Estimating Demand for Differentiated Products with Zeroes in Market Share Data*

Amit Gandhi
UW-Madison
Microsoft

Zhentong Lu
SUFE

Xiaoxia Shi †
UW-Madison

March 9, 2017

Abstract

In this paper we introduce a new approach to estimating differentiated product demand systems that allows for products with zero sales in the data. Zeroes in demand are a common problem in product differentiated markets, but fall outside the scope of existing demand estimation techniques. Our solution to the zeroes problem is based on constructing bounds for the conditional expectation of the inverse demand. These bounds can be translated into moment inequalities that are shown to yield consistent and asymptotically normal point estimator for demand parameters under natural conditions for differentiated product markets. In Monte Carlo simulations, we demonstrate that the new approach works well even when the fraction of zeroes is as high as 95%. We apply our estimator to supermarket scanner data and find price elasticities become on the order of twice as large when zeroes are properly controlled.

Keywords: Demand Estimation, Differentiated Products, Profile, Measurement Error, Moment Inequality.

JEL: C01, C12, L10, L81.

1 Introduction

In this paper we introduce a new approach to differentiated product demand estimation that allows for zeroes in empirical market share data. Such zeroes are a highly prevalent feature of demand in a variety of empirical settings, ranging from workhorse scanner retail data, to data as diverse as

*Previous version of this paper was circulated under the title "Estimating Demand for Differentiated Products with Error in Market Shares."

†We are thankful to Steven Berry, Jean-Pierre Dubé, Philip Haile, Bruce Hansen, Ulrich Müller, Aviv Nevo, Jack Porter, and Chris Taber for insightful discussions and suggestions; We would also like to thank the participants at the MIT Econometrics of Demand Conference, Chicago-Booth Marketing Lunch, the Northwestern Conference on "Junior Festival on New Developments in Microeconometrics," the Cowles Foundation Conference on "Structural Empirical Microeconomic Models," 3rd Cornell - Penn State Econometrics & Industrial Organization Workshop, as well as seminar participants at Wisconsin-Madison, Wisconsin-Milwaukee, Cornell, Indiana, Princeton, NYU, Penn and the Federal Trade Commission for their many helpful comments and questions.

homicide rates and international trade flows (we discuss these examples in further depth below). Zeroes naturally arise in "big data" applications which allow for increasingly granular views of consumers, products, and markets (see for example Quan and Williams (2015), Nurski and Verboven (2016)). Unfortunately, the standard estimation procedures following the seminal Berry, Levinsohn, and Pakes (1995) (BLP for short) cannot be used in the presence of zero empirical shares - they are simply not well defined when zeroes are present. Furthermore, ad hoc fixes to market zeroes that are sometimes used in practice, such as dropping zeroes from the data or replacing them with small positive numbers, are subject to biases which can be quite large (discussed further below). This has left empirical work on demand for differentiated products without satisfying solutions to the zero shares problem, which is the key void our paper aims to fill.

In this paper we provide an approach to estimating differentiated product demand models that provides consistency (and asymptotic normality) for demand parameters despite a possibly large presence of market zeroes in the data. We first isolate the econometric problem caused by zeroes in the data. The problem we show is driven by the wedge between *choice probabilities*, which are the theoretical outcome variables predicted by the demand model, and *market shares*, which are the empirical revealed preference data used to estimate choice probabilities. Although choice probabilities are strictly positive in the underlying model, market shares are often zero if choice probabilities are small. The root of the zeroes problem is that substituting market shares (or some other consistent estimate) for choice probabilities in the moment conditions that identify the model, which is the basis for the traditional estimators, will generally lead to asymptotic bias. While this bias is assumed away in the traditional approach, it cannot be avoided whenever zeroes are prevalent in the data.

Our solution to this problem is to construct a set of moment *inequalities* for the model, which are by design robust to the sampling error in market shares - our moment inequalities will hold at the true value of the parameters regardless of the magnitude of the measurement error in market shares for choice probabilities. Despite taking an inequality form, we use these moment inequalities to form a GMM-type point estimator based on minimizing the deviations from the inequalities. We show this estimator is consistent so long as there is a positive mass of observations whose latent choice probabilities are bounded sufficiently away from zero, e.g., products for whom market shares are not likely to be zero. This is natural in many applications (as illustrated in Section 2), and strictly generalizes the restrictions on choice probabilities for consistency under the traditional approach. Asymptotic normality then follows by adapting arguments from censored regression models by Kahn and Tamer (2009).

Computationally, our estimator closely resembles the traditional approach with only a slight adjustment in how the empirical moments are constructed. In particular it is no more burdensome than the usual estimation procedures for BLP and can be implemented using either the standard nested fixed point method of the original BLP, or the MPEC method as advocated more recently by Dubé, Fox, and Su (2012).

We investigate the finite sample performance of the approach in a variety of mixed logit ex-

amples. We find that our estimator works well even when the the fraction of zeros is as high as 95%, while the standard procedure with the observations with zeroes deleted yields severely biased estimators even with mild or moderate fractions of zeroes.

We apply our bounds approach to widely used scanner data from the Dominicks Finer Foods (DFF) retail chain. In particular, we estimate demand for the tuna category as previously studied by Chevalier, Kashyap, and Rossi (2003) and continued by Nevo and Hatzitaskos (2006) in the context of testing the loss leader hypothesis of retail sales. We find that controlling for products with zero demand using our approach gives demand estimates that can be more than twice as elastic than standard estimates that select out the zeroes. We also show that the estimated price elasticities do not increase during Lent, which is a high demand period for this product category, after we control for the zeroes. Both of these findings have implications for reconciling the loss-leader hypothesis with the data.

The plan of the paper is the following. In Section 2, we illustrate the stylized empirical pattern of Zipf's law where market zeroes naturally arise. In Section 3, we describe our solution to the zeroes problem using a simple logit setup without random coefficients to make the essential matters transparent. In Section 4, we introduce our general approach for discrete choice model with random coefficients. Section 5 and 6 present results of Monte Carlo simulations and the application to the DFF data, respectively. Section 7 concludes.

2 The Empirical Pattern of Market Zeroes

In this section we highlight some empirical patterns that arise in applications where the zero shares problem arises, which will also help to motivate the general approach we take to it in the paper. Here we will primarily use workhorse store level scanner data to illustrate these patterns. It is this same data that will also be used for our empirical application. However we emphasize that our focus here on scanner data is only for the sake of a concrete illustration of the market zeroes problem - the key patterns we highlight in scanner data are also present in many other economic settings where demand estimation techniques are used (discussed further below and illustrated in the Appendix).

We employ here a widely studied store level scanner data from the Dominick's Finer Foods grocery chain, which is public data that has been used by many researchers.[1] The data comprises 93 Dominick's Finer Foods stores in the Chicago metropolitan area over the years from 1989 to 1997. Like other store level scanner data sets, this data set provides demand information (price, sales, marketing) at a store/week/UPC level, where a UPC (universal product code) is a unique

[1]For a complete list of papers using this data set, see the website of Dominick's Database: http://research.chicagobooth.edu/marketing/databases/dominicks/index.aspx

bar code that identifies a product[2].

Table 1 presents information on the resulting product variety across the different product categories in data. The first column shows the number of products in an average store/week - the number of UPC's can be seen varying from roughly 50 (e.g., bath tissue) to over four hundred (e.g., soft drinks) within even these fairly narrowly defined categories. Thus there is considerable product variety in the data. The next two columns illustrate an important aspect of this large product variety: there are often just a few UPC's that dominate each product category whereas most UPC's are not frequently chosen. The second column illustrates this pattern by showing the well known "80/20" rule prevails in our data: we see that roughly 80 percent of the total quantity purchased in each category is driven by the top 20 percent of the UPC's in the category. In contrast to these "top sellers", the other 80 percent of UPC's contain relatively "sparse sellers" that share the remaining 20 percent of the total volume in the category. The third column shows an important consequence of this sparsity: many UPC's in a given week at a store simply do not sell. In particular, we see that the fraction of observations with zero sales can even be nearly 60% for some categories.

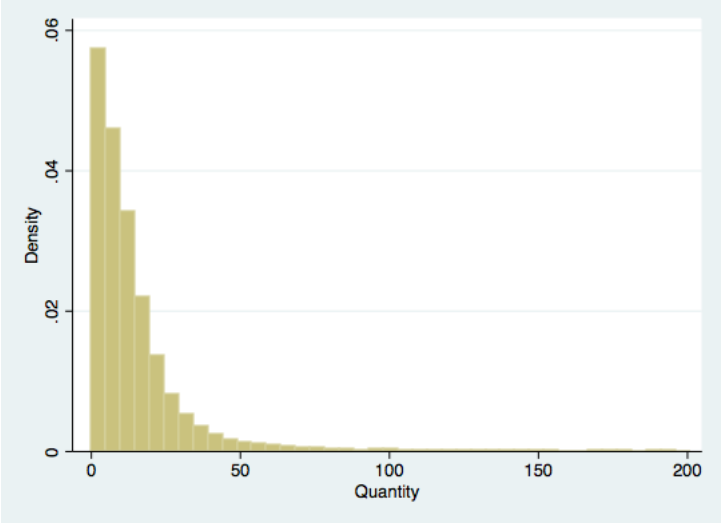
Table 1: Selected Product Categories in the Dominick's Database

Category	Average Number of UPC's in a Store/Week Pair	Percent of Total Sale of the Top 20% UPC's	Percent of Zero Sales
Beer	179	87.18%	50.45%
Cereals	212	72.08%	27.14%
Crackers	112	81.63%	37.33%
Dish Detergent	115	69.04%	42.39%
Frozen Dinners	123	66.53%	38.32%
Frozen Juices	94	75.16%	23.54%
Laundry Detergents	200	65.52%	50.46%
Paper Towels	56	83.56%	48.27%
Refrigerated Juices	91	83.18%	27.83%
Soft Drinks	537	91.21%	38.54%
Snack Crackers	166	76.39%	34.53%
Soaps	140	77.26%	44.39%
Toothbrushes	137	73.69%	58.63%
Canned Tuna	118	82.74%	35.34%
Bathroom Tissues	50	84.06%	28.14%

We can visualize this situation another way by fixing a product category (here we use canned

[2]Store level scanner data can often be augmented with a panel of household level purchases (available, for example, through IRI or Nielsen). Although the DFF data do not contain this micro level data, the main points of our analysis are equally applicable to the case where household level data is available. In fact our general choice model will accommodate the possibility of micro data. Store level purchase data is actually a special case household level data where all households are observationally identical (no observable individual level characteristics).

Figure 1: Zipf's Law in Scanner Data



tuna) and simply plotting the histogram of the volume sold for each week/UPC realization for a single store in the data. This frequency plot is given in *Figure 1*. As can be see there is a sharp decay in the empirical frequency as the purchase quantity becomes larger, with a long thin tail. In particular the bulk of UPC's in the store have small purchase volume: the median UPC sells less than 10 units a week, which is less than 1.5% of the median volume of Tuna the store sells in a week. The mode of the frequency plot is a zero share.

This power-law decay in the frequency of product demand is often associated with "Zipf's law" or the "the long tail", which has a long history in empirical economics.[3] We present further illustrations of this long-tail demand pattern found in international trade flows as well as cross-county homicide rates in Appendix A, which provides a sense of the generality of these stylized facts.

The key takeaway from these illustrations is that the presence of market zeroes in the data is closely intertwined to the prevalence of power-law patterns of demand. We will later exploit this relationship to place structure on the data generating process that underlies market zeroes.

3 A First Pass Through Logit Demand

Why do zero shares create a problem for demand estimation? In this section, we use the workhorse multinomial logit model to explain the zeroes problem and introduce our new estimation strategy. Formal treatment for general differentiated product demand models is given in the next section.

[3]See Anderson (2006) for a historical summary of Zipf's law and many examples from the social and natural sciences. See Gabaix (1999a) for an application of Zipf's law to the economics literature.

3.1 Zeroes Problem in the Logit Model

Consider a multinomial logit model for the demand of J products ($j = 1, \ldots, J$) and an outside option ($j = 0$). A consumer i derives utility $u_{ijt} = \delta_{jt} + \epsilon_{ijt}$ from product j in market t, where δ_{jt} is the mean-utility of product j in market t, and ϵ_{ijt} is the idiosyncratic taste shock that follows the type-I extreme value distribution. As is standard, the mean-utility δ_{jt} of product $j > 0$ is modeled as

$$\delta_{jt} = X'_{jt}\beta + \xi_{jt}, \tag{3.1}$$

where X_{jt} is the vector of observable (product, market) characteristics, often including price, and ξ_{jt} is the unobserved characteristic. The outside good $j = 0$ has mean utility normalized to $\delta_{0t} = 0$. The parameter of interest is β.

Each consumer chooses the product that yields the highest utility. Aggregating consumers' choices, we obtain the true choice probability of product j in market t, denoted as

$$\pi_{jt} = \Pr(\text{product } j \text{ is chosen in market } t).$$

The standard approach introduced by Berry (1994) for estimating β is to combine demand system inversion and instrumental variables.

First, for demand inversion, one uses the logit structure to find that

$$\delta_{jt} = \ln(\pi_{jt}) - \ln(\pi_{0t}), \text{ for } j = 1, \ldots, J. \tag{3.2}$$

Then, to handle the potential endogeneity of X_{jt} (correlation with ξ_{jt}), one finds a random vector z_{jt}, such that

$$E[\xi_{jt} | z_{jt}] = 0. \tag{3.3}$$

Then two stage least squares with δ_{jt} defined in terms of choice probabilities as the dependent variable becomes the identification strategy for β.

Unfortunately π_{jt} is not observed as data - it is a theoretical choice probability defined by the model but only indirectly revealed through actual consumer choices. The standard approach to this following Berry (1994), Berry, Levinsohn, and Pakes (1995), and many subsequent papers in the literature has been to substitute s_{jt}, the empirical market share of product j in market t based on the choices of n potential consumers, for π_{jt}, and run a two-stage least square with $\ln(s_{jt}) - \ln(s_{0t})$ as dependent variable, x_{jt} as covariates, and z_{jt} as instruments to obtain estimates for β.

Plugging in the estimate s_{jt} for π_{jt} appears innocuous at first glance because the number of potential consumers (n) in a market from which s_{jt} is constructed is typically large. Nevertheless problems arise when there are (jt)'s for which π_{jt} is very small. Because the slope of the natural logarithm function approaches infinity when the argument approaches zero, even small estimation error of π_{jt} may lead to large error in the plugged-in version of δ_{jt} when π_{jt} is very small. In particular, s_{jt} may frequently equal zero in this case, causing the demand inversion to fail completely. The first is the theoretical root of the small π_{jt} problem, while the second is an unmistakable symptom.

Data sets with this symptom are frequently encountered in empirical research as discussed in the Section 2. With such data, a common practice is to ignore the (jt)'s with $s_{jt} = 0$, effectively lumping those j's into the outside option in market t. This leads however to a selection problem.

