{[0,1),[1,+\infty)\}} \int_I \left(\frac{1}{1+\sum_{j'\in\mathbf{J}} y_{j'}}\right)^{J+1} dy.
\end{aligned}$$

Because there are 2^J possible I's, it then suffices to prove that for any I,

$$\int_I \left(\frac{1}{1+\sum_{j'\in\mathbf{J}} y_{j'}}\right)^{J+1} dy < \infty.$$

Denote the number of j's such that $I_j = [1, +\infty)$ by k. When $k = 0$, $\int_I (1+\sum_{j'\in\mathbf{J}} y_{j'})^{-(J+1)} dy < 1$. When $k > 0$, without loss of generality, suppose that $I_j = [1, +\infty)$ for $1 \le j \le k$, and

$I_j = [0, 1)$ for $k+1 \leq j \leq J$. Then,

$$\int_I \left(\frac{1}{1+\sum_{j' \in \mathbf{J}} y_{j'}} \right)^{J+1} dy \leq \underbrace{\int_1^{+\infty} \dots \int_1^{+\infty}}_{k} \left(\frac{1}{\sum_{j'=1}^{k} y_{j'}} \right)^{J+1} dy_1 \dots dy_k$$
$$\leq \underbrace{\int_1^{+\infty} \dots \int_1^{+\infty}}_{k} \left(\frac{1}{k \prod_{j'=1}^{k} y_{j'}^{\frac{1}{k}}} \right)^{J+1} dy_1 \dots dy_k$$
$$= \frac{1}{k^{J+1}} \left(\int_1^{\infty} y^{-\frac{J+1}{k}} dy \right)^k$$
$$= \frac{1}{k^{J+1-k}} \left(\frac{1}{J+1-k} \right)^k.$$

The transition from the first to the second line is obtained by using $\sum_{j'=1}^{k} y_{j'} \geq k(\prod_{j'=1}^{k} y_{j'})^{1/k}$.

$\square$

Because of Lemma 2, $\phi \in L^1(\mathbb{R}^J)$ and hence its Fourier transformation is well defined. Moreover, note that the right-hand side of (A.1) is a convolution of ϕ and $dG_{\mu|X^{(2)}} - dG'_{\mu|X^{(2)}}$.[52] Consequently,

$$\mathscr{F}(\phi)(v)[\psi_{G_{\mu|X^{(2)}}}(v) - \psi_{G'_{\mu|X^{(2)}}}(v)] = 0 \tag{A.2}$$

for any $v \in \mathbb{R}^J$, where $\mathscr{F}(.)$ denotes Fourier transformation and ψ_G is the characteristic function of distribution G.

Step 3.

Lemma 3. *The set* $\{v \in \mathbb{R}^J : \mathscr{F}(\phi)(v) = 0\}$ *is of zero Lebesgue measure.*

Combining (A.2) and Lemma 3, we obtain that $\psi_{G_{\mu|X^{(2)}}} = \psi_{G'_{\mu|X^{(2)}}}$ almost everywhere. Because characteristic functions are continuous, then we obtain $\psi_{G_{\mu|X^{(2)}}} = \psi_{G'_{\mu|X^{(2)}}}$ every where and hence $G_{\mu|X^{(2)}} = G'_{\mu|X^{(2)}}$. In the remaining part, we prove Lemma 3.

Proof. Note that it suffices to prove that the real (or the imaginary) part of $\mathscr{F}(\phi)$ is real analytic and not constantly zero. As long as this result is proved, according to Mityagin (2015), the zero set of the non-constant real (imaginary) part of $\mathscr{F}(\phi)$ is of zero Lebesgue measure. As a consequence, the zero set of $\mathscr{F}(\phi)$ is also of zero Lebesgue measure.

We first prove the real and imaginary parts of $\mathscr{F}(\phi)$ are real analytic. It suffices to evaluate

[52] Here $dG_{\mu|X^{(2)}} - dG'_{\mu|X^{(2)}}$ is defined as a distribution.

$\left|\frac{\partial^L \mathscr{F}(\phi)(y)}{\prod_{j=1}^J \partial y_j^{l_j}}\right|$, where $\sum_{j=1}^J l_j = L$. Note that:

$$\frac{\partial^L \mathscr{F}(\phi)(y)}{\prod_{j=1}^J \partial y_j^{l_j}} = \mathscr{F}(\prod_{j=1}^J(-\mathrm{i}\lambda_j^{l_j})\phi)(y),$$

where i is the imaginary unit. We now show that for any $y \in \mathbb{R}^J$,

$$\left|\mathscr{F}(\prod_{j=1}^J(-\mathrm{i}\lambda_j^{l_j})\phi)(y)\right| \leq 2^J J^L \prod_{j=1}^J l_j!$$

First,

$$\begin{aligned}
\left|\mathscr{F}(\prod_{j=1}^J(-\mathrm{i}\lambda_j^{l_j})\phi)(y)\right| &\leq \int \prod_{j=1}^J |\lambda_j|^{l_j}\phi(\lambda)d\lambda \\
&= \int_{\mathbb{R}_+^J} \frac{\prod_{j=1}^J |\ln y_j|^{l_j}}{(1+\sum_{j=1}^J y_j)^{J+1}} dy \\
&= \sum_{I=\otimes_{j=1}^J I_j, I_j \in \{[0,1),[1,+\infty)\}} \int_I \frac{\prod_{j=1}^J |\ln y_j|^{l_j}}{(1+\sum_{j=1}^J y_j)^{J+1}} dy.
\end{aligned}$$

Similar to the proof of Lemma 2, we evaluate $\int_I \frac{\prod_{j=1}^J |\ln y_j|^{l_j}}{(1+\sum_{j=1}^J y_j)^{J+1}} dy$ for each I. Denote the number of j's such that $I_j = [1, +\infty)$ by k. When $k = 0$,

$$\begin{aligned}
\int_I \frac{\prod_{j=1}^J |\ln y_j|^{l_j}}{(1+\sum_{j=1}^J y_j)^{J+1}} dy &\leq \prod_{j=1}^J \int_0^1 |\ln y_j|^{l_j} dy_j \\
&= \prod_{j=1}^J \int_0^{+\infty} \lambda_j^{l_j} e^{-\lambda_j} d\lambda_j \\
&= \prod_{j=1}^J l_j!
\end{aligned}$$

