Estimating Demand for Differentiated Products with Error in Market Shares

Amit Gandhi, Zhentong Lu, and Xiaoxia Shi*
University of Wisconsin-Madison

January 31, 2013

Abstract

In this paper we introduce a new approach to estimating differentiated product demand system that allows for error in market shares as measures of choice probabilities. In particular, our approach allows for products with zero sales in the data, which is a frequent phenomenon that arises in product differentiated markets but lies outside the scope of existing demand estimation techniques. Although we find that error in market shares generally undermine the standard point identification of discrete choice models of demand, we exploit shape restrictions on demand implied by discrete choice to generate a system of moment inequalities that partially identify demand parameters. These moment inequalities are fully robust to the variability in market shares yet are also adaptive to the information revealed by market shares in a way that allows for informative inferences. In addition, we construct a profiling approach for parameter inference with moment inequalities, making it feasible to study models with a large number of parameters (as typically required in demand applications) by focusing attention on a profile of the parameters, such as the price coefficient. We use our approach to study consumer demand from scanner data using the Dominick's Finer Foods database, and find that even for the baseline logit model, demand elasticities nearly double when the full error in market shares is taken into account.

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

JEL: C01, C12, L10, L81.

*We are thankful to Steven Berry, Jean-Pierre Dube, 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", as well as seminar participants at Wisconsin-Madison, Wisconsin-Milwaukee, Cornell, Indiana, Princeton, NYU and the Federal Trade Commission for their many helpful comments and questions.

1 Introduction

In this paper we introduce a new approach to demand estimation that allows for market shares to be measured with sampling error of any magnitude. We show that the presence of sampling error in market shares generally undermines the point identification of the popular demand estimation techniques developed in Berry (1994), Berry, Levinsohn, and Pakes (1995) and Berry, Linton, and Pakes (2004) (we use "BLP" to refer to these techniques). A severe form of this problem that arises frequently in applications is the presence of zero market shares in the data, which has remained outside the scope of aggregate discrete choice demand analysis to date. We show that discrete choice demand models are informative enough to imply a set of conditional moment inequalities which are fully robust to sampling error in market shares. We use these moment inequalities as a basis for partial identification and inference of demand parameters and counterfactuals. We apply our approach to widely used scanner data, and find that accounting for the sampling error nearly doubles price elasticities relative to existing techniques that must assume it away.

The key to our approach is that we relax the asymptotic framework used in BLP so as to allow sampling error in market shares to remain present in the limit. The consistency of the BLP estimator relies on letting both the number of products/markets *and* the number of consumers with each market grow infinitely large in the asymptotic limit. However when choice probabilities are close to zero, which is often the case in demand analysis, the number of consumers n in the data will be too small for their asymptotic approximation to apply (even if n is tens of thousands). Market shares that are zero in the data are an important special case of this more general failure.

Our approach on the other hand only assumes the number of products/markets to grow large but not the number of consumers within markets. Since the number of consumer draws are allowed to remain finite, our asymptotics can readily explain zeroes in the data: there is always a positive probability of a zero demand when the number of consumers is finite. However the difficulty with this new asymptotic framework is that the sampling error causes a loss of point identification of the model.

One main contribution is to show that the structure of the discrete choice model can be used to construct moment inequalities that partially identify the model and are fully robust to the sampling error in market shares. In addition to being fully robust to error in market shares, the other key advantages of our empirical strategy are:

1. We only use the the standard instrumental variable (IV) assumptions that BLP also use to address price endogeneity. In particular, we do not invoke alternative assumptions from the literature on nonlinear measurement error, which can be hard to justify

in the context of demand estimation;[1]

2. Our approach provides informative inferences because the moment inequalities we construct are adaptive to the revealed information in market shares; and

3. Our approach allows for arbitrary dependence among product unobservables within a market, which permits any type of strategic dependence in the the design and promotions of products competing in a market.

Another main contribution of our paper is to provide a profiling approach for inference with moment inequality models. This procedure makes it feasible in practice to perform inference in moment inequality models with many parameters and is critically needed for our demand estimation problem. The existing approach to inference in a moment inequality setting such as ours – for example, Andrews and Shi (2013) – requires exhaustive grid search over the parameter space to compute confidence sets. However, such computation is infeasible for demand studies because at least a moderate number of control variables are needed to ensure validity of the instrument for price, resulting in a moderate to large dimensional parameter space.[2] We circumvent this computational burden by performing inference directly on a *profile* of the parameters, i.e., a function of the parameters that capture the policy relevant objects of interest, such as elasticity and welfare. Although our profiling procedure can be seen as the traditional profile likelihood idea applied to conditional moment inequality (CMI) models, there is a critical difference: the profiled quasi-likelihood ratio statistic in CMI models has highly nonstandard asymptotic behavior due to the partial identification of the parameter as well as the moment inequalities structure. An asymptotic approximation of this statistic depends crucially not only on the unknown slackness of the moment inequalities, but also on the unknown shape of the identified set of the nuisance parameter. We overcome this difficulty and design a bootstrap-based critical value that is robust to both sources of nonpivotalness and leads to uniformly valid confidence sets for the true value of the profile.[3]

We apply our inference strategy to the Dominick's Finer Foods (DFF) database which

[1]These alternative assumptions usually involve the classical measurement error assumption (Abrevaya and Hausman (2004)) which does not hold in our context and the existence of a control function (Gutknecht (2012)) which is hard to justify in our context.

[2]The dimension of the parameter vector can easily exceed 30 for standard specifications used in empirical work.

[3]Two papers in the literature of partially identified models touch upon the idea of profiling without using the term: Romano and Shaikh (2008) and Santos (2012). The former proposes a sub-sampling-based confidence set for a point-identified profile of the parameters under high-level conditions, while we design a bootstrap-based confidence set for a potentially partially-identified profile of the model parameter under low-level conditions. The latter deals with a partially-identified nonparametric IV model and proposes a method that can be extended to deliver inference for a profile of the nonparametric parameters of the model. But his model involves no inequalities and his results are based on point-wise asymptotics.

is a publicly available and heavily studied scanner data set. Scanner data has become a central source for demand information in consumer product markets and is routinely used by antitrust agencies to estimate demand elasticities in merger investigations (see e..g, Hosken, O'Brien, Scheffman, and Vita (2002)). Scanner data exhibits a pattern that is very commonly found in product differentiated industries: there exists a small number of popular products and a much larger "long tail" of slower selling products that often exhibit a periods of zero sales (see e.g., Anderson (2006)). The sparse demand for the large mass of products in the long tail gives rises to a serious problem of error in market shares. In fact zeroes in demand are quite rampant in the data: many products on the shelves of supermarkets don't actually sell in a given week. However it is the weekly variation in prices that is the critical variation that identifies price elasticities. To date, the only empirical strategy for resolving this tension is to simply "drop" the products in the long tail (or impute data for them) and then apply standard BLP, which has now become standard in practice. But this "selection on outcomes" induces a selection problem that can be can quite severe which we illustrate with Monte Carlo simulations. In contrast zeroes do not pose a selection problem for our empirical strategy because they are a predicted outcome of the demand model itself. We apply our approach to the DFF data and find that demand becomes almost twice as elastic when we instead include all the observations in the data and take the error in market shares into account. This direction and magnitude of our results have significant implications for policy analysis in consumer good industries.

The plan of the paper is the following. In Section 2, we describe the econometric problem using a simplified binary choice setting without random coefficients to make the essential matters transparent. In Section 3, we introduce the general multinomial discrete choice model with random coefficients. In Section 4, we present our partial identification solution. In Section 5, we present our profiling approach to inference with moment inequalities. In Section 6, we systematically develop the application of our approach to the DFF data. Section 7 concludes.

2 Discussion of Problem in a Simple Binary Choice Model

In this section we provide a discussion of the basic empirical problem we address in this paper using a simplified binary choice model. This simplified setting avoids the notational burden of the more general random coefficients multinomial choice model and thus makes the key issue transparent. We then introduce the more general setup that is the focus of this paper in the next section.

The discrete choice approach assumes that individuals have preferences over the characteristics of products (observed and unobserved) and each individual chooses the product

that maximizes utility. Market demand is then the aggregation of the individual discrete choices. This approach provides a parsimonious representation of market demand that has a link to microeconomic foundations. Herein however lies the key econometric problem – the market level demand errors become non-separable and this frustrates the standard application of instrumental variables to control for the endogeneity of prices.

To see the problem, consider a simple binary choice setting

$$u_{it} = \beta x_t + \xi_t - v_{it}$$

where x_t are the observed characteristics of the product under consideration in market t (such as the price of a product), ξ_t is an unobserved choice characteristic (potentially correlated with x_t across markets) and v_{it} is a random utility shock to consumer i in market t. Consumer i in market t purchases the product if $u_{it} \geq 0$. A standard random utility approach is that the random utility shock v_{it} is independent of the characteristics (x_t, ξ_t) and follows a distribution $v_{it} \sim G$ for some continuous and strictly increasing CDF G. Thus, the probability π_t that an individual drawn at random from the population G purchases the product in market t is given by

$$\pi_t = G\left(\beta x_t + \xi_t\right). \tag{2.1}$$

As can be immediately seen, the unobserved characteristics ξ_t, potentially correlated with x_t, is nested inside the non-linear function G. This non-separability prevents the direct application of instrumental variables methods (which we reference as IV for short) to estimate β in (2.1).[4]

The key insight of BLP was to see that the model itself can be used to eliminate this specific source of non-separability. In particular, this non-separability can be eliminated by transforming both sides of (2.1) with G^{-1} to express the demand relationship equivalently as

$$G^{-1}\left(\pi_t\right) = \beta x_t + \xi_t. \tag{2.2}$$

If G is known (or alternatively known up to finite dimensional parameters), then the existence of instruments z_t such that $E\left[\xi_t \mid z_t\right] = 0$ allows standard instrumental variables methods to identify β. Specifically, β is identified by

$$\beta = \frac{E\left[G^{-1}\left(\pi_t\right) z_t\right]}{E\left[x_t z_t\right]}.$$

[4]See Blundell and Powell (2003) for deeper discussion of the failure of instrumental variable methods for correcting endogeneity concerns in non-separable models and the contrast with control functions.

The standard IV estimator replaces these expectations by their sample analogues

$$\hat{\beta}_T = \left(\sum_{t=1}^{T} G^{-1}(\pi_t) z_t \right) / \left(\sum_{t=1}^{T} x_t z_t \right) \tag{2.3}$$

and thus $\hat{\beta}_T \to_p \beta$ by standard law of large numbers.

However there is a critical problem with this solution: choice probabilities π_t cannot actually be observed in the data, but rather only market shares s_t are observed. The market share s_t is constructed as an average of the choices of a sample of i.i.d. individuals in market t, i.e.,

$$s_t = \frac{\sum_{i=1}^{n_t} d_{it}}{n_t} \tag{2.4}$$

and $d_{it} = 1$ if the sampled individual i in market t consumes the product, and 0 otherwise. The empirical strategy that BLP employed, which has become universal in the literature, is to replace market shares s_t for choice probabilities π_t in (2.3) and thus define the BLP estimator as:

$$\beta_T^{BLP} = \left(\sum_{t=1}^{T} G^{-1}(s_t) z_t \right) / \left(\sum_{t=1}^{T} x_t z_t \right). \tag{2.5}$$

However for the estimator (2.5) to be consistent, we would need

$$T^{-1} \sum_{t=1}^{T} G^{-1}(s_t) z_t \to_p E\left[G^{-1}\left(\pi_t\right) z_t \right] = \beta E[x_t z_t].$$

This requires that a new term introduced by market shares tends to zero in the limit, namely:

$$T^{-1} \sum_{t=1}^{T} \left[G^{-1}(s_t) - G^{-1}(\pi_t) \right] z_t \to_p 0. \tag{2.6}$$

To help understand what (2.6) means, observe that $E\left[s_t \mid \pi_t \right] = \pi_t$ or equivalently $E\left[\left(s_t - \pi_t \right) z_t \right] = 0$, i.e., the deviation $(s_t - \pi_t)$ is pure sampling error, and hence the law of large numbers would imply that $T^{-1} \sum_t \left(s_t - \pi_t \right) z_t \to_p 0$. However this does not imply that (2.6) holds because of the non-linearity of G^{-1}. Indeed, $E\left[\left(G^{-1}(s_t) - G^{-1}(\pi_t) \right) z_t \right]$ does not even exist because G^{-1} is not defined at 0 and 0 is always an outcome of s_t with positive probability mass. Thus standard law of large numbers arguments *cannot* justify (2.6).

Instead, consistency of the BLP estimator requires taking (2.6) as a high level asymptotic assumption.[5] This asymptotic condition is not a standard one, and its applicability depends

[5]The only theoretical discussion of this sampling error problem is provided in Berry, Linton, and Pakes

on whether it provides a good approximation to the data. In particular, its applicability requires that the left-hand-side of (2.6) be reasonably close to zero at the actual T and n_t in the data. This is tantamount to assuming that the number of consumers n_t is so large in every market t that $G^{-1}(s_t) - G^{-1}(\pi_t)$ is nearly zero uniformly across all markets t.

While there are many settings where this assumption may be sensible, there are many others where it is not and the left hand side of (2.6) is quite far from zero in the data. In these cases, the demand estimates derived from using the BLP estimator (2.5) will no longer be close to the BLP asymptotic limit (which is the true value) and can be economically rather misleading. Two settings that arise frequently in practice and where this bias poses a serious concern are:

1. When the data on demand arises from a national survey or national sales of many consumers, but this demand information is broken into local markets. This creates a small sample problem of consumers within markets, and hence sampling variability in s_t tends to be large and makes the left-hand-side of (2.6) large. Many industries give rise to this problem, such as demand for airlines (see e.g., Berry, Carnall, and Spiller (1996); Berry and Jia (2010)), telecommunications (see e.g., Goolsbee and Petrin (2004); Goolsbee and Klenow (2006)), and healthcare (see e.g., Brand, Gowrisankaran, Nevo, and Town (2012)).[6]

2. When the data on demand arises from a large sample of consumers within a narrow market, but the market is studied at the disaggregated product level. At this disaggregated level, the narrowly defined product categories often exhibit a well known "long tail" pattern where most products have very small choice probabilities (i.e, slow moving items) relative to the top few sellers in the category (see Anderson (2006)). These small choice probabilities cause $G^{-1}(\pi_t)$ to be incredibly sensitive to replacing π_t with s_t, even when the sampling error $\epsilon_t := \pi_t - s_t$ is quite small.[7] Thus we will

(2004) (BLintonP for short). Strictly speaking BLintonP focuses on the case of a large number of products within a single market. But their key intermediate condition implies the convergence condition (2.6) in the many market binary choice model without simulation error. See condition (i) on page 10 as well as the first line of page 35 in BLintonP. Their primitive condition Assumption A3 is sufficient for and thus stronger than this condition.

[6]In the case of airlines, the standard demand data comes from the Department of Transportation's 10 percent sample of all ticket sales. While this national survey is quite large, when broken down to the local market level, i.e., a particular origin-destination market, it is well known that it leaves a very small number of observations within smaller market routes which typically have to be dropped from the analysis. Likewise in the case of telecommunications, the national surveys that are used (such as the well known Forrester surveys) are large at the national level but becomes incredibly thin at the local market level that demand is studied. In the case of demand for health insurance plans and hospitals, the standard data come from patient discharge records within a state that when broken down to the zip code level give rise to a small number of consumers problem, which can be readily seen by the "zeroes" in demand for many hospitals.

[7]This is because the derivative of $G^{-1}(z)$ approaches infinity when z approaches zero for typical choices of G, and thus very small differences in s_t and π_t will translate into large differences between $G^{-1}(s_t)$ and

have a large departure of the left-hand-side of (2.6) from zero for even a relatively small sampling error in shares s_t.

3. A more serious manifestation of this latter problem is when some products exhibit zero sales in a market, i.e., $s_t = 0$, in which case the left-hand-side of (2.6) is $-\infty$ for standard models (i.e. logit, probit, etc) of G and is thus clearly no where close to zero. Scanner data, which has been a central source of information for demand studies, has long been recognized to pose exactly this challenge for existing demand estimation techniques. See e.g., Briesch, Dillon, and Blattberg (2008) and Park and Gupta (2009) for a discussion. This severe form of the error in market shares problem has been met with a variety of "tricks" in the applied literature, ranging from ignoring the zeroes altogether from the data (and thus inducing a selection problem) to imputing non-zero values for the zero observations.[8] However, none of these tricks address the actual source of the zeroes, which is the sampling error in market shares, and thus none delivers consistent estimators.[9]

The contribution of this paper is to provide an approach that treats the sampling error in market shares in a fully general way and thereby allows us to extend the domain of demand estimation to the above environments that are important for applied work. That is, while we maintain the standard asymptotic in the number of markets T, we impose no asymptotic approximation involving the number of consumers within a market, i.e., we relax the asymptotic assumption (2.6). This relaxation allows us to construct an asymptotic theory that can address data with zeroes and error in shares more generally. Observe that once we impose relax all restrictions on market shares beyond the sampling process (2.4), the estimating equation becomes

$$G^{-1}\left(s_t + \epsilon_t\right) = \beta x_t + \xi_t. \tag{2.7}$$

As can be seen in (2.7), the sampling error in market shares generates a non-separable error ϵ_t, which once again undermines IV estimation.[10]

As we show, this new source of non-separability causes fundamental difficulties for identification and inference. Nevertheless, we show that we can address these difficulties using the same instrumental variables assumptions that form the basis of BLP. We now detail these developments in the subsequent sections using the general model.

$G^{-1}(\pi_t)$.

[8] The quantile regression also has been suggested to us to address the "zero" problem.

[9] We illustrate the poor performance of these tricks in Section 6.

[10] Interestingly, ϵ_t would enter as separable in the direct representation of demand (2.1), but of course ξ_t would still be non-separable in that case.

3 Identification

3.1 The Basic Environment

In this section, we describe the general demand model for product differentiated goods and the basic identification problem.

Consider T markets. In each market, say t, has a set of $J_t + 1$ differentiated products. The product labeled $j = 0$ in each market t is referred to as the "outside option", and the goods labeled $j = 1, \ldots, J_t$ are the "inside goods". The inside goods in market t are characterized by a vector of observable demand shifters $x_t = (x_{1t}, \ldots, x_{J_t t}) \in X$, where each $x_{jt} \in \mathbb{R}^K$ for $j = 1, \ldots, J_t$ is a vector of product attributes (typically including price) corresponding to the inside products. Let $\xi_t = (\xi_{1t}, \ldots, \xi_{J_t t}) \in \mathbb{R}^{J_t}$ denote a vector of demand shocks, where each ξ_{jt} for $j = 1, \ldots, J_t$ is typically interpreted as the unobservable (to the econometrician) attribute of each inside product.

The demand of a randomly drawn consumer i from market t is described by a random utility model. For simplicity, we use the standard random coefficients model employed by Berry (1994), but the ideas we present extend in a straightforward way to more general specifications. The utility to consumer i for product $j = 0, \ldots, J_t$ in market t is

$$u_{ijt} = \delta_{jt} + v_{ijt}, \tag{3.1}$$

where

1. $\delta_{jt} = x_{jt}\beta_0 + \xi_{jt}$ is the mean utility of product $j > 0$ in market t, and mean utility of the outside good $j = 0$ is normalized to $\delta_{0t} = 0$. Let $\delta_t = (\delta_{1t}, \ldots, \delta_{J_t t})$ denote the vector of mean utilities of the "inside" goods $j > 0$.

2. The vector $v_{i\cdot t} = (v_{i0t}, \ldots, v_{iJ_t t}) \sim F(\cdot \mid x_t; \lambda_0)$ is the random vector of tastes in market t. Notice that allowing x_t and a parameter to enter F make our specification encompass general random coefficients because one can then view β_0 as the mean of the random coefficients and v_{ijt} as the product of the error from the random coefficients and the product characteristic x_{jt}. We will assume for simplicity that the random vector $v_{i\cdot t}$ has full support on $\mathbb{R}^{J_t+1}$, which is a property exhibited by all the standard random utility models, For example, if one component of each random utility term v_{ijt} is an idiosyncratic preference shock with full support (as in the logit, mixed logit or probit models), then full support of $v_{i\cdot t}$ holds.[11]

[11] The main role of the full support assumption is for expositional and computational convenience . We could in principle proceed instead under the weaker "connected substitutes" structure of Berry, Gandhi, and Haile (2011).

3. The vector $\theta_0 = (\beta_0, \lambda_0) \in \Theta$ denotes the true value of the parameters, where $\Theta \subset \mathbb{R}^{d_\theta}$ where d_θ is a positive integer is the parameter space.

Each consumer i in market t chooses product j if $u_{ijt} \geq u_{ij't}$ for all $j' = 0, 1, ..., J_t$. Then the random utility model can be aggregated to yield a system of choice probabilities

$$\pi_{jt} = \sigma_j(\delta_t, x_t; \lambda_0) \quad j = 1, \ldots, J_t, \tag{3.2}$$

where $\sigma_j,\ j = 1, ..., J_t$ are known functions. Let $\pi_t = (\pi_{1t}, \ldots, \pi_{J_t t})'$ denote the vector of inside good choice probabilities predicted by the random utility model in market t. The choice probability system can be inverted under general conditions as shown in Berry, Gandhi and Haile (2011) to obtain

$$\delta_{jt} = \sigma_j^{-1}(\pi_t, x_t; \lambda_0)\ j = 1, ..., J_t. \tag{3.3}$$

We refer to $\sigma_j^{-1}(\cdot, x_t; \lambda_0)$ as the inverse share function of product j.

For later use, we define $\vec{\pi}_t = (\pi_{0t}, \pi_t')'$ to denote the vector of choice probability for all $J_t + 1$ goods. Clearly, $\pi_{0t} = 1 - \pi_t' 1_{J_t}$ and hence π_t uniquely determines $\vec{\pi}_t$ and vice versa.

We observe the aggregate demand of n_t consumers who are sampled in market t,[12] which can be represented as the market share s_{jt} for $j = 0, 1, \ldots, J_t$ where

$$s_{jt} = \frac{\sum_{i=1}^{n_t} d_{ijt}}{n_t} \tag{3.4}$$

and

$$d_{ijt} = \begin{cases} 1 & i^{th} \text{ consumer in market } t \text{chooses product } j \\ 0 & \text{otherwise.} \end{cases}$$

Given that all consumers in the market are observationally identical (i.e., there are no individual specific covariates to distinguish different consumers in the sample), each observed consumer in the market has identical choice probabilities π_t. Thus the vectors of empirical shares $s_t = (s_{1t}, ..., s_{J_t t})'$ and $\vec{s}_t = (s_{0t}, s_t')$ are the sample analogue of the underlying population choice probabilities π_t and $\vec{\pi}_t$, respectively. In particular, conditional on π_t and n_t, the vector $n_t \vec{s}_t$ follows a multinomial distribution $MN(n_t, \vec{\pi}_t)$.

Finally we impose the instrumental variable condition in the form of a conditional mean restriction

$$E[\xi_{jt} \mid z_{jt}] = 0 \quad \forall j = 1, \ldots, J_t \ a.s.\ [z_t, J_t] \tag{3.5}$$

[12]The number of consumers n_t can equal the population size of a city or the number of consumers in a survey from a city (where the city is defined as the market), or the number of consumers who enter a store in a given week (where the store/week unit is defined as a market), among a variety of other possibilities depending on the empirical context.

where z_{jt} is a vector of instruments for x_{jt} and $z_t = (z_{1t}, ..., z_{J_t t})'$. We note here that standard empirical work typically assumes that entry and exit of products across markets is exogenous. This implies that J_t can join the instruments and gives us the conditional mean restriction

$$E[\xi_{jt} \mid z_{jt}, J_t] = 0 \quad \forall j = 1, \ldots, J_t \ a.s. \ [z_t, J_t]. \tag{3.6}$$

3.2 Identification Problem

We now consider what features of the model can be identified if we let the number of markets $T \to \infty$ but allow the observed number of consumers n_t within each market t to be fixed. In this new asymptotic limit, the distribution of $(n_t, s_t, x_t, z_t, J_t)$ rather than (π_t, x_t, z_t, J_t) is identified, i.e., we learn the joint distribution of *observed market shares* and the other market observables rather than choice probabilities and market observables in the limit as the number of markets grows large.

As this new asymptotic allows for sampling variability in market shares in the limit, a fundamental problem arises: point identification is lost due to the sampling error and the nonlinearity of the model. What we now show is that this problem in our setting is rather severe – the identifying content of the conditional mean restriction (3.5) completely vanishes in the absence of further restrictions. That is, without more information, the standard instrumental variables restriction loses *all* of its empirical content when we treat market shares s_t rather than choice probabilities π_t as the relevant observable in the asymptotic limit.

For the sake of clarity, we show this negative result in the context the simplest random utility of demand, i.e., the simple logit model with a single product and no covariates. In this simple model,

$$u_{it} = c + \xi_t + v_{it},$$

where $v_{it} \overset{iid}{\sim} EV$ is a standardized (type I) extreme value random variate and $E[\xi_t] = 0$. The constant c is the only parameter to identify. In this case it is straightforward to see that

$$c + \xi_t = \sigma^{-1}(\pi_t, x_t, \lambda_0) \equiv \log\left(\frac{\pi_t}{1 - \pi_t}\right),$$

and hence the constant c equals

$$E\left[\log\left(\frac{\pi_t}{1 - \pi_t}\right)\right]. \tag{3.7}$$

Suppose that instead of observing π_t we only observe market share s_t constructed from the choices of n i.i.d. consumers in each market t.[13] Since the consumers are i.i.d., we

[13]Here we treat n_t as a fixed constant for all market but the results can be understood as conditional on

have $ns_t \mid \pi_t \sim BN(n, \pi_t)$, the binomial distribution with parameters (n, π_t). Clearly, the expectation of s_t identifies $E[\pi_t]$ because $E[s_t] = E[E[s_t \mid \pi_t]] = E[\pi_t]$. However we now show that the parameter c is completely not identified even with identification of the entire distribution of the random variable s_t. This is because the distribution of s_t severely under-identifies the distribution of π_t, i.e., for a given distribution of s_t found in the data there are a large number of distributions of π_t that are compatible with it and these multiple distributions of π_t allow for all kinds of outcomes for the expectation of interest $E\left[\log\left(\frac{\pi_t}{1-\pi_t}\right)\right]$.

To see this more precisely, first observe that the probability mass function (pmf) of ns_t given $\pi_t \in [0,1]$ is

$$p_{ns_t|\pi_t}(l|\pi) = \left(\begin{array}{c} n \\ l \end{array}\right) \pi^l (1-\pi)^{n-l}, \ \forall l = 0, ..., n. \tag{3.8}$$

Suppose that the true distribution of π_t is $F_\pi : [0,1] \to [0,1]$. Then the unconditional pmf $p_{ns_t}(\cdot; F_\pi)$ of ns_t implied by F_π is

$$p_{ns_t}(l; F_\pi) = \int \left(\begin{array}{c} n \\ l \end{array}\right) \pi^l (1-\pi)^{n-l} dF_\pi(\pi), \ \forall l = 0, \ldots n. \tag{3.9}$$

That is, the distribution of ns_t is a mixture of binomials with mixing probability F_π. The distribution $p_{ns_t}(\cdot; F_\pi)$ is identified from the data, but because p_{ns_t} is a discrete distribution with $n+1$ support points, and F_π is potentially continuous, F_π is underidentified by the equation (3.9). We show below that the knowledge of $p_{ns_t}(\cdot)$ in general can only give trivial bounds ($(-\infty, \infty)$) for $E\left[\log\left(\frac{\pi_t}{1-\pi_t}\right)\right]$.

To state the result precisely, observe that the set of all possible mixture distributions p_{ns_t} that can be potentially observed in the data is geometrically constructed as follows. Let $\vec{p}_s(F_\pi) = (p_{ns_t}(0; F_\pi), ..., p_{ns_t}(n; F_\pi))'$, and let [14]

$$P_s = \{\vec{p}_s(F_\pi) : F_\pi(t) = 1\{t \geq x\} \text{ for some } x \in [0,1]\}. \tag{3.10}$$

Let $\vec{p}_s^* = (p_s^*(0), ..., p_s^*(n))'$ where $p_s^*(\cdot)$ is the true pmf of ns_t. Then the convex hull of P_s, $co(P_s)$, is the set of all possible values that $\vec{p}_s^*$ can take. A point $\vec{p}_s^*$ in the interior of $co(P_s)$ is called a generic point in the set because the boundary of $co(P_s)$ has Lebesgue measure zero and can be approximated arbitrarily closely by points in the interior.