To see this, suppose $s_{jt} = 0$ for some (j, t) and one drops these observations from the analysis - effectively one is using a selected sample where the selection criterion is $s_{jt} > 0$. In this selected sample, the conditional mean of ξ_{jt} is no longer zero, i.e.,

$$E[\xi_{jt}|x_{jt}, s_{jt} > 0] \neq 0. \tag{3.4}$$

This is the well-known selection-on-unobservables problem and with such sample selection, an attenuation bias ensues.[4] The attenuation bias generally leads to demand estimates that appear to be too inelastic.[5]

Another commonly adopted empirical "trick" is to add a small positive number $\epsilon > 0$ to the s_{jt}'s that are zero, and use the resulting modified shares $s^{\epsilon}_{jt} > 0$ in place of π_{jt}. [6] However, this trick only treats the symptom, i.e., $s_{jt} = 0$, but overlooks the nature of the problem: the true choice probability π_{jt} is small. And in this case, small estimation error in any estimator $\widehat{\pi}_{jt}$ of π_{jt} would lead to large error in the plugged-in version of δ_{jt} and the estimation of β. This problem manifests itself directly because the estimate $\widehat{\beta}$ can be incredibly sensitive to the particular choice of the small number being added with little guidance on what is the "right" choice of small number. In general, like selecting away the zeroes, the "adding a small number trick" is also a biased estimator for β. We illustrate both biases in the Monte Carlo section (Section 5).

Despite their failure as general solutions, these "ad hoc zero fixes" have in them what could be a useful idea – Perhaps the variation among the non-zero share observations can be used to estimate the model parameters, while at the same time the presence of zeroes is controlled in such a way that avoids bias. We now present a new estimator that formalizes this possibility by using moment *inequalities* to control for the zeroes in the data while using the variation in the remaining part of the data to consistently estimate the demand parameters. We continue in this section to illustrate our approach within the logit model before treatment of the general case in the next section.

3.2 A Bounds Estimator

Our bounds approach turns the selection-on-unobservable problem into a selection-on-observable strategy, with the key features that the selection is not based on market share but on exogenous vari-

[4] In fact,

$$E[\xi_{jt}|x_{jt}, s_{jt} > 0] > 0 \tag{3.5}$$

in the homoskedastic case. This is because the criterion $s_{jt} > 0$ selects high values of ξ_{jt} and leaves out low values of ξ_{jt}.

[5] It is easy to see that the selection bias is of the same direction if the selection criterion is instead $s_{jt} > 0$ for all t, as one is effectively doing when focusing on a few top sellers that never demonstrate zero sales in the data. The reason is that the event $s_{jt} > 0$ for all t contains the event $s_{jt} > 0$ for a particular t. If the markets are weakly dependent, the particular t part of the selection dominates.

[6] Berry, Linton, and Pakes (2004)and Freyberger (2015) study the biasing effect of plugging in s_{jt} for π_{jt}. Their bias corrections do not apply when there are zeroes in the empirical shares.

ables, and is not determined ex-ante by the econometrician but rather automatically performed by the estimator. Specifically, we assume that there exist a set of "safe product/market" (j,t), identified by the instrumental variable z_{jt}, with inherently thick demand such that s_{jt} has a small chance of being zero. In particular, we assume a partition on the support of z_{jt}: $\text{supp}(z_{jt}) = \mathcal{Z} = \mathcal{Z}_0 \cup \mathcal{Z}_1$ that separates the safe product/markets ($z_{jt} \in \mathcal{Z}_0$) from the remaining "risky product/markets" ($z_{jt} \in \mathcal{Z}_1$). [7] The safe products have inherently desirable characteristics that often make them the "top sellers" described in Section 2, while the risky products have less attractive characteristics that often yield sparse demand. If we knew $\mathcal{Z}_0$ and focused on the observations such that $z_{jt} \in \mathcal{Z}_0$, the standard estimator would be consistent. The key challenge in the data is that the econometricians will not know $\mathcal{Z}_0$ in advance. Our bounds estimator automatically utilizes the variation in $\mathcal{Z}_0$, but at the same time safely controls for the observations in $\mathcal{Z}_1$, to consistently estimate β without requiring the researcher either to know or to estimate the underlying partition $(\mathcal{Z}_0, \mathcal{Z}_1)$.

Our approach first uses two mean-utility estimators: δ^u_{jt} and δ^ℓ_{jt} that are functions of empirical market shares (rather than the true choice probability), to form bounds on $E[\delta_{jt}|\, z_{jt}]$:

$$E\left[\delta^u_{jt}\middle|\, z_{jt}\right] \geq E\left[\delta_{jt}|\, z_{jt}\right] \geq E\left[\delta^\ell_{jt}\middle|\, z_{jt}\right], \forall j,t \ a.s. \tag{3.6}$$

where δ_{jt} is the true mean-utility in (3.1). Next, the inequalities (3.6) combined with (3.3) imply

$$E\left[\delta^u_{jt} - x'_{jt}\beta\middle|\, z_{jt}\right] \geq 0 \geq E\left[\delta^\ell_{jt} - x'_{jt}\beta\middle|\, z_{jt}\right] \ a.s. \tag{3.7}$$

Observe that the moment restriction (3.3) implies that

$$E\left[\left(\delta_{jt} - x'_{jt}\beta\right) g\left(z_{jt}\right)\right] = 0 \ \forall g \in \mathcal{G},$$

where $\mathcal{G}$ is a set of instrumental variable functions. Using instead our upper and lower mean utility estimators in place of the true mean utility we have the following moment inequalities

$$E\left[\left(\delta^u_{jt} - x'_{jt}\beta\right) g(z_{jt})\right] \geq 0 \geq E\left[\left(\delta^\ell_{jt} - x'_{jt}\beta\right) g(z_{jt})\right] \ \forall g \in \mathcal{G}. \tag{3.8}$$

Following Andrews and Shi (2013), we take each $g \in \mathcal{G}$ to be an indicator function for a hypercube $B_g \subseteq \text{supp}\,(z)$, i.e.,

$$g(z_{jt}) = 1\left(z_{jt} \in B_g\right),$$

and as long as $\mathcal{G}$ is rich enough, identification information in (3.7) is preserved by the moment equalities (3.8).

[7] We will formalize the requirement on the partition in Section 4.

To form our estimator, define

$$\bar{\rho}_T^u(\beta, g) = (TJ)^{-1} \sum_{t=1}^{T} \left(\sum_{j=1}^{J} \left(\delta_{jt}^u - x_{jt}^{'} \beta \right) g(z_{jt}) \right),$$
$$\bar{\rho}_T^\ell(\beta, g) = (TJ)^{-1} \sum_{t=1}^{T} \left(\sum_{j=1}^{J} \left(x_{jt}^{'} \beta - \delta_{jt}^\ell \right) g(z_{jt}) \right).$$

Let $[a]_-$ denote $|\min\{0, a\}|$. Our estimator is then

$$\widehat{\beta}^{BD} = \arg\min_{\theta} \sum_{g \in \mathcal{G}} \mu(g) \left\{ [\bar{\rho}_T^u(\beta, g)]_-^2 + \left[\bar{\rho}_T^\ell(\theta, g) \right]_-^2 \right\}, \tag{3.9}$$

where $\mu(g)$ is a probability density function on $\mathcal{G}$, that is $\mu(g) > 0$ for all $g \in \mathcal{G}$, and $\sum_{g \in \mathcal{G}} \mu(g) = 1$. The function $\mu(g)$ is used to ensure summability of the terms, and the choice of $\mu(\cdot)$ is discussed in the next section.

Why is $\hat{\beta}^{BD}$ consistent? A heuristic proof is as follows. Let us define the partition $\mathcal{G} = \mathcal{G}_0 \cup \mathcal{G}_1$ where each $g \in \mathcal{G}_0$ has support inside $\mathcal{Z}_0$. This partition does not need to be explicitly formed by the econometrician (only the flexible set of instrumental variable functions $\mathcal{G}$ over the entire support of z_{jt} in the observed data is needed as an input), but only needs to exist in the underlying DGP. We can then separate the objective function underlying (3.9) into two additive pieces

$$\sum_{g \in \mathcal{G}_0} \mu(g) \left\{ [\bar{\rho}_T^u(\beta, g)]_-^2 + \left[\bar{\rho}_T^\ell(\beta, g) \right]_-^2 \right\} + \sum_{g \in \mathcal{G}_1} \mu(g) \left\{ [\bar{\rho}_T^u(\beta, g)]_-^2 + \left[\bar{\rho}_T^\ell(\beta, g) \right]_-^2 \right\}. \tag{3.10}$$

Notice that at the true parameter value β^0, each of these sums in (3.10) converges in probability to 0 because of the validity of the moment inequalities (3.8) at the true value β^0. What happens away from the true value $\beta^* \neq \beta^0$? Observe that the second sum over $\mathcal{G}_1$ is by construction nonnegative regardless of the value of β. The first sum on the other hand approaches for each $g \in \mathcal{G}_0$ the square of

$$\sum_{g \in \mathcal{G}_0} \mu(g) E\left[\left(\delta_{jt} - x_{jt}^{'} \beta^* \right) g(z_{jt}) \right]$$

because $\bar{\rho}_T^u(\beta, g)$ and $\bar{\rho}_T^l(\beta, g)$ converge as $T \to \infty$ for $g \in \mathcal{G}_0$ (this is, for products whose z_{jt} lies in the safe set $\mathcal{Z}_0$). Then so long as the instruments $\{g(z_{jt})\}_{g \in \mathcal{G}_0}$ have sufficient variation for IV rank condition with x_{jt} to hold (the standard logit identifying condition), we are ensured that for at least a positive mass of $g \in \mathcal{G}_0$ we have that

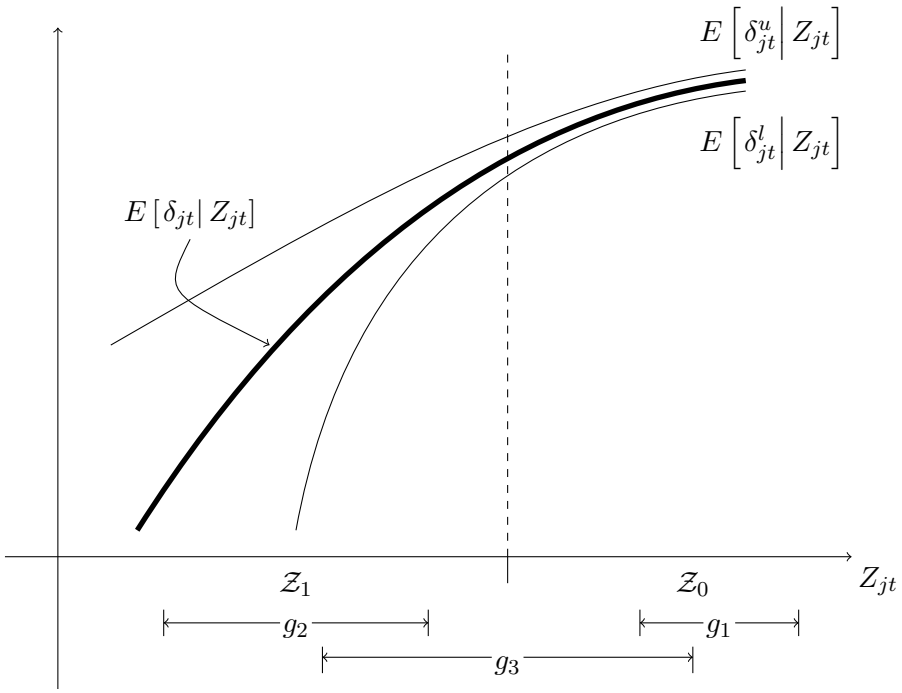
$$E\left[\left(\delta_{jt} - x_{jt}^{'} \beta^* \right) g(z_{jt}) \right] \neq 0.$$

Thus the first sum in (3.10) will converge in probability to a *strictly* positive number. Hence the limiting value of the objective function (3.9) attains a minimum at the true value β^0 and thus

by standard arguments $\widehat{\beta}^{BD} \to_p \beta_0$.

Figure 2 provides a graphical illustration of the above arguments. In the safe products region $\mathcal{Z}_0$, the bounds are tight and provide identification power, while in $\mathcal{Z}_1$, the bounds may be uninformative but still valid. So instrumental functions such as $g_1 \in \mathcal{G}_0$ will form moment equalities that point identify the model. Other instrumental functions, such as $g_2, g_3 \in \mathcal{G}_1$, are associated with slack moment inequalities so they do not undermine the identification.

Figure 2: Illustration of Bounds Approach



The bounds estimator thus controls for the zeroes in the data while using the variation among the safe products to consistently estimate the model parameters. We now generalize this logic and formalize it to the general differentiated product demand context with general error distribution for the random utility model. We will show both consistency and asymptotic normality of the estimator in this general case.

4 The General Model and Estimator

The researcher has data on a sample of markets $t = 1 \ldots, T$, and for each market t, there is a sample of individuals $i = 1, \ldots n_t$ choosing from the $j = 0, \ldots, J_t$ products in the market. A product j in market t is characterized by a vector of characteristics $x_{jt} \in \mathbb{R}^{d_x}$ that are observed to the researcher, and a scalar unobserved product attribute ξ_{jt}. We will refer to the bundle (x_{jt}, ξ_{jt}) as $j's$ product characteristics (observed and unobserved). Note that to better match the feature of popular data sets, we allow a t subscript for J, that is, different markets can have different number of products. We will also allow a t subscript for n, the number of potential consumers.

In discrete choice models each consumer $i = 1, \ldots, n_t$ in market t is assumed to make a single choice from the product varieties $j = 0, \ldots, J_t$ in the market, where $j = 0$ denotes the outside option of not purchasing. This choice is determined by maximizing a utility function that is random from the perspective of the researcher. Specifically, the utility consumer i derives from consuming product j in market t is given by

$$u_{ijt} = \delta_{jt} + \epsilon_{ijt},$$

where:

1. δ_{jt} is the mean-utility of product j in market t. Normalize $\delta_{0t} = 0$. As is standard, δ_{jt} is modeled as

$$\delta_{jt} = x_{jt}^{'}\beta + \xi_{jt}, \tag{4.1}$$

   where x_{jt} is the vector of observable (product, market) characteristics, often including price, and ξ_{jt} is the vector of unobservable characteristics;

2. ϵ_{ijt} is the idiosyncratic taste shock governed by the following distribution,

$$\epsilon_{it} = (\epsilon_{i0t}, \ldots, \epsilon_{iJ_tt}) \sim F\left(\cdot \,|x_t; \lambda\right), \tag{4.2}$$

   where x_t stands for $(x_{1t}^{'}, \ldots, x_{J_tt}^{'})'$, and $F(\cdot|x_t, \lambda)$ is a conditional cumulative distribution function known up to the finite dimensional unknown parameter λ. Thus, the unknown parameter in the model is $\theta = (\beta', \lambda')'$. For clarity, we use $\theta_0 \equiv (\beta_0', \lambda_0')'$ to denote the true value of the unknown parameter.