When $k > 0$, without loss of generality, suppose that $I_j = [1, +\infty)$ for $1 \leq j \leq k$, and $I_j = [0, 1)$

for $k+1 \leq j \leq J$. Then,

$$\begin{aligned}
\int_I \frac{\prod_{j=1}^J |\ln y_j|^{l_j}}{(1+\sum_{j=1}^J y_j)^{J+1}} dy &\leq \int_1^{+\infty} \frac{\prod_{j=1}^k |\ln y_j|^{l_j}}{(1+\sum_{j=1}^k y_j)^{J+1}} dy_1 ... dy_k \prod_{j=k+1}^J \int_0^1 |\ln y_j|^{l_j} dy_j \\
&\leq \frac{1}{k^{J+1}} \int_1^{+\infty} \prod_{j=1}^k (\ln y_j)^{l_j} y_j^{-\frac{J+1}{k}} dy_1 ... dy_k \prod_{j=k+1}^J l_j! \\
&= \frac{1}{k^{J+1}} \prod_{j=k+1}^J l_j! \prod_{j=1}^k \int_0^{+\infty} \lambda_j^{l_j} e^{-\frac{J+1-k}{k}\lambda_j} d\lambda_j \\
&= \frac{1}{k^{J+1}} \left(\frac{k}{J+1-k}\right)^L \prod_{j=1}^J l_j! \\
&\leq J^L \prod_{j=1}^J l_j!.
\end{aligned}$$

The transition from the first to the second line is obtained by using $\sum_{j=1}^k y_j \geq k(\prod_{j=1}^k y_j)^{1/k}$. Then, summing over 2^J integrals, we obtain:

$$\left| \frac{\partial^L \mathscr{F}(\phi)(y)}{\prod_{j=1}^J \partial y_j^{l_j}} \right| \leq 2^J J^L \prod_{j=1}^J l_j!.$$

Denote the real part of $\mathscr{F}(\phi)(y)$ by $\mathrm{Re}[\mathscr{F}(\phi)](y)$. Then, for any $y \in \mathbb{R}^J$,

$$\left| \frac{\partial^L \mathrm{Re}[\mathscr{F}(\phi)](y)}{\prod_{j=1}^J \partial y_j^{l_j}} \right| \leq 2^J J^L \prod_{j=1}^J l_j!.$$

Note that for y such that $|y - y_0| < J^{-2}$, the Taylor expansion of $\mathrm{Re}[\mathscr{F}(\phi)](y)$ around $y = y_0$ can be controlled by

$$\begin{aligned}
\left| \sum_{L=0}^{\infty} \frac{1}{L!} \left[\sum_{j=1}^J (y_j - y_{0j}) \frac{\partial}{\partial y_j} \right]^L \mathrm{Re}[\mathscr{F}(\phi)](y_0) \right| &\leq \sum_{L=0}^{\infty} \frac{1}{L!} d^L \sum_{\sum l_j = L} \frac{L!}{\prod_{j=1}^J l_j!} \left| \frac{\partial^L \mathrm{Re}[\mathscr{F}(\phi)](y)}{\prod_{j=1}^J \partial y_j^{l_j}} \right| \\
&\leq 2^J \sum_{L=0}^{\infty} (dJ^2)^L \\
&\leq 2^J \sum_{L=0}^{\infty} \frac{1}{2^L}.
\end{aligned}$$

The transition from the first to the second line uses $\sum_{\sum_{j=1}^J l_j = L} 1 \leq J^L$. As a result, the Taylor expansion of $\mathrm{Re}[\mathscr{F}(\phi)](y)$ converges for $|y-y_0|<J^{-2}$. Finally, for $|y-y_0|<0.5J^{-2}$,

$$\begin{aligned}
&\left|\mathrm{Re}[\mathscr{F}(\phi)](y)-\sum_{L=0}^{R}\frac{1}{L!}\left[\sum_{j=1}^{J}(y_j-y_{0j})\frac{\partial}{\partial y_j}\right]^L\mathrm{Re}[\mathscr{F}(\phi)](y_0)\right| \\
&\leq\left[\frac{1}{2J^2}\right]^{R+1}\sum_{\sum l_j=R+1}\frac{1}{\prod_{j=1}^{J}l_j!}\sup_{|y-y_0|<\frac{1}{2J^2}}\left|\frac{\partial^L\mathrm{Re}[\mathscr{F}(\phi)](y)}{\prod_{j=1}^{J}\partial y_j^{l_j}}\right| \\
&\leq 2^J\left[\frac{1}{2J^2}\right]^{R+1}J^{2(R+1)} \\
&\to 0.
\end{aligned}$$

Consequently, $\mathrm{Re}[\mathscr{F}(\phi)]$ is equal to its Taylor expansion and therefore real analytic. Similarly, we can prove that the imaginary part of $\mathscr{F}(\phi)$ is also real analytic. Moreover, because ϕ is not zero functional, then $\mathscr{F}(\phi)$ is not zero functional. As a result, either the real or the imaginary part of $\mathscr{F}(\phi)$ is not constantly zero. The proof is completed. □

A.2 Identification of F

Because the distribution of $\mu_i=X^{(2)}\beta_i$ given $X^{(2)}$ is identified, then we identify the distribution of $X^{(2)\mathrm{T}}X^{(2)}\beta_i$. According to Condition 2 of Theorem 1, $X^{(2)\mathrm{T}}$ is of full column rank and therefore $X^{(2)\mathrm{T}}X^{(2)}$ is invertible. We then identify the distribution of β_i. The proof is completed.

B Identification of F in the presence of random intercepts

In this appendix, we prove Theorem 1 in the presence of random intercepts. To simplify the exposition, we denote such random intercepts by $\eta_i=(\eta_{ij})_{j\in\mathbf{J}}$ and define μ_i as

$$\mu_i=X^{(2)}\beta_i^{(2)}+\eta_i.$$

We propose conditions under which the joint distribution of $(\beta_i^{(2)},\eta_i)$, F, is identified.

Theorem 2. *Suppose that*

1. *For any $X^{(2)}\in\mathscr{X}$ and any $j\in\mathbf{J}$, $\sigma_j(\delta;X^{(2)},F)$ is identified in $\mathscr{D}\ni\delta$, where $\mathscr{D}$ is an open set in $\mathbb{R}^J$.*
2. *There exists $X^{(2)},Y^{(2)}\in\mathscr{X}$ and $M\in\mathbb{R}^{J\times J}$, such that $X^{(2)}$ and $Y^{(2)}$ are of full-column rank, $Y^{(2)}=MX^{(2)}$, and the absolute values of the eigenvalues of M are strictly smaller than 1.*
3. *For any $X^{(2)}\in\mathscr{X}$, $\beta_i^{(2)}$ and η_i are independent conditional on $X^{(2)}$.*

Then, F is identified.