Theorem 1. *The distribution $\vec{p}_s^*$ generically produces no informative restrictions on the expectation* (3.7)*, i.e., for any generic $\vec{p}_s^*$ and any $M > 0$ there exists distributions of π_t, n_t.*

[14] The set P_s is the set of $\vec{p}_s(F_\pi)$ vectors generated by all F_π that is degenerate at one point in $[0,1]$.

$F_\pi^M : (0,1) \to [0,1]$ *and* $F_\pi^{-M} : (0,1) \to [0,1]$ *that are consistent with* $\vec{p}_s^* - \vec{p}_s^* = \vec{p}_s\left(F_\pi^M\right)$ *and* $\vec{p}_s^* = \vec{p}_s\left(F_\pi^{-M}\right)$ *– and satisfy*

$$E_{F_\pi^M}\left[\log\left(\frac{\pi_t}{1-\pi_t}\right)\right] > M \quad \textit{and} \quad E_{F_\pi^{-M}}\left[\log\left(\frac{\pi_t}{1-\pi_t}\right)\right] < -M$$

The proof is given in Appendix A and easily extends to the general demand model discussed above. The result has important implications for empirical work with differentiated products. When only empirical shares rather than choice probabilities are identified in the data, the conditional mean restriction (3.5) has zero empirical content, i.e., empirical shares provide no restrictions on the expectation of interest

$$E\left[\sigma_j^{-1}\left(\pi_t, x_t; \lambda_0\right) \mid z_{jt}\right]. \tag{3.11}$$

Thus the instrumental variable assumptions applied to the data imposes no informative restrictions on the model parameters.

4 Identification using Moment Inequalities

4.1 Constructing Product Level Moment Inequalities

The negative conclusion of Theorem 1 is driven by the fact that the distribution of empirical shares s_t observed in the data can be rationalized by an underlying distribution of choice probabilities F_π that contains support points arbitrarily close to 0 or 1. Thus to avoid this conclusion and restore empirical content to the instrumental variable assumption, it is necessary to restrict the support of the true underlying (but un-observed) F_π ex-ante so that it is bounded away from zero. In the context of the general product differentiated demand model, this restriction takes the form of Assumption 1 below. Letting Δ_{J_t} be the J_t dimensional unit simplex (the set of all possible choice probability vectors over the J_t+1 products), we assume the following:

Assumption 1. *There exists* $\varepsilon\left(z_t, J_t\right) = \left(\varepsilon_0\left(z_t, J_t\right), \ldots, \varepsilon_{J_t}\left(z_t, J_t\right)\right)$ *such that* $\varepsilon_j(z_t, J_t) > 0$ *for all* j, z_t, J_t *and the support of* $\pi_t \mid z_t, J_t$ *is contained in* $\Delta_{J_t}^{\varepsilon(z_t,J_t)} = \{\vec{\pi} = (\pi_0, \pi_1, ..., \pi_{J_t})' \in \Delta_{J_t} : \pi_j \geq \varepsilon_j\left(z_t, J_t\right)$ *for* $j = 0, \ldots, J_t\}$.

That is, there exists a vector of lower bounds $\varepsilon_t := \varepsilon\left(z_t, J_t\right) > 0$ which bound from below the value that the vector of choice probabilities π_t can possibly take (conditional on the instrument and number of products (z_t, J_t)).[15] This is a necessary assumption in light

[15] We can think of $F_{\pi|z_t,J_t}$ as the "ideal" reduced form that would point identify the model, but cannot itself be identified. What the above Theorem 1 generally shows is that we must restrict the support of this

of Theorem 1. The same assumption is also imposed in BLintonP to ensure estimation consistency,[16] but in our context we see that it is necessary even for identification. In practice, one can set $\varepsilon_{jt} := \varepsilon_j(z_t, J_t)$ to be the same across all observations j, t and to be equal, for example, to machine precision. Alternatively one can also set it according to ones prior on the minimum choice probability to sustain the fixed costs of making the product available in the market.

Our identification strategy is to exploit Assumption 1 along with the structure of the discrete choice model to construct new inversion mappings $\sigma_{j,l}^{-1}(s_t, n_t, x_t; \lambda_0)$ and $\sigma_{j,u}^{-1}(s_t, n_t, x_t; \lambda_0)$ that bound the expectation of interest (3.11):

$$E\left[\sigma_{j,l}^{-1}(s_t, n_t, x_t; \lambda_0) \mid z_{jt}\right] \leq E\left[\sigma_j^{-1}(\pi_t, x_t; \lambda_0) \mid z_{jt}\right] \leq E\left[\sigma_{j,u}^{-1}(s_t, n_t, x_t; \lambda_0) \mid z_{jt}\right] \quad (4.1)$$

We then can use the fact that $\sigma_j^{-1}(\pi_t, x_t; \lambda_0) = \beta_0 x_{jt} + \xi_{jt}$ along with the conditional mean restriction $E[\xi_{jt} \mid z_{jt}] = 0$ to express the empirical content of the model as a system of conditional moment inequalities

$$\begin{aligned} E\left[\sigma_{j,u}^{-1}(s_t, n_t, x_t; \lambda_0) - \beta_0 x_{jt} \mid z_{jt}\right] &\geq 0 \\ E\left[\beta_0 x_{jt} - \sigma_{j,l}^{-1}(s_t, n_t, x_t; \lambda_0) \mid z_{jt}\right] &\geq 0 \end{aligned} \quad (4.2)$$

that holds for all $j = 1, \ldots, J_t$ almost surely with respect to $[z_t, J_t]$. Letting $y_t = (s_t, n_t, x_t, J_t)$, we can express this system more succinctly as

$$E[m_j(y_t; \theta_0) \mid z_{jt}] \geq 0 \quad \forall j = 1, \ldots, J_t, \ a.s.\ [z_t, J_t] \quad (4.3)$$

where $m_j(y_t; \theta_0)$ is a stacked vector of the two moments in (4.2). This system of conditional moment inequalities partially identifies the true parameter vector θ_0 and forms the basis for our inference strategy.

To motivate our approach to constructing the bounds (4.1), observe that although s_t is an unbiased estimator for π_t, plugging the unbiased estimator into the inverse share function, i.e., $\sigma_j^{-1}(\cdot, x_t; \lambda_0)$, causes the expectation

$$E\left[\sigma_j^{-1}(s_t, x_t; \lambda_0) \mid z_{jt}\right] \quad (4.4)$$

to no longer equal the expectation of interest (3.11) because the function σ^{-1} is nonlinear. In fact, the situation is more complicated because the expectation (4.4) does not even exist.

true underlying reduced form $F_{\pi|z_t,J_t}$ to be bounded away from the boundary of the simplex in order for the reduced form we observe $F_{s|z_t,J_t}$ to have empirical content for the model parameters.

[16] Condition S of BLintonP.

This is because there is always some positive probability that the empirical shares s_{jt} can be zero for any $\pi_{jt} \in (0,1)$, but σ^{-1} is not defined on the boundary of the simplex (see Berry, Gandhi, and Haile (2011) for further discussion of this latter fact). We solve the problem in the following steps.

1. In the first step, we transform empirical shares so they move strictly to the interior of the unit simplex. We do so using a natural transformation: the Laplace's rule of succession, which takes the form

$$\tilde{s}_t = \frac{n_t s_t + 1}{n_t + J_t + 1}.$$

   The transformed shares $\tilde{s}_t$ can be interpreted as a Bayesian posterior point estimate of π_t under a uniform prior on the J_t-dimensional unit simplex.[17] Such a transformed share will also be useful for counterfactuals when we must form such a point estimate. Using $\tilde{s}_t$ in $\sigma_j^{-1}(\cdot, x_t; \lambda_0)$ in place of s_t solves the existence problem of the expectation (4.4). However, we still have that $E\left[\sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda_0\right) \mid z_{jt}\right] \neq E\left[\sigma_j^{-1}\left(\pi_t, x_t; \lambda_0\right) \mid z_{jt}\right]$.

2. In the second step, which is the most important step, we exploit a monotonicity feature of demand that was recently shown by Berry, Gandhi, and Haile (2011) but has not yet been applied to empirical work. In particular, for a given product j and parameters λ, we can show there exists a unique real valued function $\eta_j(n_t, \pi_t, x_t, J_t; \lambda)$ defined as the unique η that solves

$$E\left[\sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right) | n_t, \pi_t, x_t, J_t\right] = \sigma_j^{-1}(\pi_t, x_t; \lambda), \tag{4.5}$$

   where e_j is a vector whose jth element is one and all other elements are zeros, and the expectation is taken with respect to the randomness in s_t. We show in Lemma B.1 of Appendix B the existence and uniqueness of such a solution. The proof exploits the monotonicity feature of inverse demand that Berry, Gandhi, and Haile (2011) showed is satisfied in discrete choice models quite generally.

3. In the last step, we let

$$\eta_{jt}^u := \eta_j^u(n_t, z_t, x_t, J_t; \lambda) := \sup_{\pi_t : \vec{\pi}_t \in \Delta_{J_t}^{\varepsilon(z_t, J_t)}} \eta_j(n_t, \pi_t, x_t, J_t; \lambda). \tag{4.6}$$

   As the proof of Lemma B.1 makes clear, the function σ_j^{-1} is monotone in the jth

[17] See e.g. Chapter 9.4 of Good (1983).

share, which gives us the inequality

$$E\left[\sigma_j^{-1}\left(\tilde{s}_t + \eta_{jt}^u \cdot e_j, x_t; \lambda\right) | n_t, \pi_t, x_t, J_t\right] \geq \sigma_j^{-1}(\pi_t, x_t; \lambda), \tag{4.7}$$

for all π_t such that $\vec{\pi}_t \equiv (1 - \pi_t' 1_{J_t}, \pi_t')' \in \Delta_{J_t}^{\varepsilon(z_t, J_t)}$. Hence taking expectations of both sides of this inequality conditional on the instruments z_{jt} and using Assumption 1 along with the the law of iterated expectation, we have[18]

$$E\left[\sigma_j^{-1}\left(\tilde{s}_t + \eta_{jt}^u \cdot e_j, x_t; \lambda\right) \mid z_{jt}\right] \geq E\left[\sigma_j^{-1}(\pi_t, x_t; \lambda) \mid z_{jt}\right]. \tag{4.8}$$

Similarly, we let

$$\eta_{jt}^l := \eta_j^l(n_t, z_t, x_t, J_t; \lambda) := \inf_{\pi_t : \vec{\pi}_t \in \Delta_{J_t}^{\varepsilon(z_t, J_t)}} \eta_j(n_t, \pi_t, x_t, J_t; \lambda). \tag{4.9}$$

and we have:

$$E\left[\sigma_j^{-1}\left(\tilde{s}_t + \eta_{jt}^l \cdot e_j, x_t; \lambda\right) \mid z_{jt}\right] \leq E\left[\sigma_j^{-1}(\pi_t, x_t; \lambda) \mid z_{jt}\right].$$

In Appendix B, we show how η_{jt}^u and η_{jt}^l are explicitly computed in the case of the work-horse logit and nested logit demand models. There we also provide computational guidance for more general models.

Now, letting

$$\begin{aligned}
\sigma_{j,l}^{-1}(s_t, n_t, x_t; \lambda_0) &:= \sigma_j^{-1}\left(\tilde{s}_t + \eta_{jt}^l \cdot e_j, x_t; \lambda\right) \\
\sigma_{j,u}^{-1}(s_t, n_t, x_t; \lambda_0) &:= \sigma_j^{-1}\left(\tilde{s}_t + \eta_{jt}^u \cdot e_j, x_t; \lambda\right)
\end{aligned}$$

we have thus constructed the new inversion mappings that satisfy the bounding inequality (4.1).[19]

Remark 1. Our bounds (4.1) have the key property of being "adaptive" to the observed shares s_t in a way that makes them especially useful for applied work. To appreciate this adaptivity property, consider instead a "naive bounding" approach that uses Assumption 1, but does not exploit the monotonicity implied by the underlying discrete choice model. Such

[18] Formally we first take conditional expectations conditional on (z_t, J_t) to establish that $E\left[\sigma_j^{-1}\left(\tilde{s}_t + \eta_{jt}^u \cdot e_j, x_t; \lambda\right) \mid z_t, J_t\right] \geq E\left[\sigma_j^{-1}(\pi_t, x_t; \lambda) \mid z_t, J_t\right]$ using Assumption 1. We then take conditional expectations of both sides of this inequality conditional on just z_{jt}.

[19] Note that in principle $\sigma_{j,l}^{-1}$ and $\sigma_{j,u}^{-1}$ also depend upon (z_t, J_t) because the computation of η_{jt}^u and η_{jt}^l in principle can depend upon (z_t, J_t) (because we allow $\varepsilon_t := \varepsilon(z_t, J_t)$ to depend upon this vector). We suppress this dependence both for notational simplicity and also, as noted above, in practice it is common to simply let $\varepsilon_{jt} = \varepsilon$ in which case the arguments (z_t, J_t) drops out. We will slightly abuse notation in what follows and leave out the dependence of a few other functions on (z_t, J_t) for notational simplicity.

a strategy would directly correct the difference between $E\left[\sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda\right) \mid n_t, \pi_t, x_t, J_t\right]$ and $\sigma_j^{-1}\left(\pi_t, x_t; \lambda\right)$ by constructing the linear correction factors:

$$\begin{aligned}
\mu_{j,naive}^{u}\left(x_t, n_t; \lambda\right) &= \sup_{\pi_t : \vec{\pi}_t \in \Delta_{J_t}^{\varepsilon(z_t, J_t)}} \left\{\sigma_j^{-1}\left(\pi_t, x_t; \lambda\right) - E\left[\sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda\right) \mid n_t, \pi_t, x_t, J_t\right]\right\} \\
\mu_{j,naive}^{l}\left(x_t, n_t; \lambda\right) &= \inf_{\pi_t : \vec{\pi}_t \in \Delta_{J_t}^{\varepsilon(z_t, J_t)}} \left\{\sigma_j^{-1}\left(\pi_t, x_t; \lambda\right) - E\left[\sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda\right) \mid n_t, \pi_t, x_t, J_t\right]\right\}.
\end{aligned} \tag{4.10}$$

The expectations $E[\sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda\right) + \mu_{j,naive}^{l} | z_{jt}]$ and $E[\sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda\right) + \mu_{j,naive}^{u} | z_{jt}]$ by construction bound the expectation of interest (3.11). However these bounds will typically be quite loose because the difference $\sigma_j^{-1}\left(\pi_t, x_t; \lambda\right) - E\left[\sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda\right) \mid n_t, \pi_t, x_t, J_t\right]$ is large for π_t such that $\vec{\pi}_t$ is close to the boundary of $\Delta_{J_t}^{\varepsilon(z_t, J_t)}$ and thus the correction factors $\mu_{j,naive}^{l}\left(x_t, n_t; \lambda\right)$ and $\mu_{j,naive}^{u}\left(x_t, n_t; \lambda\right)$ will also be large as a result. These large correction factors are then applied indiscriminately to all realizations of $\sigma_j^{-1}(\tilde{s}_t, x_t; \lambda)$, even when $\vec{s}_t \equiv (1 - s_t' 1_{J_t}, s_t')'$ is far from the boundary.

The key novelty of our approach is taking advantage of the monotonicity implied by the structure of the discrete choice model and correcting $\tilde{s}_{jt}$ rather than $\sigma_j^{-1}(\tilde{s}_t, x_t; \lambda)$ directly. To see the advantage we gain, observe that the "linear correction factors" on $\sigma_j^{-1}(\tilde{s}_t, x_t; \lambda)$ implied by our construction are

$$\begin{aligned}
\mu_j^{l}\left(s_t, x_t, n_t; \lambda\right) &= \sigma_{j,l}^{-1}\left(s_t, n_t, x_t; \lambda\right) - \sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda\right) \\
\mu_j^{u}\left(s_t, x_t, n_t; \lambda\right) &= \sigma_{j,u}^{-1}\left(s_t, n_t, x_t; \lambda\right) - \sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda\right).
\end{aligned}$$

These implied factors $\mu_j^{u}\left(s_t, x_t, n_t; \lambda\right)$ and $\mu_j^{l}\left(s_t, x_t, n_t; \lambda\right)$ are functions of s_t, unlike the naive correction factors above. The implied factors are large for markets with $\vec{s}_t$ near the boundary of the unit simplex (i.e., when the correction is most needed) and are negligible for markets with $\vec{s}_t$ being well inside the interior of the unit simplex (i.e., when the correction is least needed). The adaptiveness comes from the fact that our correction η_{jt}^{u} and η_{jt}^{l} enter in the same way as the noise $(\tilde{s}_t - \pi_t)$ and thus has large effect on $\sigma_j^{-1}(\tilde{s}_t, x_t; \lambda)$ when and only when the noise does. This adaptiveness allows our approach to deliver informative inferences as our empirical study to follow shall illustrate.

4.2 Aggregating Moment Inequalities to the Market Level

In (4.3), it is shown that the aggregate demand model can be written as

$$E\left[m_j\left(y_t; \theta_0\right) \mid z_{jt}\right] \geq 0 \quad \forall j = 1, \ldots, J_t \text{ a.s. } [z_t, J_t]. \tag{4.11}$$

The model (4.11) appears almost the same as the conditional moment inequality model discussed extensively in, e.g., Andrews and Shi (2013) and Chernozhukov, Lee, and Rosen (2008). However there is one essential difference that we need now address. The existing methods of inference are designed for generic problems in which observations are independent, or at least can be assumed to satisfy a special form of weak dependence (e.g. mixing). Such assumptions are not readily satisfied in the aggregate demand model because the observations $\{x_{jt}, z_{jt}, \xi_{jt}\}$ tend to be correlated across j *within* the same market t in non-standard ways due to the strategic interaction between products in a market. [20]

Instead of treating each (j,t) as an observation, we propose to aggregate the moments up to the market level and use the pure market level variation as the basis for inference and use the more standard independence or weak dependence assumption on the market level variation.

The aggregation we seek needs to be done properly to preserve all the identification information there is in (4.11) under acceptable assumptions on the data generating process. The first step is to transform (4.11) into moments not conditioning on product level variables – moments that can be aggregated. Let $g(z_{jt})$ be a real-valued function where the function g lies in the collection $\mathcal{G}$. We take $\mathcal{G}$ to be a collection of indicator functions:

$$\mathcal{G} = \{1(z \in C) : C \in \mathcal{C}\}, \tag{4.12}$$

where $\mathcal{C}$ is a collection of subsets of $\mathcal{Z}$, the support of z_{jt}. The following Lemma shows the equivalent form of (4.11). The proof is the same as that of Lemma 3 in Andrews and Shi (2013) and is omitted.

Lemma 1. *Suppose that $\mathcal{C} \cup \{\emptyset\}$ is a semi-ring of subsets of $\mathcal{Z}$. Also suppose that $\mathcal{Z}$ can be written as the union of countable disjoint sets in $\mathcal{C}$ and the sigma field generated by $\mathcal{C} \cup \{\emptyset\}$ equals $\mathcal{B}(\mathcal{Z})$ – the Borel sigma field on $\mathcal{Z} \subseteq R^{d_z}$.*[21] *Then,* (4.11) *holds if and only if*

$$E[m_j(y_t, \theta_0) g(z_{jt})] \geq 0, \ \forall g \in \mathcal{G}, \ \forall j = 1, ..., J_t \ a.s. \ [J_t]. \tag{4.13}$$

[20] The dependence of $m_j(y_t, \theta_0)$ on other products' characteristics cannot be captured by a market level fixed effect. Also it is not possible to stack up the $m_j : j = 1, ..., J_t$ and treat the model as a market level model with a multi-dimensional moment condition because J_t varies across markets.

[21] A semi-ring, $\mathcal{R}$, of subsets of a universal set $\mathcal{Z}$ is defined by three properties: (i) $\emptyset \in \mathcal{R}$, (ii) $A, B \in \mathcal{R} \Rightarrow A \cap B \in \mathcal{R}$ and (iii) if $A \subset B$ and $A, B \in \mathcal{R}$, then there exists disjoint sets $C_1, ..., C_N \in \mathcal{R}$ such that $B - A = \cup_{i=1}^{N} C_i$. An example of a $\mathcal{C}$ that satisfies the assumptions in Lemma 1 when $\mathcal{Z}$ is discrete is $\mathcal{C}_d = \{\{z\} : z \in \mathcal{Z}\}$. An example when $\mathcal{Z} = [0,1]$ is $\mathcal{C}_c = \{[a,b) : a, b \in [0,1]\} \cup \{\{b\}\}$.

The second step is to aggregate up the moments in (4.13) to market level:

$$E\left[J_t^{-1}\sum_{j=1}^{J_t} m_j(y_t,\theta_0)g(z_{jt})\right] \geq 0,\ \forall g \in \mathcal{G}. \tag{4.14}$$

The aggregated moment condition contains exactly the same information as (4.13) because the model is agnostic about how products with different t but the same j index are linked to each other, i.e. there is no ex ante information in the product index j. As a result of this agnostic stand, we have for all $j^{'} = 1, 2, ..., J_t$,

$$\begin{aligned} E\left[J_t^{-1}\sum_{j=1}^{J_t} m_j(y_t,\theta_0)g(z_{jt})\right] &= E\left(E\left[J_t^{-1}\sum_{j=1}^{J_t} m_j(y_t,\theta_0)g(z_{jt})|J_t\right]\right) \\ &= E\left[m_{j'}(y_t,\theta_0)g(z_{j't})\right]. \end{aligned} \tag{4.15}$$

It is then immediate that the market level moment condition (4.14) holds if and only if (4.13) does.

Let $w_t = (y_t, z_t)$ and let

$$\rho(w_t,\theta,g) = J_t^{-1}\sum_{j=1}^{J_t} m_j(y_t,\theta)g(z_{jt}). \tag{4.16}$$

Then the model (4.14) can be written as

$$E[\rho(w_t,\theta_0,g)] \geq 0,\ \forall g \in \mathcal{G}. \tag{4.17}$$

The next section takes the model in (4.17) as the starting point and develop a profiling method for the inference of any parameter that is identified through a (possibly set valued) function of θ_0.

The above aggregation allows J_t to be endogenous (or exogenous). However this aggregation is different from the aggregation implied by the traditional BLP approach that treats each (j,t) pair as an observation (and implicitly treats J_t as exogenous). To be more comparable with the literature, we can define an alternative aggregation as

$$\rho(w_t,\theta,g) = \sum_{j=1}^{J_t} m(y_t,\theta)g(z_{jt}). \tag{4.18}$$

Similar arguments as those above can be used to show that (4.17) holds with $\rho(w_t,\theta,g)$ defined as in (4.18) provided that J_t is exogenous (i.e. (3.6) holds instead of (3.5)). Because

the focus of this paper is on the bound construction rather than on the endogeneity of entry and exit of products, in our empirical application, we use the $\rho(w_t, \theta, g)$ in (4.18), which is in direct analogue to the traditional BLP estimation to which our inference results are compared.

5 Estimation and Inference

In this section, we introduce the estimation and inference procedure. To focus attention on our main contribution – the profiling approach, we assume the markets are drawn from an i.i.d. distribution. In Appendix G, we discuss how to allow for dependence among markets.

5.1 Profiling Confidence Set

The model (4.17) is a moment inequality model with many moment conditions. One could use the method developed in Andrews and Shi (2013) to construct a confidence set for θ_0. However, Andrews and Shi (2013)'s 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. Computing the level set of a non smooth function is essentially a grid-search problem which is only feasible *if* d_θ is small. However, in demand estimation, d_θ cannot be small because at least a moderate number of covariates have to be controlled for the assumption $E(\xi_{jt}|z_{jt}) = 0$ to be reasonable.

On the other hand, in demand estimation the coefficients of the control variables are nuisance parameters that often are of no particular interest. The 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. The profiling approach is an old approach that has led to the well-known QLR (quasi-likelihood ratio) test for nonlinear profiles of parameters in extremum estimation problems (see Newey and McFadden (1994)). In conditional moment inequality (CMI) models, however, profile QLR-based inference has not been explored much. On the other hand, such an inference approach is more crucially needed in CMI models than in standard extremum estimation problems because Wald-based confidence sets are no longer appropriate due to partial identification and inequality constraints. Our contribution of this section therefore is to provide such an approach. We show that the QLR test statistic (appropriately formed for the CMI models) combined with a carefully designed simulated critical value overcomes the complicacy of conditional moment inequality models and delivers uniformly asymptotically valid confidence sets.

The profiling approach applies to general moment inequality models with many moment inequalities. Thus from this point on, we treat $\rho(w_t, \theta, g)$ as a generic moment function with dimension k. In the demand model above, $k = 2$.

The parameter of interest, γ_0, is related to θ_0 through:

$$\gamma_0 \in \Gamma(\theta_0) \subseteq R^{d_\gamma}, \tag{5.1}$$

where $\Gamma : \Theta \to 2^{R^{d_\gamma}}$ is a known mapping where $2^{R^{d_\gamma}}$ denotes the collection of all subsets of R^{d_γ}. Three examples of Γ are given below:

Example. $\Gamma(\theta) = \{\alpha\}$: γ_0 is the price coefficient α_0. In the simple logit model, the price coefficient is all one needs to know to compute the demand elasticity.

Example. $\Gamma(\theta) = \{e_j(p, \pi, \theta, x) = (\alpha p_j/\pi_j)(\partial\sigma_j(\sigma^{-1}(\pi, x, \lambda), x, \lambda)/\partial\delta_j)\}$: γ_0 is the own-price demand elasticity of product j at a given value of the price vector p, the choice probability vector π and the covariates x.

Example. $\Gamma(\theta) = \{e_j(p, \pi, \theta, x) : \pi \in [\pi^l, \pi^u]\}$: γ_0 is the demand elasticity of product j at a given value of the price vector p, the covariates x and at the choice probability vector that is known to lie between π^l and π^u. This example is particularly useful when the elasticity depends on the choice probability but the choice probability is only known to lie in an interval.

Let Γ_0 be the identified set of γ_0: $\Gamma_0 = \{\gamma \in R^{d_\gamma} : \exists\theta \in \Theta_0 \ s.t. \ \Gamma(\theta) \ni \gamma\}$, where $\Theta_0 = \{\theta \in \Theta : E\rho(w_t, \theta, g) \geq 0\}$. The profiling approach constructs a confidence set for γ_0 by inverting a test of the hypothesis:

$$H_0 : \gamma \in \Gamma_0, \tag{5.2}$$

for each parameter value γ. The confidence set is the collection of values that are not rejected by the test.

Let $\Gamma^{-1}(\gamma) = \{\theta \in \Theta : \Gamma(\theta) \ni \gamma\}$. The test to be inverted uses the *profiled* test statistic:

$$\hat{T}_T(\gamma) = T \times \min_{\theta \in \Gamma^{-1}(\gamma)} \hat{Q}_T(\theta), \tag{5.3}$$

where $\hat{Q}_T(\theta)$ is an empirical measure of the violation to the moment inequalities. The confidence set of confidence level p is the set of all points for which the test statistic does not exceed a critical value $c_T(\gamma, p)$:

$$CS_T = \{\gamma \in R^{d_\gamma} : \hat{T}_T(\gamma) \leq c_T(\gamma, p)\}. \tag{5.4}$$

Notice that the new confidence set only involves computing a d_γ-dimensional level set, where d_γ is often 1. The profiling transfers the burden of searching (for low values) over the surface of the non smooth function $T(\theta) - c(\theta)$ to searching over the surface of the typically smooth and often convex function $\hat{Q}_T(\theta)$.

We choose a critical value, $c_T(\gamma, p)$, of significance level $1 - p \in (0, 0.5)$, to satisfy

$$\lim_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \Pr{}_F(\hat{T}_T(\gamma) > c_T(\gamma, p)) \leq 1 - p, \tag{5.5}$$

where F is the distribution on $(w_t)_{t=1}^T$ and $\mathcal{H}_0$ is the null parameter space of (γ, F). The definition of $\mathcal{H}_0$ along with other technical assumptions are given in Appendix C.[22]

As a result of (5.5), the confidence set asymptotically has the correct minimum coverage probability:

$$\liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \Pr_F(\gamma \in CS_T) \geq p. \tag{5.6}$$

The left hand side is called the "asymptotic size" of the confidence set in Andrews and Shi (2013). We achieve the asymptotic size control by deriving an asymptotic approximation for the distribution of the profiled test statistic $\hat{T}_T(\gamma)$ that is uniformly valid over $(\gamma, F) \in \mathcal{H}_0$ and simulating the critical value from the approximating distribution through either a subsampling or a bootstrapping procedure.

In the rest of the section, we describe the test statistic and the critical value in detail and show that (5.6) holds.