It is worth noting that allowing x_t and the parameter λ to enter F makes this specification encompass random coefficient specifications $u_{ijt} = x_{jt}^{'}\beta_i + \xi_{jt}$, where β_i follows some distribution (e.g., joint normal), because one can then view β as the mean of the random coefficients and ϵ_{ijt} as the sum of the products of the de-meaned random coefficients and the product characteristic x_{jt}.[8]

We assume consumers demand the product that maximizes utility. Thus integrating out ϵ_{it} yields a system of choice probabilities for agents in the market

$$\sigma(\delta_t, x_t, \lambda) \equiv (\sigma_1(\delta_t, x_t, \lambda), \ldots, \sigma_{J_t}(\delta_t, x_t, \lambda))',$$

where $\delta_t = (\delta_{1t}, \ldots, \delta_{J_tt})'$. Then we obtain the demand system

$$\pi_t \equiv (\pi_{1t}, \ldots, \pi_{J_t})' = \sigma(\delta_t, x_t, \lambda), \tag{4.3}$$

where $\pi_{jt} = \Pr(\text{product } j \text{ is chosen in market } t)$ represents the true choice probability of product j in market t. Let $\sigma^{-1}(\pi_t, x_t, \lambda) \equiv (\sigma_1^{-1}(\pi_t, x_t, \lambda), \ldots, \sigma_{J_t}^{-1}(\pi_t, x_t, \lambda))'$ denote the inverse demand function such that:

$$\delta_t = \sigma^{-1}(\pi_t, x_t, \lambda). \tag{4.4}$$

[8] Requiring $F(\cdot|x_t, \lambda)$ to be known up to a finite dimensional parameter rules out the vertical model because for the vertical model, ϵ_{it} is a function of the unobservable product characteristics (quality).

Note that in the simple logit model, $\sigma_j(\delta_t, x_t, \lambda)$ reduces to $\sigma_j(\delta_t) = \frac{\exp(\delta_{jt})}{1+\sum_{j'=1}^{J_t}\exp(\delta_{j't})}$, and $\sigma_j^{-1}(\pi_t, x_t, \lambda)$ reduces to $\sigma^{-1}(\pi_t) = \ln(\pi_{jt}) - \ln(\pi_{0t})$.

Inverting the demand system allows for the use of instrumental variables to identify θ. In particular, instruments for the model are a random vector z_{jt} that satisfies

$$E\left[\xi_{jt}\left|z_{jt}\right.\right] = 0. \tag{4.5}$$

Combining (4.4) and (4.5), the model yields the following moment restriction:

$$E\left[\sigma_j^{-1}(\pi_t, x_t, \lambda) - x_{jt}'\beta \middle| z_{jt}\right] = 0. \tag{4.6}$$

If π_t is observed, identification can be stated as follows. The model is identified if and only if for any $\theta = (\beta, \lambda) \neq \theta_0$,

$$\Pr_F\left(m_F^*(\theta, z_{jt}) \neq 0 \text{ and } z_{jt} \in \mathcal{Z} = supp\left(z_{jt}\right)\right) > 0,$$

where

$$m_F^*\left(\theta, z_{jt}\right) = E\left[\sigma_j^{-1}\left(\pi_t, x_t; \lambda\right) - x_{jt}'\beta \middle| z_{jt}\right] \tag{4.7}$$

Primitive conditions for identification are given in Berry and Haile (2014).

4.1 Bounds Estimator in the General Case

Like in the logit case, we construct a pair of inverse demand functions: $\delta_{jt}^u(\lambda)$ and $\delta_{jt}^\ell(\lambda)$, to form bounds on $E\left[\sigma_j^{-1}\left(\pi_t, x_t, \lambda\right) \middle| z_{jt}\right]$, i.e.,

$$E\left[\delta_{jt}^u(\lambda) \middle| z_{jt}\right] \geq E\left[\sigma_j^{-1}\left(\pi_t, x_t, \lambda\right) \middle| z_{jt}\right] \geq E\left[\delta_{jt}^\ell(\lambda) \middle| z_{jt}\right], \ a.s. \tag{4.8}$$

These inequalities combined with (4.5) form the moment inequalities that our estimation of θ is based upon:

$$E\left[\delta_{jt}^u(\lambda) - x_{jt}'\beta \middle| z_{jt}\right] \geq 0 \geq E\left[\delta_{jt}^\ell(\lambda) - x_{jt}'\beta \middle| z_{jt}\right], \ a.s. \tag{4.9}$$

To construct these upper and lower mean utility estimates $\delta_{jt}^l(\lambda), \delta_{jt}^u(\lambda)$, we start by applying the Laplace rule of succession to obtain an initial choice probability estimator that does not have zeros: $\tilde{s}_{jt} = \frac{ns_{jt}+1}{n+J+1}$.[9] We call this the Laplace share estimator. It is a good estimator for the choice probabilities when the prior information is only that these probabilities should be positive, as argued in Jaynes (2003, Chap. 18), and thus provides a good starting point for our construction.

[9]The Laplace rule of succession was proposed by Pierre-Simon Laplace in the early 19th century to predict the probability of an event happening given n independent past observations and the prior knowledge that the probability must be strictly between 0 and 1. It is a concept fundamental to modern probability theory despite being widely misunderstood and criticized. See Jaynes (2003, Chap. 18) for a thorough discussion.

We do not use the Laplace share estimator directly in place of π_t, but use it to construct bounds $\delta^u_{jt}(\lambda)$ and $\delta^\ell_{jt}(\lambda)$. Specifically, we define

$$\delta^u_{jt}(\lambda) = \Delta_{jt}(\tilde{s}_t, x_t; \lambda) + \log\left(\frac{\tilde{s}_{jt} + \eta_t}{\tilde{s}_{0t} - \eta_t}\right) \tag{4.10}$$

$$\delta^\ell_{jt}(\lambda) = \Delta_{jt}(\tilde{s}_t, x_t; \lambda) + \log\left(\frac{\tilde{s}_{jt} - \eta_t}{\tilde{s}_{0t} + \eta_t}\right), \tag{4.11}$$

where

$$\Delta_{jt}(\tilde{s}_t, x_t; \lambda) \equiv \sigma_j^{-1}(\tilde{s}_t, x_t; \lambda) - \log\left(\frac{\tilde{s}_{jt}}{\tilde{s}_{0t}}\right), \tag{4.12}$$

and η_t is a scalar in $(0, 1/(n_t + J_t + 1))$.

It is instructive to consider the simple logit case, where $\sigma_j^{-1}(\tilde{s}_t, x_t; \lambda) = \log\left(\frac{\tilde{s}_{jt}}{\tilde{s}_{0t}}\right)$, the term $\Delta_{jt}(\tilde{s}_t, x_t; \lambda) = 0$ so the bounds boil down to

$$\delta^u_{jt} = \log\left(\frac{\tilde{s}_{jt} + \eta_t}{\tilde{s}_{0t} - \eta_t}\right) \text{ and } \delta^\ell_{jt} = \log\left(\frac{\tilde{s}_{jt} - \eta_t}{\tilde{s}_{0t} + \eta_t}\right). \tag{4.13}$$

Thus the tuning term η_t perturbs (both in a positive and negative direction) the Laplace share for each product one at a time, and $\delta^u_{jt}, \delta^l_{jt}$ are then formed by applying the logit inversion to this perturbed share.

Remark 1. Observe that the distance between δ^u_{jt} and δ^ℓ_{jt} is large when $\tilde{s}_{jt}$ is small (i.e., respecting the large error caused by the noise in s_t) and is negligible when $\tilde{s}_{jt}$ is large. Thus intuitively, for observations such that $z_{jt} \in \mathcal{Z}_0$, which defines the safe products in a market, $\tilde{s}_{jt}$ is large with a high probability so that $E\left[\delta^u_{jt} \middle| z_{jt}\right]$ and $E\left[\delta^\ell_{jt} \middle| z_{jt}\right]$ closely resemble $E[\delta_{jt} | z_{jt}]$. On the other hand, the difference between $E\left[\delta^u_{jt} \middle| z_{jt}\right]$ and $E\left[\delta^\ell_{jt} \middle| z_{jt}\right]$ may be large for risky products (i.e., $z_{jt} \in \mathcal{Z}_1$) because $\tilde{s}_{jt}$ has a high probability being close to zero. This feature of the construction is a key to the consistency result to be discussed later.

We now formally establish the validity of the bounds defined by (4.10) and (4.11).

Assumption 1. *The conditional distribution of* $(n_t s_{jt})_{j=0}^{J_t}$ *given* $(\pi_{jt}, x_{jt}, z_{jt})_{j=1}^{J_t}$ *is multinomial with parameters* n_t *and* $(\pi_{jt})_{j=0}^{J_t}$.

Assumption 2. *The inverse demand function* $\sigma_j^{-1}(\cdot, x_t, \lambda)$ *is well-defined and continuous on the probability simplex* $\Delta^{J_t} \equiv \{(p_1, \ldots, p_{J_t}) \in (0,1)^{J_t} : 1 - \sum_{j=1}^{J_t} p_j > 0\}$ *for any* x_t *and any* λ.

Lemma 1. *Suppose that Assumptions 1 and 2 hold. Then, there exists* $\eta_t \in (0, 1/(n_t + J_t + 1))$ *such that the inequalities in (4.8) hold at* $\lambda = \lambda_0$ *with* $\delta^u_{jt}(\lambda)$ *and* $\delta^\ell_{jt}(\lambda)$ *defined in (4.10) and (4.11).*

Remark 2. The scalar η_t is chosen to guarantee equation (4.8). The η_t satisfying (4.8) may depend on π_t, x_t, and n_t, and thus may itself be a random variable, which makes it appear difficult to choose. However, we find that a rule of thumb works very well in both our Monte Carlo and

empirical exercises. The rule is to choose, for example, $\eta_t = \frac{1-10^{-3}}{n_t+J_t+1}$ to start with, and increasing it to $\eta_t = \frac{1-10^{-4}}{n_t+J_t+1}$, and to $\eta_t = \frac{1-10^{-5}}{n_t+J_t+1}$, and so on, until the estimates stabilize. To see why this rule of thumb is reasonable, it is useful to note that if one choice, say, η_t^1, satisfies (4.8), another choice, say η_t^2, that lies between η_t^1 and $1/(n_t + J_t + 1)$ also satisfies (4.8). This is so due to the monotonicity of right-hand-side of (4.10) and (4.11) in η_t. On the other hand, using η_t's that are closer to the boundary $1/(n_t + J_t + 1)$ will generally not hurt estimation precision much because identification is based on the safe products, for which even the upper bound $1/(n_t + J_t + 1)$ is negligible relative to $\tilde{s}_{jt}$ with high probability. This suggests that we do not need to know the precise range of η_t's that work, but can afford to make a conservative choice, as our rule of thumb does.

In order to estimate θ based on the moment inequalities (4.9), we first transform the conditional moments inequalities into unconditional ones, following Andrews and Shi (2013), using a set $\mathcal{G}$ of instrumental functions, where an instrumental function is a function of z_{jt}. The set $\mathcal{G}$ that we use is given below, and it guarantees that (4.9) is equivalent to

$$E\left[(\delta_{jt}^u(\lambda) - x_{jt}'\beta)g(z_{jt})\right] \geq 0 \geq E\left[(\delta_{jt}^\ell(\lambda) - x_{jt}'\beta)g(z_{jt})\right]. \tag{4.14}$$

Andrews and Shi (2013) discussed many different choices of $\mathcal{G}$ including uncountable sets and countable sets. We only consider countable $\mathcal{G}$ sets. Thus, given a data set of J products from T markets, we can construct a sample criterion function as:

$$\widehat{Q}_T(\theta) = \sum_{g\in\mathcal{G}} [\bar{\rho}_T^u(\theta, g)]_-^2 \, \mu(g) + \sum_{g\in\mathcal{G}} \left[\bar{\rho}_T^\ell(\theta, g)\right]_-^2 \mu(g), \tag{4.15}$$

where

$$\begin{aligned}
\bar{\rho}_T^u(\theta, g) &= (T\bar{J})^{-1} \sum_{t=1}^{T} \left(\sum_{j=1}^{J_t} (\delta_{jt}^u(\lambda) - x_{jt}'\beta)g(z_{jt}) \right), \\
\bar{\rho}_T^\ell(\theta, g) &= (T\bar{J})^{-1} \sum_{t=1}^{T} \left(\sum_{j=1}^{J_t} (x_{jt}'\beta - \delta_{jt}^\ell(\lambda))g(z_{jt}) \right),
\end{aligned} \tag{4.16}$$

where $\bar{J} = T^{-1}\sum_{t=1}^{T} J_t$ is the average number of products on a market. The function $\mu(\cdot)$ is a probability distribution on $\mathcal{G}$, which gives weights to each unconditional moment inequality. Our choice for $\mu(\cdot)$ is given below after the choice for $\mathcal{G}$ is introduced.

Our bound estimator for $\theta = (\beta', \lambda')'$ is defined as [10]

$$\widehat{\theta}_T^{BD} = \arg\min_{\theta} \widehat{Q}_T(\theta). \tag{4.17}$$

Numerically solving for $\widehat{\theta}_T^{BD}$ is not much different from solving for the standard BLP estimator. As in the standard procedure, the criterion function is convex in β [11]. Thus, it is useful to separate the minimization problem into two steps:

$$\min_{\lambda} \min_{\beta} \widehat{Q}_T(\beta, \lambda). \tag{4.18}$$

The β minimization can be solved efficiently and accurately even when many control variables are included in x_{jt}. The λ minimization typically is a low-dimensional problem. One point worth noting is that the inverse demand functions involved in the quantities $\delta_{jt}^u(\lambda)$ and $\delta_{jt}^\ell(\lambda)$ can be solved by the same contraction mapping algorithm used in the standard BLP procedure. Alternatively, the optimization problem (4.18) can be formulated and solved as a MPEC problem using the machinery of Dubé, Fox, and Su (2012).

Now we define the instrumental function collection $\mathcal{G}$ and the weight on it $\mu(\cdot)$ that we use in the simulation and the empirical application of this paper.[12] For $\mathcal{G}$, we divide the instrument vector z_{jt} into discrete instruments, $z_{d,jt}$, and continuous instruments $z_{c,jt}$. Let the set $\mathcal{Z}_d$ be the discrete set of values that $z_{d,jt}$ can take. Normalize the continuous instruments to lie in $[0,1]$: $\tilde{z}_{c,jt} = F_{N(0,1)}\left(\widehat{\Sigma}_{z_c}^{-1/2} z_{c,jt}\right)$, where $F_{N(0,1)}(\cdot)$ is the standard normal cdf and $\widehat{\Sigma}_{z_c}$ is the sample covariance matrix of $z_{c,jt}$. The set $\mathcal{G}$ is defined as

$$\begin{aligned}
\mathcal{G} &= \{g_{a,r,\zeta}(z_d, z_c) = 1((\tilde{z}_c', z_d')' \in C_{a,r,\zeta}) : C_{a,r,\zeta} \in \mathcal{C}\}, \text{ where} \\
\mathcal{C} &= \{(\times_{u=1}^{d_{z_c}}((a_u - 1)/(2r), a_u/(2r)]) \times \{\zeta\} : \ a_u \in \{1, 2, ..., 2r\}, \text{ for } u = 1, ..., d_{z_c}, \\
&\qquad r = r_0, r_0 + 1, ..., \text{ and } \zeta \in \mathcal{Z}_d\}.
\end{aligned} \tag{4.19}$$

In practice, we truncate r at a finite value $\bar{r}_T$. This does not affect the first order asymptotic property of our estimator as long as $\bar{r}_T \to \infty$. For $\mu(\cdot)$, we use

$$\mu(\{g_{a,r,\zeta}\}) \propto (100 + r)^{-2} (2r)^{-d_{z_c}} K_d^{-1} \text{ for } g \in \mathcal{G}_{d,cc}, \tag{4.20}$$

where K_d is the number of elements in $\mathcal{Z}_d$. The same μ measure is used and works well in Andrews and Shi (2013).