Assumption 1 of Theorem 2 is the same as that in Theorem 1. Assumption 2 of Theorem 2 is an extension of Assumption 2 of Theorem 1. Besides the full column rank requirement, Assumption 2 further requires a *single* variation in $X^{(2)}$. It is motivated by the practical issue that product characteristics may not always change much (or in a limited way) across markets. In particular, it is remarkably weaker than usual support conditions in the literature.[53] In the case of $K_2 = 1$, i.e., $\beta_i^{(2)}$ is a scalar, this assumption is implied by any continuous support $\mathscr{X}$, or non-singleton discrete support $\mathscr{X} \subset \mathbb{R}_+^J$ (e.g., two different price vectors). In the case of $K_2 > 1$, this assumption requires that characteristics matrices $X^{(2)}$ and $Y^{(2)}$ be of full-column rank and distinct in such a way that $X^{(2)}$ is "closer" to the origin than $Y^{(2)}$ in all directions defined by the eigenvectors of M. Notably, this is implied by local support condition, i.e., $\mathscr{X}$ is an open neighborhood of $X^{(2)}$: one can find a $\lambda \in (0, 1)$ and $Y^{(2)} = \lambda \mathbf{I}_{J\times J} X^{(2)} \in \mathscr{X}$ such that Assumption 2 is satisfied. Assumption 3 requires the conditional independence between $\beta_i^{(2)}$ and η_i. This independence condition is already employed in theoretical literatures (see Chernozhukov et al. (2019) and Allen and Rehbeck (2020) for examples). Assumption 3 is not necessary when variation in $X^{(2)}$ is rich enough.[54] However, when the variation in $X^{(2)}$ is limited, e.g., a single variation as in Assumption 2, F may not be identified without Assumption 3. See Appendix B.1 for more details.

Proof. First, applying the same arguments in Appendix A.1, we identify the distribution of $\mu_i = X^{(2)}\beta_i^{(2)} + \eta_i$ conditional on $X^{(2)}$, for any $X^{(2)} \in \mathscr{X}$. Using Assumption 3 of Theorem 2, we identify the following two characteristic functions: for any $\nu \in \mathbb{R}^J$,

$$\psi_{\mu|X^{(2)}}(\nu) = \psi_{\beta^{(2)}}(X^{(2)\mathrm{T}}\nu)\psi_\eta(\nu),$$
$$\psi_{\mu|Y^{(2)}}(\nu) = \psi_{\beta^{(2)}}(Y^{(2)\mathrm{T}}\nu)\psi_\eta(\nu),$$

for $X^{(2)}$ and $Y^{(2)}$ satisfying Assumption 2 of Theorem 2. We then identify the ratio $r(\nu) = \psi_{\beta^{(2)}}(X^{(2)\mathrm{T}}\nu)/\psi_{\beta^{(2)}}(Y^{(2)\mathrm{T}}\nu)$ for any $\nu \in \mathbb{R}^J$. Because $X^{(2)}$ is of full column rank, then $X^{(2)\mathrm{T}}$ is of full row rank K_2 and $K_2 \leq J$. As a consequence, for any $v \in \mathbb{R}^{K_2}$, there exists some ν such

[53] For papers using special regressor with large support, see Lewbel (2000), Berry and Haile (2009), Fox and Gandhi (2016), Fox and Lazzati (2017), Lewbel and Pendakur (2017), Dunker et al. (2017), Masten (2018) among others. For those using limited support condition together with restrictions on the location of the support or/and on the moments of the random coefficients, see, for example, Lewbel (2010), Fox et al. (2012), Masten (2018), Chernozhukov et al. (2019), Gaillac and Gautier (2019) and Allen and Rehbeck (2020).

[54] See the identification analysis of Ichimura and Thompson (1998) and Gautier and Kitamura (2013) for arguments that do not require this independence condition.

that $X^{(2)\mathrm{T}}\nu = v$. Then,

$$\begin{aligned}\psi_{\beta^{(2)}}(v) &= \psi_{\beta^{(2)}}(X^{(2)\mathrm{T}}\nu)\\ &= r(\nu)\psi_{\beta^{(2)}}(Y^{(2)\mathrm{T}}\nu)\\ &= r(\nu)\psi_{\beta^{(2)}}(X^{(2)\mathrm{T}}M\nu)\\ &= r(\nu)r(M\nu)\psi_{\beta^{(2)}}(Y^{(2)\mathrm{T}}M\nu)\\ &= \prod_{l=0}^{L} r(M^l\nu)\psi_{\beta^{(2)}}(Y^{(2)\mathrm{T}}M^L\nu).\end{aligned}$$

Because the absolute values of the eigenvalues of M is strictly smaller than 1, then $|M^L\nu| \to 0$ as $L \to \infty$. Consequently, $\psi_{\beta^{(2)}}(v) = \prod_{l=0}^{\infty} r(M^l\nu)$ and therefore identified. This implies the identification of the distribution of $\beta_i^{(2)}$. The identification of the characteristic function (and therefore the distribution) of η_i follows from

$$\psi_\eta(\nu) = \psi_{\mu_1|X^{(2)}}(\nu)/\psi_{\beta^{(2)}}(X^{(2)\mathrm{T}}\nu).$$

Because of Assumption 3 of Theorem 2, the joint distribution of $(\beta_i^{(2)}, \eta_i)$ is then identified. $\square$

B.1 Non-Identification of F without the Independence Condition

We provide an example in which F is not identified when $\mathscr{X}$ only has two points and $\beta_i^{(2)}$ and η_i are not independent. Suppose that $J = 1$, i.e., there is only one inside product, and $(\beta_i^{(2)}, \eta_i)$ follows a centered normal distribution with a covariance matrix Ω. Suppose that Ω has 3 unknowns: the variance of $\beta_i^{(2)}$, the variance of η_i, and their correlation $r \neq 0$, the support $\mathscr{X}$ has only two points: $\mathscr{X} = \{x, y\}$, and the distribution of $\mu_i = x^{(2)}\beta_i^{(2)} + \eta_i$ is identified conditional on $x^{(2)} = x, y$. Then, conditional on $x^{(2)}$, μ_i follows a centered normal distribution with the variance being $(x^{(2)}, 1)\Omega(x^{(2)}, 1)^{\mathrm{T}}$. We can identify $(x, 1)\Omega(x, 1)^{\mathrm{T}}$ and $(y, 1)\Omega(y, 1)^{\mathrm{T}}$. Without further assumptions, we obtain 2 equations with 3 unknowns. Then, Ω cannot be uniquely determined and the distribution of $(\beta_i^{(2)}, \eta_i)$ is not identified.

C Consistency of the sieve estimator

In this section, we provide a set of low-level sufficient conditions for the consistency of $(\hat{\beta}^{(1)}, \hat{F})$ using sieve spaces (9) and (10). In the setting of model (2), we will show how the proposed conditions verify those in Theorem 4.1 of Newey and Powell (2003) (or Lemma 3.1 of Ai and Chen (2003)) to prove the consistency.[55]

[55]The conditions for the consistency are similar in the two papers. As stated in Ai and Chen (2003), the proof of Lemma 3.1 relies on Theorem 4.1 of Newey and Powell (2003).