5.2 Test Statistic

The test statistic is the QLR statistic (i.e. a criterion-function-based statistic)[23]

$$\begin{aligned}\hat{T}_T(\gamma) &= T \times \min_{\theta\in\Gamma^{-1}(\gamma)} \hat{Q}_T(\theta) \text{ with} \\ \hat{Q}_T(\theta) &= \int_{\mathcal{G}_T} S(\bar{\rho}_T(\theta, g), \hat{\Sigma}_T^\iota(\theta, g)) d\mu(g),\end{aligned} \tag{5.7}$$

[22]Note that we use F to denote the distribution of the full observed data vector and thus (γ, F) captures everything unknown in the expression $\Pr_F(\hat{T}_T(\gamma) > c_T(\gamma, p))$. This notation differs from the traditional literature where the true distribution of the data is often indicated by the true value of θ, but is standard in the recent partial identification literature. See Romano and Shaikh (2008) and Andrews and Shi (2013).

[23]Note that we do not follow the traditional QLR test exactly to define $\hat{T}_T(\gamma) = T \times \min_{\theta\in\Gamma^{-1}(\theta)} \hat{Q}_T(\theta) - T \times \min_{\theta\in\Theta} \hat{Q}_T(\theta)$. This is because the validity of our critical value depends on certain monotonicity of the asymptotic approximation of the test statistic and the monotonicity does not hold with this alternative test statistic due to the subtraction of $T \times \min_{\theta\in\Theta} \hat{Q}_T(\theta)$.

where $\mathcal{G}_T$ is a truncated/simulated version of $\mathcal{G}$ such that $\mathcal{G}_T \uparrow \mathcal{G}$ as $T \to \infty$, $\mu(\cdot)$ is a probability measure on $\mathcal{G}$, $S(m, \Sigma)$ is a real-valued function that measures the discrepancy of m from the inequality restriction $m \geq 0$, and

$$\begin{aligned}
\bar{\rho}_T(\theta, g) &= T^{-1} \sum_{t=1}^{T} \rho(w_t, \theta, g), \\
\hat{\Sigma}_T^{\iota}(\theta, g) &= \hat{\Sigma}_T(\theta, g) + \iota \times \hat{\Sigma}_T(\theta, 1) \\
\hat{\Sigma}_T(\theta, g) &= T^{-1} \sum_{t=1}^{T} \rho(w_t, \theta, g)\rho(w_t, \theta, g)' - \bar{\rho}_T(\theta, g)\bar{\rho}_T(\theta, g)'.
\end{aligned} \tag{5.8}$$

In the above definition, ι is a small positive number which is used because in some form of S defined in Appendix C, the inverse of $\hat{\Sigma}_T^{\iota}(\theta, g)$'s diagonal elements enter, and the ι prevents us from taking inverse of zeros. In some other forms of S, e.g. the one defined below and used in the simulation and empirical section of this paper, the ι does not enter the test statistic because $S(m, \Sigma)$ does not depend on Σ.

Appendix C gives the assumptions that the user-chosen quantities S, μ, $\mathcal{G}$ and $\mathcal{G}_T$ should satisfy. Under those assumptions, we can show that $\min_{\theta \in \Gamma^{-1}(\gamma)} \hat{Q}_T(\theta)$ consistently estimate $\min_{\theta \in \Gamma^{-1}(\gamma)} Q_F(\theta)$ where

$$\begin{aligned}
Q_F(\theta) &= \int_{\mathcal{G}} S(\rho_F(\theta, g), \Sigma_F^{\iota}(\theta, g)) d\mu(g), \text{ with} \\
\rho_F(\theta, g) &= E_F(\rho(w_t, \theta, g)), \\
\Sigma_F(\theta, g) &= Cov_F(\rho(w_t, \theta, g)) \text{ and } \Sigma_F^{\iota}(\theta, g) = \Sigma_F(\theta, g) + \iota \Sigma_F(\theta, 1).
\end{aligned} \tag{5.9}$$

The symbols "E_F" and "Cov_F" denote expectation and covariance under the data distribution F respectively. Notice that Γ_0 depends on F. We make this explicit by changing the notation Γ_0 to $\Gamma_{0,F}$ for the rest of this paper.

We can also show that $\min_{\theta \in \Gamma^{-1}(\gamma)} Q_F(\theta) = 0$ if and only if $\gamma \in \Gamma_{0,F}$. This result combined with the convergence result implies that $\hat{T}_T(\gamma)$ diverges to infinity at $\gamma \notin \Gamma_{0,F}$. That implies that there is no information loss in using such a test statistic.

Lemma 2 summarizes those two results. The parameter space $\mathcal{H}$ of (γ, F) appearing in the lemma is defined in Assumption C.2 in the appendix.

Lemma 2. *Suppose that the conditions in Lemma* 1 *and Assumptions* C.1,C.2, C.4, C.5(a) and C.6 (a) and (d) *hold. Then for any* $(\gamma, F) \in \mathcal{H}$,

(a) $\min_{\theta \in \Gamma^{-1}(\gamma)} \hat{Q}_T(\theta) \to_p \min_{\theta \in \Gamma^{-1}(\gamma)} Q_F(\theta)$ *under* F, *and*

(b) $\min_{\theta \in \Gamma^{-1}(\gamma)} Q_F(\theta) \geq 0$ *and* $= 0$ *if and only if* $\gamma \in \Gamma_{0,F}$.

In the simulation and the empirical application of this paper, the following choices of

S, $\mathcal{G}$, $\mathcal{G}_T$ and μ are used mainly for computational convenience. 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]: $\tilde{z}_{c,jt} = F_{N(0,1)}(\hat{\Sigma}_{z_c}^{-1/2} z_{c,jt})$, where $F_{N(0,1)}(\cdot)$ is the standard normal cdf and $\hat{\Sigma}_{z_c}$ is the sample covariance matrix of $z_{c,jt}$. The set $\mathcal{G}$ is defined as

$$\begin{aligned}\mathcal{G} &= \{g_{a,r,\zeta}(z_d, z_c) = 1(\tilde{z}_c \in C_{a,r}, z_d = \zeta) : C_{a,r} \in \mathcal{C}_{cc}, \zeta \in \mathcal{Z}_d\}, \text{ where} \\ \mathcal{C}_{cc} &= \{\times_{u=1}^{d_{z_c}}((a_u - 1)/(2r), a_u/(2r)] : \ a_u \in \{1, 2, ..., 2r\}, \text{ for } u = 1, ..., d_{z_c} \\ &\qquad \text{and } r = r_0, r_0 + 1, ...\}\end{aligned} \tag{5.10}$$

where "cc" stands for "countable hyper-cube." For $\mathcal{G}_T$, it is a truncated version of $\mathcal{G}$. It is defined the same as $\mathcal{G}$ except that in the definition of $\mathcal{C}_{cc}$, we let r run from r_0 to $\bar{r}_T$ where $\bar{r}_T \to \infty$ as $T \to \infty$.

For S , we use

$$S(m, \Sigma) = \sum_{j=1}^{k} [m_j]_-^2, \tag{5.11}$$

where m_j is the jth coordinate of m and $[x]_- = |\min\{x, 0\}|$. There may be efficiency loss from not weighting the moments using the variance matrix, but this S function brings great computational convenience because it makes the minimization problem in (5.3) a convex one. For $\mu(\cdot)$, we use

$$\mu(\{g_{a,r,\zeta}\}) \propto (100 + r)^{-2}(2r)^{-d_{z_c}} K_d^{-1} \text{ for } g \in \mathcal{G}_{d,cc}, \tag{5.12}$$

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

5.3 Critical Value

We propose two types of critical values, one based on standard subsampling and the other based on a bootstrapping procedure with moment shrinking. Both are simple to compute. The bootstrap critical value may have better small sample properties, and is the procedure we use in the empirical section.[24] It is worth noting that we resample at the market level for both the subsampling and the bootstrap.

Let us formally define the subsampling critical value first. It is obtained through the standard subsampling steps: [1] from $\{1, ..., T\}$, draw without replacement a subsample of market indices of size b_T; [2] compute $\hat{T}_{T,b_T}(\gamma)$ in the same way as $\hat{T}_T(\gamma)$ except using the

[24]The bootstrap procedure here, like in most problems with partial identification, does not lead to higher-order improvement.

subsample of markets corresponding to the indices drawn in [1] rather than the original sample; [3] repeat [1]-[2] S_T times obtain S_T independent (conditional on the original sample) copies of $\hat{T}_{T,b_T}(\gamma)$; [4] let $c^*_{sub}(\gamma, p)$ be the p quantile of the S_T independent copies. Let the subsampling critical value be

$$c_T^{sub}(\gamma, p) = c^*_{sub}(\gamma, p + \eta^*) + \eta^*, \tag{5.13}$$

where $\eta^* > 0$ is an infinitesimal number. The infinitesimal number is used to avoid making hard-to-verify uniform continuity and strict monotonicity assumptions on the distribution of the test statistic. It can be set to zero if one is willing to make the continuity assumptions. Such infinitesimal numbers are also employed in Andrews and Shi (2013). One can follow their suggestion of using $\eta^* = 10^{-6}$.

Let us now define the bootstrap critical value. It is obtained through the following steps: [1] from the original sample $\{1, ..., T\}$, draw with replacement a bootstrap sample of size T; denote the bootstrap sample by $t_1, ..., t_T$, [2] let the bootstrap statistic be

$$T_T^*(\gamma) = \min_{\theta \in \Theta : \gamma \in \Gamma(\theta)} \int_{\mathcal{G}} S(\hat{\nu}_T^*(\theta, g) + \kappa_T^{1/2} \bar{\rho}_T(\theta, g), \hat{\Sigma}_T^{\iota}(\theta, g)) d\mu(\mathcal{G}), , \tag{5.14}$$

where $\hat{\nu}_T^*(\theta, g) = \sqrt{T}(\bar{\rho}_T^*(\theta, g) - \bar{\rho}_T(\theta, g))$, $\bar{\rho}_T^*(\theta, g) = T^{-1} \sum_{\tau=1}^{T} \rho(X_{t_\tau}, \theta, g)$, and κ_T is a sequence of moment shrinking parameters: $\kappa_T / T + \kappa_T^{-1} \to 0$; [3] repeat [1]-[2] S_T times and obtain S_T independent (conditional on the original sample) copies of $T_T^*(\gamma)$; [4] let $c^*_{bt}(\gamma, p)$ be the p quantile of the S_T copies. Let the bootstrap critical value be

$$c_T^{bt}(\gamma, p) = c^*_{bt}(\gamma, p + \eta^*) + \eta^*, \tag{5.15}$$

where $\eta^* > 0$ is an infinitesimal number which has the same function as in the subsampling critical value above.

Critical values that are not based on resampling are possible, too. For example, one can define a critical value similar to the bootstrap one, except with $\hat{\nu}_T^*(\theta, g)$ replaced by a Gaussian process with covariance kernel that equals the sample covariance of $\rho(w_t, \theta^{(1)}, g^{(1)})$ and $\rho(w_t, \theta^{(2)}, g^{(2)})$ for $(\theta^{(j)}, g^{(j)}) \in \Theta \times \mathcal{G}, j = 1, 2$. For lack of space, we do not discuss such critical values in detail.

5.4 Coverage Probability

We show that the confidence sets defined in (5.4) using either $c_T^{sub}(\gamma, p)$ and $c_T^{bt}(\gamma, p)$ have asymptotically correct coverage probability uniformly over $\mathcal{H}_0$ under appropriate assumptions. The assumptions are given in the appendix for brevity.

Theorem 2 (CP). *Suppose that the conditions for Lemma 1 and Assumptions* C.1-C.3 *and* C.5-C.7 *hold, then*

(a) (5.6) *holds with* $c_T(\gamma, p) = c_T^{sub}(\gamma, p)$*, and*

(b) (5.6) *holds with* $c_T(\gamma, p) = c_T^{bt}(\gamma, p)$.

The proof of Theorem 2 is quite lengthy and is given in Appendix E. Here we provide some intuition how it works and why it is lengthy. To start, rewrite the test statistic as:

$$\begin{aligned}\hat{T}_T(\gamma) &= \min_{\theta \in \Gamma^{-1}(\gamma)} \int_{\mathcal{G}_T} S(\sqrt{T}\bar{\rho}_T(\theta, g), \hat{\Sigma}_T^{\iota}(\theta, g)) d\mu(g) \\ &= \min_{\theta \in \Gamma^{-1}(\gamma)} \int_{\mathcal{G}_T} S(\hat{\nu}_T(\theta, g) + \sqrt{T}\rho_F(\theta, g), \hat{\Sigma}_T^{\iota}(\theta, g)) d\mu(g),\end{aligned}$$

where

$$\hat{\nu}_T(\theta, g) = \sqrt{T}(\bar{\rho}_T(\theta, g) - \rho_F(\theta, g)). \tag{5.16}$$

The asymptotic distribution of $\hat{T}_T(\gamma)$ is difficult to derive because of the term $\sqrt{T}\rho_F(\theta, g)$, which typically does not converge as $T \to \infty$. We note that the issue is more complicated here than in Andrews and Shi (2013) and other existing papers dealing with moment inequality models where a similar term (i.e. the slackness parameter) also presents. In those papers, one can fix a (sequence of) θ in the identified set Θ_0. At that θ, the slackness parameter is known to have a lower bound: *zero*. Thus, one can either replace it by zero to obtain conservative (but valid) inference, or replace it by something asymptotically no greater for less conservative inference. Those techniques work because the test statistic is non-increasing in the slackness parameter. However, in the present context, we have to minimize over $\theta \in \Gamma^{-1}(\gamma)$, where $\Gamma^{-1}(\gamma)$ contains both points in Θ_0 and points outside. Points outside Θ_0 are relevant for the asymptotic behavior of $\hat{T}_T(\gamma)$ and thus cannot be ignored. However, at those points, $\sqrt{T}\rho_F(\theta, g)$ does not have a known lower bound — it may diverge to $-\infty$. As a result, the techniques in the literature do not guarantee valid inference.

Nonetheless, we show that subsampling is a uniformly valid inference procedure, and we also propose a bootstrap procedure that is in shape similar to that in Andrews and Soares (2010) for a certain choice of their moment selection function. The essence of both our subsampling and bootstrap procedures is to replace $\hat{\nu}_T(\theta, g)$ by a random process $\hat{\nu}_T^*(\theta, g)$ that asymptotically approximate its distribution and to effectively replace $\sqrt{T}\rho_F(\theta, g)$ by a discounted version of it: $\sqrt{\kappa_T}\rho_F(\theta, g)$ where $\kappa_T \to \infty$ and $\kappa_T^{-1}T \to \infty$. Intuitively, the procedure works for two reasons: (1) for $\theta \in \Theta_0$, the discounting makes the term smaller and thus the statistic bigger, and more importantly, (2) for $\theta \notin \Theta_0$, $\sqrt{\kappa_T}\rho_{F,j}(\theta, g)$ might be bigger than $\sqrt{T}\rho_{F,j}(\theta, g)$ making $\int_{\mathcal{G}_T} S(\hat{\nu}_T^*(\theta, g) + \sqrt{\kappa_T}\rho_F(\theta, g), \hat{\Sigma}_T^{\iota}(\theta, g)) d\mu(g)$ smaller, but under appropriate conditions, there must be a $\theta^{\dagger}$ closer to Θ_0 than θ, such that $\sqrt{T}\rho_{F,j}(\theta^{\dagger}, g)$

is similar to $\sqrt{\kappa_T}\rho_{F,j}(\theta, g)$. In other words, the discounting does not create any new small values $\int_{\mathcal{G}_T} S(\hat{\nu}_T^*(\theta, g) + \sqrt{\kappa_T}\rho_F(\theta, g), \hat{\Sigma}_T^\iota(\theta, g))d\mu(g)$ that $\min_{\theta\in\Gamma^{-1}(\gamma)}$ has not taken account of before the discounting. As a result, the discounting does not make the overall test statistic smaller. The proof of Theorem 2 can be seen as formalization of (1) and (2). The formal arguments involve detailed characterization and careful control of the behavior of various components of $\hat{T}_T(\gamma)$ in a (not necessarily $T^{-1/2}$-) neighborhood of $\Theta_0 \cap \Gamma^{-1}(\gamma)$ and thus is quite lengthy.

6 Monte Carlo Simulation

In this section we present a Monte Carlo experiment that illustrates two key points: (1) the large biases that the existing techniques experience even when n_t is seemingly large and (2) the performance of our inference strategy. We present another Monte Carlo in Appendix F that is specifically designed to help us calibrate tuning parameters of the bootstrap for the empirical application below. In that Appendix we also perform a repeated simulation study which demonstrates that our profiling procedure attains the correct asymptotic coverage probabilities and also exhibits good power for a data generating process that resembles the structure of the data in our application. In this section, we focus on a simple yet informative data generating process to highlight the contrast of controlling for error in market shares versus ignoring it. We set T to be a large number, 5000, and only report results from one simulation repetition.

We consider a simple binary choice logit model. Assume that each market has only one inside good and an outside good. The utility of consumer i in market t from consuming the inside good is $u_{it} = \delta_t + v_{i1t}$ and that from consuming the outside good is v_{i0t}, where v_{i0t} and v_{i1t} are independent type-I extremum value random variables that are independent of δ_t. Let δ_t depend only on one single observed product characteristic x_t in the form $\delta_t = \alpha_0 + \beta_0 x_t + \xi_t$. Then

$$\log\left(\frac{\pi_t}{1-\pi_t}\right) = \alpha_0 + \beta_0 x_t + \xi_t. \tag{6.1}$$

For simplicity we take x_t and ξ_t to be independent so that x_t is its own instrument. Let the true value of the parameters be $\alpha_0 = -1$, $\beta_0 = -1$. Let x_t take three values 1, 4 and $\tilde{x}$ with probabilities .3, .5 and .2, respectively, where $\tilde{x}$ is a real number larger than 4. Let ξ_t exhibit a simple form of heteroskedasticity: $\xi_t \sim N\left(0, \sigma_\xi^2 1\{x_t \leq 4\} + 4\sigma_\xi^2 1\{x_t > 4\}\right)$. We vary $\tilde{x}$ and σ_ξ to get different proportions of zero s_t's in the generated data.

The number of consumer draws in each market that determine market shares s_t are taken to be 10,000, which is a very large number of consumer observations per product

relative to most applications. The data only produce a small fraction of zero observations - roughly 15 percent of the markets experience a zero share in the simulation for the range of $\tilde{x}$ and σ_ξ we consider.

We first implement our bounds estimator that formally deals with the problem of sampling error in market shares, which is the underlying source for the zeroes in the data. The confidence intervals we obtain for β_0 (when we profiled out α_0) are shown in the first two rows of Table 1.[25]. As can be seen all the intervals cover the true value of -1, and our fifty percent confidence interval (which is the partially identified analogue of a point estimate) remains quite tight across treatments. [26]

We next compare our approach with three potential alternatives that ignore the sampling error in market shares but instead use informal "tricks" for dealing with zero shares in the data. The first approach is to drop the zeroes from the data and using the standard BLP estimator, which is a very common approach in practice as we discuss further below in the application Section 7 (this strategy is labeled as "BLP Logit" in Table 1). The second alternative is to modify the shares to eliminate the zeros and use the standard BLP estimator on the resulting sample. In particular we add one consumer to each product in the data and re-compute shares, which is the equivalence of using our Laplacian shares $\tilde{s}_t$ in BLP directly (this is labeled below as "+1 Logit" in Table 1). The third informal strategy is to ignore the real source of the zeros, treat the zeros as truncations in the dependent variable, and then run quantile regressions (this is labeled as "QReg" below). The idea behind this third trick is the general impression that quantile regressions are more robust to certain kinds of truncation problems than mean regressions. [27]Four different quantiles are considered: 50% (median), 80%, 90% and 95%. For these alternative approaches, both point estimates and 95% confidence intervals are reported. The confidence intervals for the simple logit and the +1 logit approaches use the heteroskedasticity robust formula.

The results are shown in the remaining rows of Table 1. As can be seen, all of these informal attempts yield serious biases. This is a fairly natural outcome because none of them deal with the fundamental source of sampling variability that drive the presence of zero shares in the first place. In particular, the "Simple Logit" has a parameter value

[25]We follow the implementation instructions given in Section 5.2 and set $r_0 = 1$ and $\bar{r}_T = 50$. Note that we do not assume that the support of x_t is known in the implementation. For the tuning parameter κ_T, we use $1/\sqrt{0.5 \log T}$, which is a recommended choice from the Monte Carlo experiment in Appendix F.

[26]As mentioned above, what we report are results from one simulation repetition, but because we take T to be 5000, which is quite large, changing the seed in the random number generating process varies the results very little.

[27]When running the quantile regressions, we replace the zero shares by the minimum positive share in the simulated data. Using some nonzero small numbers to replace the zeros is necessary because the invert demand function evaluated at zero will be $-\infty$ and for such markets, the quantile regression cannot give finite estimates in our example. We find that the quantile regression results are very sensitive to what we use to replace the zeros, suggesting our problem (which clearly is not a truncation problem) is not approximated well by a truncation problem that the quantile regression can handle.

attenuated towards zero, which is expected given the "selection on outcomes" that results from dropping the zeroes from the sample. What is more interesting is that the other alternative approaches that try to rescue the zeroes from being selected out of the sample continue to suffer from the same attenuation bias. Thus we see that even in this simple setting with a large number of consumers per product in each market, properly controlling for the error in market shares is essential for recovering the true demand parameters in the presence of zero shares in the data.

Table 1: Monte Carlo Results for the Binary Logit Example

	$\sigma_\xi = .5$, $\tilde{x} = 9.5$	$\sigma_\xi = 1$, $\tilde{x} = 10.5$	$\sigma_\xi = 1.5$, $\tilde{x} = 11.5$	$\sigma_\xi = 2$, $\tilde{x} = 12.5$
50% CS	[-1.00, -.99]	[-1.00, -.99]	[-1.03, -.98]	[-1.14, -.96]
95% CS	[-1.01, -.98]	[-1.03, -.96]	[-1.08, -.93]	[-1.24, -.89]
BLP Logit	-.86 [-.87, -.85]	-.77 [-.79, -.75]	-.67 [-.70, -.65]	-.58 [-.61, -.55]
"+1" Logit	-.81 [-.81, -.80]	-.71 [-.72, -.71]	-.63 [-.64, -.62]	-.56 [-.57, -.55]
50% QReg	-.81 [-.82, -.81]	-.71 [-.72, -.70]	-.63 [-.64, -.62]	-.57 [-.58, -.55]
80% QReg	-.88 [-.88, -.87]	-.82 [-.83, -.81]	-.77 [-.78, -.76]	-.73 [-.74, -.72]
90% QReg	-.91 [-.91, -.90]	-.87 [-.89, -.85]	-.78 [-.81, -.76]	-.75 [-.78, -.72]
95% QReg	-.84 [-.86, -.83]	-.83 [-.85, -.81]	-.75 [-.77, -.72]	-.69 [-.73, -.66]
% zeros overall	13.54%	15.12%	15.80%	16.90%

Note: True value = -1, $T = 5,000$, $\kappa_T = T/(0.5 \cdot \log(T))$, ε_t =minimum true share $\forall t$.

7 Empirical Application

7.1 Introduction

We now turn to applying our inference approach to real data. We use supermarket scanner data to examine the effects of taking sampling errors in market shares into account for demand elasticity estimation and thus for policy analysis. Scanner data is an attractive area of application to illustrate the usefulness of our approach for two key reasons. First, scanner data from supermarkets have become a major source of information for demand estimation in the literature, and is now routinely used by the DOJ and FTC in merger cases (see e.g., Hosken, O'Brien, Scheffman, and Vita (2002)). Second, scanner data very clearly exhibit the problem of error in market shares: within even narrowly defined product categories, there exists a very large selection of products stocked by the retailer, but only a small handful of

these products reliably sell from week to week. The bulk of products are sparsely demanded. This phenomenon has more generally been coined the "long tail" phenomenon (see e.g., Anderson (2006)). The "long tail" refers to the large mass of slower selling products that typically drives characterizes most of the product variety in an industry, and has become especially pronounced with internet retailing (and associated with the success of companies such as Amazon and eBay). [28]

Products in the long tail have especially small choice probabilities and thus naturally produce many zero sales in the data, i.e., many products in a given store do not sell in a given week. This indicates a significant sampling error in the market shares. Our approach is the first solution to demand estimation with scanner data that does not require either selecting out products that experience zero sales or aggregating across products to remove the zeros. Selecting out the zeros generates a selection problem whereas aggregating products generates an aggregation bias well known in the scanner data literature (see Hosken, O'Brien, Scheffman, and Vita (2002)). Our approach on the other hand allows researchers to analyze the full range of product variety in the data, which is consistent with the choice problem consumers actually face. As our results indicate, studying the scanner data at this fully disaggregated product level and taking the sampling error in market shares into full account lead to interesting and new insights about consumer demand elasticities.

7.2 The Dominick's Finer Foods Data

We obtain data from Dominick's Database through the Kilts center at the University of Chicago, which covers weekly store-level scanner data at Dominick's Finer Foods (DFF) and has been used by many researchers as the basis of demand studies, e.g., Chintagunta and Vishal (2003), Chen and Yang (2007), etc.[29] The data comprises all Dominick's Finer Foods chain stores in the Chicago metropolitan area over the years from 1989 to 1997. Like other scanner data sets, this data set provides information on demand at store/week/UPC level, where a UPC is a unique bar code that identifies a product.

We follow the prevailing literature (see Nair and Chintagunta (2005), Chintagunta and Vishal (2003)) and define a market in the data as a store/week pair, and define products within a market to be the UPC's that a store places on its shelf in a given week.[30] These are

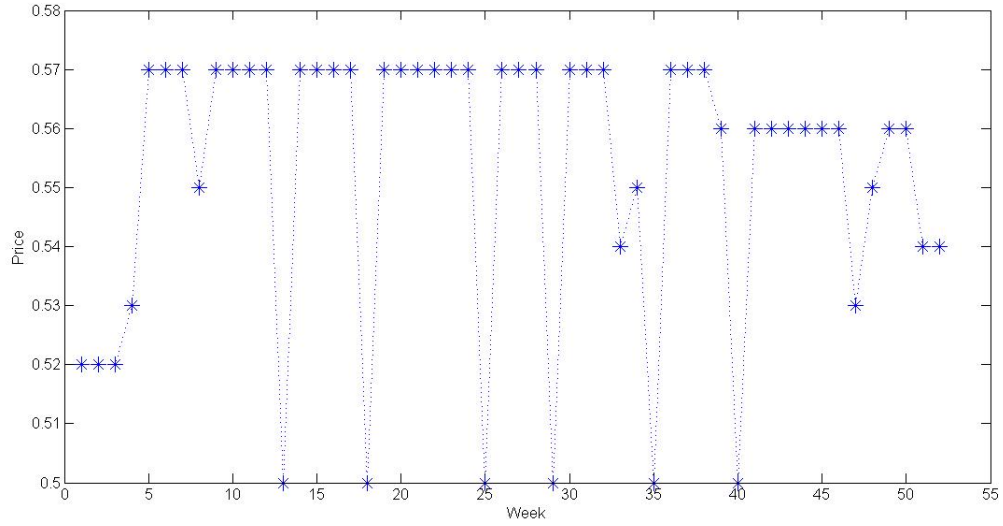
[28] Although the demand for any individual product in the tail is small, which may make it appear a "waste of space" for retailers to carry, the tail as a whole is a significant fraction of total demand and has become an increasingly profitable segment for retailers as inventory costs have fallen. For further discussion see Anderson (2006).

[29] 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

[30] An ideal feature of the data is that a UPC is listed for a given store/week market if it actually is a UPC the store carries that week. Thus the data enable us to identify true "zero sales" – no consumer who entered the store demanded the product that week, and these are not confounded by the possibility that the

the natural units of a market and product in the data, particularly because the key source of the price variation in the data comes from the decision by an individual store to put an individual UPC on sale.[31] This is illustrated in Figure 1, which shows the time series of price for a representative UPC at a representative store in the data. As can be seen, this price variation largely takes the form of the product going on a temporary sale and then reverting back to an "everyday" price.

Figure 1: Price Variation of a UPC



The data in principle provide about 40,000 store/week markets and nearly 180 million store/week/UPC product level observations. We follow the literature and take the number of consumers in a market to be given by the "Customer Count" variable, which gives us the number of customers that entered the given store during the given week and made a purchase.[32]

We display summaries of the different product categories in Table 2. The first column shows the extent of products variety that is present in an average store/week - the number of UPC's can be seen varying from roughly 50 (e.g., oatmeal and bath tissue) to over four hundred (e.g., cookies) within even these fairly narrowly defined categories. The second column shows that, within each category, there are relatively few "hit" products and a

product simply was not stocked that week.

[31] Nair and Chintagunta (2005) shows that the time series variation rather than across store variation accounts for the bulk of the price variation in the data.

[32] The mean of the "Customer Count" variable is 18,350 and its standard deviation is 5,226. The fact that the literature takes the sample of consumers to be those who enter the store as the market size means that what is being implicitly estimated is a "store" level demand system, i.e., part of the outside option is to buy the same product at another store. This understanding will help us with the interpretation of our findings.