[10] When there is not a partition in the space of z_{jt} that distinguishes the safe products out, the moment inequalities (4.9) partially identify θ. In that case, the confidence set procedure in Andrews and Shi (2013), as well as the profiling approach in an early version of this paper, may be used for inference. However, in the current version of this paper, we focus on the point identification case, which is much more computationally tractable.

[11] The convexity can be seen by examining the second order derivative of $\widehat{Q}_T(\theta)$ with respect to β.

[12] We note that appropriate choices of $\mathcal{G}$ and μ are not unique. For other possible choices, see Andrews and Shi (2013).

4.2 Consistency and Asymptotic Normality

In the asymptotic framework, we let the number of markets T go to infinity, and let the number of consumers in each market, n_t, be a function of T that also goes to infinity as T does. The number of products J_t may also be a function of T that goes to infinity as the latter; it may also stay finite.

The key concept behind our approach is the notion of safe products. We define the safe products according to the value that z_{jt} takes. Let $\mathcal{Z}_0$ be a subset of R^{d_z}, where d_z is the dimension of z_{jt}. The product j is said to be a safe product in market t if $z_{jt} \in \mathcal{Z}_0$. Thus, the instrumental variable not only induces exogenous variation of the explanatory variables as in standard setup, but also serves as an identifier of the safe products. The requirements on the set $\mathcal{Z}_0$ is listed below.

If j is a safe product in market t, its market share π_{jt} tends to be sufficiently different from zero, so that the slope of $\sigma_j^{-1}(\pi_t, x_t; \lambda)$ at the true choice probability π_t tends not to be huge. As a result the inverse demand function $\sigma_j^{-1}(\widehat{\pi}_t, x_t; \lambda)$ should be sufficiently close to $\sigma_j^{-1}(\pi_t, x_t; \lambda)$ for a consistent estimator $\widehat{\pi}_t$ of π_t. Thus, the first requirement is as follows.

Assumption 3. *For any estimator $\widehat{\pi}_t$ of π_t such that $\sup_{j=0,\ldots,J_t,t=1,\ldots,T} |\widehat{\pi}_{jt} - \pi_{jt}| \to_p 0$, we have*

(a) $\sup_{t=1,\ldots,T;j=1,\ldots,J_t:z_{jt}\in\mathcal{Z}_0} \sup_\lambda |(\sigma_j^{-1}(\widehat{\pi}_t, x_t; \lambda) - \sigma_j^{-1}(\pi_t, x_t; \lambda))| \to_p 0$.

(b) $\sup_{t=1,\ldots,T;j=0,\ldots,J_t:z_{jt}\in\mathcal{Z}_0} |\ln \widehat{\pi}_{jt} - \ln \pi_{jt}| \to_p 0$.

Remark 3. Assumption 3 is a strict generalization of the key consistency requirement for the standard estimator as formalized by Assumption A.8 in Freyberger (2015) or Assumption A5 in Berry, Linton, and Pakes (2004). Our Assumption 3 relaxes their approach by not placing any restriction on the size of π_{jt} for the risky products ($z_{jt} \notin \mathcal{Z}_0$). For those products, π_{jt} can be very small (asymptotically, it can approach zero very fast), and $\sigma_j^{-1}(s_t, x_t; \lambda)$ can be very different from $\sigma_j^{-1}(\pi_t, x_t; \lambda)$ (asymptotically, the two may not converge to each other), causing standard estimators to fail.

In order to leverage the mass of safe product/market realizations in $\mathcal{Z}_0$ to achieve consistency, we need to ensure that the variation in $\mathcal{Z}_0$ alone is enough to point identify θ_0. This is the analogue of the general identification condition (4.7) holding if we hypothetically selected the sample so that $z_{jt} \in \mathcal{Z}_0$.

Assumption 4. *For any $\theta \neq \theta_0$, $\Pr_F\left(m_F^*(\theta, z_{jt}) \neq 0 \text{ and } z_{jt} \in \mathcal{Z}_0\right) > 0$, where $m_F^*(\theta, z_{jt})$ is defined in (4.7).*

Note that if the econometrician ex-ante knew $\mathcal{Z}_0$, then it would be straightforward to implement the standard GMM estimator on the subsample $\mathcal{Z}_0$. But, we do not know $\mathcal{Z}_0$ ex-ante. The main idea behind the design of our bound estimator is to automatically utilizes the identification information in $\mathcal{Z}_0$ while safely controlling for the presence of the risky mass $\mathcal{Z}_1$, without requiring the researcher either to know or to estimate the partition ex-ante.

Assumption 5 below is a regularity condition that guarantees that the identification in the assumption above can be achieved through the instrumental functions defined in (4.19) in the fashion of Andrews and Shi (2013). Part (a) is a moment condition that is stronger than needed for

obtaining the Andrews and Shi (2013) type result, but the extra strength is used for the consistency of $\widehat{\theta}_T^{BD}$ later.

Assumption 5. (a) $E_F[\sup_{\theta\in\Theta}|\sigma_j^{-1}(\pi_t, x_t; \lambda) - x_{jt}'\beta|1\{z_{jt} \in \mathcal{Z}_0\}] < \infty$.

(b) *The set $\mathcal{Z}_0$ is a countable disjoint union of elements in $\mathcal{C}$, where $\mathcal{C}$ is defined in (4.19).*

Lemma 2. *Under Assumptions 4 and 5, we have for any $\theta \neq \theta_0$, there exists a $g_{a,r,\zeta} \in \mathcal{G}$ such that $\mathcal{C}_{a,r,\zeta} \subseteq \mathcal{Z}_0$ and*

$$E_F[(\sigma_j^{-1}(\pi_t, x_t; \lambda) - x_{jt}'\beta)g_{a,r,\zeta}(z_{jt})] \neq 0. \tag{4.21}$$

A few more standard assumptions are also needed for consistency. These are given next. Let

$$\rho_F^*(\theta, g) = E_F[(\sigma_j^{-1}(\pi_t, x_t; \lambda) - x_{jt}'\beta)g(z_{jt})]. \tag{4.22}$$

Assumption 6. (a) *At any point λ, $\sigma_j^{-1}(\pi_t, x_t; \lambda)$ is continuous in λ with probability one.*

(b) $\sup_{\theta\in\Theta}\left|(T\bar{J})^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}((\sigma_j^{-1}(\pi_t, x_t; \lambda) - x_{jt}'\beta)g(z_{jt}) - \rho_F^*(\theta, g))\right| \to_p 0$ *for all* $g \in \mathcal{G}_0$.

Assumption 7 (a) below is the same as the analogous condition in Freyberger (2015) and (b) is weaker than requiring J_t to be bounded.

Assumption 7. (a) $\max_{j,t}|s_{jt} - \pi_{jt}| \to_p 0$ *as* $T \to \infty$.

(b) $\min_{t=1,\ldots,T}\{n_t\} \to \infty$ *and* $\max_{t=1,\ldots,T}\{J_t/n_t\} \to 0$.

Define:

$$\begin{aligned}\rho_{F,T}^u(\theta, g) &= E_F[(\delta_{jt}^u(\lambda) - x_{jt}'\beta)g(z_{jt})] \\ \rho_{F,T}^\ell(\theta, g) &= E_F[(x_{jt}'\beta - \delta_{jt}^\ell(\lambda))g(z_{jt})].\end{aligned} \tag{4.23}$$

These functions have the T subscript because the η_t that enters $\delta_{jt}^u(\lambda)$ and $\delta_{jt}^\ell(\lambda)$ depends on n_t and J_t, which depend on T. Assumption 8 is a uniform law of large number type requirement, which is implied by some mild moment existence conditions if the markets are independent from each other.

Assumption 8. *For $k = u, \ell$, the functions $\rho_{F,T}^k$ is well defined, and*

$$\sup_{g\in\mathcal{G}}\left|(T\bar{J})^{-1}\sum_{t=1}^{T}\sum_{j=1}^{J_t}((\delta_{jt}^k(\lambda_0) - \beta_0'x_{jt})g(z_{jt}) - \rho_{F,T}^k(\theta_0, g))\right| \to_p 0.$$

The following theorem shows the consistency of the bound estimator.

Theorem 1. *Suppose that Assumptions 1-8 hold. Then*

$$\|\widehat{\theta}_T^{BD} - \theta_0\| \to_p 0. \tag{4.24}$$

More Assumptions are need to derive the asymptotic normality of the bound estimator. These conditions are technical rather than illuminating. Thus, we relegate them to Appendix C.1.

Theorem 2. *Suppose that Assumptions 1-8 and C.1-C.7 hold. Then*

$$\sqrt{T\bar{J}}(\widehat{\theta}_T^{BD} - \theta_0) \to_d N(0, \Gamma V \Gamma),$$

where $\Gamma = \left[\sum_{g\in\mathcal{G}_0} \frac{\partial \rho_F^*(\theta_0,g)}{\partial \theta} \frac{\partial \rho_F^*(\theta_0,g)}{\partial \theta'} \mu(g)\right]^{-1}$, *and*

$$V = \sum_{g,g^*\in\mathcal{G}_0} \Sigma(g,g^*) \frac{\partial \rho_F^*(\theta_0,g)}{\partial \theta} \frac{\partial \rho_F^*(\theta_0,g^*)}{\partial \theta'} \mu(g)\mu(g^*), \text{ with} \tag{4.25}$$

$\mathcal{G}_0 = \{g_{a,r,\zeta} \in \mathcal{G} : \Pr((\tilde{z}_c', z_d)' \in C_{a,r,\zeta}) = \Pr((\tilde{z}_c', z_d)' \in C_{a,r,\zeta} \cap \mathcal{Z}_0)\}$, *and* $\Sigma(g,g^*)$ *being the limit of of*

$$Cov\left((T\bar{J})^{-1/2} \sum_{t=1}^{T} \sum_{j=1}^{J_t} (\sigma_j^{-1}(\pi_t, x_t; \lambda) - \beta' x_{jt}) g(z_{jt}), (T\bar{J})^{-1/2} \sum_{t=1}^{T} \sum_{j=1}^{J_t} (\sigma_j^{-1}(\pi_t, x_t; \lambda) - \beta' x_{jt}) g^*(z_{jt}) \right).$$

A standard bootstrap procedure can be used to estimate the standard deviation of the estimator in practice and we shall discuss the implementation details of this procedure in the empirical section.

4.3 Partial Identification as an Alternative

The approach above provides a consistent point estimator based on an underlying set of moment inequalities. Point estimation relies on Assumptions 3 and 4, which allows for using variation among safe products for consistency. This is natural in many applications where the long tail pattern is present and we illustrate its performance in the Monte Carlo below. Nevertheless in settings where these Assumptions are questionable, we can still use the underlying moment inequalities (4.14) as a basis for partial identification and inference.

The model (4.14) is a moment inequality model with many moment conditions. One can use the method developed in Andrews and Shi (2013) to construct a joint confidence set for the full vector θ_0. This confidence set is constructed by inverting an Anderson-Rubin test: $CS = \{\theta : T(\theta) \leq c(\theta)\}$ for some test statistic $T(\theta)$ and critical value $c(\theta)$. Computing this set amounts to computing the 0-level set of the function $T(\theta) - c(\theta)$, where $c(\theta)$ typically is simulated quantiles and thus a non-smooth function of θ. This is feasible if the dimension of θ_0 is moderate, especially if one has access to parallel computing technology. If the dimension is high, however, the computational cost gets exponentially higher, and methods for it have not been well developed.

On the other hand, in demand estimation, θ_0 is high dimensional mainly because of many control variables included in x_{jt}. The coefficients of the control variables are nuisance parameters that often are of no particular interest. The typical parameters of interest are the price coefficient or the price elasticities, which are small dimensional. Based on this observation, we propose a *profiling* method to profile out the nuisance parameters and only construct confidence sets for a

parameter of interest. Since this part of the discussion is rather technical and tangential to our main contribution, we relegate it to Appendix D. Also, readers are referred to the early version of this paper (Gandhi, Lu, and Shi (2013)) for Monte Carlo simulations and empirical results using the profiling approach under partial identification.

5 Monte Carlo Simulations

In this section, we present two sets of Monte Carlo experiments with random coefficient logit models. The first experiment investigates the performance of our approach with moderate fractions of zero shares, which should cover most of the empirical scenarios. In the second experiment, we test our estimator with a data generating process that produces extremely large fractions of zeros; the purpose is to further illustrate the key idea of our estimator in exploiting the long tail pattern that is naturally present in the data.

Both experiments use the a random coefficient logit model, where the utility of consumer i for product j in market t is

$$u_{ijt} = \alpha_0 + x_{jt}\beta_0 + \lambda_0 x_{jt} v_i + \xi_{jt} + \epsilon_{ijt},$$

where $v_i \sim N(0,1)$, λ_0 is the standard deviation of the random coefficients on x_{jt}, ϵ_{ijt}'s are i.i.d. across i, j and t following Type I extreme value distribution. The parameters of interest are β_0 and λ_0, while α_0 is a nuisance parameter. In both experiments, we set $\lambda_0 = .5$, $\beta_0 = 1$ and vary α_0 for different designs. We simulate T markets, each with J products.

5.1 Moderately Many Zeroes

In the first experiment, the observed and unobserved characteristics are generated as $x_{jt} = \frac{j}{10} + N(0,1)$ and $\xi_{jt} \sim N\left(0, .1^2\right)$ for each product j in market t. Thus one feature of the design is that the x_{jt} has some persistence across markets - products with larger index tend to have higher value of x (which respects the nature of the variation in the scanner data shown in Section 2. Finally, the vector of empirical shares in market t, $(s_{0t}, s_{1t}, ..., s_{Jt})$, is generated from Multinomial $\left(n, [\pi_{0t}, \pi_{1t}, ..., \pi_{Jt}]'\right) \Big/ n$, where n represents the number of consumers in each market.[13]

With the simulated data set $\{(s_{jt}, x_{jt}) : j = 1, ..., J\}_{t=1}^{T}$, we compute our bound estimator (bound), the standard BLP estimator using s_t in place of π_t and discarding observations with $s_{jt} = 0$ (ES), the standard BLP estimator using $\tilde{s}_t$ (no zeros) in place of π_t (LS).

All the estimators require simulating the market shares and solving demand systems for each trial of λ in optimizing the objective function for estimation. We use the same set of random draws

[13] The π_t has no closed form solution in the random coefficient model, and thus, we compute them via simulation, i.e.,

$$\pi_{jt} = \frac{1}{s}\sum_{i=1}^{s} \frac{\exp\left(\alpha_0 + x_{jt}\beta_0 + \lambda_0 x_{jt} v_i + \xi_{jt}\right)}{1 + \sum_{k=1}^{J} \exp\left(\alpha_0 + x_{kt}\beta_0 + \lambda_0 x_{kt} v_i + \xi_{kt}\right)},$$

where $s = 1000$ is the number of consumer type draws (v_i).

of v_i as in the data generating process to eliminate simulation error as it is not the focus of this paper. BLP contraction mapping method is employed to numerically solve the demand systems.

We simulate 1000 datasets $\{(s_t^r, x_t^r) : t = 1, ..., T\}_{r=1}^{1000}$ and implement all the estimators mentioned above on each for a repeated simulation study. For the instrumental functions, we use the countable hyper-cubes defined in (4.19), and set $\bar{r}_T = 50$. We let $\eta = \frac{1-\iota}{n+J+1}$ with $\iota = 10^{-6}$ in constructing the bounds on the conditional expectation of the inverse demand function. Setting smaller ι, e.g., 10^{-10} gives virtually the same results as reported in the following tables. For the BLP estimator, we use $\left(1, x_{jt}, x_{jt}^2 - 1, x_{jt}^3 - 3x_{jt}\right)$ (the first three Hermite polynomials) as instruments to construct the GMM objective function. Alternative transformations of x_{jt} as instruments yield effectively the same results.