To simplify exposition, we include exogenous product characteristics X_t in Z_t and remove the notation t. Moreover, we assume that random coefficients $\beta_i^{(2)}$ are continuous random variables. Denote by f their joint density function in $\mathbb{R}^{K_2}$ whose support is denoted by Ω. We abuse the notation Θ_F to refer to the space of density functions and $\hat{f}$ to refer to the estimator of f.

Assumption 1.

(i). For any k_{1T}, the basis functions $\boldsymbol{\phi}^{k_{1T}}(Z) = (\phi_k(Z))_{k=1}^{k_{1T}} \in \mathbb{R}^{k_{1T}\times 1}$ satisfy:

1. *$k_{1T} \to \infty$ and $k_{1T}/T \to 0$*
2. *For any g such that $\mathbb{E}[g(Z)^2] < \infty$, there exists $\boldsymbol{\pi} \in \mathbb{R}^{1\times k_{1T}}$ such that $\mathbb{E}\left[\left(g(Z) - \boldsymbol{\pi}\boldsymbol{\phi}^{k_{1T}}(Z)\right)^2\right] = o(1)$.*

(ii). $\hat{\Sigma}(Z) \xrightarrow{p} \Sigma(Z)$ uniformly over $Z \in \mathbf{D}_Z$. Moreover, $\Sigma(Z)$ is finite positive definite uniformly over $Z \in \mathbf{D}_Z$.

(iii). There is a norm $|.|_s$ defined on $\Theta_{\beta^{(1)}} \times \Theta_F$ such that $\Theta_{\beta^{(1)}} \times \Theta_F$ is compact under $|.|_s$.

(iv). For any $k_{2T} > 0$, $\Theta_{k_{2T},F} \subset \Theta_F$. Moreover, for any $(\beta^{(1)}, F) \in \Theta_{\beta^{(1)}} \times \Theta_F$, there exists $(\beta^{(1)}_{k_{2T}}, F_{k_{2T}}) \in \Theta_{\beta^{(1)}} \times \Theta_{k_{2T},F}$, such that $|(\beta^{(1)}_{k_{2T}}, F_{k_{2T}}) - (\beta^{(1)}, F)|_s \to 0$.

(v). $J \times k_{1T} \geq k_{2T} + K_1$, $k_{2T} \to \infty$.

(vi). There exists $M > 0$ such that $|X|, |\xi| \leq M$ almost surely. Moreover, there exists $r > 0$ and $p > 0$, such that $\Pr(|\beta_i^{(2)}| \leq r) > p > 0$.

Theorem 1 implies Assumption 1 of Newey and Powell (2003), i.e., the identification assumption. Assumptions 1(i) states that the sieve space generated by $\boldsymbol{\phi}^{k_{1T}}$ can approximate any function with finite mean square. Together with Assumptions 1(ii), it is essentially the same as Assumption 2 of Newey and Powell (2003). Assumption 1(iii) requires the existence of a norm under which the parameter space is compact. Assumption 1(iv) requires the (asymptotic) denseness of sieve space $\Theta_{\beta^{(1)}} \times \Theta_{k_{2T},F}$ under norm $|.|_s$. Assumption 1(iii) and (iv) imply Assumptions 4 and 5 of Newey and Powell (2003), respectively. Assumption 1(v) simply requires that the number of moment conditions exceeds the number of unknowns in any finite sample (see the GMM interpretation in equations (7) and (8)). Assumption (vi) focuses on bounded covariates X and demand shocks ξ, but still allows for arbitrarily large (and given) bounds of X and ξ. Moreover, Assumption 1(vi) requires the distribution of random coefficients to have non-zero density around the origin. This requirement is mild and satisfied by most centered continuous distributions. Crucially, Assumption 1(vi) is specific to model (2) and key to the proof of consistency. It implies the Hölder continuity of σ^{-1} required by Assumption 3 of Newey and Powell (2003).

We now construct a norm $|.|_s$ that satisfies Assumption 1(iii) and show that Assumption 1(vi) implies the Hölder continuity property under this norm.

Construction of $|.|_s$ We follow the construction in section 2 of Gallant and Nychka (1987). Let m denote the number of derivatives of the unknown density defined on $\mathbb{R}^{K_2}$. Define

$$\begin{aligned}\Theta_{\beta^{(1)}} &= \{\beta^{(1)} : \|\beta^{(1)}\|_2 \le B_\beta\} \\ \Theta_F &= \{f : f(v) = h^2(v) + \epsilon h_0(v), \int_\Omega v h(v) dv = \mathbf{0}_{K_2\times 1}\},\end{aligned} \tag{C.1}$$

where $\|.\|_2$ is the Euclidean norm in $\mathbb{R}^{K_1}$, $\epsilon > \epsilon_0$, B_0 is a constant, h_0 is a known density function with mean zero, and both h and h_0 satisfy $\|h\|_{m_0+m,2,\mu_0} < B_0$, $\|h_0\|_{m_0+m,2,\mu_0} < B_0$, with $\|.\|_{m_0+m,2,\mu_0}$ defined as Sobolev norm:

$$\|g\|_{m,p,\mu} = \left(\sum_{|\lambda|\le m} \int \left|D^\lambda g(u)\right|^p \mu(u) du\right)^{1/p},$$

$m_0, \delta_0 > K_2/2$, and $\mu_0(u) = (1 + u^{\mathrm{T}} u)^{\delta_0}$. The corresponding $\Theta_{k_{2T},F}$ is defined as:

$$\Theta_{k_{2T},F} = \{f : f(v) = P^2_{k_{2T}}(u)\phi^2(u; \sigma_1^2, ..., \sigma_{K_2}^2) + \epsilon h_0(v), \int_\Omega v h(v) dv = \mathbf{0}_{K_2\times 1}\}$$

with $\|P_{k_{2T}}(u)\phi(u; \sigma_1^2, ..., \sigma_{K_2}^2)\|_{m_0+m,2,\mu_0} < B_0$. Then, the norm $|.|_s$ is defined as: for some $\delta \in (K_2/2, \delta_0)$,

$$|(\beta^{(1)}, f)|_s = \|\beta^{(1)}\|_2 + \|f\|_{m,\infty,\delta},$$

Using Theorem 1 of Gallant and Nychka (1987), one can show that the closure of $\Theta_{\beta^{(1)}} \times \Theta_F$ is compact under $|.|_s$. Note that $\Theta_{k,F} \subset \Theta_{k',F}$ for any $k \le k'$. Then, Theorem 2 of Gallant and Nychka (1987) implies that $\Theta_{\beta^{(1)}} \times \Theta_{k_{2T},F}$ is asymptotically dense in this closure under $|.|_s$.

When Ω is bounded, it suffices to restrict the definition of Sobolev space and the corresponding norms on Ω and the arguments above still hold.