"long tail" of slow moving products. In particular, the data appears consistent with the well known "80/20" rule, which is a classic empirical regularity that finds roughly 80 percent of sales in a market being driven by the top 20 percent of products (see Anderson (2006)). The third column shows the key difficulty associated with this long tail pattern of sales for demand estimation - many of the slow moving products are often not sold at all in a given store/week market. In particular, we see that the fraction of observations with zero sales can even exceed 60% for some categories.

Table 2: 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
Analgesics	224	80.12%	58.02%
Bath Soap	72	87.80%	74.51%
Beer	179	87.18%	50.45%
Bottled Juices	187	74.40%	29.87%
Cereals	212	72.08%	27.14%
Cheeses	283	80.41%	27.01%
Cigarettes	161	96.43%	66.21%
Cookies	407	81.40%	42.57%
Crackers	112	81.63%	37.33%
Canned Soup	218	76.25%	19.80%
Dish Detergent	115	69.04%	42.39%
Front-end-candies	199	76.24%	32.37%
Frozen Dinners	123	66.53%	38.32%
Frozen Entrees	346	74.65%	37.30%
Frozen Juices	94	75.16%	23.54%
Fabric Softeners	123	65.74%	43.74%
Grooming Products	461	80.04%	62.11%
Laundry Detergents	200	65.52%	50.46%
Oatmeal	51	71.37%	26.15%
Paper Towels	56	83.56%	48.27%
Refrigerated Juices	91	83.18%	27.83%
Soft Drinks	537	91.21%	38.54%
Shampoos	820	83.69%	69.23%
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%
Toothpastes	187	74.19%	51.93%
Bathroom Tissues	50	84.06%	28.14%

7.3 Demand Estimation on the Bath Tissue Data

The preponderance of zero sales in the data poses a serious problem for demand estimation because they are a direct reflection of a significant error in market shares as measures of choice probabilities as has already been discussed above. In particular, as we have already discussed above both theoretically and illustrated with the Monte Carlo, the standard BLP estimator is no longer consistent when the error in market shares is not sufficiently small. We now turn to applying our inference strategy to the DFF data.

We focus on the bathroom tissue category for our current analysis. Our choice is based on a few different considerations. First, several authors have previously considered the bathroom tissue category in the DFF data e.g., Israilevich (2004), Romeo (2005), Misra and Mohanty (2008), and further the bathroom tissue industry has been a source of some policy interest, see e.g., Hausman and Leonard (2002). These papers provide a point of comparison for our demand estimates. Second, this category has a smaller fraction of zeroes as compared to some other product categories (as can be seen in Table 2), and thus is far from a "worst case" scenario for the selection bias that zero sales pose for the BLP approach.

As has been mentioned, markets are naturally formed by store/week pairs. A number of papers analyzing the DFF data have focused on the interaction between market demographics and consumer preferences because there is rich demographic variation associated with the locations of different stores (see e.g., Hoch, Kim, Montgomery, and Rossi (1995)). We respect this concern by focusing attention on a single store.[33] Given our choice of bathroom tissue, we will focus on the first two years of data from this store, which are 1991-1992. This choice reflects the fact that a major change in the bathroom tissue industry took place in 1993 when one of the major brands Charmin brand introduced its "ultra" line of products (see Hausman and Leonard (2002) for a discussion), which very likely had a large impact on brand preferences due at the very least to the big changes in advertising campaigns across brands that ensued. The period 1991-1992 thus represents a more stable demand period. This leaves us with a sample size of 4428 observations (UPC/week), which consists of 104 weeks with an average number of UPC's in each week being 43.

To gain a better picture of the long tail pattern in our subsample of the data, we separate products into two separate groups, namely those that experience a zero sale during some period in the data (we shall refer to this first group as the "zeroes") and those that did not (we shall refer to this second group as the "non-zeroes"). Out of 51 total UPC's in our data, 25 UPC's fall into the "zeroes" group, i.e., roughly half the UPC's experience a zero during

[33]The store we select is number 134 in the dataset, located at the city of West Chicago. It is the only store in a pricing zone (zone number is 13), which belongs to "medium" price tier and is one of the 16 DFF's reported zones. Our results are in no way dependent upon the selection of this particular store.

at least one week of sales. Demand for an average product in the "non-zeroes" group is roughly four times as high as for the "zeroes" category overall. Even conditional on selling in a given week, the average product in the "non-zeroes" group has roughly three times higher demand than an average product in the "zeroes" group, which is consistent with the fact that products in the "zeroes" group have significantly smaller choice probabilities. However the "zeroes" products as a whole are a significant component of total sales: roughly 20 percent of the total sales arise from sales from these UPC's.

We will focus on estimating the logit demand model. The logit remains the workhorse of demand analysis for differentiated products both because of its computational simplicity and the transparency of its policy implications (see e.g., Werden and Froeb (1994)). It is also a fundamental starting point that serves to motivate potentially richer specifications. Our strategy in isolating the logit is also intentional: we wish to demonstrate that even for this widely recognized and seemingly well understood model, and even using workhorse data (i.e., the DFF data), the problem of error in market shares still poses serious empirical problems and that our inference strategy can actually reveal new insights in this context.[34]

Observe that in the logit case, the price elasticity of a set of products I_t (possibly a single UPC, or a brand) is simply $\epsilon_{I_t t} = \alpha p_{I_t t} (1 - \pi_{I_t t})$ (where $p_{I_t t}$ is a price index of the set and $\pi_{I_t t}$ is the choice probability of the set), and hence inference on α is sufficient to construct price elasticities.

We define the covariates x_{jt} to include indicator variables for package size, brand, promotion, holiday, year and a flexible set of interactions between these variables. There are 11 brands, 9 package sizes, and promotion of UPC indicates that the store is marketing a promotion on the UPC. The nature of the price variation in the data was depicted in Figure 1, which draws attention to the potential endogeneity problem between price p_{jt} and the unobservable ξ_{jt}, where the latter could reflect unobserved shelving and/or advertising choice by the store. In particular, 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 might expect a negative correlation between price and the unobservable.[35] We construct instruments for price by inverting DFF's data on gross margin to calculate the chain's wholesale costs,

[34] Our general approach of course allows for random coefficients.

[35] The usual concern that price is positively correlated with the intrinsic product quality is being offset in this scanner data environment by the fact we have a rich way to proxy for a UPC's intrinsic product quality in the form of brand, package, and brand/package interactions. An alternative strategy for controlling the intrinsic product quality is to use UPC fixed effects. However given that we are including all the UPC's in this analysis (indeed one of our main empirical points is to highlight the importance of not selecting out UPC's when doing demand studies on scanner data), UPC fixed effects regression exhibit high degree of instability and sensitivity stemming from colinearity among the UPC dummies and the time varying covariates. Our current strategy appears to control for much of what a UPC fixed effect strategy seems to empirically offer.

which is the standard price instrument in the literature that has studied the DFF data.[36]

7.4 Implementation

We now move to performing inference on demand elasticities using our moment inequality approach. An important goal is to quantify the importance of accounting for the sampling error in market shares that is evident from the sparse demand we observe in the data. There are a few implementation details we briefly discuss.

First, in comparison to the Monte Carlos, we now must profile out many more coefficients besides the constant term because the specification in Section 7.3 includes many more control variables. Indeed, to our knowledge, such a high dimensional model as the one we here consider has not been empirically examined in the moment inequality literature, and our ability to do so is due to the focus on a subset of parameters (namely the price coefficient) that the profiling procedure allows. This expanded set of covariates also requires a larger set $\mathcal{G}_t$ of instrumental variable functions. [37] [38]

Second, because our markets unfold over time there is potential time dependence across different markets. Our profiling inference theory presented in Section 5 was derived under the assumption that different markets are independent. Nevertheless, we can modify the theory in a relatively straightforward way to account for such serial dependence between markets by generalizing our bootstrap procedure to a block bootstrap. We presented this modification in Appendix G, where a formal analogue of Theorem 2 is given for the block bootstrap. We use this block bootstrap procedure (for different choices of the block length) to profile out the nuisance parameter α and employ obtain confidence sets for β_0. Observe that a special case of the block bootstrap with block length of 1 is identical to the independence bootstrap we presented in Section 5.3.

Third, the bootstrapping procedure for both the independence and dependence cases requires the choice of the tuning parameter κ_T, which balances the power and the size. For any $\kappa_T = o(T)$, the asymptotic power of our test increases with κ_T. However, for the

[36] The gross margin is defined as (retail price - wholesale cost)/retail price, so we get wholesale cost using retail price$\times$(1 - gross margin).

[37] We construct the set $\mathcal{G}_T$ as described in (5.10). Our discrete instruments are "brand", "size", "promotion", "holiday", and "year" each taking 11, 9, 2, 2 and 2 values, respectively. Our only continuous instrument is whole sale cost and we use $r_0 = 1$ and $\bar{r}_T = 5$. The $\mathcal{G}_T$ thus constructed potentially contain a total number of $(2 + 4 + ... + 10) \times 11 \times 9 \times 2 \times 2 \times 2 = 23760$ g functions and following (5.12), the weight for a g function indexed by r is $(100 + r)^{-2}(2r \times 11 \times 9 \times 2 \times 2 \times 2)^{-1}$.

[38] Determining the number of hypercubes (g functions) will depend upon the empirical application - too few g functions leads to information loss while too many of them increases sample noise. And we haven't found a general theoretical rule for choosing it. From our own (somewhat limited) Monte Carlo and empirical experience, choosing the number such that, on average, each smallest cube contains 10 to 50 sample points usually "works". In this example, the number of smallest cubes is $10 \times 11 \times 9 \times 2 \times 2 \times 2 = 7920$. But we find most of them contains no sample points and only 401 of them are "nonempty". So, on average, each of the 401 nonempty cubes contains about $11 \approx 4438/401$ sample points.

asymptotic theory to provide a good approximation, $(\kappa_T/T)^{1/2}$ needs to be reasonably small in order to kill a non-estimable (asymptotically Gaussian) term in the bootstrap statistic.[39] We choose $\kappa_T = T/(c \log T)$ because $(\kappa_T/T)^{1/2} = 1/\sqrt{c \log T}$ goes to zero reasonably fast. The shrinking rate $\log T$ is the same as its counterpart suggested in Andrews and Soares (2010) and Andrews and Shi (2013). We choose the constant c by designing a simulation study with a data generating process that matches many key elements of our data and studying the power properties of different c. The details are described in Appendix F. We find there that $c = .5$ and $c = .6$ had the most desirable power properties (all values of c we tried leads confidence intervals that cover the true value with probability greater than the significance level under the DGP's we consider) and these are the values we use for the empirical analysis. Note that we use a block length of 5 for the block bootstrap, which is a commonly accepted level in the literature. Given the evidence shown in Appendix F, this represents a conservative choice.

Finally, we take the minimum bound on choice probabilities ε_{jt} to be the same across j, t and to equal to the smallest observed share in the data. Given how shares are constructed, this is simply one over the maximum customer count observed in the data. This has a simple interpretation as a weak supply side constraint that the retailer makes its marketing decision so that each product is expected to sell at least one unit (on its busiest week).

7.5 Results

The results of our inference is shown in Table 3. As can be seen the tuning parameter $c = .6$ gives a tighter confidence interval than $c = .5$. Our Monte Carlo analysis found that for both choices of the tuning parameter the coverage probability of the confidence interval was respected, suggesting both are valid confidence intervals. Choosing either set of results will not affect our analysis below.

Table 3: 95% Confidence Intervals of Price Coefficient and Average Own Price Elasticity

c	Price Coefficient	Average Own Price Elasticity
0.5	[-5.60, -2.63]	[-10.38, -4.86]
0.6	[-4.14, -3.62]	[-7.67, -6.70]

Note: $\kappa_T = T/(c \cdot \log(T))$ and ε_{jt} =1/maximum customer count $\forall j, t$.

We compare our approach to a BLP estimation of the parameters with the same data. Our Monte Carlo analysis in both Section 6 and Section F suggests that BLP will produce biased estimator that gives the impression of too inelastic demand. This prediction is

[39] see e.g. (E.73) in the proof of Theorem 2(b).

confirmed by the results in Table 4. Notice that the IV estimate leads to an even more inelastic demand than the OLS estimates, which is consistent with the direction of the endogeneity problem in the scanner data as discussed above.

Table 4: Point Estimates and 95% Confidence Intervals of the Logit Models

	IV	OLS
Price Coefficient	-1.50	-2.17
	[-2.25, -0.75]	[-2.55, -1.80]
Average Own Price Elasticity	-2.40	-3.47
	[-3.60, -1.20]	[-4.08, -2.88]

Note: The confidence intervals are constructed using the Driscoll-Kraay standard error (see Driscoll and Kraay (1998)) which is robust to very general forms of cross-sectional as well as temporal dependence.

To better understand our estimates in comparison with the BLP estimates above, we translate our price coefficient from the UPC level demand system into average brand level elasticities (where the average is taken for each brand with respect to all weeks in the data). The results are given in Table 5. This allows us to compare our findings against the brand level elasticities estimated by Hausman and Leonard (2002) (HL for short) using city wide aggregate data from a different source for this industry. Because the HL estimates were formed using aggregate city wide data on brand monthly consumption with a representative agent model of aggregate demand, the elasticities we derive should be at least as large as the HL estimates because: (1) our data reflects store level purchases (hence there can be substitution to other stores), and (2) our data is at a weekly level and hence captures purchase behavior rather than consumption behavior.[40] Observe that the brand elasticities derived under the BLP logit are all considerably less elastic than the HL estimates, which is contrary to the economic intuition. We note that these BLP elasticities at the brand level are similar in magnitude to the brand elasticities derived in other papers for other product categories that start from a UPC level demand system see for example Chintagunta (2000)).

We also note that the BLP logit approach still generate elasticities that are too low if instead of dropping the UPC with zero demand, we form "aggregate" products from the UPC level data, i.e., brands, and estimate a BLP brand level logit (this is exhibited in the last column of the table). Similar biases emerge from other aggregation strategies, consistent with the aggregation bias discussed by Hosken, O'Brien, Scheffman, and Vita (2002).

A key advantage of our approach is the ability to produce consistent inferences directly on the disaggregated data. Our estimates reveal elasticities that are at least as elastic,

[40] See e.g., Hosken, O'Brien, Scheffman, and Vita (2002) for a further discussion of these distinctions.

and often times clearly more elastic than the HL estimates. Our finding of more elastic demand when the error in market shares is fully taken into account has some significant policy implications. A standard "complaint" against logit-type models (including mixed logit models) for demand for differentiated products is that it tends to produce elasticities that are unrealistically inelastic compared to standard intuitions about an industry. Our empirical exercise suggests that error in market shares could be a general source of this problem and that our moment inequality approach offers a practical solution.

Table 5: Own Price Elasticity Comparison

Brand	Our 95% CI 1 ($c = 0.5$)	Our 95% CI 2 ($c = 0.6$)	BLP IV Logit 95% CI	Hausman and Leonard	Brand-Level BLP IV Logit 95% CI
Angel Soft	[-6.56, -3.08]	[-4.85, -4.24]	[-2.23, -1.30]	-4.07	[-1.89, -1.48]
Charmin	[-10.65, -5.00]	[-7.88, -6.89]	[-3.61, -2.11]	-2.29	[-3.21, -2.52]
Cottonelle	[-9.21, -4.32]	[-6.81, -5.95]	[-3.12, -1.83]	-3.29	[-2.65, -2.08]
Kleenex	[-6.26, -2.94]	[-4.63, -4.05]	[-2.12, -1.24]	-3.29	[-1.80, -1.42]
Quilted Northern	[-8.03, -3.77]	[-5.94, -5.19]	[-2.73, -1.59]	-3.08	[-2.31, -1.82]
Scott	[-3.65, -1.71]	[-2.70, -2.36]	[-1.24, -0.72]	-1.80	[-1.05, -0.83]

8 Conclusion

We have shown that when there are errors in market shares in general, the standard conditional mean restriction that BLP exploit as the basis for their empirical strategy lacks any identifying power. When the true underlying choice probabilities can be bounded away from zero, we show that the consumer choice model has enough content to construct a system of moment inequalities that have the property of being *adaptive* to the information revealed by the observed market shares. We also construct a profiling approach to inference with moment inequalities; this allows us to study demand models with a potentially high dimensional parameter vector because the counterfactual implications of such models typically rely on a lower dimensional profile of the parameters, such as the price coefficient or a price elasticity. Our application to scanner data reveals that taking the error in market shares in the data into account has economically important implications for inferences about price elasticities.

A key message from our analysis is that it is critical to not ignore the sampling variability in shares when estimating discrete choice models with disaggregated market data. In many empirical settings, such as airlines (see e.g., Berry, Carnall, and Spiller (1996)), television (see e.g., Goolsbee and Petrin (2004)), and scanner data (Chintagunta, Dube, and Goh (2005)), sampling error in shares is a first order concern since the number of consumers

sampled in each market is not large enough for the sparsity of demand.

A potentially fruitful area for future research is the application of our approach is to 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 sampling variability in market shares when we aggregate over limited sample of households 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 offers a novel solution to address the joint problem of endogenous prices and flexible consumer heterogeneity with micro data that we plan to pursue in future work.

Appendix

In this appendix, we give proofs for the results in the main text and other supporting material. In Section A, we prove Theorem 1 given in Section 3.2. In Section B, we give and prove a lemma that ensures that the correction factors η^l and η^u in Section 4.1 are well-defined. We also describe a way for computing them. Sections C–E are devoted to establishing the formal results in Section 5. Section C collects all the assumptions. Section D proves Lemmas 2 and C.1. Section E proves Theorem 2. Section F discusses the details of our Monte Carlo design. Finally, Section G shows how to handle weak dependence across markets.

A Proof of Theorem 1

Let $int(A)$ denote the interior points of a subset A of Δ_n – the n dimensional unit simplex – and let $br(A)$ denote the boundary points of A and $br(A) = A \backslash int(A)$. Theorem 1 is immediately implied by the following lemma:

Lemma A.1. (a) If $\vec{p}_s^* \in br(co(P_s))$, then F_π is point identified and F_π has discrete support which contains at most $(n+2)/2$ points.

(b) If $\vec{p}_s^* \in int(co(P_s))$, then there exists a sequence $\{F^-_{\pi,1/i}\}_{i=1}^\infty$ such that $\vec{p}_s^* = \vec{p}_s(F^-_{\pi,1/i})$ and $\lim_{i\to\infty} \int [\log x - \log(1-x)] dF^-_{\pi,1/i}(x) = -\infty$ and a sequence $\{F^+_{\pi,1/i}\}_{i=1}^\infty$ such that $\vec{p}_s^* = \vec{p}_s(F^+_{\pi,1/i})$ and $\lim_{i\to\infty} \int [\log x - \log(1-x)] dF^+_{\pi,1/i}(x) = \infty$.

(c) Any point in $br(co(P_s))$ is arbitrarily close to a point in $int(co(P_s))$.

Lemma A.2 is useful for proving Lemma A.1 and its own proof is given at the end of this section. For Lemma A.2, define P_s: $P_s^\zeta = \{\vec{p}_s(F_\pi) : F_\pi(t) = 1\{t \geq x\} \text{ for some } x \in (\zeta, 1-\zeta)\}$ for $\zeta \in [0, 1/2)$.

Lemma A.2. (a) For any $\zeta \in [0, 1/2)$, $int(co(P_s^\zeta)) \neq \emptyset$.

(b) $co(P_s) \subseteq cl(co(P_s^0))$.

(c) $int(co(P_s)) \subseteq \cup_{m=1}^\infty int(co(P_s^{1/m}))$.

(d) For any m, there exists a constant B such that, for any $\vec{p}_s \in co(P_s^{1/m})$, there is a F_π such that $\vec{p}_s = \vec{p}_s(F_\pi)$ and $|\int [\log(\pi) - \log(1-\pi)] dF_\pi(\pi)| \leq B$.

Proof of Lemma A.1. (a) Part (a) is a corollary of Theorem 1(ii) of Wood (1999).

(b) Suppose that $\vec{p}_s^* \in int(co(P_s))$; then there exists a positive integer m^* such that $\vec{p}_s \in int(co(P_s^{1/m^*}))$ by Lemma A.2(c). Then there exists $\epsilon_1 > 0$ small enough such that for any $\vec{p}_s \in \Delta_n$ such that $||\vec{p}_s - \vec{p}_s^*|| \leq \epsilon_1$, we have $\vec{p}_s \in int(co(P_s^{1/m^*}))$.

Let ϵ_2 be a small positive number and $F_\pi^{\epsilon_2}(\pi) = 1\{\pi \geq \epsilon_2\}$ — $F_\pi^{\epsilon_2}$ puts all probability mass on the point ϵ_2. Let $\vec{p}_s^{\epsilon_2} = \vec{p}_s(F_\pi^{\epsilon_2})$ and

$$\vec{p}_s^{\dagger} = (1 + \epsilon_1/\sqrt{4n}) \times \vec{p}_s^{*} - (\epsilon_1/\sqrt{4n}) \times \vec{p}_s^{\epsilon_2}. \tag{A.1}$$

Then $\vec{p}_s^{\dagger} \in int(co(P_s^{1/m^*}))$ because $\|\vec{p}_s^{\dagger} - \vec{p}_s^{*}\| = (\epsilon_1/\sqrt{4n})\|(\vec{p}_s^{*} - \vec{p}_s^{\epsilon_2})\| \leq (\epsilon_1/\sqrt{4n})\|1_{n+1}\| = \epsilon_1\sqrt{n+1}/\sqrt{4n} < \epsilon_1$. By definition, we have

$$\vec{p}_s^{*} = \frac{1}{1 + \epsilon_1/\sqrt{4n}}\vec{p}_s^{\dagger} + \frac{\epsilon_1/\sqrt{4n}}{1 + \epsilon_1/\sqrt{4n}}\vec{p}_s^{\epsilon_2}. \tag{A.2}$$

Because $\vec{p}_s^{\dagger} \in int(co(P_s^{1/m^*}))$, there exists $F_\pi^{\dagger}$ such that $\vec{p}_s^{\dagger} = \vec{p}_s(F_\pi^{\dagger})$ and $|\int[\log(\pi) - \log(1-\pi)]dF_\pi^{\dagger}(\pi)| < B$ by Lemma A.2(d). Let

$$F_{\pi,\epsilon_2}^{-}(\pi) = \frac{1}{1 + \epsilon_1/\sqrt{4n}}F_\pi^{\dagger}(\pi) + \frac{\epsilon_1/\sqrt{4n}}{1 + \epsilon_1/\sqrt{4n}}F_\pi^{\epsilon_2}(\pi).$$

Then clearly, $\vec{p}_s^{*} = \vec{p}_s(F_{\pi,\epsilon_2}^{-})$ and

$$\int[\log(\pi) - \log(1-\pi)]dF_{\pi,\epsilon_2}^{-}(\pi) = \frac{1}{1 + \epsilon_1/\sqrt{4n}}\int[\log(\pi) - \log(1-\pi)]dF_\pi^{\dagger}(\pi) + \frac{\epsilon_1/\sqrt{4n}}{1 + \epsilon_1/\sqrt{4n}}\int[\log(\pi) - \log(1-\pi)]dF_\pi^{\epsilon_2}(\pi).$$

We know that $\frac{1}{1+\epsilon_1/\sqrt{4n}}|\int[\log(\pi) - \log(1-\pi)]dF_\pi^{\dagger}(\pi)| < \frac{B}{1+\epsilon_1/\sqrt{4n}}$ and it does not depend on ϵ_2. Also, $\lim_{\epsilon_2 \downarrow 0}\int[\log(\pi) - \log(1-\pi)]dF_\pi^{\epsilon_2}(\pi) = \lim_{\epsilon_2 \downarrow 0} -\log((1-\epsilon_2)/\epsilon_2) = -\infty$. Thus,

$$\lim_{\epsilon_2 \downarrow 0}\int[\log(\pi) - \log(1-\pi)]dF_{\pi,\epsilon_2}^{-}(\pi) = -\infty. \tag{A.3}$$

Similarly, we can find $F_{\pi,\epsilon_2}^{+}(\pi)$ that is consistent with $\vec{p}_s^{*}$ and

$$\lim_{\epsilon_2 \downarrow 0}\int[\log(\pi) - \log(1-\pi)]dF_{\pi,\epsilon_2}^{+}(\pi) = \infty. \tag{A.4}$$

Thus part (b) is proved.

(c) By Lemma A.2(a), $int(co(P_s)) \neq \emptyset$. This combined with the convexity of $co(P_s)$ implies that $co(P_s) \subseteq cl(int(co(P_s)))$. Thus, part (c) is proved. □

Proof of Lemma A.2. (a) To show that $int(co(P_s^{\zeta})) \neq \emptyset$, it suffices to show that the dimension of $co(P_s^{\zeta})$ is n. To show the later, it suffices to find n independent vectors in P_s^{ζ}. Let

$x_1, ..., x_n$ be n distinct points in $(\zeta, 1-\zeta)$. For $j = 1, ..., n$, define the vector $\vec{p}_s^{\,j} = (\vec{p}_s^{\,j}(j'))_{j'=0}^{n}$, where

$$p_s^j(j') = \left(\begin{array}{c} n \\ j' \end{array} \right) (x_j)^{j'} (1 - x_j)^{n-j'}. \tag{A.5}$$

The matrix formed by the n vectors are:

$$\left(\begin{array}{cccc} \left(\begin{array}{c} n \\ 0 \end{array} \right) (x_1)^0 (1-x_1)^n & \left(\begin{array}{c} n \\ 1 \end{array} \right) (x_1)^1 (1-x_1)^{n-1} & ... & \left(\begin{array}{c} n \\ n \end{array} \right) (x_1)^n (1-x_1)^0 \\ \left(\begin{array}{c} n \\ 0 \end{array} \right) (x_2)^0 (1-x_2)^n & \left(\begin{array}{c} n \\ 1 \end{array} \right) (x_2)^1 (1-x_2)^{n-1} & ... & \left(\begin{array}{c} n \\ n \end{array} \right) (x_2)^n (1-x_2)^0 \\ ... & ... & ... & ... \\ \left(\begin{array}{c} n \\ 0 \end{array} \right) (x_n)^0 (1-x_n)^n & \left(\begin{array}{c} n \\ 1 \end{array} \right) (x_n)^1 (1-x_n)^{n-1} & ... & \left(\begin{array}{c} n \\ n \end{array} \right) (x_n)^n (1-x_n)^0 \end{array} \right).$$

The matrix has same rank as

$$\left(\begin{array}{cccc} ((1-x_1)/x_1)^n & ((1-x_1)/x_1)^{n-1} & ... & 1 \\ ((1-x_2)/x_2)^n & ((1-x_2)/x_2)^{n-1} & ... & 1 \\ ... & ... & ... & ... \\ ((1-x_n)/x_n)^n & ((1-x_n)/x_n)^{n-1} & ... & 1 \end{array} \right). \tag{A.6}$$

We know that the matrix in (A.6) has full rank by the property of polynomial sequences. Therefore, the matrix formed by the vectors $\vec{p}_s^{\,1}, ..., \vec{p}_s^{\,n}$ has rank n, implying that the vectors are independent. Thus, part (a) is proved.

(b) Consider an arbitrary point in $\vec{p}_s \in co(P_s)$. Then there exists F_π such that

$$\vec{p}_s = \vec{p}_s(F_\pi).$$

Let $F_{\pi,\epsilon}(x)$ be defined as:

$$F_{\pi,\epsilon}(x) = F_\pi((x - \epsilon)/(1 - 2\epsilon)).$$

Let

$$\vec{p}_{s,\epsilon} = \vec{p}_s(F_{\pi,\epsilon})$$

Then $\vec{p}_{s,\epsilon} \in co(P_s^0)$ because $F_{\pi,\epsilon}(x)$'s support is a subset of $(0, 1)$. However, because $\lim_{\epsilon \downarrow 0} F_{\pi,\epsilon}(x) = F_\pi(x)$ for any x that is a continuity point of $F_\pi(x)$, we have

$$\lim_{\epsilon \downarrow 0} \vec{p}_{s,\epsilon} = \vec{p}_s.$$

Thus, $\vec{p}_s \in cl(co(P_s^0))$. Because $\vec{p}_s$ is an arbitrary point in $co(P_s)$, this implies that $co(P_s) \subseteq cl(co(P_s^0))$.