The bias and standard deviation of the estimators are presented in *Table* 2. As we can see from the table, The standard estimator with s_t shows large bias for both β and λ. Replacing the empirical share s_t with the Laplace share $\tilde{s}_t$ (and thus not discarding the observations with $s_{jt} = 0$) increases the bias for β although reducing the bias for λ. Our bound estimators are the least biased, and its bias is very small for both parameters, especially when the sample size (T) is larger.

Table 2: Monte Carlo Results: Random-Coefficient Logit Model

DGP	T	Ave. % of Zeros		ES		Bound		LS	
				β	λ	β	λ	β	λ
I	25	9.53%	Bias	-.1936	.3706	-.0443	.0436	-.2380	.2938
			SD	.0185	.0354	.0348	.0474	.0189	.0296
	50	9.46%	Bias	-.1940	.3717	-.0236	.0195	-.2353	.2916
			SD	.0150	.0271	.0294	.0399	.0146	.0229
	100	9.48%	Bias	-.1939	.3706	-.0081	.0018	-.2347	.2901
			SD	.0126	.0215	.0235	.0315	.0118	.0191
II	25	18.58%	Bias	-.6104	.6730	-.0329	.0169	-.4900	.3994
			SD	.0664	.0841	.0534	.0525	.0319	.0388
	50	18.55%	Bias	-.6036	.6648	-.0040	-.0069	-.4867	.3970
			SD	.0528	.0662	.0403	.0399	.0242	.0300
	100	18.53%	Bias	-.6018	.6613	.0037	-.0120	-.4865	.3960
			SD	.0394	.0489	.0299	.0298	.0199	.0250
III	25	41.16%	Bias	-1.3199	.7299	.0253	-.0344	-1.0112	.3830
			SD	.3056	.2201	.0725	.0487	.0564	.0476
	50	41.12%	Bias	-1.2937	.7099	.0263	-.0299	-1.0060	.3794
			SD	.2003	.1418	.0550	.0375	.0430	.0367
	100	41.07%	Bias	-1.2903	.7051	.0112	-.0171	-1.0044	.3762
			SD	.1435	.1028	.0394	.0282	.0342	.0282
IV	25	52.41%	Bias	-1.1039	.4041	.0453	-.0461	-1.1613	.2857
			SD	.2467	.1381	.0939	.0549	.0551	.0416
	50	52.38%	Bias	-1.0969	.3973	.0260	-.0297	-1.1564	.2829
			SD	.1804	.1017	.0665	.0415	.0422	.0318
	100	52.35%	Bias	-1.0901	.3922	.0104	-.0175	-1.1548	.2805
			SD	.1335	.0761	.0493	.0327	.0335	.0246

Note: 1. $J = 50$, $N = 10,000$, $\beta_0 = 1$, $\lambda_0 = .5$, Number of Repetitions = 1000.

2. "ES": Empirical Shares; "LS": Laplace Shares.

3. DGP: I, II, III and IV correspond to$\alpha_0 = -9$, -10, -12 and -13, respectively.

5.2 Extremely Many Zeroes

Next we pressure test our bound estimator by pushing the fraction of zeroes in empirical shares toward the extreme. We modify the DGP slightly to produce very high fraction of zeros. Specifically, we generate x_{jt} from the following discrete distribution

x	1	12	15
$\Pr(x_{jt} = x)$	.99	.005	.005

and

$$\xi_{jt} \sim 1(x_{jt} = 1) \times N(0, 2^2) + 1(x_{jt} \neq 1) \times N(0, .1^2).$$

All the other aspects of the DGP is the identical to the previous DGP.

The fractions of zeroes are made very high: 82%-96% by choosing the α_0 parameter. With such high fractions of zeroes, the vast majority of observations are uninformative. Thus, we need

larger sample size for any estimator to perform well. We consider $T = 100, 200, 400$. For simplicity of presentation and to reduce computational burden, we will here fix λ at its true value, and only investigate the behaviors of the estimators for β .

The results are reported in *Table* 3, and they are very encouraging for the bound approach. The ES estimator is severely biased toward 0, so is the LS estimator. The bound estimator is remarkably accurate in these extreme cases. The performance highlights the key idea of identification behind our estimator: utilizing the information in safe products with inherently thick demand to identify the model while controlling the risky products with small/zero sales properly.

Table 3: Monte Carlo Results: Very Large Fraction of Zeros

DGP	T	Ave. % of Zeros		β		
				ES	Bound	LS
I	100	82.91%	Bias	-.3222	-.0072	-.2643
			SD	.0272	.0342	.0240
	200	82.92%	Bias	-.3219	-.0072	-.2633
			SD	.0142	.0095	.0041
	400	82.94%	Bias	-.3194	-.0060	-.2633
			SD	.0267	.0068	.0031
II	100	89.59%	Bias	-.3777	-.0059	-.3311
			SD	.0129	.0133	.0063
	200	89.57%	Bias	-.3777	-.0066	-.3308
			SD	.0125	.0095	.0045
	400	89.55%	Bias	-.3759	-.0060	-.3308
			SD	.0230	.0066	.0033
III	100	96.35%	Bias	-.5613	-.0060	-.5499
			SD	.0090	.0139	.0090
	200	96.36%	Bias	-.5615	-.0064	-.5498
			SD	.0069	.0097	.0064
	400	96.35%	Bias	-.5605	-.0061	-.5495
			SD	.0102	.0071	.0046

Note: 1. $T = 100$, $J = 50$, $N = 10,000$, $\beta_0 = 1$, $\lambda_0 = .5$, Number of Repetitions = 1000.
2. We fix $\lambda = \lambda_0$ (at the true value) without estimating it.
3. DGP: I, II, III correspond to $\alpha_0 = -13, -14, -17$.

6 Empirical Application

In this section, we apply our estimator on the same DFF scanner data previewed in Section 2. In particular, we focus on the canned tuna category, as previously studied by Chevalier, Kashyap, and Rossi (2003) (CKR for short) and Nevo and Hatzitaskos (2006) (NH for short). CKR observed using the DFF data discussed in Section 2 that the share weighted price of tuna fell by 15 percent during Lent (which we replicate below in our sample from the same data source), which is a high demand period for this product. They attributed the outcome to loss-leading behavior on the part

of retailers. NH on the other hand suggest that this pricing pattern in the tuna data could instead be explained by increased price sensitivity of consumers (consistent with an increase in search) which causes a re-allocation of market shares towards less expensive products in the Lent period, and hence a fall in the observed share weighted price index. They test this hypothesis directly in the data by estimating demand parameters separately in the Lent and Non-Lent periods, and find that demand becomes more elastic in the high demand (Lent) period.

Here we revisit the groundwork laid by NH to examine the difference in price elasticity between Lent and non-Lent periods. The main difference in our analysis is that we use data on all products in the analysis, while NH restrict the sample to include only the top 30 UPCs and thus automatically drop products with small/zero sales. There are two main questions we seek to address are: a) Does the selection of UPC's with only positive shares significantly bias the estimates of price elasticity and b) Does the difference in price elasticities between the Lent and Non-Lent period persist after properly controlling for zeroes.

To make the comparison clear, we use largely the same specification of the model used in NH. In particular we consider a logit specification

$$u_{ijt} = \alpha p_{jt} + \beta x_{jt} + \xi_{jt} + \epsilon_{ijt},$$

where the control variables x_{jt} consist of UPC fixed effects and a time trend.[14] Thus the week to week variation in the product-/market-level unobserved demand shock ξ_{jt} largely captures the short-term promotional efforts, e.g., in-store advertising and shelving choices, because the UPC fixed effects control the intrinsic product quality that is likely to be stable over short time horizon. Because stores are likely to advertise or shelf the product in a more prominent way during weeks when the product is on a price sale, we expect a negative correlation between price and the unobservable. We construct instruments for price by inverting DFF's data on gross margin to calculate the chain's wholesale costs, which is the standard price instrument in the literature that has studied the DFF data.[15]

We implement our bound estimator defined by (4.17) to obtain point estimate of (α, β) in the model. And the 95% confidence interval for the parameters are obtained using a standard bootstrap procedure[16].

[14] Empirical market shares are constructed using quantity sales and the number of people who visited the store that week (the customer count) as the relevant market size.

[15] The gross margin is defined as (retail price - wholesale cost)/retail price, so we get wholesale cost using retail price×(1 - gross margin). The instrument defensible in the store disaggregated context we consider here because it has been shown that price sales in retail price primarily reflect a reduction in retailer margins rather than a reduction in marginal costs (see e.g., Chevalier, Kashyap, and Rossi (2003) and Hosken and Reiffen (2004)). Thus sales (and hence promotions) are not being driven by the manufacturer through temporary reduction in marginal costs.

[16] The procedure contains the following steps: 1) draw with replacement a bootstrap sample of *markets*, denoted as $\{t_1, ..., t_T\}$; 2) compute the bound estimator $\widehat{\theta}_T^{BD*}$ using the bootstrap sample; 3) repeat 1)-2) for B_T times and obtain B_T independent (conditional on the original sample) copies of $\widehat{\theta}_T^{BD*}$; 4) $q_T^*(\tau)$ is the τ-th quantile of the B_T copies of $\left(\widehat{\theta}_T^{BD*} - \widehat{\theta}_T^{BD}\right)$, then the 95% bootstrap confidence interval is $\left[\widehat{\theta}_T^{BD} - q_T^*(.975), \widehat{\theta}_T^{BD} - q_T^*(.025)\right]$.

The estimation results are presented in *Table* 4 and 5. [17] *Table* 4 shows that standard logit estimator that inverts empirical shares to recover mean utilities (and hence drops zeroes) has a significant selection bias towards zero. The UPC level elasticities for the logit model are small in economic magnitude, with the average elasticity in the data being -.572. Furthermore, over 90% percent of products having inelastic demand. Using our bounds approach instead to control for zeroes has a major effect on the estimated elasticities. Average demand elasticity for UPC's becomes -1.362 and less than 35% percent of observations have inelastic demand. This change in the direction of elasticities is consistent with the attenuation bias effects of dropping products with small/zero market shares.

Table 4: Demand Estimation Results

	BLP	Bound
Price Coefficient	-.390	-.910
95% CI	[-.40, -.38]	[-1.06, -.81]
Ave. Own Price Elasticity	-.572	-1.362
Fraction of Inelastic Products	90.04%	33.79%
No. of Obs.	862,683	959,331

Table 5: Demand in Lent vs. Non-Lent

	BLP		Bound	
	Lent	Non-Lent	Lent	Non-Lent
Price Coefficient	-.518	-.371	-.743	-.911
95% CI	[-55, -.48]	[-.38, -.36]	[-.84, -.45]	[-1.01, -.65]
Ave. Own Price Elasticity	-.757	-.544	-1.09	-1.302
Fraction of Inelastic Products	84.02%	92.84%	43.65%	35.00%
No. of Obs.	70,496	792,187	78,838	880,493

Our second result is that we do not find evidence to suggest demand is becoming more elastic in the high demand period, as shown in *Table* 5. Using the standard logit estimator with zeroes being dropped shows findings consistent with Nevo and Hatzitaskos (2006) - demand appears more elastic in the high demand Lent period. On the contrary, this effect disappears, and marginally changes signs, under our bounds estimator that controls for the zeroes. Thus we do not see evidence in our estimation of price elasticity being higher during the high demand period.

This finding can be rationalized if the magnitude of the selection problem with dropping zeroes is different across the two periods. Such a change in the distribution of the unobservable ξ_{jt} in the Lent period is indeed consistent with several features of the data. To see this, let us first recall that the main reduced form fact in the data documented Nevo and Hatzitaskos (2006) that suggested

[17] In principle we can estimate our model separately for each store, letting preferences change freely over stores depending on local preferences. These results are available upon request. Here we present for the results of demand pooling together all stores together as was done by Nevo and Hatzitaskos (2006). The store level regressions results are very similar to the pooled store regression and the latter is a more concise summary of demand behavior that we present here.

a change in price sensitivity in the Lent period. We replicate this reduced form finding in *Table* 6, which shows that although the price index of tuna during Lent appears to be approximately 15 percent less expensive than other weeks (as previously underscored by CKR), the average price of tuna is virtually unchanged between the Lent versus non-Lent period. Hence it is a re-allocation of demand towards less expensive products during Lent that drives the change in the aggregate price index.

Table 6: Regression of Price Index on Lent

	P	$\bar{P}$
	(Price Index)	(Average Price)
Lent	-.150	-.009
s.e.	(.0005)	(.0003)

We take this decomposition one step further than NH, and examine the price index separately for products "on sale" and "regularly priced" during these periods.[18] As can be seen in *Table* 7, it is the sales price index that is the key driver of the aggregate price index being cheaper during Lent. However the average price of an "on-sale" product is not cheaper in the Lent period. This shows that it is a re-allocation towards more steeply discounted "on-sale" product during Lent that is driving this change in the aggregate price index. But we do not see a corresponding such reallocation for "regularly priced" products.

Table 7: Regression of Sales Price Index on Lent

	P		$\bar{P}$	
	(Price Index)		(Average Price)	
	Sale	Regular	Sale	Regular
Lent	-.199	.035	.010	.001
s.e.	(.0017)	(.0003)	(.0016)	(.0003)

This suggests a tighter coordination of promotional effort and discounting in the high demand period. In effect more steeply discounted products are receiving larger promotional effort on the part of the retailer during the high demand, which is closer in spirit to the loss-leader hypothesis originally advanced for this data by CKR. Because promotional effort in the model is largely captured through the unobservable ξ_{jt}, this change in behavior of the unobservable would also account for the selection effect due to dropping zeroes changing across the two periods. This hypothesis is also consistent with our estimated model: the correlation between p_{jt} and ξ_{jt} among products that are flagged as being on sale (having at least a 5% reduction from highest price of previous 3 weeks) increases from -.16 to -.24 between the Non-Lent and Lent periods.

[18] We flag an observation in the data as being on sale if that particular UPC in that particular store in that particular week has at least a 5% reduction from highest price of previous 3 weeks.

7 Conclusion

We have shown that differentiated product demand models have enough content to construct a system of moment inequalities that can be used to consistently estimate demand parameters despite a possibly large presence of observations with zero market shares in the data. We construct a GMM-type estimator based on these moment inequalities that is consistent and asymptotically normal under assumptions that are a reasonable approximation to the DGP in many product differentiated environments. Our application to scanner data reveals that taking the market zeroes in the data into account has economically important implications for price elasticities.

A key message from our analysis is that it is critical to not ignore the zero shares when estimating discrete choice models with disaggregated market data. And a potentially fruitful area for future research is the application of our approach is individual level choice data, such as a household panel. Aggregating over households is still necessary to control for price endogeneity, such as described by Berry, Levinsohn, and Pakes (2004) and Goolsbee and Petrin (2004), and thus zero market shares when we aggregate over limited sample of households in the data is a clear problem for many contexts. Nevertheless the demographic richness in the household panel provides additional identifying power for random coefficients. The approach we describe can offer a novel solution to the joint problem of endogenous prices and flexible consumer heterogeneity with micro data, which we plan to pursue in future work.