Hölder continuity of σ^{-1}. It suffices to prove

- $\mathbb{E}[|\xi_t|^2 | Z_t = Z] < M^2$, for any $Z \in \mathbf{D}_Z$.
- There exists $A > 0$ such that for $(\beta^{(1)}, f), (\beta^{(1)'}, f') \in \Theta_{\beta^{(1)}} \times \Theta_F$,

$$\left|\left[\sigma^{-1}(s; X^{(2)}, f) - X^{(1)}\beta^{(1)}\right] - \left[\sigma^{-1}(s; X^{(2)}, f') - X^{(1)}\beta^{(1)'}\right]\right| \le A \times \left|(\beta^{(1)}, f) - (\beta^{(1)'}, f')\right|_s$$

Proof. The first statement is a direct result of $|\xi_t| < M$ almost surely. To prove the second result, note that

$$\left|\left[\sigma^{-1}(s;X^{(2)},f)-X^{(1)}\beta^{(1)}\right]-\left[\sigma^{-1}(s;X^{(2)},f')-X^{(1)}\beta^{(1)'}\right]\right|\leq\left|\sigma^{-1}(s;X^{(2)},f)-\sigma^{-1}(s;X^{(2)},f')\right| \\ +\left|X^{(1)}(\beta^{(1)}-\beta^{(1)'})\right|. \tag{C.2}$$

Define $s'=\sigma(\sigma^{-1}(s;X^{(2)},f');X^{(2)},f)$. Then, $\sigma^{-1}(s';X^{(2)},f)=\sigma^{-1}(s;X^{(2)},f')$ and

$$\begin{aligned}\left|\sigma^{-1}(s;X^{(2)},f)-\sigma^{-1}(s;X^{(2)},f')\right| &= \sigma^{-1}(s;X^{(2)},f)-\sigma^{-1}(s';X^{(2)},f) \\ &= \left|\int_0^1 \frac{\partial\sigma^{-1}(\tau s+(1-\tau)s';X^{(2)},f)}{\partial s}d\tau\ (s-s')\right| \qquad (C.3) \\ &\leq \bar{\lambda}|s-s'|,\end{aligned}$$

where $\bar{\lambda}$ is the maximum of the maximal eigenvalue of $\frac{\partial\sigma^{-1}(\tau s+(1-\tau)s';X^{(2)},f)}{\partial s}$, or equivalently, the maximum of the reciprocal of minimal eigenvalue of $\frac{\partial\sigma(\delta_\tau;X^{(2)},f)}{\partial\delta}$ for $\tau\in[0,1]$, where $\sigma(\delta_\tau;X^{(2)},f)=\tau s+(1-\tau)s'$. Note that for any δ,

$$\frac{\partial\sigma(\delta;X^{(2)},f)}{\partial\delta}=\int\left[\mathrm{Diag}(s_i)-s_is_i^{\mathrm{T}}\right]f(\beta_i^{(2)})d\beta_i^{(2)},$$

where s_i is the individual-i specific choice probability given $\beta_i^{(2)}$. Then, for any unit vector $v\in\mathbf{R}^J$, using Cauchy-Schwarz inequality, we have:

$$\begin{aligned}v^{\mathrm{T}}\int\left[\mathrm{Diag}(s_i)-s_is_i^{\mathrm{T}}\right]f(\beta_i^{(2)})d\beta_i^{(2)}\ v &= \int\left[\sum_{j\in\mathbf{J}}s_{ij}v_j^2-(\sum_{j\in\mathbf{J}}s_{ij}v_j)^2\right]f(\beta_i^{(2)})d\beta_i^{(2)} \\ &= \int\left[s_{i0}\sum_{j\in\mathbf{J}}s_{ij}v_j^2+\sum_{j\in\mathbf{J}}s_{ij}\sum_{j\in\mathbf{J}}s_{ij}v_j^2-(\sum_{j\in\mathbf{J}}s_{ij}v_j)^2\right]f(\beta_i^{(2)})d\beta_i^{(2)} \\ &\geq \int\left[s_{i0}\sum_{j\in\mathbf{J}}s_{ij}v_j^2\right]f(\beta_i^{(2)})d\beta_i^{(2)} \\ &\geq \int\left[s_{i0}\min_{j\in\mathbf{J}}s_{ij}\right]f(\beta_i^{(2)})d\beta_i^{(2)} \\ &= \int\frac{\min_{j\in\mathbf{J}}\{\exp\{\delta_j+X_j^{(2)}\beta_i^{(2)}\}\}}{\left(1+\sum_{j\in\mathbf{J}}\exp\{\delta_j+X_j^{(2)}\beta_i^{(2)}\}\right)^2}f(\beta_i^{(2)})d\beta_i^{(2)}.\end{aligned}$$

Because of the boundedness of $|X|,|\xi|$ in Assumption 1(vi), then there exists $M_1>0$ such that

$$\frac{\min_{j\in\mathbf{J}}\{\exp\{\delta_j+X_j^{(2)}\beta_i^{(2)}\}\}}{\left(1+\sum_{j\in\mathbf{J}}\exp\{\delta_j+X_j^{(2)}\beta_i^{(2)}\}\right)^2}\geq M_1$$

uniformly for X, ξ, and $|\beta_i^{(2)}| \le r$. Then,

$$\int \frac{\min_{j\in\mathbf{J}}\{\exp\{\delta_j + X_j^{(2)}\beta_i^{(2)}\}\}}{\left(1+\sum_{j\in\mathbf{J}}\exp\{\delta_j + X_j^{(2)}\beta_i^{(2)}\}\right)^2} f(\beta_i^{(2)})d\beta_i^{(2)} \ge \int_{|\beta_i^{(2)}|\le r} M_1 f(\beta_i^{(2)})d\beta_i^{(2)} \\ \ge M_1 \times p,$$

and the minimal eigenvalue of $\frac{\partial\sigma(\delta;X^{(2)},f)}{\partial\delta}$ is greater or equal to $M_1 \times p$ for any δ. As a consequence,

$$\bar{\lambda} \le \frac{1}{M_1 \times p}. \tag{C.4}$$

Moreover, let $\delta' = \sigma^{-1}(s'; X^{(2)}, f) = \sigma^{-1}(s; X^{(2)}, f')$. Then, for any $j \in \mathbf{J}$,

$$\begin{aligned}\left|s_j - s_j'\right| &= \left|\int \frac{\exp\left(\delta_j' + X_j^{(2)}\beta_i^{(2)}\right)}{1+\sum_{j'\in\mathbf{J}}\exp\left(\delta_{j'} + X_{j'}^{(2)}\beta_i^{(2)}\right)}\left[f'(\beta_i^{(2)}) - f(\beta_i^{(2)})\right]d\beta_i^{(2)}\right| \\ &\le \|f - f'\|_{m,\infty,\delta}\end{aligned}$$

and consequently $|s - s'| = \sqrt{J}\|f - f'\|_{m,\infty,\delta}$. Then, combining this with (C.4), (C.3), and (C.2), we obtain:

$$\begin{aligned}&\left|\left[\sigma^{-1}(s;X^{(2)},f) - X^{(1)}\beta^{(1)}\right] - \left[\sigma^{-1}(s;X^{(2)},f') - X^{(1)}\beta^{(1)'}\right]\right| \\ &\le \frac{\sqrt{J}}{M_1 \times p}\|f - f'\|_{m,\infty,\delta} + |X^{(1)}|\|\beta^{(1)} - \beta^{(1)'}\|_2 \\ &\le A \times \left|(\beta^{(1)}, f) - (\beta^{(1)'}, f')\right|_s,\end{aligned}$$

with $A = \max\{\frac{\sqrt{J}}{M_1 \times p}, M\}$. The proof is completed. □

D Moment restrictions using (9) and (10)