(c) Given part (b), in order to show part (c), it suffices to show

$$co(P_s^0) \subseteq cl\left(\cup_{m=1}^{\infty} int(co(P_s^{1/m}))\right), \tag{A.7}$$

because $\cup_{m=1}^{\infty} int(co(P_s^{1/m}))$ is an open set and interior of the closure of an open set is the open set itself. To show (A.7), it suffices to show

$$co(P_s^0) \subseteq \cup_{m=1}^{\infty} co(P_s^{1/m}) \text{ and} \tag{A.8}$$

$$\cup_{m=1}^{\infty} co(P_s^{1/m}) \subseteq cl\left(\cup_{m=1}^{\infty} int(co(P_s^{1/m}))\right). \tag{A.9}$$

Next, we show (A.8) and (A.9).

Consider an arbitrary point $\vec{p_s} \in co(P_s^0)$. By the Carathéodory's theorem for convex hull (see, for example, (Rockafellar (1970, p.155)), there exists $n+1$ points in P_s^0 such that $\vec{p}_s$ is a mixture of these $n+1$ points. In other words, there exists F_π^* that has at most $n+1$ support points in $(0,1)$ such that $\vec{p}_s = \vec{p}_s(F_\pi^*)$. Let ζ_1^* be the minimum of the $n+1$ support points of F_π^*, let ζ_2^* be one minus the maximum of the $n+1$ support points of F_π^* and let $\zeta^* = \min\{\zeta_1^*, \zeta_2^*\}/2$. Then $\zeta^* > 0$. This implies that $\vec{p}_s$ is also a mixture of $n+1$ points in $P_s^{\zeta^*}$. Or in other words:

$$\vec{p}_s \in co(P_s^{\zeta^*}). \tag{A.10}$$

Then, $\vec{p}_s \in co(P_s^{1/m})$ for all $m > 1/\zeta^*$. Thus, $\vec{p}_s \in \cup_{m=1}^{\infty} co(P_s^{1/m})$. This shows (A.8).

Consider an arbitrary point $\vec{p}_s \in \cup_{m=1}^{\infty} co(P_s^{1/m})$. Then there exists m^* such that $\vec{p}_s \in co(P_s^{1/m^*})$. Because $co(P_s^{1/m^*})$ is convex by definition and has nonempty interior by part (a), every point in $co(P_s^{1/m^*})$ is the limit of a sequence of points in $int(co(P_s^{1/m^*}))$. Thus, $\vec{p}_s$ is the limit of a sequence of points in $\cup_{m=1}^{\infty} int(co(P_s^{1/m}))$. This shows that $\vec{p}_s = cl\left(\cup_{m=1}^{\infty} int(co(P_s^{1/m}))\right)$ and (A.9) is proved.

(d) For any $\vec{p}_s \in co(P_s^{1/m})$, there exists F_π with support on $[1/m, 1-1/m]$ such that $\vec{p}_s = \vec{p}_s(F_\pi)$. Therefore,

$$\int [\log(x) - \log(1-x)] dF_\pi(x) \leq \sup_{x \in Supp(F_\pi)} [\log(x) - \log(1-x)] = \log(m-1)$$

and

$$\int [\log(x) - \log(1-x)] dF_\pi(x) \geq \inf_{x \in Supp(F_\pi)} [\log(x) - \log(1-x)] = -\log(m-1).$$

Thus, part (d) holds with $B = \log(m-1)$. □

B Existence and Example of the Implicit Function $\eta_j(n_t, \pi_t, x_t; \lambda)$

Lemma B.1. The function $f(\eta) := E\left[\sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right) | n_t, \pi_t, x_t, J_t\right]$ is continuous and strictly increasing in η. Furthermore, $f(\eta) \to -\infty$ as $\eta \to -1/(n_t + J_t + 1)$ and $f(\eta) \to \infty$ as $\eta \to 1/(n_t + J_t + 1)$.

Proof. Recall that $\tilde{s}_t = \frac{n_t s_t + 1}{n_t + J_t + 1}$. Consider any $j \in \{1, \ldots, J_t\}$ and any given realization of $\tilde{s}_t$. Observe that $\tilde{s}_t + \eta' \cdot e_j \geq \tilde{s}_t + \eta \cdot e_j$ for $\eta' > \eta$. Thus using the fact $\sigma^{-1}(\cdot, x_t; \lambda)$ is an inverse isotone mapping as shown by Theorem 1 in Berry, Gandhi, and Haile (2011), we have that $\sigma_j^{-1}\left(\tilde{s}_t + \eta' \cdot e_j, x_t; \lambda\right) \geq \sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right)$. Also because the j^{th} element of $\tilde{s}_t + \eta' \cdot e_j$ is strictly greater than the j^{th} element of $\tilde{s}_t + \eta \cdot e_j$ and all the other elements of these two vectors are equal, which combined with the fact $\sigma^{-1}\left(\tilde{s}_t + \eta' \cdot e_j, x_t; \lambda\right) \neq \sigma^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right)$ (because inverse isotone implies σ is invertible), implies via Lemma 2 in Berry, Gandhi, and Haile (2011) that $\sigma_j^{-1}\left(\tilde{s}_t + \eta' \cdot e_j, x_t; \lambda\right) > \sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right)$. Because this holds for all possible realizations of $\tilde{s}_t$, strict monotonicity also hold for the expectation taken with respect to realizations of $\tilde{s}_t$, i.e., $f(\eta)$ is strictly monotone in η.

Observe finally that as $\eta \to -1/(n_t + J_t + 1)$ the share of good j in the vector $\tilde{s}_t + \eta \cdot e_j$ approaches 0 for a realization of s_t such that $s_{jt} = 0$, and thus $\sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right)$ must approach $-\infty$ for this realization of s_t as a consequence of the full support assumption on v_{ijt}. However, because all other realizations of s_t are such that $\sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right)$ is decreasing as established above, then the expectation taken with respect to realizations of s_t approaches $-\infty$. A similar argument can be made for $\eta \to 1/(n_t + J_t + 1)$ based on the recognition that the share of good 0 in the share vector $\tilde{s}_t + \eta \cdot e_j$ is approaching zero at the realization of s_t such that $s_{jt} = 1$. □

Example: Logit Demand

For each j and t, η_{jt}^u can be computed by solving the following constraint optimization problem:

$$\max_{\vec{\pi}_t : \pi_t \in \Delta_{J_t}^{\varepsilon_t}, \eta \in [-1/(n_t+J_t+1), 1/(n_t+J_t+1)]} \eta$$
$$s.t.\ E\left[\sigma_j^{-1}\left(\frac{n_t s_t + 1}{n_t + J_t + 1} + \eta \cdot e_j, x_t; \lambda_0\right) | n_t, \pi_t, x_t, z_t, J_t\right] = \sigma_j^{-1}\left(\pi_t, x_t; \lambda_0\right), \qquad \text{(B.1)}$$

where s_t satisfies $n_t \vec{s}_t | n_t, \pi_t, x_t, z_t, J_t \sim MN(n_t, \vec{\pi}_t)$. Similarly, η_{jt}^l can be computed by solving the same optimization problem but with max replaced by min.

In the case of simple logit model: $\sigma_j^{-1}(\pi_t, x_t, \lambda_0) = \log(\pi_{jt}/\pi_{0t})$, where $\pi_{0t} = 1 - \pi_t' 1_{J_t}$.

Then, the constraint in the above problem is simplified to

$$E\left[\log\left(\frac{n_t s_{jt}+1+(n_t+J_t+1)\eta}{n_t s_{0t}+1-(n_t+J_t+1)\eta}\right)|\pi_t, n_t, J_t\right] = \log\left(\frac{\pi_{jt}}{\pi_{0t}}\right), \tag{B.2}$$

where $n_t(s_{jt}, s_{0t}, 1-s_{jt}-s_{0t})|n_t, \pi_t, J_t \sim MN(n_t, (\pi_{jt}, \pi_{0t}, 1-\pi_{jt}-\pi_{0t}))$. This is a major simplification numerically because (1) the constraint in the above optimization problem only depends on the three dimensional parameter $(\pi_{jt}, \pi_{0t}, \eta)$ regardless of J_t and thus the dimension of the optimization problem does not increase with J_t; and (2) $\eta_{jt}^u = -\eta_{jt}^l$ because π_{jt} and π_{0t} appear symmetrically in the equation and thus there is no need to solve both the max and the min problems. The numerical simplicity of the simple logit model easily extends to nested logit models.

C Assumptions for Profiling Inference

In this section, we list all the technical assumptions required for the profiling approach. The assumptions are grouped into seven categories. Assumption C.1 restricts the space of θ; Assumption C.2 restricts the space of (γ, F), i.e. the parameters that determines the true data generating process. Assumption C.3 further restricts the space of (γ, F) to satisfy the null hypothesis $\gamma \in \Gamma_0$. Assumption C.4 is the full support condition on the measure μ on $\mathcal{G}$. Assumption C.5 regulates how $\mathcal{G}_T$ approaches $\mathcal{G}$ as T increases. Assumption C.6 restricts the function $S(m, \Sigma)$ to satisfy certain continuity, monotonicity and convexity conditions. Assumption C.7 regulates the subsample size b_T and the moment shrinking parameter κ_T in the bootstrap procedure. Throughout, we let E^* and E_* denote outer and inner expectations respectively and $\Pr^*$ and $\Pr_*$ denote outer and inner probabilities.

Assumption C.1. (a) Θ is compact, (b) Γ is upper hemi-continuous, and (c) $\Gamma^{-1}(\gamma)$ is either convex or empty for any $\gamma \in R^{d_\gamma}$.

To introduce Assumption C.2 we need the following extra notation. Let $\nu_F(\theta, g) : (\theta, g) \in \Theta \times \mathcal{G}$ denote a tight Gaussian process with covariance kernel

$$\Sigma_F(\theta^{(1)}, g^{(1)}, \theta^{(2)}, g^{(2)}) = Cov_F\left(\rho(w_t, \theta^{(1)}, g^{(1)}), \rho(w_t, \theta^{(2)}, g^{(2)})\right). \tag{C.1}$$

Notice that $\Sigma_F(\theta, g) = \Sigma_F(\theta, g, \theta, g)$.

Let the derivative of $\rho_F(\theta, g)$ with respect to θ be $G_F(\theta, g)$.

For any $\gamma \in R^{d_\gamma}$, let the set $\Theta_{0,F}(\gamma)$ be

$$\Theta_{0,F}(\gamma) = \{\theta \in \Theta : Q_F(\theta) = 0 \ \& \ \Gamma(\theta) \ni \gamma\}, \tag{C.2}$$

We call $\Theta_{0,F}(\gamma)$ the zero-set of $Q_F(\theta)$ under (γ, F). Note that for any $\gamma \in R^{d_\gamma}$, $\gamma \in \Gamma_{0,F}$ if and only if $\Theta_{0,F}(\gamma) \neq \emptyset$.

Let the distance from a point to a set be the usual mapping:

$$d(a, A) = \inf_{a^* \in A} \|a - a^*\|, \tag{C.3}$$

where $\|\cdot\|$ is the Euclidean distance.

Let $\mathcal{F}$ denote the set of all probability measures on $(w_t)_{t=1}^T$. Let $\bar{\mathcal{G}} = \mathcal{G} \cup \{1\}$. Let $\mathcal{M}$ denote the set of all positive semi-definite $k \times k$ matrices. The following assumption defines the parameter space $\mathcal{H}$ for the pair (γ, F).

Assumption C.2. *The parameter space $\mathcal{H}$ of the pairs (γ, F) is a subset of $R^{d_\gamma} \times \mathcal{F}$ that satisfies:*

(a) *under every F such that $(\gamma, F) \in \mathcal{H}$ for some $\gamma \in R^{d_\gamma}$, the markets are independent and ex ante identical to each other, i.e. $\{\rho(w_t, \theta, g)\}_{t=1}^T$ is an i.i.d. sample for any θ, g;*

(b) $\lim_{M\to\infty} \sup_{(\gamma,F)\in\mathcal{H}} E_F^*[\sup_{(\theta,g)\in\Gamma^{-1}(\gamma)\times\bar{\mathcal{G}}} ||\rho(w_t, \theta, g)||^2 1\{||\rho(w_t, \theta, g)||^2 > M\}] = 0$*;*

(c) *the class of functions $\{\rho(w_t, \theta, g) : (\theta, g) \in \Gamma^{-1}(\gamma) \times \bar{\mathcal{G}}\}$ is F-Donsker and pre-Gaussian uniformly over $\mathcal{H}$;*

(d) *the class of functions $\{\rho(w_t, \theta, g)\rho(w_t, \theta, g)' : (\theta, g) \in \Gamma^{-1}(\gamma) \times \bar{\mathcal{G}}\}$ is Glivenko-Cantelli uniformly over $\mathcal{H}$;*

(e) *$\rho_F(\theta, g)$ is differentiable with respect to $\theta \in \Theta$, and there exists constants C and $\delta_1 > 0$ such that, for any $(\theta^{(1)}, \theta^{(2)})$, $\sup_{(\gamma,F)\in\mathcal{H}, g\in\bar{\mathcal{G}}} ||vec(G_F(\theta^{(1)}, g)) - vec(G_F(\theta^{(2)}, g))|| \leq C \times ||\theta^{(1)} - \theta^{(2)}||^{\delta_1}$, and*

(f) *$\Sigma_F^\iota(\theta, g) \in \Psi$ for all $(\gamma, F) \in \mathcal{H}$ and $\theta \in \Gamma^{-1}(\gamma)$ where Ψ is a compact subset of $\mathcal{M}$, and $\{vec(\Sigma_F(\cdot, g^{(1)}, \cdot, g^{(2)})) : (\Gamma^{-1}(\gamma))^2 \to R^{k^2} : (\gamma, F) \in \mathcal{H}, g^{(1)}, g^{(2)} \in \bar{\mathcal{G}}\}$ are uniformly bounded and uniformly equicontinuous.*

Remark. Part (a) is the i.i.d. assumption, which can be replaced with appropriate weak dependence conditions at the cost of more complicated derivation in the uniform weak convergence of the bootstrap empirical process. Part (b) is standard uniform Lindeberg condition. Part (c)-(d) imposes restrictions on the complexity of the set $\mathcal{G}$ as well as on the shape of $\rho(w_t, \theta, g)$ as a function of θ. A sufficient condition is (i) $\rho(w_t, \theta, g)$ is Lipschitz continuous in θ with the Lipschitz coefficient being integrable and (ii) the set $\mathcal{C}$ in the definition of $\mathcal{G}$ forms a Vapnik-Červonenkis set and J_t is bounded. The Lipschitz continuity is also a sufficient condition of part (f).

The following assumptions defines the null parameter space, $\mathcal{H}_0$, for the pair (γ, F).

Assumption C.3. *The null parameter space $\mathcal{H}_0$ is a subset of $\mathcal{H}$ that satisfies:*

(a) *for every $(\gamma, F) \in \mathcal{H}_0$, $\gamma \in \Gamma_{0,F}$, and*

(b) *there exists* $C,\ c > 0$ *and* $2 \leq \delta_2 < 2(\delta_1+1)$ *such that* $Q_F(\theta) \geq C \cdot (d(\theta, \Theta_{0,F}(\gamma))^{\delta_2} \wedge c)$ *for all* $(\gamma, F) \in \mathcal{H}_0$ *and* $\theta \in \Gamma^{-1}(\gamma)$.

Remark. Part (b) is an identification strength assumption. It requires the criterion function to increase at certain minimum rate as θ is perturbed away from the identified set. This assumption is weaker than the quadratic minorant assumption in Chernozhukov, Hong, and Tamer (2007) if $\delta_2 > 2$ and as strong as the latter if $\delta_2 = 2$. Putting part (b) and Assumption C.2(e) together, we can see that there is a trade-off between the minimum identification strength required and the degree of Hölder continuity of the first derivative of $\rho_F(\cdot, g)$. If $\rho_F(\cdot, g)$ is linear, δ_2 can be arbitrarily large – the criterion function can increase very slowly as θ is perturbed away from the identified set.

The following assumption is on the measure μ. For any θ, let a pseudo-metric on $\mathcal{G}$ be: $||g^{(1)} - g^{(2)}||_{\theta,F} = ||\rho_{F,j}(\theta, g^{(1)}) - \rho_{F,j}(\theta, g^{(2)})||$. This assumption is needed for Lemma 2 and not needed for the asymptotic size result Theorem 2.

Assumption C.4. *For any* $\theta \in \Theta$, $\mu(\cdot)$ *has full support on the metric space* $(\mathcal{G}, ||\cdot||_{\theta,F})$.

Remark. Assumption C.4 implies that for any $\theta \in \Theta$, F and j, if $\rho_{F,j}(\theta, g_0) < 0$ for some $g_0 \in \mathcal{G}$, then there exists a neighborhood $\mathcal{N}(g_0)$ with positive μ-measure such that $\rho_{F,j}(\theta, g) < 0$ for all $g \in \mathcal{N}(g_0)$.

The following assumption is on the set $\mathcal{G}_T$.

Assumption C.5. (a) $\mathcal{G}_T \uparrow \mathcal{G}$ *as* $T \to \infty$ and

(b)$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \sup_{\theta\in\Gamma^{-1}(\gamma)} \int_{\mathcal{G}\setminus\mathcal{G}_T} S(\sqrt{T}\rho_F(\theta, g), \Sigma_F(\theta, g)) d\mu(g) = 0.$

The following assumptions are imposed on the function S. For a $\xi > 0$, let the ξ-expansion of Ψ be $\Psi^\xi = \{\Sigma \in \mathcal{M} : \inf_{\Sigma_1\in\Psi} ||vech(\Sigma) - vech(\Sigma_1)|| \leq \xi\}$.

Assumption C.6. (a) $S(m, \Sigma) : (-\infty, \infty]^k \times \Psi^\xi \to R$ *is continuous for some* $\xi > 0$.

(b) *There exists a constant* $C > 0$ *and* $\xi > 0$ *such that for any* $m_1, m_2 \in R^k$ *and* $\Sigma_1, \Sigma_2 \in \Psi^\xi$, *we have* $|S(m_1, \Sigma_1) - S(m_2, \Sigma_2)| \leq C\sqrt{(S(m_1, \Sigma_1) + S(m_2, \Sigma_2))(S(m_2, \Sigma_2) + 1)\Delta}$, *where* $\Delta = ||m_1 - m_2||^2 + ||vech(\Sigma_1 - \Sigma_2)||$.

(c) S *is non-increasing in* m.

(d) $S(m, \Sigma) \geq 0$ *and* $S(m, \Sigma) = 0$ *if and only if* $m \in [0, \infty]^k$.

(e) S *is homogeneous in* m *of degree* 2.

(f) S *is convex in* $m \in R^{d_m}$ *for any* $\Sigma \in \Psi^\xi$.

Remark. We show in the lemma below that Assumption C.6 is satisfied by the example in (5.11) (which is used in our empirical section) as well as the SUM and MAX functions in

Andrews and Shi (2013):

$$\begin{aligned} \text{SUM: } S(m,\Sigma) &= \sum_{j=1}^{k} [m_j/\sigma_j]_-^2, \text{ and} \\ \text{MAX: } S(m,\Sigma) &= \max_{1\le j\le k} [m_j/\sigma_j]_-^2, \end{aligned} \tag{C.4}$$

where σ_j^2 is the jth diagonal element of Σ. Assumptions C.6(b) and (f) rule out the QLR function in Andrews and Shi (2013): $S(m,\Sigma) = \min_{t\ge 0}(m-t)'\Sigma^{-1}(m-t)$.

Lemma C.1. (a) *Assumption* C.6 *is satisfied by the S function in* (5.11) *for any set Ψ.*

(b) *Assumption* C.6 *is satisfied by the* SUM *and the* MAX *functions in* (C.4) *if Ψ is a compact subset of the set of positive semi-definite matrix with diagonal elements bounded below by some constant $\xi_2 > 0$.*

The following assumptions are imposed on the tuning parameters in the subsampling and the bootstrap procedures.

Assumption C.7. (a) *In the subsampling procedure, $b_T^{-1} + b_T T^{-1} \to 0$ and $S_T \to \infty$, and*

(b)*In the bootstrap procedure, $\kappa_T^{-1} + \kappa_T T^{-1} \to 0$ and $S_T \to \infty$.*

D Proof of Lemmas 2 and C.1

Proof of Lemma 2. (a) Assumptions C.2(c)-(d) imply that under F,

$$\begin{aligned} \Delta_{\rho,T} &\equiv \sup_{\theta\in\Gamma^{-1}(\gamma),g\in\bar{\mathcal{G}}} ||\bar{\rho}_T(\theta,g) - \rho_F(\theta,g)|| \to_p 0, \text{ and} \\ &\sup_{\theta\in\Gamma^{-1}(\gamma),g\in\bar{\mathcal{G}}} ||vech(\hat{\Sigma}_T(\theta,g) - \Sigma_F(\theta,g))|| \to_p 0. \end{aligned} \tag{D.1}$$

The second convergence implies that

$$\Delta_{\Sigma,T} \equiv \sup_{\theta\in\Gamma^{-1}(\gamma),g\in\mathcal{G}} ||vech(\hat{\Sigma}_T^{\iota}(\theta,g) - \Sigma_F^{\iota}(\theta,g))|| \to_p 0. \tag{D.2}$$

By Assumption C.2(b), $\sup_{\theta\in\Gamma^{-1}(\gamma),g\in\mathcal{G}} ||\rho_F(\theta,g)|| < M^*$ for some $M^* < \infty$. Thus, $\{(\rho_F(\theta,g),\Sigma_F^{\iota}(\theta,g)) : (\theta,g)\in\Gamma^{-1}(\gamma)\times\mathcal{G}\}$ is a subset of the compact set $[-M^*,M^*]^k\times\Psi$. By Assumption C.2(f) and Equations (D.1) and (D.2), we have $\{(\bar{\rho}_T(\theta,g),\hat{\Sigma}_T^{\iota}(\theta,g)) : (\theta,g)\in\Gamma^{-1}(\gamma)\times\mathcal{G}\} \subseteq [-M^*-\xi,M^*+\xi]^k\times\Psi^{\xi}$ with probability approaching one for any $\xi > 0$. By Assumption C.6(a), $S(m,\Sigma)$ is uniformly continuous on $[-M^*,M^*]^k\times\Psi$. Therefore,

for any $\epsilon > 0$,

$$\begin{aligned}
&\Pr_F\left(\left|\min_{\theta\in\Gamma^{-1}(\gamma)}\hat{Q}_T(\theta) - \min_{\theta\in\Gamma^{-1}(\gamma)}\int_{\mathcal{G}_T} S(\rho_F(\theta,g),\Sigma_F^\iota(\theta,g))d\mu(g)\right| > \epsilon\right)\\
\leq &\Pr_F\left(\sup_{\theta\in\Gamma^{-1}(\gamma),g\in\mathcal{G}}|S(\bar{\rho}_t(\theta,g),\hat{\Sigma}_t^\iota(\theta,g)) - S(\rho_F(\theta,g),\Sigma_F^\iota(\theta,g))| > \epsilon\right)\\
\to &0. \qquad \text{(D.3)}
\end{aligned}$$

Now it is left to show that $\min_{\theta\in\Gamma^{-1}(\gamma)}\int_{\mathcal{G}_T} S(\rho_F(\theta,g),\Sigma_F^\iota(\theta,g))d\mu(g) \to \min_{\theta\in\Gamma^{-1}(\gamma)} Q_F(\theta)$ as $T\to\infty$. Observe that

$$\begin{aligned}
0 \leq &\min_{\theta\in\Gamma^{-1}(\gamma)} Q_F(\theta) - \min_{\theta\in\Gamma^{-1}(\gamma)}\int_{\mathcal{G}_T} S(\rho_F(\theta,g),\Sigma_F^\iota(\theta,g))d\mu(g)\\
\leq &\sup_{\theta\in\Gamma^{-1}(\gamma)}\int_{\mathcal{G}/\mathcal{G}_T} S(\rho_F(\theta,g),\Sigma_F^\iota(\theta,g))d\mu(g)\\
\leq &\int_{\mathcal{G}/\mathcal{G}_T}\sup_{\theta\in\Gamma^{-1}(\gamma)} S(\rho_F(\theta,g),\Sigma_F^\iota(\theta,g))d\mu(g). \qquad \text{(D.4)}
\end{aligned}$$

We have $\sup_{\theta\in\Gamma^{-1}(\gamma)} S(\rho_F(\theta,g),\Sigma_F^\iota(\theta,g)) < \infty$, because $\rho_F(\theta,g)\in[-M^*,M^*]^k$ and $\Sigma_F^\iota(\theta,g)\in\Psi$ and Assumption C.6(a). Thus the last line of (D.4) converges to zero under Assumption C.5(a). This and (D.3) together show part (a).

(b) The first half of part (b), $\min_{\theta\in\Gamma^{-1}(\gamma)} Q_F(\theta)\geq 0$, is implied by Assumption C.6(d).

Suppose $\gamma\in\Gamma_{0,F}$. Then there exists a $\theta^*\in\Gamma^{-1}(\gamma)$ such that $\rho_F(\theta^*,g)\geq 0$ for all $g\in\mathcal{G}$ by Lemma 1. This implies that $S(\rho_F(\theta^*,g),\Sigma_F(\theta^*,g)) = 0$ for all $g\in\mathcal{G}$ by Assumption C.6(d). Thus, $Q_F(\theta^*) = 0$. Because $\min_{\theta\in\Gamma^{-1}(\gamma)} Q_F(\theta)\leq Q_F(\theta^*) = 0$, this shows the "if" part of the second half.

Suppose that $\min_{\theta\in\Gamma^{-1}(\gamma)} Q_F(\theta) = 0$. By Assumptions C.1(a)-(b), $\Gamma^{-1}(\gamma)$ is compact. By Assumptions C.2(e) and (f), $Q_F(\theta)$ is continuous in θ. Thus, there exists a $\theta^*\in\Gamma^{-1}(\gamma)$ such that $Q_F(\theta^*) = \min_{\theta\in\Gamma^{-1}(\gamma)} Q_F(\theta) = 0$. We show by contradiction that this implies $\gamma\in\Gamma_{0,F}$. Suppose that $\gamma\notin\Gamma_{0,F}$. Then for any $\theta\in\Gamma^{-1}(\gamma)$, in particular, for θ^*, $\rho_{F,j}(\theta^*,g^*) < 0$ for some $g^*\in\mathcal{G}$ and some $j\leq d_m$ by Lemma 1. Then by Assumption C.4, there exists a neighborhood $\mathcal{N}(g^*)$ with positive μ-measure, such that $\rho_{F,j}(\theta^*,g) < 0$ for all $g\in\mathcal{N}(g^*)$. This implies that $Q_F(\theta^*) > 0$, which contradicts $Q_F(\theta^*) = 0$. Thus, the "only if" part is proved. □

Proof of Lemma C.1. We prove part (b) only. Part (a) follows from the arguments for part (b) because the S function in part (a) is the same as the SUM S function with $\Sigma = I$. Let ξ be any positive number less than ξ_2. Then the diagonal elements of all matrices in Ψ^ξ are

bounded below by $\xi_2 - \xi$.

We prove the SUM part first. Assumptions C.6(a), (c)-(f) are immediate. It suffices to verify Assumptions C.6(b). To verify Assumption C.6(b), observe that

$$\begin{aligned}
|S(m_1,\Sigma_1) - S(m_2,\Sigma_2)| &= \left|\sum_{j=1}^{k}([m_{1,j}/\sigma_{1,j}]_- - [m_{2,j}/\sigma_{2,j}]_-)([m_{1,j}/\sigma_{1,j}]_- + [m_{2,j}/\sigma_{2,j}]_-)\right| \\
&\leq \left\{2\sum_{j=1}^{k}([m_{1,j}/\sigma_{1,j}]_- - [m_{2,j}/\sigma_{2,j}]_-)^2(S(m_1,\Sigma_1) + S(m_2,\Sigma_2))\right\}^{1/2} \\
&\equiv \{2A(S(m_1,\Sigma_1) + S(m_2,\Sigma_2))\}^{1/2}, \qquad \text{(D.5)}
\end{aligned}$$

where the inequality holds by the Cauchy-Schwartz inequality and the inequality $(a + b)^2 \leq 2(a^2 + b^2)$, and the $\equiv$ holds with $A := \sum_{j=1}^{k}([m_{1,j}/\sigma_{1,j}]_- - [m_{2,j}/\sigma_{2,j}]_-)^2$. Now we manipulate A in the following way:

$$\begin{aligned}
A &= \sum_{j=1}^{k}([m_{1,j}/\sigma_{1,j}]_- - [m_{2,j}/\sigma_{1,j}]_- + [m_{2,j}/\sigma_{1,j}]_- - [m_{2,j}/\sigma_{2,j}]_-)^2 \\
&\leq 2\sum_{j=1}^{k}([m_{1,j}/\sigma_{1,j}]_- - [m_{2,j}/\sigma_{1,j}]_-)^2 + 2\sum_{j=1}^{k}([m_{2,j}/\sigma_{1,j}]_- - [m_{2,j}/\sigma_{2,j}]_-)^2 \\
&= 2\sum_{j=1}^{k}([m_{1,j}/\sigma_{1,j}]_- - [m_{2,j}/\sigma_{1,j}]_-)^2 + 2\sum_{j=1}^{k}(\sigma_{2,j} - \sigma_{1,j})^2[m_{2,j}/\sigma_{2,j}]_-^2/\sigma_{1,j}^2 \\
&\leq 2\|m_1 - m_2\|^2/(\xi_2 - \xi) + 2\{\|vech(\Sigma_1 - \Sigma_2)\|/(\xi_2 - \xi)\}S(m_2,\Sigma_2) \\
&\leq 2(\xi_2 - \xi)^{-1}(S(m_2,\Sigma_2) + 1)(||m_1 - m_2||^2 + ||vech(\Sigma_1 - \Sigma_2)||), \qquad \text{(D.6)}
\end{aligned}$$

where the first inequality holds by the inequality $(a + b)^2 \leq 2(a^2 + b^2)$ and the second inequality holds because $(\sigma_{2,j} - \sigma_{1,j})^2 \leq |\sigma_{2,j}^2 - \sigma_{1,j}^2| \leq ||vech(\Sigma_1 - \Sigma_2)||$ and because $\sigma_{1,j}^2, \sigma_{2,j}^2 \geq \xi_2 - \xi$. Plug (D.6) in (D.5), and we obtain Assumptions C.6(b).