A Further Illustrations of Zipf's Law

In *Figure 3* we illustrate this regularity using data from the two other applications that were mentioned in *Section 2*: homicide rates and international trade flows. The left hand graph shows the annual murder rate (per 10,000 people) for each county in the US from 1977-1992 (for details about the data see Dezhbakhsh, Rubin, and Shepherd (2003)). The right hand side graph shows the import trade flows (measured in millions of US dollars) among 160 countries that have a regional trade agreement in the year 2006 (for details about the data see Head, Mayer, et al. (2013)). In each of these two cases we see the characteristic pattern of Zipf's law - a sharp decay in the frequency for large outcomes and a large mass near zero (with a mode at zero in each case).

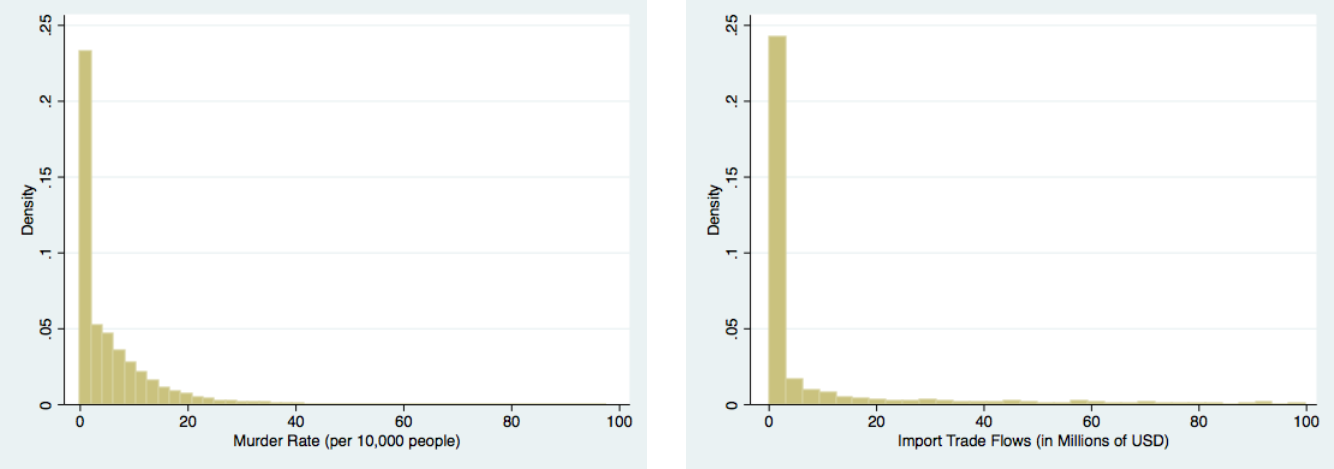


Figure 3: Zipf's Law in Crime and Trade Data

B Proofs Lemma 1 and Theorem 1

B.1 Proof of Lemma 1

Proof of Lemma 1. First consider the derivation:

$$
\begin{aligned}
&E\left[\ln\left(\frac{\tilde{s}_{jt}+\eta}{\tilde{s}_{0t}-\eta}\right)\middle|\,\pi_t,x_t\right]\\
&=E\left[\ln\left(\frac{n_t s_{jt}+1}{n_t+J_t+1}+\eta\right)\middle|\,\pi_t,x_t\right]-E\left[\ln\left(\frac{n_t s_{0t}+1}{n_t+J_t+1}-\eta\right)\middle|\,\pi_t,x_t\right]\\
&\geq \ln\left(\frac{1}{n_t+J_t+1}+\eta\right)-E\left[\ln\left(\frac{n s_{0t}+1}{n_t+J_t+1}-\eta\right)\middle|\,\pi_t,x_t\right]\\
&\geq \ln\left(\frac{1}{n_t+J_t+1}+\eta\right)-\ln\left(\frac{n_t+1}{n_t+J_t+1}-\eta\right)\Pr(n s_{0t}\geq 1|\pi_t)-\\
&\quad \ln\left(\frac{1}{n_t+J_t+1}-\eta\right)\Pr(n_t s_{0t}=0|\pi_t)\\
&\geq \ln\left(\frac{1}{n_t+J_t+1}+\eta\right)-\ln\left(\frac{n_t+1}{n_t+J_t+1}-\eta\right)-\ln\left(\frac{1}{n_t+J_t+1}-\eta\right)(1-\pi_{0t})^n\\
&\geq \ln\left(\frac{1+\eta(n_t+J_t+1)}{n_t+1+\eta(n_t+J_t+1)}\right)-\ln\left(\frac{1}{n_t+J_t+1}-\eta\right)(1-\pi_{0t})^{n_t},
\end{aligned}
\tag{B.1}
$$

where the first inequality holds because $n_t s_{jt}\geq 0$, the second inequality holds because $n_t s_{0t}\leq n_t$, the third inequality holds by $Pr(n_t s_{0t}\geq 1|\pi_t)\leq 1$ and Assumption 1. As η approaches $1/(n_t+J_t+1)$ from below, the right-hand-side diverges to positive infinity. Therefore, for any finite (π_t,x_t)-measuable quantity, there exists an $\eta_t\in(0,1/(n_t+J_t+1))$ such that $E\left[\ln\left(\frac{\tilde{s}_{jt}+\eta}{\tilde{s}_{0t}-\eta}\right)\middle|\,\pi_t,x_t\right]$ is greater than this quantity when $\eta=\eta_t$.

Next, define the ϵ-shrinkage of the J dimensional simplex be $\Delta_J^\epsilon=\{(p_1,\ldots,p_J)\in(0,1)^J : p_j\geq\epsilon, 1-\sum_{j=1}^J p_j\geq\epsilon\}$. By the definition of the Laplace share, $\Delta_{jt}(\tilde{s}_t,x_t,\lambda_0)$ lies in the interval

$$
\left[\min_{\pi\in\Delta_{J_t}^{1/(n_t+J_t+1)}}\Delta_{jt}(\pi,x_t,\lambda_0),\ \max_{\pi\in\Delta_{J_t}^{1/(n_t+J_t+1)}}\Delta_{jt}(\pi,x_t,\lambda_0)\right].
\tag{B.2}
$$

The interval is well-defined and finite by Assumption 2. Similarly, $\delta_{jt}(\lambda_0)$ is finite. Therefore, there exists η_t such that

$$
\begin{aligned}
E\left[\ln\left(\frac{\tilde{s}_{jt}+\eta_t}{\tilde{s}_{0t}-\eta_t}\right)\middle|\,\pi_t,x_t\right]&\geq-\min_{\pi\in\Delta_{J_t}^{1/(n_t+J_t+1)}}\Delta_{jt}(\pi,x_t,\lambda_0)+\delta_{jt}(\lambda_0)\\
&\geq -E[\Delta_{jt}(\tilde{s}_t,x_t,\lambda_0)|\pi_t,x_t]+\delta_{jt}(\lambda_0).
\end{aligned}
\tag{B.3}
$$

This shows that $E[\delta_{jt}^u(\lambda_0)|\pi_t,x_t]\geq\delta_{jt}(\lambda_0)$, which implies that

$$
E[\delta_{jt}^u(\lambda_0)|z_{jt}]\geq E[\delta_{jt}(\lambda_0)|z_{jt}].
\tag{B.4}
$$

This proves the upper bound part of (4.8). The lower bound part is analogous and thus omitted. □

B.2 Proof of Theorem 1

Next, we prove Theorem 1. To do so, we present three lemmas first. Proofs of these lemmas are presented after that of Theorem 1. Consider the subset of the instrumental function collection:

$$\mathcal{G}_0 = \{g_{a,r,\zeta} \in \mathcal{G} : \Pr((\tilde{z}_c', z_d)' \in C_{a,r,\zeta}) = \Pr((\tilde{z}_c', z_d)' \in C_{a,r,\zeta} \cap \mathcal{Z}_0)\}. \tag{B.5}$$

Let

$$Q_0^*(\theta) = \sum_{g \in \mathcal{G}_0} (\rho_F^*(\theta, g))^2 \mu(g). \tag{B.6}$$

Lemma 3. *Suppose that Assumptions 1-6 hold. Then for any $c > 0$,*

$$\inf_{\theta \in \Theta : \|\theta - \theta_0\| > c} Q_0^*(\theta) > 0. \tag{B.7}$$

Let $\widehat{Q}_{0,T}(\theta) = \sum_{g \in \mathcal{G}_0} \left\{ \left([\bar{\rho}_T^u(\theta, g)]_-^2 + [\bar{\rho}_T^\ell(\theta, g)]_-^2 \right) \mu(g) \right\}$.

Lemma 4. *Suppose that Assumptions 3-7 hold. Then, $\sup_{\theta \in \Theta} |\widehat{Q}_{0,T}(\theta) - Q_0^*(\theta)| \to_p 0$.*

Lemma 5. *Suppose that Assumption 8 hold. Then,*

$$\widehat{Q}_T(\theta_0) = o_p(1). \tag{B.8}$$

Proof of Theorem 1. Consider an arbitrary $c > 0$. Let $q = \inf_{\theta \in \Theta : \|\theta - \theta_0\| > c} Q_0^*(\theta)$. Then $q > 0$. The theorem is implied by the following derivation:

$$\begin{aligned}
\Pr\left(\|\widehat{\theta}_T^{BD} - \theta_0\| > c \right) &\leq \Pr\left(Q_{0,T}^*(\widehat{\theta}_T^{BD}) \geq q \right) \\
&= \Pr\left(Q_0^*(\widehat{\theta}_T^{BD}) - \widehat{Q}_{0,T}(\widehat{\theta}_T^{BD}) + \widehat{Q}_{0,T}(\widehat{\theta}_T^{BD}) \geq q \right) \\
&\leq \Pr\left(\sup_{\theta \in \Theta} |Q_0^*(\theta) - \widehat{Q}_{0,T}(\theta)| + \widehat{Q}_{0,T}(\widehat{\theta}_T^{BD}) \geq q \right) \\
&\leq \Pr\left(\sup_{\theta \in \Theta} |Q_0^*(\theta) - \widehat{Q}_{0,T}(\theta)| + \widehat{Q}_T(\widehat{\theta}_T^{BD}) \geq q \right) \\
&\leq \Pr\left(\sup_{\theta \in \Theta} |Q_0^*(\theta) - \widehat{Q}_{0,T}(\theta)| + \widehat{Q}_T(\theta_0) \geq q \right) \\
&\leq \Pr\left(\sup_{\theta \in \Theta} |Q_0^*(\theta) - \widehat{Q}_{0,T}(\theta)| \geq q/2 \right) + \Pr\left(\widehat{Q}_T(\theta_0)) \geq q/2 \right) \\
&\to 0,
\end{aligned} \tag{B.9}$$

where the first inequality holds by Lemma 3, the third inequality holds because $\widehat{Q}_T(\theta)$ differs from $\widehat{Q}_{0,T}(\theta)$ only in that the forms takes the integral over a larger range, and because the common integrant of both are non-negative, the fourth inequality holds because $\widehat{Q}_T(\widehat{\theta}_T^{BD}) \leq \widehat{Q}_T(\theta_0)$ by the definition of $\widehat{\theta}_T^{BD}$ and the convergence holds by Lemmas 4 and 5. □

Proof of Lemma 3. Consider a sequence $\{\theta_m\}_{m=1}^{\infty}$ such that

$$\lim_{m\to\infty} Q_0^*(\theta_m) = \inf_{\theta\in\Theta:\|\theta-\theta_0\|>c} Q_0^*(\theta). \tag{B.10}$$

Because Θ is compact, it is without loss of generality to assume that $\lim_{m\to\infty}\theta_m = \theta^*$ for some $\theta^* \in \Theta$.

Because $\|\theta_m - \theta_0\| > c$ for all m, we have

$$\|\theta^* - \theta_0\| \geq c. \tag{B.11}$$

That is, $\theta^* \neq \theta_0$. By Lemma 2, there exists a $g^* \in \mathcal{G}_0$ such that $\rho_F^*(\theta^*, g^*) \neq 0$. Next, we show that

$$\lim_{m\to\infty} \rho_F^*(\theta_m, g^*) = \rho_F^*(\theta^*, g^*). \tag{B.12}$$

Once this is established, the result of the lemma is implied by

$$\inf_{\theta\in\Theta:\|\theta-\theta_0\|>c} Q_0^*(\theta) = \lim_{m\to\infty} Q_0^*(\theta_m) \geq \lim_{m\to\infty} \mu(g^*)\rho_F^*(\theta_m, g^*)^2 = \mu(g^*)\rho_F^*(\theta^*, g^*)^2 > 0. \tag{B.13}$$

We show (B.12) using the dominated convergence theorem (DCT). By Assumption 6(a), we have with probability one,

$$(\sigma_j^{-1}(\pi_t, x_t, \lambda_m) - \beta_m' x_{jt})g^*(z_{jt}) \to (\sigma_j^{-1}(\pi_t, x_t, \lambda^*) - \beta^{*\prime} x_{jt})g^*(z_{jt}). \tag{B.14}$$

Also observe that $|(\sigma_j^{-1}(\pi_t, x_t, \lambda_m) - \beta_m' x_{jt})g^*(z_{jt})| \leq \sup_{\theta\in\Theta}|(\sigma_j^{-1}(\pi_t, x_t, \lambda) - \beta' x_{jt})|1\{z_{jt} \in \mathcal{Z}_0\}$. The right-hand-side is integrable by Assumption 5(a). Therefore, the DCT applies and yields (B.12) □

Proof of Lemma 4. First, we show that,

$$\sup_{g\in\mathcal{G}_0}\sup_{\lambda}\left|(T\bar{J})^{-1}\sum_{t=1}^{T}\left(\sum_{j=1}^{J_t}(\delta_{jt}^u(\lambda) - \sigma_j^{-1}(\pi_t, x_t; \lambda))g(z_{jt})\right)\right| \to_p 0. \tag{B.15}$$

To show this, observe that with probability one, the left-hand-side is less than or equal to

$$\sup_{t,j:z_{jt}\in\mathcal{Z}_0} \sup_{\lambda} |(\sigma_j^{-1}(\tilde{s}_t, x_t; \lambda) - \sigma_j^{-1}(\pi_t, x_t; \lambda)| + \sup_{t,j:z_{jt}\in\mathcal{Z}_0} \left| \ln\left(\frac{\tilde{s}_{jt}+\eta}{\tilde{s}_{0t}-\eta}\right) - \ln\left(\frac{\tilde{s}_{jt}}{\tilde{s}_{0t}}\right)\right| \tag{B.16}$$

The in-probability convergence of this to zero is implied by Assumptions 3 and the fact that $\eta \leq 1/(n_t + J_t + 1) \to 0$, as long as we can show that $\max_{j=0,\ldots,J_t;t=1,\ldots,T} |\tilde{s}_{jt} - \pi_{jt}| \to_p 0$. The latter convergence is true by the following derivation:

$$\begin{aligned}
&\max_{j;t} |\tilde{s}_{jt} - \pi_{jt}| \\
&\leq \max_{j;t} |s_{jt} - \pi_{jt} + (n_t + J_t + 1)^{-1} - (J_t + 1)s_{jt}/(n_t + J_t + 1)| \\
&\leq \max_{j;t} |s_{jt} - \pi_{jt}| + |n_t^{-1}| + |(J_t + 1)/n_t| \\
&\to_p 0,
\end{aligned} \tag{B.17}$$

where the convergence holds by Assumption 7. Therefore, (B.15) is proved.