First, note that $\frac{\phi_\Omega^2(v)}{\int_\Omega \phi_\Omega^2(v)dv}$ defines a (known) probabilist distribution on Ω. Denote its $(l_1, ..., l_{K_2})$th moment by $m(l_1, ..., l_{K_2})$ and $P_k(v) = \sum_{\sum_{r=1}^{K_2} l_r = K_2} \alpha_{(l_1,...,l_{K_2})} \prod_{r=1} v_r^{l_r} = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{v}$, where $\boldsymbol{\alpha} = (\alpha_{(l_1,...,l_{K_2})})_{(l_1,...,l_{K_2})}$ and $\mathbf{v} = (\prod_{r=1}^{K_2} v_r^{l_r})_{(l_1,...,l_{K_2})}$. Then, the shape restriction

$$1 = \int_\Omega P_k^2(v)\phi_\Omega^2(v)dv = \left[\int_\Omega \phi_\Omega^2(v)dv\right]\mathbb{E}\left[P_k^2(v)\right] = \left[\int_\Omega \phi_\Omega^2(v)dv\right]\boldsymbol{\alpha}^{\mathrm{T}}\mathbb{E}\left[\mathbf{v}\mathbf{v}^{\mathrm{T}}\right]\boldsymbol{\alpha}$$

defines a quadratic constraint on parameters $\boldsymbol{\alpha}$. Suppose that the jth element of $\mathbf{v}$ is $\prod_{r=1}^{K_2} v_r^{l_r^{(j)}}$. Then, the (j,h) element of $\mathbb{E}\left[\mathbf{v}\mathbf{v}^{\mathrm{T}}\right]$ is $m(l_1^{(j)} + l_1^{(h)}, ..., l_{K_2}^{(j)} + l_{K_2}^{(h)})$. Analogously, higher-order restriction can be expressed in quadratic forms: the $(t_1, ..., t_{K_2})$th moment of the distribution defined by $P_k^2(v)\phi_\Omega^2(v)$ is:

$$\left[\int_\Omega \phi_\Omega^2(v)dv\right] \boldsymbol{\alpha}^{\mathrm{T}}\mathbb{E}\left[\mathbf{v}\mathbf{v}^{\mathrm{T}}\prod_{j=1}^{K_2} v_j^{t_j}\right]\boldsymbol{\alpha},$$

where the (j,h) element of $\mathbb{E}\left[\mathbf{v}\mathbf{v}^{\mathrm{T}}\prod_{j=1}^{K_2} v_j^{t_j}\right]$ is $m(l_1^{(j)}+l_1^{(h)}+t_1, ..., l_{K_2}^{(j)}+l_{K_2}^{(h)}+t_{K_2})$.

E Data Generating Process of Monte Carlo Simulations

For individual i, the indirect utility from purchasing product j in market t is

$$U_{ijt} = \delta_j - \alpha_i p_{jt} + \beta x_{jt} + \varepsilon_{ijt},$$

where

$$\begin{aligned}
&\beta = 1, \alpha_i \sim f, \mathbb{E}[\alpha_i] = \alpha = 3,\\
&x_{jt} \overset{\text{i.i.d.}}{\sim} \mathscr{U}([0,1]),\\
&\xi_{jt} \overset{\text{i.i.d.}}{\sim} \mathscr{U}([-0.5, 0.5])
\end{aligned}$$

across $j \in \mathbf{J}, t = 1, ..., 500$. Moreover, we assume that the marginal cost of production of j in market t is:

$$mc_{jt} = 1 + z_{jt} + w_{jt},$$

where $z_{jt} \overset{\text{i.i.d.}}{\sim} \mathscr{U}([-0.5, 0.5])$ is cost shifter for product j in market t, $w_{jt} \overset{\text{i.i.d.}}{\sim} \mathscr{U}([-0.25, 0.25])$ is cost shock. We also assume that x_{jt}, ξ_{jt}, z_{jt}, w_{jt} are mutually independent.

F Additional Tables and Figures

Table 5: RMSE, Parametric and Sieve Estimations, $J = 50$, $T = 500$

	Demand parameters			Density function
	$\alpha = -3$	$\beta = 1$	$\{\delta_j\}_{j\in\mathbf{J}}$	f
Specification I	$\alpha_i \sim \mathcal{N}(0, 0.64)$			
Parametric	0.0693	0.0061	0.1829	0.1425
Hermite Sieve with $\hat{\sigma}$				
$d = 2$	0.0749	0.0061	0.1209	0.1462
$d = 3$	0.0734	0.0061	0.1185	0.1468
Hermite Sieve with $\sigma_0 = 0.8$				
$d = 2$	0.0289	0.0060	0.0476	0.1428
$d = 3$	0.0308	0.0060	0.0494	0.1432
Specification II	$\alpha_i \sim \mathcal{U}([-1, 1])$			
Parametric	0.0631	0.0060	0.1058	0.1706
Polynomial Sieve with $\hat{a}$				
$d = 2$	0.0632	0.0061	0.1337	0.1435
$d = 3$	0.0638	0.0061	0.1349	0.1446
Polynomial Sieve with $a_0 = 1$				
$d = 2$	0.0113	0.0060	0.0946	0.0260
$d = 3$	0.0116	0.0060	0.0967	0.0330
Specification III	$\alpha_i \sim 0.5\mathcal{N}(-1, 0.2) + 0.5\mathcal{N}(1, 0.2)$			
Parametric	0.2888	0.0067	0.4738	0.4718
Hermite Sieve with $\hat{\sigma}$				
$d = 2$	0.2459	0.0063	0.3911	0.4354
$d = 3$	0.2677	0.0062	0.4214	0.4392
Hermite Sieve with $\sigma_0 = \sqrt{1.04}$				
$d = 2$	0.0963	0.0061	0.1599	0.4129
$d = 3$	0.1087	0.0060	0.1764	0.4141

Notes: The RMSEs are computed on the basis of 100 independently simulated samples, each with $J = 50$ products and $T = 500$ markets. For product-specific intercepts $\{\delta_j\}_{j\in\mathbf{J}}$, we report the median of the RMSEs of the estimates $\{\hat{\delta}_j\}_{j\in\mathbf{J}}$. For the density function, we report the root of median of $\mathbb{E}\left[(\hat{f} - f)^2\right]$, where f is the true density function and $\mathbb{E}$ is with respect to the true distribution f. All the sieve estimates are obtained by using identify weighting matrix, i.e., $\Sigma(Z) = \mathbf{I}$.