The proof for the MAX part is the same as the SUM part except some minor changes. The first and obvious change is to replace all $\sum_{j=1}^{k}$ involved in the above arguments by $\max_{j=1,\ldots,k}$. The second change is to replace the Cauchy-Schwartz inequality used in (D.5) by the inequality $|\max_j a_j b_j| \leq (\max_j a_j^2 \times \max_j b_j^2)^{1/2}$. The rest of the arguments stay unchanged. □

E Proof of Theorem 2

We first introduce the approximation of $\hat{T}_T(\gamma)$ that connects the distribution of $\hat{T}_T(\gamma)$ with those of the subsampling statistic and the bootstrap statistic. For any $\theta \in \Theta_{0,F}(\gamma)$, let $\Lambda_T(\theta,\gamma) = \{\lambda : \theta + \lambda/\sqrt{T} \in \Gamma^{-1}(\gamma),\ d(\theta+\lambda/\sqrt{T}, \Theta_{0,F}(\gamma)) = ||\lambda||/\sqrt{T}\}$. In words, $\Lambda_T(\theta,\gamma)$ is the set of all deviations from θ along the fastest paths away from $\Theta_{0,F}(\gamma)$. With this notation handy, we can define the approximation of $\hat{T}_T(\gamma)$ as follows:

$$T_T^{appr}(\gamma) = \tag{E.1}$$
$$\min_{\theta\in\Theta_{0,F}(\gamma)} \min_{\lambda\in\Lambda_T(\theta,\gamma)} \int_{\mathcal{G}} S(\nu_F(\theta,g) + G_F(\theta,g)\lambda + \sqrt{T}\rho_F(\theta,g), \Sigma_F^{\iota}(\theta,g))d\mu(g).$$

Theorem E.1 shows that $T_T^{appr}(\gamma)$ approximates $\hat{T}_T(\gamma)$ asymptotically.

Theorem E.1. *Suppose that the conditions in Lemma* 1 *and Assumptions* C.1-C.3 and C.5-C.6 *hold. Then for any real sequence* $\{x_T\}$ *and scalar* $\eta > 0$ *,*

$$\liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \left[\Pr{}_F(\hat{T}_T(\gamma) \le x_T + \eta) - \Pr(T_T^{appr}(\gamma) \le x_T)\right] \ge 0 \text{ and}$$
$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \left[\Pr{}_F(\hat{T}_T(\gamma) \le x_T) - \Pr(T_T^{appr}(\gamma) \le x_T + \eta)\right] \le 0.$$

Theorem E.1 is a key step in the proof of Theorem 2 and is proved in the next sub-subsection. The remaining proof of Theorem 2 is given in the subsection after that.

E.1 Proof of Theorem E.1

The following lemma is used in the proof of Theorem E.1. It is a portmanteau theorem for uniform weak approximation, which is an extension of the portmanteau theorem for (pointwise) weak convergence in Chapter 1.3 of van der Vaart and Wellner (1996). Let $(\mathbb{D}, d)$ be a metric space and let BL_1 denote the set of all real functions on $\mathbb{D}$ with a Lipschitz norm bounded by one.

Lemma E.1. (a) *Let* $(\Omega,\mathbb{B})$ *be a measurable space. Let* $\{X_T^{(1)} : \Omega \to \mathbb{D}\}$ *and* $\{X_T^{(2)} : \Omega \to \mathbb{D}\}$ *be two sequences of mappings. Let* $\mathcal{P}$ *be a set of probability measures defined on* $(\Omega,\mathbb{B})$*. Suppose that* $\sup_{P\in\mathcal{P}} \sup_{f\in BL_1} |E_P^* f(X_T^{(1)}) - E_{*,P} f(X_T^{(2)})| \to 0$*. Then for any open set* $G_0 \subseteq \mathbb{D}$ *and closed set* $G_1 \subset G_0$*, we have*

$$\liminf_{T\to\infty} \inf_P \left[\Pr{}_{*,P}(X_T^{(1)} \in G_0) - \Pr{}_P^*(X_T^{(2)} \in G_1)\right] \ge 0 \text{ and}$$

(b) *Let* $(\Omega,\mathbb{B})$ *be a product space:* $(\Omega,\mathbb{B}) = (\Omega_1 \times \Omega_2, \sigma(\mathbb{B}_1 \times \mathbb{B}_2))$*. Let* $\mathcal{P}_1$ *be a set*

of probability measures defined on $(\Omega_1, \mathbb{B}_1)$ *and* P_2 *be a probability measure on* $(\Omega_2, \mathbb{B}_2)$. *Suppose that* $\sup_{P_1 \in \mathcal{P}_1} \Pr^*_{P_1}(\sup_{f \in BL_1} |E^*_{P_2} f(X_T^{(1)}) - E_{*,P_2} f(X_T^{(2)})| > \varepsilon) \to 0$ *for all* $\varepsilon > 0$. *Then for any open set* $G_0 \subseteq \mathbb{D}$ *and closed set* $G_0 \subset G_1$, *we have for any* $\varepsilon > 0$,

$$\limsup_{T \to \infty} \sup_{P_1 \in \mathcal{P}_1} \Pr^*_{P_1}(\Pr^*_{P_2}(X_T^{(1)} \in G_1) - \Pr_{*,P_2}(X_T^{(2)} \in G_0) > \varepsilon) = 0.$$

Proof of Lemma E.1. (a) We first show that there is a Lipschitz continuous function sandwiched by $1(x \in G_0)$ and $1(x \in G_1)$. Let $f_a(x) = (a \cdot d(x, G_0^c)) \wedge 1$, where G_0^c is the complement of G_0. Then f_a is a Lipschitz function and $f_a(x) \leq 1(x \in G_0)$ for any $a > 0$. Because G_1 is a closed subset of G_0, $\inf_{x \in G_1} d(x, G_0^c) > c$ for some $c > 0$. Let $a = c^{-1} + 1$. Then $f_a(x) \geq 1(x \in G_1)$. Thus, the function $f_a(x)$ is sandwiched between $1(x \in G_0)$ and $1(x \in F_1)$. Equivalently,

$$a^{-1} 1(x \in G_1) \leq a^{-1} f_a(x) \leq a^{-1} 1(x \in G_0), \ \forall x \in \mathbb{D}. \tag{E.2}$$

By definition, $a^{-1} f_a(x) \in BL_1$. Using this fact and (E.2), we have

$$\begin{aligned}
&a^{-1} \liminf_{T \to \infty} \inf_{P \in \mathcal{P}} \left[\Pr_{*,P}(X_T^{(1)} \in G_0) - \Pr^*_P(X_T^{(2)} \in G_1) \right] \\
&= \liminf_{T \to \infty} \inf_{P \in \mathcal{P}} [a^{-1} \Pr_{*,P}(X_T^{(1)} \in G_0) - E_{*,P} a^{-1} f_a(X_T^{(1)}) + \\
&\quad E_{*,P} a^{-1} f_a(X_T^{(1)}) - E^*_P a^{-1} f_a(X_T^{(2)}) + E^*_P a^{-1} f_a(X_T^{(2)}) - a^{-1} \Pr^*_P(X_T^{(2)} \in G_1)] \\
&\geq \liminf_{T \to \infty} \inf_{P \in \mathcal{P}} \left[E_{*,P} a^{-1} f_a(X_T^{(1)}) - E^*_P a^{-1} f_a(X_T^{(2)}) \right] = 0.
\end{aligned} \tag{E.3}$$

Therefore, part (a) is established.

(b) Use the same a and $f_a(x)$ as above, we have

$$\begin{aligned}
\Pr^*_{P_2}(X_T^{(1)} \in G_1) - \Pr_{*,P_2}(X_T^{(2)} \in G_0) &\leq a \left[E^*_{P_2} a^{-1} f_a(X_T^{(1)}) - E_{*,P_2} a^{-1} f_a(X_T^{(2)}) \right] \\
&\leq a \sup_{f \in BL_1} |E_{*,P_2} f(X_T^{(1)}) - E^*_{P_2} f(X_T^{(2)})|.
\end{aligned} \tag{E.4}$$

This implies part (b). □

Proof of Theorem E.1. We only need to show the first inequality because the second one follows from the same arguments with $\hat{T}_T(\gamma)$ and $T_T^{appr}(\gamma)$ flipped.

The proof consists of four steps. In the first step, we show that the truncation of $\mathcal{G}$ has

asymptotically negligible effect: for all $\epsilon > 0$,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_F(|\hat{T}_T(\gamma) - \bar{T}_T(\gamma)| > \epsilon) = 0, \tag{E.5}$$

where $\bar{T}_T(\gamma)$ is the same as $\hat{T}_T(\gamma)$ except that the integral is over $\mathcal{G}$ instead of $\mathcal{G}_T$.

In the second step, we define a bounded version of $\bar{T}_T(\gamma)$: $\bar{T}_T(\gamma; B_1, B_2)$ and a bounded version of $T_T^{appr}(\gamma)$: $\bar{T}_T^{appr}(\gamma; B_1, B_2)$ and show that for any B_1, $B_2 > 0$ and any real sequence $\{x_T\}$,

$$\liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \left[\Pr_F(\bar{T}_T(\gamma; B_1, B_2) \le x_T + \eta) - \Pr(\bar{T}_T^{appr}(\gamma; B_1, B_2) \le x_T)\right] \ge 0. \tag{E.6}$$

In the third step, we show that $\bar{T}_T(\gamma; B_1, B_2)$ is asymptotically close in distribution to $\bar{T}_T(\gamma)$ for large enough B_1, B_2: for any $\epsilon > 0$, there exists $B_{1,\epsilon}$ and $B_{2,\epsilon}$ such that

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_F(\bar{T}_T(\gamma; B_{1,\epsilon}, B_{2,\epsilon}) \neq \bar{T}_T(\gamma)) < \epsilon. \tag{E.7}$$

In the fourth step, we show that $\bar{T}_T^{appr}(\gamma; B_1, B_2)$ is asymptotically close in distribution to $T_T^{appr}(\gamma)$ for large enough B_1, B_2: for any $\epsilon > 0$, there exists $B_{1,\epsilon}$ and $B_{2,\epsilon}$ such that

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_F(\bar{T}_T^{appr}(\gamma; B_{1,\epsilon}, B_{2,\epsilon}) \neq T_T^{appr}(\gamma)) < \epsilon. \tag{E.8}$$

The four steps combined proves the Theorem. Now we give detailed arguments of the four steps.

STEP 1. First we show a property of the function S that is useful throughout all steps: for any (m_1, Σ_1) and $(m_2, \Sigma_2) \in R^k \times \Psi^\xi$,

$$|S(m_1, \Sigma_1) - S(m_2, \Sigma_2)| \le C^2 \times (S(m_2, \Sigma_2) + 1)(\Delta + \sqrt{\Delta^2 + 8\Delta})/2, \tag{E.9}$$

for the Δ and C in Assumption C.6(b). Let $\Delta_S := |S(m_1, \Sigma_1) - S(m_2, \Sigma_2)|$. Assumption C.6(b) implies that

$$\begin{aligned} \Delta_S^2 &\le C^2 \times (S(m_1, \Sigma_1) + S(m_2, \Sigma_2))(S(m_2, \Sigma_2) + 1)\Delta \\ &\le C^2 \times (\Delta_S + 2S(m_2, \Sigma_2))(S(m_2, \Sigma_2) + 1)\Delta. \end{aligned} \tag{E.10}$$

Solve the quadratic inequality for Δ_S, we have

$$\begin{aligned}\Delta_S &\leq \frac{C^2}{2}\left[(S(m_2,\Sigma_2)+1)\Delta + \sqrt{(S(m_2,\Sigma_2)+1)^2\Delta^2 + 8S(m_2,\Sigma_2)(S(m_2,\Sigma_2)+1)\Delta}\right] \\ &\leq \frac{C^2}{2}(S(m_2,\Sigma_2)+1)(\Delta + \sqrt{\Delta^2+8\Delta}) \end{aligned} \tag{E.11}$$

This shows (E.9).

Now observe that

$$\begin{aligned} 0 &\leq \bar{T}_T(\gamma) - \hat{T}_T(\gamma) \\ &\leq \sup_{\theta\in\Gamma^{-1}(\gamma)} \int_{\mathcal{G}/\mathcal{G}_T} S(\sqrt{T}\bar{\rho}_T(\theta,g), \hat{\Sigma}_T^\iota(\theta,g))d\mu(g) \\ &\leq \sup_{\theta\in\Gamma^{-1}(\gamma)} \int_{\mathcal{G}/\mathcal{G}_T} S(\sqrt{T}\rho_F(\theta,g), \Sigma_F^\iota(\theta,g))d\mu(g) + \\ &\quad \sup_{\theta\in\Gamma^{-1}(\gamma)} \int_{\mathcal{G}/\mathcal{G}_T} |S(\sqrt{T}\rho_F(\theta,g), \Sigma_F^\iota(\theta,g)) - S(\sqrt{T}\bar{\rho}_T(\theta,g), \hat{\Sigma}_T^\iota(\theta,g))|d\mu(g) \\ &= o(1) + \sup_{\theta\in\Gamma^{-1}(\gamma)} \int_{\mathcal{G}/\mathcal{G}_T} |S(\sqrt{T}\rho_F(\theta,g), \Sigma_F^\iota(\theta,g)) - S(\sqrt{T}\bar{\rho}_T(\theta,g), \hat{\Sigma}_T^\iota(\theta,g))|d\mu(g) \\ &\leq o(1) + \sup_{\theta\in\Gamma^{-1}(\gamma)} \int_{\mathcal{G}/\mathcal{G}_T} C^2 \times \left(S(\sqrt{T}\rho_F(\theta,g), \Sigma_F^\iota(\theta,g)) + 1\right) d\mu(g) \times \\ &\qquad \sup_{\theta\in\Gamma^{-1}(\gamma), g\in\mathcal{G}/\mathcal{G}_T} c\left(||\hat{\nu}_T(\theta,g)||^2 + ||vech(\Sigma_F^\iota(\theta,g) - \hat{\Sigma}_T^\iota(\theta,g))||\right) \\ &= o(1) + o(1)\times c(O_p(1)) \\ &= o_p(1), \end{aligned} \tag{E.12}$$

where $c(x) = x + \sqrt{x^2+8x}/2$, the third inequality holds by the triangle inequality, the first equality holds by Assumption C.5(b), the fourth inequality holds by (E.9) and the second equality holds by Assumptions C.5(a)-(b) and C.2(c)-(d). The $o(1)$, $o_p(1)$ and $O_p(1)$ are uniform over $(\gamma, F) \in \mathcal{H}$. Thus, (E.5) is shown.

STEP 2. We define the bounded versions of $\bar{T}_T(\gamma)$ as

$$\begin{aligned}\bar{T}_T(\gamma; B_1, B_2) = &\min_{\theta\in\Theta_{0,F}(\gamma)} \min_{\lambda\in\Lambda_T^{B_2}(\theta,\gamma)} \\ &\int_{\mathcal{G}} S(\hat{\nu}_T^{B_1}(\theta+\lambda/\sqrt{T}, g) + G_F(\tilde{\theta}_T, g)\lambda + \sqrt{T}\rho_F(\theta,g), \hat{\Sigma}_T^\iota(\theta+\lambda/\sqrt{T}, g))d\mu(g) \end{aligned} \tag{E.13}$$

where $\Lambda_T^{B_2}(\theta,\gamma) = \{\lambda \in \Lambda_T(\theta,\gamma) : TQ_F(\theta+\lambda/\sqrt{T}) \leq B_2\}$, the process $\hat{\nu}_T^{B_1}(\cdot,\cdot) = \max\{-B_1, \min\{B_1, \hat{\nu}_T(\cdot,\cdot)\}\}$ and $\tilde{\theta}_T$ is a value lying on the line segment joining θ and

$\theta + \lambda/\sqrt{T}$ satisfying the mean value expansion:

$$\rho_F(\theta + \lambda/\sqrt{T}, g) = \rho_F(\theta, g) + G_F(\tilde{\theta}_T, g)\lambda/\sqrt{T}. \tag{E.14}$$

Define the bounded version of $T_T^{appr}(\gamma)$ as

$$\begin{aligned}&\bar{T}_T^{appr}(\gamma; B_1, B_2) = \\ &\min_{\theta \in \Theta_{0,F}(\gamma)} \min_{\lambda \in \Lambda_T^{B_2}(\theta,\gamma)} \int_{\mathcal{G}} S(\nu_F^{B_1}(\theta, g) + G_F(\theta, g)\lambda + \sqrt{T}\rho_F(\theta, g), \Sigma_F^{\iota}(\theta, g))d\mu(g),\end{aligned} \tag{E.15}$$

where $\nu_F^{B_1}(\cdot, \cdot) = \max\{-B_1, \min\{B_1, \nu_F(\cdot, \cdot)\}\}$.

First we show a useful result: there exists some constant $\bar{C} > 0$ such that for all $(\gamma, F) \in \mathcal{H}_0$ and $\lambda \in \Lambda_T^{B_2}(\theta, \gamma)$ and for the δ_2 in Assumption C.3(b), we have

$$||\lambda|| \leq \bar{C} \times T^{(\delta_2 - 2)/(2\delta_2)}. \tag{E.16}$$

This is shown by observing, for all $(\gamma, F) \in \mathcal{H}_0$ and $\lambda \in \Lambda_T^{B_2}(\theta, \gamma)$,

$$\begin{aligned}B_2 >& TQ_F(\theta + \lambda/\sqrt{T}) \\ \geq& C \cdot ((T \times d(\theta + \lambda/\sqrt{T}, \Theta_{0,F}(\gamma))^{\delta_2}) \wedge (c \times T)).\end{aligned} \tag{E.17}$$

The second inequality holds by Assumption (C.3)(b). Because $c \times T$ is eventually greater than B_2 as $T \to \infty$, we have for large enough T,

$$B_2 \geq C \times T \times (||\lambda||/\sqrt{T})^{\delta_2}. \tag{E.18}$$

This implies (E.16).

Equation (E.16) implies two results:

$$\begin{aligned}&(1) \sup_{(\gamma,F)\in\mathcal{H}_0} \sup_{\theta\in\Theta_{0,F}(\gamma)} \sup_{\lambda\in\Lambda_T^{B_2}(\theta,\gamma)} \|\lambda\|/\sqrt{T} \leq O(T^{-1/\delta_2}) = o(1) \\ &(2) \sup_{(\gamma,F)\in\mathcal{H}_0} \sup_{\theta\in\Theta_{0,F}(\gamma)} \sup_{\lambda\in\Lambda_T^{B_2}(\theta,\gamma)} \sup_{g\in\mathcal{G}} \|G_F(\theta + O(\|\lambda\|)/\sqrt{T}, g)\lambda - G_F(\theta, g)\lambda\| \\ &\qquad \leq O(1) \times \|\lambda\|^{\delta_1+1} T^{-\delta_1/2} \leq O(T^{(\delta_2 - 2(\delta_1+1))/(2\delta_2)}) = o(1).\end{aligned} \tag{E.19}$$

The first result holds immediately given (E.16) and the second result holds by Assumption C.2(e).

Define an intermediate statistic

$$\bar{T}_T^{med}(\gamma; B_1, B_2) = \min_{\theta \in \Theta_{0,F}(\gamma)} \min_{\lambda \in \Lambda_T^{B_2}(\theta,\gamma)} \int_{\mathcal{G}} S(\hat{\nu}_T^{B_1}(\theta, g) + G_F(\theta, g)\lambda + \sqrt{T}\rho_F(\theta, g), \Sigma_F^{\iota}(\theta, g))d\mu(g). \quad \text{(E.20)}$$

Then $\bar{T}_T^{med}(\gamma; B_1, B_2)$ and $\bar{T}_T^{appr}(\gamma; B_1, B_2)$ are respectively the following functional evaluated at $\nu_F(\cdot, \cdot)$ and $\hat{\nu}_T(\cdot, \cdot)$:

$$h(\nu) = \min_{\theta \in \Theta_{0,F}(\gamma)} \min_{\lambda \in \Lambda_T^{B_2}(\theta,\gamma)} \int_{\mathcal{G}} S(\nu^{B_1}(\theta, \cdot) + G_F(\theta, \cdot)\lambda + \sqrt{T}\rho_F(\theta, \cdot), \Sigma_F^{\iota}(\theta, \cdot))d\mu. \quad \text{(E.21)}$$

The functional $h(\nu)$ is uniformly bounded for all large enough T because for any fixed $\theta \in \Theta_{0,F}(\gamma)$ and $\lambda \in \Lambda_T^{B_2}(\theta, \gamma)$,

$$\begin{aligned}
h(\nu) &\leq 2\int_{\mathcal{G}} S(G_F(\theta, \cdot)\lambda + \sqrt{T}\rho_F(\theta, \cdot), \Sigma_F^{\iota}(\theta, \cdot))d\mu + 2\int_{\mathcal{G}} S(\nu^{B_1}(\theta, \cdot), \Sigma_F^{\iota}(\theta, \cdot))d\mu \\
&\leq 2\sup_{\Sigma \in \Psi} S(-B_1 1_k, \Sigma) + 2\int_{\mathcal{G}} S(G_F(\theta, \cdot)\lambda + \sqrt{T}\rho_F(\theta, \cdot), \Sigma_F^{\iota}(\theta, \cdot))d\mu \\
&\leq 2\sup_{\Sigma \in \Psi} S(-B_1 1_k, \Sigma) + 2T \times Q_F(\theta + \lambda/\sqrt{T}) + \\
&C^2 \times (T \times Q_F(\theta + \lambda/\sqrt{T}) + 1)\sup_{g \in \mathcal{G}}(\Delta_T(g) + \sqrt{\Delta_T(g)^2 + 8\Delta_T(g)}) \\
&\leq 2\sup_{\Sigma \in \Psi} S(-B_1 1_k, \Sigma) + 2B_2 + C^2(B_2 + 1) \times o(1), \quad \text{(E.22)}
\end{aligned}$$

where $\Delta_T(g) := \|G_F(\theta, g)\lambda + \sqrt{T}\rho_F(\theta, g) - \sqrt{T}\rho_F(\theta_T, g)\|^2 + \|vech(\Sigma_F^{\iota}(\theta, g) - \Sigma_F^{\iota}(\theta_T, g))\|$ and $\theta_T = \theta + \lambda/\sqrt{T}$. The first inequality holds by Assumptions C.6(e)-(f), the second inequality holds by Assumptions C.2(f) and Assumptions C.6(c), the third inequality holds by (E.9) and the last inequality holds by (E.19).

The functional $h(\nu)$ is Lipschitz continuous for all large enough T with respect to the uniform metric because

$$\begin{aligned}
|h(\nu_1) - h(\nu_2)| &\leq 2C \sup_{\theta \in \Theta_{0,F}(\gamma)} \sup_{\lambda \in \Lambda_T^{B_2}(\theta,\gamma)} \sup_{g \in \mathcal{G}} \|\nu_1(\theta, g) - \nu_2(\theta, g)\| \cdot (1 + h(\nu_1) + 2h(\nu_2)) \\
&\leq \bar{C} \sup_{\theta \in \Gamma^{-1}(\gamma), g \in \mathcal{G}} \|\nu_1(\theta, g) - \nu_2(\theta, g)\|, \quad \text{(E.23)}
\end{aligned}$$

where $\bar{C}$ is any constant such that $\bar{C} > 2C \times (6\sup_{\Sigma \in \Psi} S(-B_1 1_k, \Sigma) + 6B_2 + 1)$, the first inequality holds by Assumption C.6(b) and the second holds by (E.22).

Therefore, for any $f \in BL_1$ and any real sequence $\{x_T\}$, the composite function $f \circ$

$(\bar{C}^{-1}h(\cdot)+x_T)\in BL_1$. By AssumptionC.2(c), we have

$$\limsup_{T\to\infty}\sup_{(\gamma,F)\in\mathcal{H}_0}\sup_{f\in BL_1}|E_F f(\bar{T}_T^{med}(\gamma;B_1,B_2)+x_T)-Ef(\bar{T}_T^{appr}(\gamma;B_1,B_2)+x_T)|=0. \quad \text{(E.24)}$$

This combined with Lemma E.1(a) (with $G_0=(-\infty,\eta)$ and $G_1=(-\infty,0]$) gives

$$\liminf_{T\to\infty}\inf_{(\gamma,F)\in\mathcal{H}_0}\left[\Pr{}_F(\bar{T}_T^{med}(\gamma;B_1,B_2)\le x_T+\eta)-\Pr(\bar{T}_T^{appr}(\gamma;B_1,B_2)\le x_T)\right]\ge 0. \quad \text{(E.25)}$$

Now it is left to show that $\bar{T}_T^{med}(\gamma;B_1,B_2)$ and $\bar{T}_T(\gamma;B_1,B_2)$ are close. First, we have

$$\begin{aligned}
&|\bar{T}_T(\gamma;B_1,B_2)-\bar{T}_T^{med}(\gamma;B_1,B_2)|\\
\le&\sup_{\theta\in\Theta_{0,F}(\gamma),\lambda\in\Lambda_T^{B_2}(\theta,\gamma)}\int_{\mathcal{G}}\Big|S(\hat{\nu}_T^{B_1}(\theta+\lambda/\sqrt{T},g)+G_F(\tilde{\theta}_T,g)\lambda+\sqrt{T}\rho_F(\theta,g),\hat{\Sigma}_T^{\iota}(\theta+\lambda/\sqrt{T},g))\\
&-S(\hat{\nu}_T^{B_1}(\theta,g)+G_F(\theta,g)\lambda+\sqrt{T}\rho_F(\theta,g),\Sigma_F^{\iota}(\theta,g))\Big|\,d\mu(g)\\
\le&C^2\times\sup_{\theta\in\Theta_{0,F}(\gamma),\lambda\in\Lambda_T^{B_2}(\theta,\gamma)}\max_{g\in\mathcal{G}}c(\Delta_T(\theta,\lambda,g))\times\int_{\mathcal{G}}(1+M_T(\theta,\lambda,g))d\mu(g), \quad \text{(E.26)}
\end{aligned}$$

where $c(x)=(x+\sqrt{x^2+8x})/2$, C is the constant in (E.9),

$$\begin{aligned}
\Delta_T(\theta,\lambda,g)=&\|\hat{\nu}_T^{B_1}(\theta+\lambda/\sqrt{T},g)-\hat{\nu}_T^{B_1}(\theta,g)+G_F(\tilde{\theta}_T,g)\lambda-G_F(\theta,g)\lambda\|^2+\\
&\|vech(\hat{\Sigma}_T(\theta+\lambda/\sqrt{T},g)-\Sigma_F(\theta,g))\|\text{ and}\\
M_T(\theta,\lambda,g)=&S(\hat{\nu}_T^{B_1}(\theta,g)+G_F(\theta,g)\lambda+\sqrt{T}\rho_F(\theta,g),\Sigma_F^{\iota}(\theta,g)). \quad \text{(E.27)}
\end{aligned}$$

Below we show that for any $\epsilon>0$, and some universal constant $\bar{C}>0$,

$$\sup_{(\gamma,F)\in\mathcal{H}_0}\Pr{}_F\left(\sup_{\theta\in\Theta_{0,F}(\gamma),\lambda\in\Lambda_T^{B_2}(\theta,\gamma),g\in\mathcal{G}}\Delta_T(\theta,\lambda,g)>\epsilon\right)\to 0\text{ and} \quad \text{(E.28)}$$

$$\sup_T\sup_{(\gamma,F)\in\mathcal{H}_0}\sup_{\theta\in\Theta_{0,F}(\gamma),\lambda\in\Lambda_T^{B_2}(\theta,\gamma)}\int_{\mathcal{G}}M_T(\theta,\lambda,g)d\mu(g)<\bar{C}. \quad \text{(E.29)}$$

Once (E.28) and (E.29) are shown, it is immediate that for any $\epsilon>0$,

$$\sup_{(\gamma,F)\in\mathcal{H}_0}\Pr{}_F\left(|\bar{T}_T(\gamma;B_1,B_2)-\bar{T}_T^{med}(\gamma;B_1,B_2)|>\epsilon\right)\to 0. \quad \text{(E.30)}$$

This combined with (E.25) shows (E.6).