Equation (B.15) and Assumption 6(b) together shows that for any $g \in \mathcal{G}_0$,

$$\sup_{g\in\mathcal{G}_0} \sup_{\theta\in\Theta} |\bar{\rho}_T^u(\theta, g) - \rho_F^*(\theta, g)| \to_p 0 \tag{B.18}$$

Similarly, for any $g \in \mathcal{G}_0$,

$$\sup_{g\in\mathcal{G}_0} \sup_{\theta\in\Theta} \left|\bar{\rho}_T^l(\theta, g) + \rho_F^*(\theta, g)\right| \to_p 0 \tag{B.19}$$

Then we can show that

$$\begin{aligned}
\sup_{\theta\in\Theta} |\widehat{Q}_{0,T}(\theta) - Q_0^*(\theta)| &\leq \sum_{g\in\mathcal{G}_0} \sup_{\theta\in\Theta} |[\bar{\rho}_T^u(\theta, g)]_-^2 + [\bar{\rho}_T^l(\theta, g)]_-^2 - \rho_F^*(\theta, g)^2|\mu(g) \\
&\leq \sum_{g\in\mathcal{G}_0} \sup_{\theta\in\Theta} |[\bar{\rho}_T^u(\theta, g)]_-^2 - [\rho_F^*(\theta, g)]_-^2|\mu(g) + \\
&\quad \sum_{g\in\mathcal{G}_0} \sup_{\theta\in\Theta} |[\bar{\rho}_T^\ell(\theta, g)]_-^2 - [-\rho_F^*(\theta, g)]_-^2|\mu(g) \\
&= o_p(1).
\end{aligned} \tag{B.20}$$

This concludes the proof of the lemma. □

Proof of Lemma 5. The lemma is immediately implied by Assumption 8 and equation (4.14). □

C Assumptions and Proof of Asymptotic Normality

C.1 Additional Assumptions for Asymptotic Normality

We derive the asymptotic normality of our bound point estimator using similar techniques as Khan and Tamer (2009).

Additional assumptions are needed. We divide the assumptions in to two groups. The first group, Assumptions C.1-C.3 are needed for deriving the convergence rate. On top of those, Assumptions C.4-C.7 are needed for the asymptotic normality.

Assumption C.1. For an $\epsilon > 0$ and an open ball, $B_\epsilon(\theta_0)$, of radius ϵ around θ_0,

$$\sup_{g\in\mathcal{G},\theta\in B_\epsilon(\theta_0)} \left\{ \left|\bar{\rho}_T^u(\theta,g) - \rho_{F,T}^u(\theta,g)\right| + \left|\bar{\rho}_T^\ell(\theta,g) - \rho_{F,T}^\ell(\theta,g)\right| \right\} = O_p((T\bar{J})^{-1/2}).$$

Assumption C.2. (a) $\sigma_j^{-1}(\pi_t, x_t; \lambda)$ is continuously differentiable in λ in $B_\epsilon(\lambda_0)$—an open ball around λ_0, and $E\|x_{jt}\| < \infty$.
(b) $E \sup_{\lambda\in B_\epsilon(\lambda_0)} \|\partial\sigma_j^{-1}(\pi_t, x_t; \lambda)/\partial\lambda\| < \infty$ and $E\|x_{jt}\| < \infty$.
(c) $\sum_{g\in\mathcal{G}_0} \frac{\partial\rho_F^*(\theta_0,g)}{\partial\theta}\frac{\partial\rho_F^*(\theta_0,g)}{\partial\theta'}\mu(g)$ is positive definite.

Assumption C.3. For an $\epsilon > 0$, we have

$$\sup_{g\in\mathcal{G}_0,\theta\in B_\epsilon(\theta_0)} \left\{ \left|\rho_{F,T}^u(\theta,g) - \rho_F^*(\theta,g)\right| + \left|\rho_{F,T}^\ell(\theta,g) + \rho_F^*(\theta,g)\right| \right\} = O((T\bar{J})^{-1/2}).$$

Assumption C.4. (a) $(T\bar{J})^{1/2}\rho_{F,T}^u(\theta_0,g) \to \infty$ and $(T\bar{J})^{1/2}\rho_{F,T}^\ell(\theta_0,g) \to \infty$ for all $g \in \mathcal{G}\backslash\mathcal{G}_0$.
(b) $\{\rho_{F,T}^u(\theta,g) : g \in \mathcal{G} : T = 1,2,3,\dots\}$ and $\{\rho_{F,T}^\ell(\theta,g) : g \in \mathcal{G}, T = 1,2,3,\dots\}$ are equi-continuous in θ at θ_0,
(c) $\rho_{F,T}^u(\theta,g)$ and $\rho_{F,T}^\ell(\theta,g)$ are differentiable in θ for every $g \in \mathcal{G}$ and T.

Assumption C.5. (a) For any $\epsilon > 0$, we have $\sup_{\theta\in B_\epsilon(\theta_0),g\in\mathcal{G}} \left\| \frac{\partial\bar{\rho}_T^u(\theta,g)}{\partial\theta} - \frac{\partial\rho_{F,T}^u(\theta,g)}{\partial\theta} \right\| \to_p 0$, and also $\sup_{\theta\in B_\epsilon(\theta_0),g\in\mathcal{G}} \left\| \frac{\partial\bar{\rho}_T^\ell(\theta,g)}{\partial\theta} - \frac{\partial\rho_{F,T}^\ell(\theta,g)}{\partial\theta} \right\| \to_p 0$,
(b) $\{\frac{\partial\rho_{F,T}^u(\theta,g)}{\partial\theta} : g \in \mathcal{G}, T = 1,2,3,\dots,\}$ and $\{\frac{\partial\rho_{F,T}^\ell(\theta,g)}{\partial\theta} : g \in \mathcal{G}, T = 1,2,3,\dots\}$ are equi-continuous in θ at θ_0.

Define the infeasible sample moment function:

$$\bar{\rho}_T^*(\theta,g) = (T\bar{J})^{-1}\sum_{t=1}^{T}\left(\sum_{j=1}^{J_t}(\sigma_j^{-1}(\pi_t,x_t;\lambda) - \beta' x_{jt})g(z_{jt})\right). \tag{C.1}$$

Assumption C.6. (a) $\sup_{\theta\in B_\epsilon(\theta),g\in\mathcal{G}_0} \|\bar{\rho}_T^u(\theta,g) - \bar{\rho}_T^*(\theta,g)\| = o_p((T\bar{J})^{-1/2})$, and $\sup_{\theta\in B_\epsilon(\theta),g\in\mathcal{G}_0} \|\bar{\rho}_T^\ell(\theta,g) + \bar{\rho}_T^*(\theta,g)\| = o_p((T\bar{J})^{-1/2})$.
(b) $\sup_{\theta\in B_\epsilon(\theta),g\in\mathcal{G}_0} \left\| \frac{\partial\bar{\rho}_T^u(\theta,g)}{\partial\theta} - \frac{\partial\bar{\rho}_T^*(\theta,g)}{\partial\theta} \right\| = o_p(1)$, and $\sup_{\theta\in B_\epsilon(\theta),g\in\mathcal{G}_0} \left\| \frac{\partial\bar{\rho}_T^\ell(\theta,g)}{\partial\theta} + \frac{\partial\bar{\rho}_T^*(\theta,g)}{\partial\theta} \right\| = o_p(1)$.

Assumption C.7. (a) $(T\bar{J})^{1/2}\bar{\rho}_T^*(\theta_0,\cdot) \to_d \nu_\Sigma(\cdot)$, where $\nu_\Sigma(g) : g \in \mathcal{G}_0$ is a tight Gaussian process with variance covariance kernel $\Sigma(g,g^*) : (g,g^*) \in \mathcal{G}_0^2$.
(b) $\sup_{g\in\mathcal{G}_0,\theta\in B_\epsilon(\theta_0)} \|\frac{\partial\bar{\rho}_T^*(\theta,g)}{\partial\theta} - \frac{\partial\rho_F^*(\theta,g)}{\partial\theta}\| = o_p(1)$.

C.2 Proof of Asymptotic Normality

We now prove Theorem 2. The proof uses the following lemma. The proof of the lemma is given after that of Theorem 2.

Lemma C.1. Suppose that Assumptions 1-8 and C.1-C.3 are satisfied. Then

$$\|\widehat{\theta}_T^{BD} - \theta_0\| = O_p((T\bar{J})^{-1/2}).$$

Proof of Theorem 2. The criterion function is differentiable. Thus, we have the first-order-condition:

$$\begin{aligned} 0 &= \frac{\partial \widehat{Q}_T(\widehat{\theta}_T^{BD})}{\partial\theta} \\ &= 2\sum_{g\in\mathcal{G}}[\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g)]_- \frac{\partial\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g)}{\partial\theta}\mu(g) + 2\sum_{g\in\mathcal{G}}[\bar{\rho}_T^\ell(\widehat{\theta}_T^{BD},g)]_- \frac{\partial\bar{\rho}_T^\ell(\widehat{\theta}_T^{BD},g)}{\partial\theta}\mu(g). \end{aligned} \tag{C.2}$$

Consider the derivation: for all $g \in \mathcal{G}\backslash\mathcal{G}_0$,

$$\begin{aligned} \Pr(\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g) < 0) &= \Pr((T\bar{J})^{1/2}(\bar{\rho}_T^u((\widehat{\theta}_T^{BD},g) - \rho_{F,T}^u(\theta_0,g)) < -T^{1/2}\rho_{F,T}^u(\theta_0,g)) \\ &\to 0, \end{aligned} \tag{C.3}$$

where the convergence holds by Assumptions C.1 and C.4(a). Similarly, we have, for every $g \in \mathcal{G}\backslash\mathcal{G}_0$

$$\Pr(\bar{\rho}_T^\ell(\widehat{\theta}_T^{BD},g) < 0) \to 0. \tag{C.4}$$

Thus, for every $g \in \mathcal{G}\backslash\mathcal{G}_0$

$$\begin{aligned} &\Pr\left([\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g)]_- \frac{\partial\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g)}{\partial\theta} = \mathbf{0}\right) \to 0 \text{ and} \\ &\Pr\left([\bar{\rho}_T^\ell(\widehat{\theta}_T^{BD},g)]_- \frac{\partial\bar{\rho}_T^\ell(\widehat{\theta}_T^{BD},g)}{\partial\theta} = \mathbf{0}\right) \to 0. \end{aligned} \tag{C.5}$$

Because $\mathcal{G}\backslash\mathcal{G}_0$ is a countable set, the above convergence implies that, for any subsequence of $\{T\}_{T=1}^\infty$, there exists a further subsequence $\{a_T\}_{T=1}^\infty$ such that

$$(a_T\bar{J}_{a_T})^{1/2}[\bar{\rho}_{a_T}^u(\widehat{\theta}_{a_T}^{BD},g)]_- \frac{\partial\bar{\rho}_{a_T}^u(\widehat{\theta}_{a_T}^{BD},g)}{\partial\theta} \to \mathbf{0} \text{ and } (a_T\bar{J}_{a_T})^{1/2}[\bar{\rho}_{a_T}^\ell(\widehat{\theta}_{a_T}^{BD},g)]_- \frac{\partial\bar{\rho}_{a_T}^\ell(\widehat{\theta}_{a_T}^{BD},g)}{\partial\theta} \to \mathbf{0}, \tag{C.6}$$

almost surely for every $g \in \mathcal{G}\backslash\mathcal{G}_0$, where $\bar{J}_{a_T} = \sum_{t=1}^{a_T} J_t$. By the bounded convergence theorem

(applied sample path by sample path), we have

$$\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0}\mu(g)\left((a_T\bar{J}_{a_T})^{1/2}[\bar{\rho}^u_{a_T}(\widehat{\theta}^{BD}_{a_T},g)]_-\frac{\partial\bar{\rho}^u_{a_T}(\widehat{\theta}^{BD}_{a_T},g)}{\partial\theta}+(a_T\bar{J}_{a_T})^{1/2}[\bar{\rho}^\ell_{a_T}(\widehat{\theta}^{BD}_{a_T},g)]_-\frac{\partial\bar{\rho}^\ell_{a_T}(\widehat{\theta}^{BD}_{a_T},g)}{\partial\theta}\right)\to 0 \tag{C.7}$$

almost surely. Thus,

$$\sum_{g\in\mathcal{G}\setminus\mathcal{G}_0}\mu(g)\left((T\bar{J})^{1/2}[\bar{\rho}^u_T(\widehat{\theta}^{BD}_T,g)]_-\frac{\partial\bar{\rho}^u_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta}+(T\bar{J})^{1/2}[\bar{\rho}^\ell_T(\widehat{\theta}^{BD}_T,g)]_-\frac{\partial\bar{\rho}^\ell_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta}\right)\to_p 0. \tag{C.8}$$

This implies that

$$\begin{aligned}&\frac{\partial\widehat{Q}_T(\widehat{\theta}^{BD}_T)}{\partial\theta}\\&=o_p((T\bar{J})^{-1/2})+2\sum_{g\in\mathcal{G}_0}\mu(g)\left([\bar{\rho}^u_T(\widehat{\theta}^{BD}_T,g)]_-\frac{\partial\bar{\rho}^u_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta}+[\bar{\rho}^\ell_T(\widehat{\theta}^{BD}_T,g)]_-\frac{\partial\bar{\rho}^\ell_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta}\right).\end{aligned} \tag{C.9}$$

Next consider the following derivation:

$$\begin{aligned}&\sum_{g\in\mathcal{G}_0}[\bar{\rho}^u_T(\widehat{\theta}^{BD}_T,g)]_-\frac{\partial\bar{\rho}^u_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta}\mu(g)-\sum_{g\in\mathcal{G}_0}[\bar{\rho}^*_T(\widehat{\theta}^{BD}_T,g)]_-\frac{\partial\bar{\rho}^*_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta}\mu(g)\\&=\sum_{g\in\mathcal{G}_0}([\bar{\rho}^u_T(\widehat{\theta}^{BD}_T,g)]_--[\bar{\rho}^*_T(\widehat{\theta}^{BD}_T,g)]_-)(\frac{\partial\bar{\rho}^u_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta}-\frac{\partial\bar{\rho}^*_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta})\mu(g)\\&\quad+\sum_{g\in\mathcal{G}_0}([\bar{\rho}^u_T(\widehat{\theta}^{BD}_T,g)]_--[\bar{\rho}^*_T(\widehat{\theta}^{BD}_T,g)]_-)\frac{\partial\bar{\rho}^*_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta}\mu(g)\\&\quad+\sum_{g\in\mathcal{G}_0}[\bar{\rho}^*_T(\theta_0,g)+\frac{\partial\bar{\rho}^*_T(\tilde{\theta}_T,g)}{\partial\theta}(\widehat{\theta}_T-\theta_0)]_-(\frac{\partial\bar{\rho}^u_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta}-\frac{\partial\bar{\rho}^*_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta})\mu(g).\end{aligned} \tag{C.10}$$