Table 6: Bootstrap Confidence Intervals, Length

Quantile interval (%)	$[2.5, 97.5]$	$[4, 99]$	$[0.5, 99.5]$
Hermite Sieve with $\hat{\sigma}$, $d = 2$			
α	0.4148	0.3973	0.5587
β	0.0383	0.0396	0.0496
$\{\delta_j\}_{j\in\mathbf{J}}$	0.5522	0.6207	0.7505
$\{\epsilon_{jj}\}_{j\in\mathbf{J}}$	0.1016	0.1059	0.1414
Hermite Sieve with $\hat{\sigma}$, $d = 3$			
α	0.6621	0.6483	0.8390
β	0.0379	0.0393	0.0496
$\{\delta_j\}_{j\in\mathbf{J}}$	0.8218	0.9038	1.0682
$\{\epsilon_{jj}\}_{j\in\mathbf{J}}$	0.1157	0.1206	0.1568

Table 7: Descriptive Statistics, Products

Brand	Flavor	Ave. Price $ per lb	Ave. Market Share(%)	Sugar Content g per 100g
GENERAL MILLS CHEERIOS	TOASTED	4.3036	2.7477	3.57
GENERAL MILLS CINNAMON TST CR	CINNAMON TOAST	3.7109	1.4132	32.43
GENERAL MILLS COCOA PUFFS	COCOA	4.0961	0.4813	33.33
GENERAL MILLS FIBER ONE	REGULAR	4.2465	0.4810	0
GENERAL MILLS GOLDEN GRAHAMS	GRAHAM	4.1487	0.3791	30
GENERAL MILLS HONEY NUT CHEER	HONEY NUT	4.1888	2.4553	32.43
GENERAL MILLS KIX	REGULAR	4.9134	0.3869	10
GENERAL MILLS LUCKY CHARMS	TOASTED	4.4398	0.9760	33.33
GENERAL MILLS MULTI GRAIN CHE	REGULAR	5.3236	0.6912	20.51
GENERAL MILLS REESES PUFFS	COCOA PEANUT BUTTER	3.5277	0.3650	31.03
GENERAL MILLS RICE CHEX	MISSING	4.6346	0.3102	7.5
GENERAL MILLS TOTAL	REGULAR	5.0783	0.4113	15
GENERAL MILLS TRIX	FRUIT	4.5205	0.3433	32.26
GENERAL MILLS WHEATIES	TOASTED	4.7670	0.3241	13.89
KASHI GO LEAN	HONEY & GRAHAM	4.1372	0.3574	15.69
KASHI GO LEAN CRUNCH	REGULAR	3.8851	0.5264	24.53
KASHI HEART TO HEART	TOASTED HONEY	4.6371	0.4256	15.15
KASHI ORGANC PROMIS CINNMN HR	CINNAMON	3.7059	0.3980	17.65
KELLOGGS APPLE JACKS	APPLE CINNAMON	4.7492	0.5372	33.33
KELLOGGS CORN FLAKES	REGULAR	3.6375	0.6274	9.524
KELLOGGS CORN POPS	REGULAR	4.5628	0.4219	37.5
KELLOGGS FROOT LOOPS	FRUIT	4.9432	0.6203	30.77
KELLOGGS FROSTED FLAKES	REGULAR	3.5951	1.3468	35.29
KELLOGGS FROSTED MINI WHEATS	REGULAR	3.2842	1.5702	20
KELLOGGS RAISIN BRAN	REGULAR	2.7333	1.3348	28.81
KELLOGGS RAISIN BRAN CRUNCH	TOASTED HONEY	3.4974	0.3781	35.19
KELLOGGS RICE KRISPIES	TOASTED	4.6794	0.8318	10
KELLOGGS SPECIAL K	TOASTED	4.6798	0.7604	11.11
KELLOGGS SPECIAL K	FRUIT & YO	4.6026	0.4581	30.95
KELLOGGS SPECIAL K	RED BERRIE	4.8065	0.8717	28.21
KELLOGGS SPECIAL K	VANILLA AL	4.2567	0.3805	27.5
POST FRUITY PEBBLES	FRUIT	4.4412	0.3963	33.33
POST GRAPE NUTS	REGULAR	2.9885	0.5975	8.621
POST HONEY BUNCHES OF OATS	HONEY	4.2647	0.6726	21.43
POST HONEY BUNCHES OF OATS	HONEY ALMOND	3.9454	0.5836	21.43
POST HONEY BUNCHES OF OATS	HONEY ROASTED	3.9592	0.8041	21.95
POST RAISIN BRAN	RAISIN BRAN	2.7087	0.6136	32.79
POST SELECTS GREAT GRAINS	REGULAR	3.8610	0.4923	24.07
POST SHREDDED WHEAT	ORIGINAL	3.8837	0.5574	0
POST SHREDDED WHEAT	REGULAR	3.4462	0.4422	0
PRIVATE LABEL	COCOA	2.4617	0.2934	32.18
PRIVATE LABEL	RAISIN BRAN	1.9667	0.3962	32.79
PRIVATE LABEL	REGULAR	2.6111	3.6331	21.08
PRIVATE LABEL	TOASTED	2.9129	0.7568	14.38
PRIVATE LABEL	TOASTED HONEY NUT	2.6020	0.3977	25.17
QUAKER CAP N CRUNCH	REGULAR	3.4078	0.3798	44.74
QUAKER CAP N CRUNCH CRUNCH BE	BERRY	3.6990	0.4028	43.24
QUAKER CINNAMON LIFE	CINNAMON	3.4098	0.8027	23.81
QUAKER LIFE	REGULAR	3.4070	0.9984	19.05
QUAKER OATMEAL SQUARES	REGULAR	4.1077	0.3788	16.07

Notes: Price is measured as $ per lb. Sugar content is measured as gram per 100g and is mostly obtained from its nutrition label from the producer's website. The sugar content of a private label product with flavor characteristics f is defined as the average of the sugar content of non-private label products with the same flavor f.

References

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Allcott, H., Lockwood, B. B., and Taubinsky, D. (2019a). Regressive sin taxes, with an application to the optimal soda tax. *The Quarterly Journal of Economics*, 134(3):1557–1626.

Allcott, H., Lockwood, B. B., and Taubinsky, D. (2019b). Should we tax sugar-sweetened beverages? an overview of theory and evidence. *Journal of Economic Perspectives*, 33(3):202–27.