Now we show (E.28) and (E.29). The convergence result (E.28) is implied by the fol-

lowing results: for any $\epsilon > 0$,

$$\sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_F \left(\sup_{\theta\in\Theta_{0,F}(\gamma),\lambda\in\Lambda_T^{B_2}(\theta,\gamma),g\in\mathcal{G}} ||\hat{\nu}_T^{B_1}(\theta+\lambda/\sqrt{T},g) - \hat{\nu}_T^{B_1}(\theta,g)|| > \epsilon \right) \to 0$$

$$\sup_{(\gamma,F)\in\mathcal{H}_0} \sup_{\theta\in\Theta_{0,F}(\gamma),\lambda\in\Lambda_T^{B_2}(\theta,\gamma),g\in\mathcal{G}} ||G_F(\tilde{\theta}_T,g)\lambda - G_F(\theta,g)\lambda|| \to 0 \text{ and}$$

$$\sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_F \left(\sup_{\theta\in\Theta_{0,F}(\gamma),\lambda\in\Lambda_T^{B_2}(\theta,\gamma),g\in\mathcal{G}} ||vech(\hat{\Sigma}_T(\theta+\lambda/\sqrt{T},g) - \Sigma_F(\theta,g))|| > \epsilon \right) \to 0. \tag{E.31}$$

The first result in the above display holds by the first result in equation (E.19) and the uniform stochastic equicontinuity of the empirical process $\hat{\nu}_T(\cdot,g) : \Gamma^{-1}(\gamma) \to R^{d_m}$ with respect to the Euclidean metric. The uniform equicontinuity is implied by Assumptions C.2(b), (c) and (f) by Theorem 2.8.2 of van der Vaart and Wellner (1996). The second result in the above display holds by the second result in (E.19). The third result in (E.31) holds by Assumption C.2(d) and (f).

Result (E.29) holds because for any $\theta \in \Theta_{0,F}(\gamma)$ and $\lambda \in \Lambda_T^{B_2}(\theta,\gamma)$,

$$\begin{aligned}
&\int_{\mathcal{G}} M_T(\theta,\lambda,g)d\mu(g) \\
\leq& 2\int_{\mathcal{G}} S(\hat{\nu}_T^{B_1}\theta,g),\Sigma_F^{\iota}(\theta,g))d\mu(g) + 2\int_{\mathcal{G}} S(G_F(\theta,g)\lambda + \sqrt{T}\rho_F(\theta,g),\Sigma_F^{\iota}(\theta,g))d\mu(g) \\
\leq& \sup_{\Sigma\in\Psi} S(-B_1 1_k,\Sigma) + 2\int_{\mathcal{G}} S(G_F(\theta,g)\lambda + \sqrt{T}\rho_F(\theta,g),\Sigma_F^{\iota}(\theta,g))d\mu(g) \\
\leq& \sup_{\Sigma\in\Psi} S(-B_1 1_k,\Sigma) + 2B_2 + C^2(B_2+1)\times o(1),
\end{aligned} \tag{E.32}$$

where the first inequality holds by Assumptions C.6(f), the second inequality holds by Assumption C.6(c) and the last inequality holds by the second and third inequality in (E.22) and the $o(1)$ is uniform over (θ,λ).

STEP 3. In order to show (E.7), first extend the definition of $\bar{T}_T(\gamma;B_1,B_2)$ from Step 1 to allow B_1 and B_2 to take the value ∞ and observe that $\bar{T}_T(\gamma;\infty,\infty) = \bar{T}_T(\gamma)$.

Assumptions C.2 (c) and Lemma E.1 imply that for any $\epsilon > 0$, there exists $B_{1,\epsilon}$ large enough such that

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_F \left(\sup_{\theta\in\Theta,g\in\mathcal{G}} \|\hat{\nu}_T(\theta,g)\| > B_{1,\epsilon} \right) < \epsilon. \tag{E.33}$$

Therefore we have for all B_2,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_F\left(\bar{T}_T(\gamma,\infty,B_2) \neq \bar{T}_T(\gamma;B_{1,\epsilon},B_2)\right) < \epsilon. \tag{E.34}$$

To show that $\bar{T}_T(\gamma)$ and $\bar{T}_T(\gamma;\infty,B_2)$ are close for B_2 large enough, first observe that:

$$\begin{aligned}
\bar{T}_T(\gamma) &\leq \sup_{\theta\in\Theta_{0,F}(\gamma)} \int_{\mathcal{G}} S(\hat{\nu}_T(\theta,g) + \sqrt{T}\rho_F(\theta,g), \hat{\Sigma}_T^{\iota}(\theta,g))d\mu(g) \\
&\leq \sup_{\theta\in\Theta_{0,F}(\gamma)} \int_{\mathcal{G}} S(\hat{\nu}_T(\theta,g), \hat{\Sigma}_T^{\iota}(\theta,g))d\mu(g) \\
&= O_p(1)
\end{aligned} \tag{E.35}$$

where the first inequality holds because $0 \in \Lambda_T(\theta,\gamma)$, the second inequality holds because $\rho_F(\theta,g) \geq 0$ for $\theta \in \Theta_{0,F}(\gamma)$ and by Assumption C.6(c), the equality holds by Assumption C.6(a)-(c) and Assumptions C.2 (c), (d) and (f). The $O_p(1)$ is uniform over $(\gamma,F) \in \mathcal{H}_0$.

For any T, γ, B_2, if $\bar{T}_T(\gamma) \neq \bar{T}_T(\gamma;\infty,B_2)$, then there must be a $\theta^* \in \Gamma^{-1}(\gamma)$ such that $T \times Q_F(\theta^*) > B_2$ and

$$\int_{\mathcal{G}} S(\hat{\nu}_T(\theta^*,g) + \sqrt{T}\rho_F(\theta^*,g), \hat{\Sigma}_T^{\iota}(\theta^*,g))d\mu(g) < O_p(1). \tag{E.36}$$

But

$$\begin{aligned}
&\int_{\mathcal{G}} S(\hat{\nu}_T(\theta^*,g) + \sqrt{T}\rho_F(\theta^*,g), \hat{\Sigma}_T^{\iota}(\theta^*,g))d\mu(g) \\
\geq& 2^{-1}\int_{\mathcal{G}} S(\sqrt{T}\rho_F(\theta^*,g), \hat{\Sigma}_T^{\iota}(\theta^*,g))d\mu(g) - \int_{\mathcal{G}} S(-\hat{\nu}_T(\theta^*,g), \hat{\Sigma}_T^{\iota}(\theta^*,g))d\mu(g) \\
\geq& 2^{-1}\int_{\mathcal{G}} S(\sqrt{T}\rho_F(\theta^*,g), \hat{\Sigma}_T^{\iota}(\theta^*,g))d\mu(g) - O_p(1) \\
\geq& 2^{-1}\left[TQ_F(\theta^*) - \int_{\mathcal{G}} |S(\sqrt{T}\rho_F(\theta^*,\cdot), \hat{\Sigma}_T^{\iota}(\theta^*,\cdot)) - S(\sqrt{T}\rho_F(\theta^*,\cdot), \Sigma_F^{\iota}(\theta^*,\cdot))|d\mu\right] - O_p(1) \\
\geq& 2^{-1}\left[TQ_F(\theta^*) - C^2 \sup_{g\in\mathcal{G}} c(||vech(\hat{\Sigma}_T^{\iota}(\theta^*,g) - \Sigma_F^{\iota}(\theta^*,g))||) \times (1 + TQ_F(\theta^*))\right] - O_p(1) \\
=& B_2/2 - o(1) - o_p(1) \times C^2 \times B_2/4 - O_p(1),
\end{aligned} \tag{E.37}$$

where $c(x) = (x + \sqrt{x^2+8x})/2$ and C is the constant in (E.9). The first inequality holds by Assumptions C.6(e)-(f), the second inequality holds by Assumption C.6(c) and Assumptions C.2(c)-(d) and (f), the third inequality holds by the triangle inequality, the fourth inequality holds by (E.9) and the equality holds by Assumption C.2(d). The terms $o(1)$, $o_p(1)$ and $O_p(1)$ terms are uniform over $\theta^* \in \Gamma^{-1}(\gamma)$ and $(\gamma,F) \in \mathcal{H}_0$.

Then

$$\begin{aligned}
&\sup_{(\gamma,F)\in\mathcal{H}_0} \Pr{}_F\left(\hat{T}_T(\gamma) \neq \bar{T}_T(\gamma;\infty,B_2)\right) \\
\leq &\sup_{(\gamma,F)\in\mathcal{H}_0} \Pr{}_F\left(2^{-1}(1-o_p(1))\times B_2 - o(1) - O_p(1) \leq O_p(1)\right) \\
= &\sup_{(\gamma,F)\in\mathcal{H}_0} \Pr{}_F\left(O_p(1) \geq B_2\right), \qquad \text{(E.38)}
\end{aligned}$$

where the first inequality holds by (E.36) and (E.37). Then for any ϵ, there exists $B_{2,\epsilon}$ such that

$$\lim_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr{}_F(\hat{T}_T(\gamma) \neq \bar{T}_T(\gamma;\infty,B_{2,\epsilon})) < \epsilon. \qquad \text{(E.39)}$$

Combining this with (E.34), we have (E.7).

STEP 4. In order to show (E.8), first extend the definition of $\bar{T}_T^{appr}(\gamma;B_1,B_2)$ from Step 1 to allow B_1 and B_2 to take the value ∞ and observe that $\bar{T}_T^{appr}(\gamma;\infty,\infty) = T_T^{appr}(\gamma)$.

By the same arguments as those for (E.34), for any ϵ and B_2, there exists $B_{1,\epsilon}$ large enough so that

$$\limsup_{n\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr{}_F\left(\bar{T}_T^{appr}(\gamma;\infty,B_2) \neq \bar{T}_T^{appr}(\gamma;B_{1,\epsilon},B_2)\right) < \epsilon. \qquad \text{(E.40)}$$

Also by the same reasons as those for (E.35), we have

$$T_T^{appr}(\gamma) \leq \sup_{\theta\in\Theta_{0,F}(\gamma)} \int_{\mathcal{G}} S(\nu_F(\theta,g),\Sigma_F^{\iota}(\theta,g))d\mu(g), \qquad \text{(E.41)}$$

where the right hand side is a real-valued random variable.

For any T and B_2, if $T_T^{appr}(\gamma) \neq \bar{T}_T^{appr}(\gamma;\infty,B_{2,\epsilon})$, then there must be a $\theta^* \in \Theta_{0,F}(\gamma)$, a $\lambda^{**} \in \{\lambda \in \Lambda_T(\theta^*,\gamma) : T\times Q_F(\theta^*+\lambda/\sqrt{T}) > B_2\}$ such that

$$I(\lambda^{**}) < \sup_{\theta\in\Theta_{0,F}(\gamma)} \int_{\mathcal{G}} S(\nu_F(\theta,g),\Sigma_F^{\iota}(\theta,g))d\mu(g), \qquad \text{(E.42)}$$

where $I(\lambda) = \int_{\mathcal{G}} S(\nu_F(\theta^*,g) + G_F(\theta^*,g)\lambda + \sqrt{T}\rho_F(\theta^*,g),\Sigma_F^{\iota}(\theta^*,g))d\mu(g)$. Next we show that if λ^{**} exists, then there must exists a λ^* such that

$$\begin{aligned}
&\lambda^* \in \{\lambda \in \Lambda_T(\theta^*,\gamma) : T\times Q_F(\theta^*+\lambda/\sqrt{T}) \in (B_2,2B_2]\} \text{ and} \\
&I(\lambda^*) < \sup_{\theta\in\Theta_{0,F}(\gamma)} \int_{\mathcal{G}} S(\nu_F(\theta,g),\Sigma_F^{\iota}(\theta,g))d\mu(g). \qquad \text{(E.43)}
\end{aligned}$$

If $T\times Q_F(\theta^*+\lambda^{**}/\sqrt{T}) \in (B_2,2B_2]$, then we are done. If $T\times Q_F(\theta^*+\lambda^{**}/\sqrt{T}) > 2B_2$,

there must be a $a^* \in (0,1)$ such that $T \times Q_F(\theta^* + a^*\lambda^{**}/\sqrt{T}) \in (B_2, 2B_2]$ because $TQ_F(\theta^* + 0 \times \lambda^{**}/\sqrt{T}) = 0$ and $TQ_F(\theta^* + a\lambda^{**}/\sqrt{T})$ is continuous in a (by Assumptions C.2(e) and C.6(a)). By Assumption C.6(f), $I(\lambda)$ is convex. Thus $I(a^*\lambda^{**}) \leq a^* I(\lambda^{**}) + (1-a^*)I(0)$. For the same arguments as those for (E.35), $I(0) \leq \sup_{\theta \in \Theta_{0,F}(\gamma)} \int_{\mathcal{G}} S(\nu_F(\theta, g), \Sigma_F^{\iota}(\theta, g)) d\mu(g)$. Thus, $I(a^*\lambda^{**}) < \sup_{\theta \in \Theta_{0,F}(\gamma)} \int_{\mathcal{G}} S(\nu_F(\theta, g), \Sigma_F^{\iota}(\theta, g)) d\mu(g)$. Assumption (C.1)(c) and the definition of $\Lambda_T(\theta, \gamma)$ guarantee that $a^*\lambda^{**} \in \Lambda_T(\theta^*, \gamma)$. Therefore, $\lambda^* = a^*\lambda^{**}$ satisfies (E.43).

Similar to (E.19) we have

$$\begin{aligned}
&(1)\ ||\lambda^*||/\sqrt{T} \leq B_2 \times 2C \times T^{-1/\delta_2} = B_2 \times o(1) \\
&(2) \sup_{g \in \mathcal{G}} ||G_F(\theta^* + O(||\lambda^*||)/\sqrt{T}, g)\lambda^* - G_F(\theta^*, g)\lambda^*|| \\
&\qquad \leq O(1) \times B_2^{(\delta_1+1)/\delta_2} ||\lambda||^{\delta_1+1} T^{-\delta_1/2} = B_2^{(\delta_1+1)/\delta_2} o(1),
\end{aligned} \tag{E.44}$$

where the $o(1)$ terms do not depend on B_2. Then,

$$\begin{aligned}
I(\lambda^*) &\geq 2^{-1} \int_{\mathcal{G}} S(G_F(\theta^*, g)\lambda^* + \sqrt{T}\rho_F(\theta^*, g), \Sigma_F^{\iota}(\theta^*, g)) d\mu(g) - \\
&\qquad \int_{\mathcal{G}} S(-\nu_F(\theta^*, g), \Sigma_F^{\iota}(\theta^*, g)) d\mu(g) \\
&\geq TQ_F(\theta^* + \lambda^*/\sqrt{T})/2 - C^2 \times (TQ_F(\theta^* + \lambda^*/\sqrt{T}) + 1) \times c(\Delta_T)/2 + O_p(1) \\
&= TQ_F(\theta^* + \lambda^*/\sqrt{T})/2 - C^2 \times (2B_2 + 1) \times c(\Delta_T)/4 + O_p(1),
\end{aligned} \tag{E.45}$$

where the $O_p(1)$ term is uniform over $(\gamma, F) \in \mathcal{H}_0$, $c(x) = (x + \sqrt{x^2 + 8x})/2$ and

$$\begin{aligned}
\Delta_T := &\|G_F(\theta^*, g)\lambda^* + \sqrt{T}\rho_F(\theta^*, g) - \sqrt{T}\rho_F(\theta^* + \lambda^*/\sqrt{T}, g)\|^2 \\
&+ \|vech(\Sigma_F^{\iota}(\theta^* + \lambda^*/\sqrt{T}, g) - \Sigma_F^{\iota}(\theta^*, g))\|.
\end{aligned} \tag{E.46}$$

The first inequality in (E.45) holds by Assumptions C.6(e)-(f), the second inequality holds by (E.9) and the equality holds by (E.43). By (E.44) and Assumption C.2(f), for any fixed B_2, $\lim_{T \to \infty} \Delta_T = 0$. Therefore, for each fixed B_2,

$$I(\lambda^*) \geq TQ_F(\theta^* + \lambda^*/\sqrt{T})/2 - O_p(1) \geq B_2/2 - O_p(1). \tag{E.47}$$

Thus

$$\begin{aligned}
&\sup_{(\gamma,F)\in\mathcal{H}_0} \Pr(T_T^{appr}(\gamma) \neq \bar{T}_T^{appr}(\gamma;\infty,B_2)) \\
&\leq \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr\left(\sup_{\theta\in\Theta_{0,F}(\gamma)} \int_{\mathcal{G}} S(\nu_F(\theta,g),\Sigma_F^{\iota}(\theta,g))d\mu(g) \geq B_2/2 - O_p(1)\right) \\
&= \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr(O_p(1) \geq B_2). \qquad \text{(E.48)}
\end{aligned}$$

For any $\epsilon > 0$, there exists $B_{2,\epsilon}$ large enough so that $\lim_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr(O_p(1) \geq B_2) < \epsilon$. Thus,

$$\lim_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr(T_T^{appr}(\gamma) \neq \bar{T}_T^{appr}(\gamma;\infty,B_{2,\epsilon}) < \epsilon. \qquad \text{(E.49)}$$

Combining this with (E.40), we have (E.8). □

E.2 Proof of Theorem 2

The following lemma is used in the proof of Theorem 2. It shows the convergence of the bootstrap empirical process $\hat{\nu}_T^*(\theta,g)$. Let $W_{T,t}$ be the number of times that the tth observation appearing in a bootstrap sample. Then $(W_{T,1},...,W_{T,T})$ is a random draw from a multinomial distribution with parameters T and $(T^{-1},...,T^{-1})$, and $\hat{\nu}_T^*(\theta,g)$ can be written as

$$\hat{\nu}_T^*(\theta,g) = T^{-1/2}\sum_{t=1}^{T}(W_{T,t}-1)\rho(w_t,\theta,g). \qquad \text{(E.50)}$$

In the lemma, the subscripts F and W for E and Pr signify the fact that the expectation and the probabilities are taken with respect to the randomness in the data and the randomness in $\{W_{T,t}\}$ respectively.

Lemma E.2. *Suppose that Assumption* C.2 *holds. Then for any* $\epsilon > 0$,

(a)$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}} \Pr{}_F^*(\sup_{f\in BL_1} |E_W f(\hat{\nu}_T^*(\cdot,\cdot)) - Ef(\nu_F(\cdot,\cdot))| > \epsilon) = 0$,

(b) *there exists* B_ϵ *large enough such that*

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}} \Pr{}_F^*\left(\Pr{}_W\left(\sup_{\theta\in\Gamma^{-1}(\gamma),g\in\bar{\mathcal{G}}} ||\hat{\nu}_T^*(\theta,g)|| > B_\epsilon\right) > \epsilon\right) = 0, \text{ and}$$

(c) *there exists* δ_ϵ *small enough such that*

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}} \Pr{}_F^*\left(\Pr{}_W\left(\sup_{g\in\bar{\mathcal{G}}} \sup_{||\theta^{(1)}-\theta^{(2)}||\leq\delta_\epsilon} ||\hat{\nu}_T^*(\theta^{(1)},g) - \hat{\nu}_T^*(\theta^{(2)},g)|| > \epsilon\right) > \epsilon\right) = 0.$$

Proof of Lemma E.2. (a) Part (a) is proved using a combination of the arguments in Theorem 2.9.6 and Theorem 3.6.1 in van der Vaart and Wellner (1996). Take a Poisson number N_T with mean T and independent from the original sample. Then $\{W_{N_T,1}, ..., W_{N_T,T}\}$ are i.i.d. Poisson variables with mean one. Let the Poissonized version of $\hat{\nu}_T^*(\theta, g)$ be

$$\hat{\nu}_T^{poi}(\theta, g) = T^{-1/2} \sum_{t=1}^{T} (W_{N_T,t} - 1)\rho(w_t, \theta, g). \tag{E.51}$$

Theorem 2.9.6 in van der Vaart and Wellner (1996) is a multiplier central limit theorem that shows that if $\{\rho(w_t, \theta, g) : (\theta, g) \in \Theta \times \bar{\mathcal{G}}\}$ is F-Donsker and pre-Gaussian, then $\hat{\nu}_T^{poi}(\theta, g)$ converges weakly to $\nu_F(\theta, g)$ conditional on the data in outer probability. The arguments of Theorem 2.9.6 remain valid if we strengthen the F-Donsker and pre-Gaussian condition to the uniform Donsker and pre-Gaussian condition of Assumption C.2(c) and strengthen the conclusion to uniform weak convergence:

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}} \Pr_F^* \left(\sup_{f\in BL_1} |E_W f(\hat{\nu}_T^{poi}(\cdot,\cdot)) - Ef(\nu_F(\cdot,\cdot))| > \varepsilon \right) = 0, \tag{E.52}$$

In particular, the extension to the uniform versions of the first and the third displays in the proof of Theorem 2.9.6 in van der Vaart and Wellner (1996) is straightforward. To extend the second display, we only need to replace Lemma 2.9.5 with Proposition A.5.2 – a uniform central limit theorem for finite dimensional vectors.

Theorem 3.6.1 in van der Vaart and Wellner (1996) shows that, under a fixed (γ, F), the bounded Lipschitz distance between $\hat{\nu}_T^{poi}(\theta, g)$ and $\hat{\nu}_T^*(\theta, g)$ converge to zero conditional on (outer) almost all realizations of the data. The arguments remain valid if we strengthen the Glivenko-Cantelli assumption used there to uniform Glivenko-Cantelli (which is implied by Assumption C.2(c)) and strengthen the conclusion to: for all $\varepsilon > 0$

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}} \Pr_F^* \left(\sup_{f\in BL_1} |E_W f(\hat{\nu}_T^{poi}(\cdot,\cdot)) - E_W f(\hat{\nu}_T^*(\cdot,\cdot))| > \varepsilon \right) = 0, \tag{E.53}$$

Equations (E.52) and (E.53) together imply part (a).

(b) Part (b) is implied by part (a), Lemma E.1(b) and the uniform pre-Gaussianity assumption (Assumption C.2(c)). When applying Lemma E.1(b), consider $X_T^{(1)} = \hat{\nu}_T^*$, $X_T^{(2)} = \nu_F$, $G_1 = \{\nu : \sup_{\theta,g} \|\nu(\theta, g)\| \geq B_\varepsilon\}$, and $G_2 = \{\nu : \sup_{\theta,g} \|\nu(\theta, g)\| > B_\varepsilon - 1\}$ where B_ε satisfies:

$$\sup_{(\gamma,F)\in\mathcal{H}} \Pr \left(\sup_{\theta\in\Theta, g\in\bar{\mathcal{G}}} \|\nu_F(\theta, g)\| > B_\epsilon - 1 \right) < \epsilon/2. \tag{E.54}$$

Such a B_ε exists because $\{\rho(w_t, \theta, g) : (\theta, g) \in \Theta \times \bar{\mathcal{G}}\}$ is uniformly pre-Gaussian by Assumption C.2(c).

(c) Part (c) is implied by part (a), Lemma E.1(b) and the uniform pre-Gaussianity assumption (Assumption C.2(c)). When applying Lemma E.1(b), consider $X_T^{(1)} = \hat{\nu}_T^*$, $X_T^{(2)} = \nu_F$, $G_1 = \{\nu : \sup_{||\theta^{(1)}-\theta^{(2)}||\leq\Delta_\varepsilon, g} \|\nu(\theta^{(1)}, g) - \nu(\theta^{(2)}, g)\| \geq \epsilon\}$, and $G_0 = \{\nu : \sup_{||\theta^{(1)}-\theta^{(2)}||\leq\Delta_\varepsilon, g} \|\nu(\theta^{(1)}, g) - \nu(\theta^{(2)}, g)\| > \epsilon/2\}$, where Δ_ε satisfies:

$$\sup_{(\gamma,F)\in\mathcal{H}} \Pr\left(\sup_{\|\theta^{(2)}-\theta^{(2)}\|\leq\Delta_\varepsilon, g} \|\nu_F(\theta^{(1)}, g) - \nu_F(\theta^{(2)}, g)\| > \epsilon/2\right) < \epsilon/2. \tag{E.55}$$

Such a Δ_ε exists because $\{\rho(w_t, \theta, g) : (\theta, g) \in \Theta \times \bar{\mathcal{G}}\}$ is uniformly pre-Gaussian.

□

Proof of Theorem 2. (a) Let $q_{b_T}^{appr}(\gamma, p)$ denotes the p quantile of $\bar{T}_{b_T}^{appr}(\gamma)$. Let $\eta_2 = \eta^*/3$. Below we show that,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,sub}(c_T^{sub}(\gamma, p) \leq q_{b_T}^{appr}(\gamma, p) + \eta_2) = 0. \tag{E.56}$$

where $\Pr_{F,sub}^*$ signifies the fact that there are two sources of randomness in $c_T^{sub}(\gamma, p)$ one from the original sampling and the other from the subsampling. Once (E.56) is established, we have,

$$\begin{aligned}
&\liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,sub}\left(\hat{T}_T(\gamma) \leq c_T^{sub}(\gamma, p)\right) \\
&\geq \liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \Pr_F\left(\hat{T}_T(\gamma) \leq q_{b_T}^{appr}(\gamma, p) + \eta_2\right) \\
&\geq \liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \left[\Pr_F\left(\hat{T}_T(\gamma) \leq q_{b_T}^{appr}(\gamma, p) + \eta_2\right) - \Pr\left(T_T^{appr}(\gamma) \leq q_{b_T}^{appr}(\gamma, p)\right)\right] \\
&+ \liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \left[\Pr\left(T_T^{appr}(\gamma) \leq q_{b_T}^{appr}(\gamma, p)\right) - \Pr\left(T_{b_T}^{appr}(\gamma) \leq q_{b_T}^{appr}(\gamma, p)\right)\right] \\
&+ \liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \Pr\left(T_{b_T}^{appr}(\gamma) \leq q_{b_T}^{appr}(\gamma, p)\right) \\
&\geq p,
\end{aligned} \tag{E.57}$$

where the first inequality holds by (E.56). The third inequality holds because the first two lim infs after the second inequality are greater than or equal to zero and the third is greater than or equal to p. The first lim inf is greater than or equal to zero by Theorem E.1. The second lim inf is greater than or equal to zero $T_{b_T}^{appr}(\gamma) \geq T_T^{appr}(\gamma)$ for any γ and T which holds because $\sqrt{T} \geq \sqrt{b_T}$ and $\Lambda_{b_T}(\theta, \gamma) \subseteq \Lambda_T(\theta, \gamma)$ for large enough T by Assumptions

C.1(c) and C.7(c).