The first summand on the right-hand-side is $o_p((T\bar{J})^{-1/2})$ by Assumption C.6. The second summand is $o_p((T\bar{J})^{-1/2})$ by Assumption C.6(a), and

$$\sup_{g\in\mathcal{G}_0}\|\frac{\partial\bar{\rho}^*_T(\widehat{\theta}^{BD}_T,g)}{\partial\theta}-\frac{\partial\rho^*_F(\theta_0,g)}{\partial\theta}\|=o_p(1), \tag{C.11}$$

which holds by Assumptions C.5(a)-(b), C.6(b), and C.7(b). The third summand is $o_p((T\bar{J})^{-1/2})$ by Assumption C.6(a), C.7(a), Lemma C.1 and

$$\sup_{g\in\mathcal{G}_0}\|\frac{\partial\bar{\rho}^*_T(\tilde{\theta}_T,g)}{\partial\theta}-\frac{\partial\rho^*_F(\theta_0,g)}{\partial\theta}\|=o_p(1), \tag{C.12}$$

which holds similarly to (C.11). Therefore,

$$\sum_{g\in\mathcal{G}_0}[\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g)]_-\frac{\partial\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g)}{\partial\theta}\mu(g)=o_p((T\bar{J})^{-1/2})+\sum_{g\in\mathcal{G}_0}[\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)]_-\frac{\partial\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)}{\partial\theta}\mu(g). \quad \text{(C.13)}$$

Similarly, we can show that

$$\sum_{g\in\mathcal{G}_0}[\bar{\rho}_T^\ell(\widehat{\theta}_T^{BD},g)]_-\frac{\partial\bar{\rho}_T^\ell(\widehat{\theta}_T^{BD},g)}{\partial\theta}\mu(g)=o_p((T\bar{J})^{-1/2})-\sum_{g\in\mathcal{G}_0}[-\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)]_-\frac{\partial\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)}{\partial\theta}\mu(g). \quad \text{(C.14)}$$

Equations (C.2), (C.9), (C.13), and (C.14) together show that

$$0=\frac{\partial\widehat{Q}_T(\widehat{\theta}_T^{BD})}{\partial\theta}=o_p((T\bar{J})^{-1/2})+2\sum_{g\in\mathcal{G}_0}\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)\frac{\partial\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)}{\partial\theta}\mu(g). \quad \text{(C.15)}$$

Apply a mean-value expansion of $\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)$ around θ_0, and we get

$$o_p((T\bar{J})^{-1/2})=\sum_{g\in\mathcal{G}_0}\bar{\rho}_T^*(\theta_0,g)\frac{\partial\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)}{\partial\theta}\mu(g)+\left[\sum_{g\in\mathcal{G}_0}\frac{\partial\bar{\rho}_T^*(\tilde{\theta}_T,g)}{\partial\theta}\frac{\partial\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)}{\partial\theta'}\mu(g)\right](\widehat{\theta}_T^{BD}-\theta_0). \quad \text{(C.16)}$$

Therefore,

$$\begin{aligned}
&(T\bar{J})^{1/2}(\widehat{\theta}_T^{BD}-\theta_0)\\
&=o_p(1)+\left[\sum_{g\in\mathcal{G}_0}\frac{\partial\bar{\rho}_T^*(\tilde{\theta}_T,g)}{\partial\theta}\frac{\partial\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)}{\partial\theta'}\mu(g)\right]^{-1}\sum_{g\in\mathcal{G}_0}\bar{\rho}_T^*(\theta_0,g)\frac{\partial\bar{\rho}_T^*(\widehat{\theta}_T^{BD},g)}{\partial\theta}\mu(g)\\
&\to_d\left[\sum_{g\in\mathcal{G}_0}\frac{\partial\rho_F^*(\theta_0,g)}{\partial\theta}\frac{\partial\rho_F^*(\theta_0,g)}{\partial\theta'}\mu(g)\right]^{-1}\sum_{g\in\mathcal{G}_0}\nu_\Sigma(g)\frac{\partial\rho_F^*(\theta_0,g)}{\partial\theta}\mu(g)\\
&=_d N(0,\Gamma V\Gamma), \qquad \text{(C.17)}
\end{aligned}$$

where $\Gamma=\left[\sum_{g\in\mathcal{G}_0}\frac{\partial\rho_F^*(\theta_0,g)}{\partial\theta}\frac{\partial\rho_F^*(\theta_0,g)}{\partial\theta'}\mu(g)\right]^{-1}$, and

$$V=\sum_{g,g^*\in\mathcal{G}_0}\Sigma(g,g^*)\frac{\partial\rho_F^*(\theta_0,g)}{\partial\theta}\frac{\partial\rho_F^*(\theta_0,g^*)}{\partial\theta'}\mu(g)\mu(g^*). \quad \text{(C.18)}$$

This concludes the proof of the theorem. □

Proof of Lemma C.1. Below, we show the following results:

$$\widehat{Q}_{0,T}(\widehat{\theta}_T^{BD}) = O_p((T\bar{J})^{-1}), \tag{C.19}$$

$$Q_0(\widehat{\theta}_T^{BD}) \geq c\|\widehat{\theta}_T - \theta_0\|^2, \tag{C.20}$$

$$\widehat{Q}_{0,T}(\widehat{\theta}_T^{BD}) - Q_0(\widehat{\theta}_T^{BD}) = O_p((T\bar{J})^{-1}) + O_p((T\bar{J})^{-1/2})\|\widehat{\theta}_T - \theta_0\|, \tag{C.21}$$

for some $c > 0$.

Equations (C.19)-(C.21) together imply that

$$c\|\widehat{\theta}_T - \theta_0\|^2 + O_p(T^{-1/2})\|\widehat{\theta}_T - \theta_0\| = O_p((T\bar{J})^{-1}). \tag{C.22}$$

This implies that

$$(c^{1/2}\|\widehat{\theta}_T - \theta_0\| + O_p(T^{-1/2}))^2 = O_p((T\bar{J})^{-1}), \tag{C.23}$$

which then implies the conclusion of the theorem.

Now we show (C.19). Observe that $\widehat{Q}_{0,T}(\widehat{\theta}_T^{BD}) \leq \widehat{Q}_T(\widehat{\theta}_T^{BD}) \leq \widehat{Q}_T(\theta_0)$. Thus, it suffices to show that

$$\widehat{Q}_T(\theta_0) = O_p((T\bar{J})^{-1}). \tag{C.24}$$

Consider the derivation:

$$\begin{aligned}
\widehat{Q}_T(\theta_0) &= \sum_{g\in\mathcal{G}} [\bar{\rho}_T^u(\theta_0, g) - \rho_{F,T}^u(\theta_0, g) + \rho_{F,T}^u(\theta_0, g)]_-^2 \mu(g) \\
&\quad + \sum_{g\in\mathcal{G}} [\bar{\rho}_T^\ell(\theta_0, g) - \rho_{F,T}^\ell(\theta_0, g) + \rho_{F,T}^\ell(\theta_0, g)]_-^2 \mu(g) \\
&\leq \sum_{g\in\mathcal{G}} [\bar{\rho}_T^u(\theta_0, g) - \rho_{F,T}^u(\theta_0, g)]_-^2 \mu(g) \\
&\quad + \sum_{g\in\mathcal{G}} [\bar{\rho}_T^\ell(\theta_0, g) - \rho_{F,T}^\ell(\theta_0, g)]_-^2 \mu(g) \\
&= O_p((T\bar{J})^{-1}),
\end{aligned} \tag{C.25}$$

where the inequality holds because $\rho_{F,T}^\ell(\theta_0, g) \geq 0$, and $\rho_{F,T}^u(\theta_0, g) \geq 0$ for all $g \in \mathcal{G}$ by definition, and the second equality holds by Assumption C.1. This shows (C.19).

Next, we show (C.20). Consider the derivation:

$$\begin{aligned} Q_0(\widehat{\theta}_T^{BD}) &= \sum_{g\in\mathcal{G}_0} (\widehat{\theta}_T^{BD}-\theta_0)' \frac{\partial \rho_F^*(\tilde{\theta}_T,g)}{\partial\theta}\frac{\partial \rho_F^*(\tilde{\theta}_T,g)}{\partial\theta'}(\widehat{\theta}_T-\theta_0)\mu(g) \\ &= (\widehat{\theta}_T^{BD}-\theta_0)' \left[\sum_{g\in\mathcal{G}_0} \frac{\partial \rho_F^*(\theta_0,g)}{\partial\theta}\frac{\partial \rho_F^*(\theta_0,g)}{\partial\theta'}\mu(g) + o_p(1)\right](\widehat{\theta}_T^{BD}-\theta_0) \\ &\geq c\|\widehat{\theta}_T^{BD}-\theta_0\|^2, \end{aligned} \tag{C.26}$$

where the first equality holds by a mean-value expansion with $\tilde{\theta}_T$ being a point on the line segment joining $\widehat{\theta}_T^{BD}$ and θ_0, the second equality holds by Assumption C.2(a), and the inequality holds by Assumption C.2(b) where c is the smallest eigenvalue of $\sum_{g\in\mathcal{G}_0} \frac{\partial \rho_F^*(\theta_0,g)}{\partial\theta}\frac{\partial \rho_F^*(\theta_0,g)}{\partial\theta'}\mu(g)/2$. This shows (C.20).

Finally, we show (C.21). Observe that:

$$\begin{aligned} \widehat{Q}_{0,T}(\widehat{\theta}_T^{BD}) - Q_0(\widehat{\theta}_T^{BD}) &= \sum_{g\in\mathcal{G}_0} [\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g)]_-^2 - [\rho_F^*(\widehat{\theta}_T^{BD},g)]_-^2\mu(g) \\ &\quad + \sum_{g\in\mathcal{G}_0} [\bar{\rho}_T^\ell(\widehat{\theta}_T^{BD},g)]_-^2 - [-\rho_F^*(\widehat{\theta}_T^{BD},g)]_-^2\mu(g). \end{aligned} \tag{C.27}$$

Consider the derivation regarding the first summand in the right-hand-side of the equation above:

$$\begin{aligned} &\sum_{\mathcal{G}_0} [\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g)]_-^2 - [\rho_F^*(\widehat{\theta}_T^{BD},g)]_-^2 d\mu(g) \\ &= \sum_{g\in\mathcal{G}_0} ([\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g)]_- - [\rho_F^*(\widehat{\theta}_T^{BD},g)]_-)^2 + 2[\rho_F^*(\widehat{\theta}_T^{BD},g)]_-([\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g)]_- - [\rho_F^*(\widehat{\theta}_T^{BD},g)]_-)\mu(g) \\ &\leq \sum_{g\in\mathcal{G}_0} (\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g) - \rho_F^*(\widehat{\theta}_T^{BD},g))^2\mu(g) \\ &\quad + 2\left(\sum_{g\in\mathcal{G}_0} \rho_F^*(\widehat{\theta}_T^{BD},g)^2\mu(g)\right)^{1/2}\left(\sum_{g\in\mathcal{G}_0} (\bar{\rho}_T^u(\widehat{\theta}_T^{BD},g) - \rho_F^*(\widehat{\theta}_T^{BD},g))^2\mu(g)\right)^{1/2} \\ &= O_p((T\bar{J})^{-1}) + 2O_p((T\bar{J})^{-1/2})\left(\sum_{g\in\mathcal{G}_0} \rho_F^*(\widehat{\theta}_T^{BD},g)^2\mu(g)\right)^{1/2} \\ &= O_p((T\bar{J})^{-1}) + O_p((T\bar{J})^{-1/2})\left((\widehat{\theta}_T^{BD}-\theta_0)\sum_{g\in\mathcal{G}_0} \frac{\partial \rho_F^*(\tilde{\theta}_T,g)}{\partial\theta}\frac{\partial \rho_F^*(\tilde{\theta}_T,g)}{\partial\theta'}\mu(g)(\widehat{\theta}_T^{BD}-\theta_0)\right)^{1/2} \\ &= O_p((T\bar{J})^{-1}) + O_p((T\bar{J})^{-1/2})\|\widehat{\theta}_T^{BD}-\theta_0\|, \end{aligned} \tag{C.28}$$

where the first equality holds by rearranging terms, the inequality holds by the fact that $|[a]_- - [b]_-| \leq |a-b|$ for any $a, b \in R$, and by the Cauchy-Schwarz Inequality, the second equality holds by Assumptions C.1 and C.3 and Theorem 1, the third equality holds with $\tilde{\theta}_T$ being a point on

the line segment joining $\widehat{\theta}_T^{BD}$ and θ_0 by a mean-value expansion, and the last equality holds by Assumption C.2 and Theorem 1. Similarly, we can show

$$\sum_{g\in\mathcal{G}_0}[\bar{\rho}_T^{\ell}(\widehat{\theta}_T^{BD},g)]_-^2-[-\rho_F^*(\widehat{\theta}_T^{BD},g)]_-^2\mu(g)=O_p((T\bar{J})^{-1})+O_p((T\bar{J})^{-1/2})\|\widehat{\theta}_T^{BD}-\theta_0\|. \tag{C.29}$$

Therefore, (C.21) is shown, and this concludes the proof of the theorem. □

References

Anderson, C. (2006): *The Long Tail: Why the Future of Business Is Selling Less of More.* Hyperion.

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment Inequality Models," *Econometrica*, 81.

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

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

Berry, S., J. Levinsohn, and A. Pakes (2004): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Vehicle Market," *Journal of Political Economy*, 112, 68–104.

Berry, S., O. Linton, and A. Pakes (2004): "Limit theorems for estimating the parameters of differentiated product demand systems," *Review of Economic Studies*, 71(3), 613–654.

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

Chevalier, J. A., A. K. Kashyap, and P. E. Rossi (2003): "Why Don't Prices Rise During Periods of Peak Demand? Evidence from Scanner Data," *American Economic Review*, 93(1), 15–37.

Dezhbakhsh, H., P. H. Rubin, and J. M. Shepherd (2003): "Does capital punishment have a deterrent effect? New evidence from postmoratorium panel data," *American Law and Economics Review*, 5(2), 344–376.

Dubé, J.-P., J. T. Fox, and C.-L. Su (2012): "Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation," *Econometrica*, 80(5), 2231–2267.

Freyberger, J. (2015): "Asymptotic theory for differentiated products demand models with many markets," *Journal of Econometrics*, 185(1), 162–181.

Gabaix, X. (1999a): "Zipf's Law and the Growth of Cities," *The American Economic Review, Papers and Proceedings*, 89, 129–132.

Gandhi, A., Z. Lu, and X. Shi (2013): "Estimating Demand for Differentiated Products with Error in Market Shares," *CeMMAP working paper*.

Goolsbee, A., and A. Petrin (2004): "The consumer gains from direct broadcast satellites and the competition with cable TV," *Econometrica*, 72(2), 351–381.

Head, K., T. Mayer, et al. (2013): "Gravity equations: Workhorse, toolkit, and cookbook," *Handbook of international economics*, 4.

Hosken, D., and D. Reiffen (2004): "Patterns of retail price variation," *RAND Journal of Economics*, pp. 128–146.

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. Cambridge University Press, 1st edn.

Kahn, S., and E. Tamer (2009): "Inference on Randomly Censored Regression Models Using Conditional Moment Inequalities," *Journal of Econometrics*, 152, 104–119.

Nevo, A., and K. Hatzitaskos (2006): "Why does the average price paid fall during high demand periods?," Discussion paper, CSIO working paper.

Nurski, L., and F. Verboven (2016): "Exclusive Dealing as a Barrier to Entry? Evidence from Automobiles," *The Review of Economic Studies*, 83(3), 1156.

Quan, T. W., and K. R. Williams (2015): "Product Variety, Across-market Demand Heterogeneity, And The Value Of Online Retail," *Working Paper*.