Allen, R. and Rehbeck, J. (2019). Identification with additively separable heterogeneity. *Econometrica*, 87(3):1021–1054.

Allen, R. and Rehbeck, J. (2020). Identification of random coefficient latent utility models. *arXiv preprint arXiv:2003.00276*.

Berry, S., Gandhi, A., and Haile, P. (2013). Connected substitutes and invertibility of demand. *Econometrica*, 81(5):2087–2111.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890.

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

Berry, S. T. and Haile, P. A. (2009). Nonparametric identification of multinomial choice demand models with heterogeneous consumers. Technical report, National Bureau of Economic Research.

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

Berry, S. T. and Haile, P. A. (2018). Identification of nonparametric simultaneous equations models with a residual index structure. *Econometrica*, 86(1):289–315.

Berto Villas-Boas, S. (2007). Vertical relationships between manufacturers and retailers: Inference with limited data. *The Review of Economic Studies*, 74(2):625–652.

Bronnenberg, B. J., Kruger, M. W., and Mela, C. F. (2008). Database paper—the iri marketing data set. *Marketing science*, 27(4):745–748.

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

Chen, X. and Christensen, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84.

Chen, X. and Pouzo, D. (2015). Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079.

Chen, X. and Qiu, Y. J. J. (2016). Methods for nonparametric and semiparametric regressions with endogeneity: A gentle guide. *Annual Review of Economics*, 8:259–290.

Chernozhukov, V., Fernández-Val, I., and Newey, W. K. (2019). Nonseparable multinomial choice models in cross-section and panel data. *Journal of Econometrics*, 211(1):104–116.

Chesher, A. (2003). Identification in nonseparable models. *Econometrica*, 71(5):1405–1441.

Compiani, G. (2019). Market counterfactuals and the specification of multi-product demand: A nonparametric approach. *Working Paper*.

Conlon, C. and Gortmaker, J. (2020). Best practices for differentiated products demand estimation with pyblp. *The RAND Journal of Economics*, 51(4):1108–1161.

D'Haultfœuille, X. and Février, P. (2015). Identification of nonseparable triangular models with discrete instruments. *Econometrica*, 83(3):1199–1210.

Dubois, P., Griffith, R., and O'Connell, M. (2020). How well targeted are soda taxes? *American Economic Review*, 110(11):3661–3704.

Dunker, F., Hoderlein, S., and Kaido, H. (2017). Nonparametric identification of random coefficients in endogenous and heterogeneous aggregated demand models. *cemmap working paper*.

Fan, Y. (2013). Ownership consolidation and product characteristics: A study of the us daily newspaper market. *American Economic Review*, 103(5):1598–1628.

Fox, J. T. (2017). A note on nonparametric identification of distributions of random coefficients in multinomial choice models. Technical report, National Bureau of Economic Research.

Fox, J. T. and Gandhi, A. (2016). Nonparametric identification and estimation of random coefficients in multinomial choice models. *The RAND Journal of Economics*, 47(1):118–139.

Fox, J. T., il Kim, K., and Yang, C. (2016). A simple nonparametric approach to estimating the distribution of random coefficients in structural models. *Journal of Econometrics*, 195(2):236–254.

Fox, J. T., Kim, K., Ryan, S. P., and Bajari, P. (2012). The random coefficients logit model is identified. *Journal of Econometrics*, 166(2):204–212.

Fox, J. T. and Lazzati, N. (2017). A note on identification of discrete choice models for bundles and binary games. *Quantitative Economics*, 8(3):1021–1036.

Gaillac, C. and Gautier, E. (2019). Adaptive estimation in the linear random coefficients model when regressors have limited variation. *arXiv preprint arXiv:1905.06584*.

Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica: Journal of the econometric society*, pages 363–390.

Gandhi, A. and Houde, J.-F. (2019). Measuring substitution patterns in differentiated products industries. Technical report, National Bureau of Economic Research.

Gautier, E. and Hoderlein, S. (2013). Estimating the distribution of treatment effects. *arXiv preprint arXiv:1109.0362*.

Gautier, E. and Kitamura, Y. (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica*, 81(2):581–607.

Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *The American Economic Review*, 97(3):713–744.

Gruber, J. and Kőszegi, B. (2004). Tax incidence when individuals are time-inconsistent: the case of cigarette excise taxes. *Journal of Public Economics*, 88(9-10):1959–1987.

Hoderlein, S., Klemelä, J., and Mammen, E. (2010). Analyzing the random coefficient model nonparametrically. *Econometric Theory*, 26(3):804–837.

Iaria, A. and Wang, A. (2019). Identification and estimation of demand for bundles. *Working Paper*, pages 1–92.

Ichimura, H. and Thompson, T. S. (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics*, 86(2):269–295.

il Kim, K. et al. (2014). Identification of the distribution of random coefficients in static and dynamic discrete choice models. *Korean Economic Review*, 30:191–216.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*, 97(1):145–177.

Lewbel, A. (2010). Identification of limited dependent variable models using a special regressor with limited support. *unpublished manuscript*.

Lewbel, A. and Pendakur, K. (2017). Unobserved preference heterogeneity in demand using generalized random coefficients. *Journal of Political Economy*, 125(4):1100–1148.

Lu, Z., Shi, X., and Tao, J. (2019). Semi-nonparametric estimation of random coefficient logit model for aggregate demand. *Available at SSRN 3503560*.

Masten, M. A. (2018). Random coefficients on endogenous variables in simultaneous equations models. *The Review of Economic Studies*, 85(2):1193–1250.

Matzkin, R. L. (2008). Identification in nonparametric simultaneous equations models. *Econometrica*, 76(5):945–978.

McFadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics*, pages 447–470.

Mityagin, B. (2015). The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*.

Nevo, A. (2000). Mergers with differentiated products: The case of the ready-to-eat cereal industry. *The RAND Journal of Economics*, pages 395–421.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Petrin, A. (2002). Quantifying the benefits of new products: The case of the minivan. *Journal of political Economy*, 110(4):705–729.

Saito, K. (2017). Axiomatizations of the mixed logit model. *working paper*.

Torgovitsky, A. (2015). Identification of nonseparable models using instruments with small support. *Econometrica*, 83(3):1185–1197.

Vos, M. B., Kaar, J. L., Welsh, J. A., Van Horn, L. V., Feig, D. I., Anderson, C. A., Patel, M. J., Cruz Munos, J., Krebs, N. F., Xanthakos, S. A., et al. (2017). Added sugars and cardiovascular disease risk in children: a scientific statement from the american heart association. *Circulation*, 135(19):e1017–e1034.

Wang, A. (2021). A blp demand model of product-level market shares with complementarity. *Available at SSRN 3785148.*

WHO (2015). *Guideline: sugars intake for adults and children.* World Health Organization.