Now it is left to show (E.56). In order to show (E.56), we first show that the c.d.f. of $\bar{T}_{b_T}^{appr}(\gamma)$ is close to the following empirical distribution function:

$$\hat{L}_{T,b_T}(x;\gamma) = S_T^{-1}\sum_{s=1}^{S_T} 1\left(\hat{T}_{T,b_T}^{s}(\gamma) \leq x\right). \tag{E.58}$$

Define an intermediate quantity first:

$$\tilde{L}_{T,b_T}(x;\gamma) = q_T^{-1}\sum_{l=1}^{q_T} 1\left(\tilde{T}_{T,b_T}^{l}(\gamma) \leq x\right), \tag{E.59}$$

where $q_T = \binom{T}{b_T}$ and $(\tilde{T}_{T,b_T}^{l}(\gamma))_{l=1}^{q_T}$ are the subsample statistics computed using all q_T possible subsamples of size b_T of the original sample. Conditional on the original sample, $(\hat{T}_{T,b_T}^{s}(\gamma))_{s=1}^{S_T}$ is S_T i.i.d. draws from $\tilde{L}_{T,b_T}(\cdot;\gamma)$. By the uniform Glivenko-Cantelli theorem, for any $\epsilon > 0$,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,sub}\left(\sup_{x\in R}\left|\tilde{L}_{T,b_T}(x;\gamma) - \hat{L}_{T,b_T}(x;\gamma)\right| > \epsilon\right) = 0 \tag{E.60}$$

It is implied by a Hoeffding's inequality (Theorem A on page 201 of Serfling (1980)) for U-statistics that for any real sequence $\{x_T\}$, and $\epsilon > 0$,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_F\left(\tilde{L}_{T,b_T}(x_T;\gamma) - \Pr_F\left(\tilde{T}_{T,b_T}^{l}(\gamma) \leq x_T\right) > \epsilon\right) = 0. \tag{E.61}$$

Equations (E.60) and (E.61) imply that, for any real sequence $\{x_T\}$ and $\epsilon > 0$,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,sub}\left(\hat{L}_{T,b_T}(x_T;\gamma) - \Pr_F\left(\tilde{T}_{T,b_T}^{l}(\gamma) \leq x_T\right) > \epsilon\right) = 0. \tag{E.62}$$

Apply Theorem E.1 on the subsample statistic $\tilde{T}_{T,b_T}^{l}(\gamma)$, and we have for any $\epsilon > 0$ and any real sequence $\{x_T\}$,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \left[\Pr_F\left(\tilde{T}_{T,b_T}^{l}(\gamma) \leq x_T - \epsilon\right) - \Pr\left(T_{b_T}^{appr}(\gamma) \leq x_T\right)\right] < 0. \tag{E.63}$$

Equations (E.62) and (E.63) imply that for any real sequence $\{x_T\}$,

$$\sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,sub}\left(\hat{L}_{T,b_T}(x_T;\gamma) > \left(\eta_2 + \Pr\left(T_{b_T}^{appr}(\gamma) \leq x_T + \eta_2\right)\right)\right) \to 0. \tag{E.64}$$

Plug $x_T = q^{appr}_{b_T}(\gamma, p) - 2\eta_2$ into the above equation and we have:

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr^*_{F,sub}\left(\hat{L}_{T,b_T}(q^{appr}_{b_T}(\gamma,p) - 2\eta_2;\gamma) > \eta_2 + p\right) = 0. \tag{E.65}$$

However, by the definition of $c^{sub}_T(\gamma, p)$, $\hat{L}_{T,b_T}(c^{sub}_T(\gamma,p) - \eta^*;\gamma) \geq p + \eta^* > \eta_2 + p$. Therefore

$$\limsup_{n\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr^*_{F,sub}\left(\hat{L}_{T,b_T}(q^{appr}_{b_T}(\gamma,p) - 2\eta_2;\gamma) \geq \hat{L}_{T,b_T}(c^{sub}_T(\gamma,p) - \eta^*;\gamma)\right) = 0, \tag{E.66}$$

which implies (E.56).

(b) Let $q^{bt}_{\kappa_T}(\gamma, p)$ be the p quantile of $T^{appr}_{\kappa_T}(\gamma)$ conditional on the original sample. Below we show that for $\eta_2 = \eta^*/3$,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,W}(c^{bt}_T(\gamma,p) < q^{bt}_{\kappa_T}(\gamma,p) + \eta_2) = 0. \tag{E.67}$$

where $\Pr_{F,W}$ signifies the fact that there are two sources of randomness in $c^{bt}_T(\gamma, p)$, that from the original sampling and that from the bootstrap sampling. Once (E.67) is established, we have,

$$\begin{aligned}
\liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,W}\left(\hat{T}_T(\gamma) \leq c^{bt}_T(\gamma,p)\right) &\geq \liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \Pr_F\left(\hat{T}_T(\gamma) \leq q^{bt}_{\kappa_T}(\gamma,p) + \eta_2\right) \\
&\geq \liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \Pr\left(T^{appr}_T(\gamma) \leq q^{bt}_{\kappa_T}(\gamma,p)\right) \\
&\geq \liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \Pr\left(T^{appr}_{\kappa_T}(\gamma) \leq q^{bt}_{\kappa_T}(\gamma,p)\right) \\
&= p,
\end{aligned} \tag{E.68}$$

where the first inequality holds by (E.67), the second inequality holds by Theorem E.1 and the third inequality holds because $T^{appr}_{\kappa_T}(\gamma) \geq T^{appr}_T(\gamma)$ for any γ and T which holds because $\sqrt{T} \geq \sqrt{\kappa_T}$ and $\Lambda_{\kappa_T}(\theta,\gamma) \subseteq \Lambda_T(\theta,\gamma)$ for large enough T by Assumptions C.1(c) and C.7(c).

Now we show (E.67). First, we show that the c.d.f. of $T^{appr}_{\kappa_T}(\gamma)$ is close to the following empirical distribution:

$$F_{S_T}(x,\gamma) = S_T^{-1}\sum_{l=1}^{S_T} 1\{T^*_{T,l}(\gamma) \leq x\}, \tag{E.69}$$

where $\{T^*_{T,1}(\gamma), ..., T^*_{T,S_T}(\gamma)\}$ are the S_T conditionally independent copies of the bootstrap test statistics. By the uniform Glivenko-Cantelli Theorem, $F_{S_T}(x,\gamma)$ is close to conditional

c.d.f. of $T_T^*(\gamma)$: for any $\eta > 0$

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,W}\left(\sup_{x\in R} |F_{S_n}(x,\gamma) - \Pr_W(T_T^*(\gamma) \le x)| > \eta\right) = 0. \tag{E.70}$$

The same arguments as those for Theorem E.1 can be followed to show that $T_T^*(\gamma)$ is close in law to $T_{\kappa_T}^{appr}(\gamma)$ in the following sense: for any real sequence $\{x_T\}$,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_F\left(\left[\Pr_W(T_T^*(\gamma) \le x_T - \eta_2) - \Pr(T_{\kappa_T}^{appr}(\gamma) \le x_T)\right] \ge \eta_2\right) = 0. \tag{E.71}$$

When following the arguments for Theorem E.1, we simply need to observe the resemblance between $\hat{T}_T(\gamma)$ and $T_T^*(\gamma)$ in the following form:

$$T_T^*(\gamma) = \min_{\theta\in\Theta_{0,F}(\gamma)} \min_{\lambda\in\Lambda_{\kappa_T}(\theta,\gamma)} \int_{\mathcal{G}} S(\hat{\nu}_T^{*+}(\theta + \lambda/\sqrt{T}, g) + G_F(\tilde{\theta}_T, g)\lambda + \sqrt{\kappa_T}\rho_F(\theta, g), \hat{\Sigma}_n(\theta + \lambda/\sqrt{T}, g))d\mu(g), \tag{E.72}$$

where

$$\hat{\nu}_T^{*+}(\theta, g) = \hat{\nu}_T^*(\theta, g) + \kappa_T^{1/2} n^{-1/2} \hat{\nu}_T(\theta, g), \tag{E.73}$$

and use Lemma E.2 in conjunction with Assumptions C.2(c) and use Lemma E.1(b) in place of E.1(a).

Equations (E.70) and (E.71) together imply that for any real sequence $\{x_T\}$,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,W}\left(\left[F_{S_T}(x_T - \eta_2, \gamma) - \Pr(T_{\kappa_T}^{appr}(\gamma) \le x_T)\right] \ge 2\eta_2\right) = 0. \tag{E.74}$$

Plug in $x_T = q_{\kappa_T}^{appr}(\gamma, p) - \eta_2$ and we have

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,W}\left(F_{S_T}(q_{\kappa_T}^{appr}(\gamma, p) - 2\eta_2, \gamma) \ge p + 2\eta_2\right) = 0. \tag{E.75}$$

But by definition, $F_{S_T}(c_T^{bt}(\gamma, p) - \eta^*, \gamma) \ge p + \eta^* > p + 2\eta_2$. Therefore,

$$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \Pr_{F,W}\left(F_{S_T}(q_{\kappa_T}^{appr}(\gamma, p) - 2\eta_2, \gamma) \ge F_{S_T}(c_T^{bt}(\gamma, p) - \eta^*, \gamma)\right) = 0, \tag{E.76}$$

which implies (E.67). □

F Monte Carlo Analysis for Selecting Tuning Parameters in Application

In this section we conduct a Monte Carlo simulation that exhibits the same pattern of sparse demand as we observe in the scanner data with the goal of helping to guide our choice of the tuning parameters required by the bootstrap in our profiling approach for our empirical study. In particular, we design a data generating process for the simulated data that reproduces the main stylized features of interest from the data. Recall that our empirical analysis focused on data from a single stores, and hence different markets are indexed by different weeks. We simulate $J_t = 50$ products and $n_t = 15000$ consumers for each of $t = 1, \ldots, T = 100$, which closely matches the structure of the data we will study in the next subsection. Each product is characterized by a single observable product attribute x (quality or negative of price) and an unobserved attribute ξ. Consumer $i = 1, \ldots, n_t$ in market t has utility for a product j given by

$$u_{ijt} = \alpha_0 + \beta_0 x_{jt} + \xi_{jt} + v_{ijt} \quad j = 1, \ldots, J_t$$

and $u_{i0t} = v_{i0t}$, where v_{ijt} are i.i.d. type-I extreme value. The parameter of interest is β_0, which has a true value of 1.

The data generating process for the observable x_{jt} and the unobservable ξ_{jt} were designed to replicate several important patterns from the data. There are three main elements to the data generating process, which were designed to match key features of the data.

1. There are two types of dependence across *product/week* in our data set: products in the same market (i.e., products marketed on the same week at the store) are affected by common market level features, and both the market level features and product specific characteristics may persist over weeks. We capture the within market dependence by letting x_{jt} to have two components: $x_{jt} = \bar{x}_{jt} + e_t$, where $\bar{x}_{jt}$ is the product specific characteristic that is independent across j and e_t is the market level feature. We capture the intertemporal dependence by a stationary AR(2) structure on both $\bar{x}_{jt}$ and e_t with the AR coefficients being 0.5 and 0.4 for the first and the second lag terms respectively. We also impose an AR(2) structure with the same AR coefficients on a component of the unobserved product characteristics ξ_{jt} which is discussed below. The AR coefficients roughly matches the intertemporal dependence in the data.[41]

2. As discussed above, the scanner data generally exhibits two groups of products: the first group being products that have larger choice probabilities and more certain to

[41] The AR(2) coefficient estimates for the time series $\sum_{j \leq J_t : s_{jt}, s_{0t} \in (0,1)} (\log s_{jt} - \log s_{0t}) / \sum_{j \leq J_t} 1\{s_{jt}, s_{0t} \in (0,1)\}$ are roughly 0.5 and 0.4.

have positive sales across all markets; the second group being the "long tail" of products with smaller choice probabilities and are more sparsely demanded. In particular, because the data for the category we study below shows that roughly half the products are susceptible to zero sales whereas the other half is not, we incorporate this feature into the simulation by setting the stationary mean of $\bar{x}_{jt}$ for the first $j = 1, \ldots, 25$ products to be 8 and the stationary mean of the last $j = 26, \ldots, 50$ products to be 4. The innovation terms in the AR process for $\bar{x}_{jt}$ of the first 25 products are set to be $N(0, 0.2)$ and those for the last 25 are set to be $N(0, 0.4)$. The stationary mean of the common shock e_t is set at 0 and its innovation terms $N(0, 0.1)$.[42]

3. Finally, for the products in our "long tail" category, we want to control the extent of zero sales to match what we observe in the sub-sample of the data we study. We accomplish this last goal using heteroskedasticity in ξ_{jt} that is natural to our setting. We take $\xi_{jt} = \bar{\xi}_{jt}\tilde{e}_{jt}$ where $\bar{\xi}_{jt}$ is a constant representing the standard deviation and $\tilde{e}_{jt}$ is a normalized stochastic error term. The heteroskedastic standard deviation of ξ_{jt} is set to be $\bar{\xi}_{jt} = 0.1 \times 1\{x_{jt} > 5\} + \xi_c \times 1\{x_{jt} < 5\}$, where ξ_c fully control the extent of zero sales. We find that $\xi_c = 4 - 10$ generates proportion of zero shares similar to our real data. The random (conditional on x_{jt}) component, $\tilde{e}_{jt}$, of ξ_{jt} follows the same distribution as the series e_t above for each j and is independent across j. This construction of the error distribution and the distribution of x_{jt} imply that an observation of a "zero" sale among the long tail products is largely driven by a bad ξ_{jt} shock, which accords with the interpretation of ξ_{jt} in our application (namely the unobserved promotion of a product through advertising and shelf space). The persistence of $\tilde{e}_{jt}$ means these bad shocks can be persistent, i.e., bad shelf space can persist for a few weeks. It also implies that the products are segmented into "hit" products that rarely experience zero sales and the tail products that often do. Our design enables us to produce 17 to 24 products (across different specifications) that never experience zero sales, which matches the structure of our real data.

We choose the constant c in the tuning parameter $\kappa_T = T/(c \log T)$ for the bootstrap by, for different values of c, replicating the above simulation 1000 times and computing the coverage probability (CP) of the true value $\beta_0 = 1$ and the false coverage probability (FCP) of a point outside the identified set of β. We find that the CP's are always 1, showing that our confidence set does not under cover, even for the independence bootstrap (despite the fact there is temporal dependence in the simulated data). The FCP's are shown in Table 6

[42] To generate a stationary AR(2) process Y_t with stationary mean $E(Y_t)$ and innovation term u_t, we start with $Y_{-1} = Y_0 = E[Y_t]$ and generate Y_t for $t > 0$ according to the AR(2) equation. We burn the first 20,000 terms in the process and take the $20,001$th term to the $20,100$th term as our Monte Carlo data set. For repetition study, we repeat the whole process.

below. As the table shows, at $c = 0.5 - 0.6$ our confidence sets appear to exhibit the best FCP's and we will thus focus on these values of the tuning parameter for the remainder of paper.[43]

Table 6: False Coverage Probabilities (FCP) of the 95% Confidence Interval

Block Length	1 (i.i.d.)				3				5			
$c \backslash \xi_c$	4	6	8	10	4	6	8	10	4	6	8	10
.3	.908	.959	.990	.980	.939	.972	.991	.990	.953	.984	.991	.990
.4	.387	.567	.583	.593	.528	.709	.776	.779	.610	.772	.830	.839
.5	.318	.443	.407	.396	.469	.611	.593	.607	.560	.711	.703	.710
.6	.348	.474	.403	.402	.504	.635	.579	.600	.619	.753	.714	.725
.7	.432	.541	.470	.476	.584	.705	.642	.677	.715	.822	.779	.789
.8	.500	.633	.553	.575	.689	.797	.732	.753	.812	.880	.850	.846
.9	.578	.711	.635	.658	.763	.875	.810	.834	.874	.936	.910	.926

Note: The FCPs are computed at .95, .94, .91, .89 for $\xi_\sigma = 4, 6, 8, 10$, respectively. These numbers are chosen to yield nontrivial FCPs.

We produce one additional replication of the DGP above for different values of the heteroskedasticity parameter ξ_c and present the confidence intervals obtained using our moment inequalities method in the second column of Table 7 below.

Table 7: Monte Carlo Results: Point and Bound Estimates

ξ_c	Percent of Positive Shares	Bounds Estimate Using Block Bootstrap: 50% (top) and 95% (bottom) CS's			BLP Point Estimate and 95% CS
		Block Length			
		1(i.i.d.)	3	5	
4	83.20%	[.97, 1.02] [.96, 1.05]	[.97, 1.03] [.95, 1.06]	[.97, 1.03] [.95, 1.06]	.70 [.66, .74]
6	80.38%	[.96, 1.03] [.93, 1.06]	[.96, 1.03] [.93, 1.07]	[.96, 1.03] [.92, 1.07]	.50 [.44, .55]
8	78.46%	[.93, 1.10] [.89, 1.26]	[.93, 1.10] [.89, 1.26]	[.93, 1.10] [.88, 1.26]	.32 [.24, .39]
10	76.06%	[.89, 1.10] [.85, 1.29]	[.89, 1.11] [.84, 1.38]	[.89, 1.12] [.84, 1.35]	.15 [.08, .23]

Note: True value = 1, $J = 50$, $T = 100$, $\kappa_T = T/(0.5 \cdot \log(T))$, ε_{jt} =minimum true share $\forall j, t$. The confidence intervals of the BLP estimates are constructed using the Driscoll-Kraay standard errors.

[43] Another implementation details for our inference strategy are the set of g functions, We follow the suggestions in Section 5.2 and let $\bar{r}_T = 50$, which yields on average 50 product/markets in each of the smallest hyperboxes. Lastly, we use the true value of the minimum choice probability in each replication of the simulated data as our value of the lower bound on choice probabilities ε_t for our moment inequality method.

As can be seen, our confidence intervals based on the moment inequalities always contain the true value and for the whole range of ξ_c. Moreover, our confidence intervals are fairly informative even when the degree of heteroskedasticity and hence selection in the data as determined by ξ_c is large. Also these results mimic a key finding from our empirical analysis - the BLP point estimates from dropping the zeroes have price coefficients severely biased towards zero.

G Handling Dependence

The inference theory presented so far assumes that the data $\{w_t\}$ are i.i.d. We show that it is straightforward to generalize to stationary and weakly-dependent time series data. The modifications needed are as follows.

First, with time-series data, the variance estimator $\hat{\Sigma}_T(\theta, g)$ in (5.8) is no longer appropriate. One can replace it with, for example, the kernel estimator proposed in Andrews (1991):

$$\begin{aligned}\hat{\Sigma}_n(\theta, g) &= \frac{T}{T-1} \sum_{j=-T+1}^{T-1} K\left(\frac{j}{B_T}\right) \hat{\Gamma}(j), \text{ where} \\ \hat{\Gamma}(j) &= \begin{cases} \frac{1}{T}\sum_{t=j+1}^{T} (\rho(w_t, \theta, g) - \bar{\rho}_T(\theta, g))(\rho(w_{t-j}, \theta, g) - \bar{\rho}_T(\theta, g))' & j \geq 0 \\ \frac{1}{T}\sum_{t=-j+1}^{T} (\rho(w_{t+j}, \theta, g) - \bar{\rho}_T(\theta, g))(\rho(w_t, \theta, g) - \bar{\rho}_T(\theta, g))' & j < 0, \end{cases}\end{aligned} \tag{G.1}$$

where $K(\cdot)$ is a kernel function and B_T is a bandwidth parameter. The optimal choice of both are given in Andrews (1991). One can also use the prewhitening and recoloring procedure to obtain less biased estimator following Andrews and Monahan (1992). The kernel estimator $\hat{\Sigma}_n(\theta, g)$ converges to the long-run variance of $\{\rho(w_t, \theta, g)\}$:

$$\Sigma_F(\theta, g) = \lim_{T\to\infty} \text{Var}\left(T^{-1/2} \sum_{t=1}^{T} (\rho(w_t, \theta, g) - \rho_F(\theta, g))\right). \tag{G.2}$$

In addition, the Gaussian process $\nu_F(\theta, g)$ to which the process $\hat{\nu}_T(\theta, g) \equiv \sqrt{T}(\bar{\rho}_T(\theta, g) - \rho_F(\theta, g))$ now converges weakly has covariance kernel

$$\Sigma_F(\theta^{(1)}, g^{(1)}, \theta^{(2)}, g^{(2)}) := \lim_{T\to\infty} \text{Cov}\left(\sqrt{T}\bar{\rho}_T(\theta^{(1)}, g^{(1)}), \sqrt{T}\bar{\rho}_T(\theta^{(2)}, g^{(2)})\right), \tag{G.3}$$

instead of the covariance kernel defined in (C.1).

Second, the subsampling and bootstrap procedures described in Section 5.3 may not be consistent. One needs to change the resampling procedure to accommodate the dependence. For subsampling, in step [1], instead of taking draw a subsample of size b_T, one draws

randomly from the following $T-b_T+1$ possible "block" subsamples: $\{\{1,...,b_T\},\ \{2,...,b_T+1\},...,\{T-b_T+1,...,T\}\}$. The other steps stay the same. For bootstrap, one can use the block bootstrap procedure, i.e., in step [1], instead of drawing an i.i.d. sample of size T, first divide the time series into blocks of subsamples: $\{\{1,...,b_T\},\ \{2,...,b_T+1\},...,\{T-b_T+1,...,T\}\}$, draw$\lceil T/b_T\rceil$ i.i.d. blocks from this set of blocks, assemble these blocks one after another to form a sample of size $b_T\times\lceil T/b_T\rceil$ and finally truncate the last $b_T\times\lceil T/b_T\rceil-T$ terms to form a sample of size T.

Third, some of the assumptions also need to be changed to accommodate the dependence. Specifically, we delete part (a) of Assumption C.2 and change parts (c) and (d) to:

(c) $\lim_{T\to\infty}\sup_{(\gamma,F)\in\mathcal{H}}\sup_{f\in BL_1}|E_F^* f(\hat{\nu}_T(\cdot,\cdot))-Ef(\nu_F(\cdot,\cdot))|=0$,
$\sup_{(\gamma,F)\in\mathcal{H}}E^*\sup_{(\theta,g)\in\Theta\times\bar{\mathcal{G}}}\|\nu_F(\theta,g)\|<\infty$ and
$\lim_{\delta\downarrow 0}\sup_{(\gamma,F)\in\mathcal{H}}E^*\sup_{g,||\theta-\theta^*||\leq\delta}\|\nu_F(\theta,g)-\nu_F(\theta^*,g)\|=0$.

(d) $\lim_{T\to\infty}\sup_{(\gamma,F)\in\mathcal{H}}\Pr_F^*(\sup_{(\theta,g)\in\Theta\times\bar{\mathcal{G}}}\|\hat{\Sigma}_T(\theta,g)-\Sigma_F(\theta,g)\|>\epsilon)=0$ for any $\epsilon>0$.

In addition, for the subsampling critical value, the Hoeffding's inequality used to show (E.61) no longer applies. We change q_T in (E.59) to $T-b_T+1$, which now is the number of all possible subsamples and add the following assumption:

Assumption G.1. $\limsup_{T\to\infty}\sup_{(\gamma,F)\in\mathcal{H}_0}\Pr_F\left(\tilde{L}_{T,b_T}(x_T;\gamma)-\Pr_F\left(\tilde{T}^l_{T,b_T}(\gamma)\leq x_T\right)>\epsilon\right)=0$.

For the bootstrap critical value, Lemma E.2 no longer goes through. We thus add the following assumption:

Assumption G.2. (a)$\limsup_{T\to\infty}\sup_{(\gamma,F)\in\mathcal{H}}\Pr{}^*_F(\sup_{f\in BL_1}|E_W f(\hat{\nu}_T^*(\cdot,\cdot))-Ef(\nu_F(\cdot,\cdot))|>\epsilon)=0$,

(b) *there exists B_ϵ large enough such that*

$$\limsup_{T\to\infty}\sup_{(\gamma,F)\in\mathcal{H}}\Pr{}^*_F\left(\Pr{}_W\left(\sup_{\theta\in\Gamma^{-1}(\gamma),g\in\bar{\mathcal{G}}}||\hat{\nu}_T^*(\theta,g)||>B_\epsilon\right)>\epsilon\right)=0,\ \textit{and}$$

(c) *there exists δ_ϵ small enough such that*

$$\limsup_{T\to\infty}\sup_{(\gamma,F)\in\mathcal{H}}\Pr{}^*_F\left(\Pr{}_W\left(\sup_{g\in\bar{\mathcal{G}}}\sup_{||\theta^{(1)}-\theta^{(2)}||\leq\delta_\epsilon}||\hat{\nu}_T^*(\theta^{(1)},g)-\hat{\nu}_T^*(\theta^{(2)},g)||>\epsilon\right)>\epsilon\right)=0.$$

Then we replace Assumption C.2 in the statement of Theorems 2 and E.1 by its modified version and add Assumption G.1 to the conditions for Theorem 2(a) and Assumption G.2 to the conditions of Theorem 2(b).

Lastly, we adapt the proofs of Theorems 2 and E.1. The following adaptions are sufficient:

(1) In the fourth line below equation (E.31), change "Theorem 2.8.2" to "the (i)→(ii) part of the proof of Theorem 2.8.2".

(2) In the proof for Theorem 2(a), instead of using the Hoeffding's inequality to show (E.61), use Assumption G.1 directly.

(3) In the proof for Theorem 2(a), instead of using Lemma E.2, use Assumption G.2 directly.

Assumptions G.1 and G.2 appear to be high-level, but they are easy to verify under standard time series assumptions. To verify Assumption G.1, one can use Theorem 3.1 of Politis and Romano (1994) under the assumption that $\{w_t\}$ is stationary and α-mixing. To verify Assumption G.2, one can use Theorem 4.1 of Bühlmann (1996) under the assumption that $\{w_t\}$ is β-mixing and $\{\rho(w, \theta, g) : (\theta, g) \in \Theta \times \mathcal{G}\}$ is a Vapnik-Červonenkis class with an envelope function that has exponentially decaying tail. For brevity, we omit the details.

References

Abrevaya, J., and J. A. Hausman (2004): "Response error in a transformation model with an application to earnings-equation estimation," *The Econometrics Journal*, 7, 366–388.

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

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Andrews, D. W. K., and J. C. Monahan (1992): "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953–966.

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

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

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

Berry, S., M. Carnall, and P. Spiller (1996): "Airline hubs: costs, markups and the implications of customer heterogeneity," Discussion paper, National Bureau of Economic Research.

Berry, S., A. Gandhi, and P. Haile (2011): "Connected Substitutes and Invertibility of Demand," Discussion paper, National Bureau of Economic Research.

Berry, S., and P. Jia (2010): "Tracing the Woes: An Empirical Analysis of the Airline Industry," *American Economic Journal: Microeconomics*, 2, 1–43.

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.

Blundell, R., and J. Powell (2003): "Endogeneity in nonparametric and semiparametric regression models," *ECONOMETRIC SOCIETY MONOGRAPHS*, 36, 312–357.

Brand, K., G. Gowrisankaran, A. Nevo, and R. Town (2012): "Mergers When Prices Are Negotiated: Evidence from the Hospital Industry," Discussion paper, University of Arizona working paper.

Briesch, R., W. Dillon, and R. Blattberg (2008): "Treating zero brand sales observations in choice model estimation: Consequences and potential remedies," *Journal of Marketing Research*, 45(5), 618–632.

Bühlmann, P. L. (1996): "The Blockwise Bootstrap in Time Series and Empirical Processes," dissertation, Swiss Federal Institute of Technology Zürich.

Chen, Y., and S. Yang (2007): "Estimating Disaggregate Models Using Aggregate Data Through Augmentation of Individual Choice," *Journal of Marketing Research*, XLIV, 613–621.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284.

Chernozhukov, V., S. Lee, and A. Rosen (2008): "Inference with Intersection Bounds," unpublished manuscript, Department of Economics, University College London.

Chintagunta, P., J. Dube, and K. Goh (2005): "Beyond the endogeneity bias: The effect of unmeasured brand characteristics on household-level brand choice models," *Management Science*, pp. 832–849.

Chintagunta, P. K. (2000): "A Flexible Aggregate Logit Demand Model," Working Paper, University of Chicago.

Chintagunta, P. K., and J.-P. D. S. Vishal (2003): "Balancing Profitability and Customer Welfare in a Supermarket Chain," *Quantitative Marketing and Economics*, 1, 111–147.

Driscoll, J. C., and A. C. Kraay (1998): "Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data," *Review of Economics and Statistics*, 80, 549–560.

Good, I. J. (1983): *Good Thinking: the Foundations of Probability and its Applications*. the University of Minnesota Press, 1 edn.

Goolsbee, A., and P. Klenow (2006): "Valuing consumer products by the time spent using them: An application to the Internet," Discussion paper, National Bureau of Economic Research.

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

Gutknecht, D. (2012): "Nonclassical Measurement Error in the Dependent Variable a Nonlinear Model," Discussion paper, Department of Economics, University of Warwick.

Hausman, J., and G. Leonard (2002): "The Competitive Effects of a New Product Introduction: A Case Study," *The Journal of Industrial Economics*, 50, 237–263.

Hoch, S. J., B.-D. Kim, A. L. Montgomery, and P. E. Rossi (1995): "Determinants of Store-Level Price Elasticity," *Journal of Marketing Research*, 32, 17–29.

Hosken, D., D. O'Brien, D. Scheffman, and M. Vita (2002): "Demand System Estimation and its Application to Horizontal Merger Analysis," Discussion paper, Federal Trade Commission, Bureau of Economics.

Israilevich, G. (2004): "Assessing Supermarket Product-Line Decisions: The Impact of Slotting Fees," *Quantitative Marketing and Economics*, 2, 141–167.

Misra, S., and S. Mohanty (2008): "Estimating Bargaining Games in Distribution Channels," Working Paper, University of Rochester.

Nair, H., and J.-P. D. P. Chintagunta (2005): "Accounting for Primary and Secondary Demand Effects with Aggregate Data," *Marketing Science*, 24, 444–460.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. McFadden, vol. 4 of *Handbook of Econometrics*, chap. 36, pp. 2111–2245. Elsevier.

Park, S., and S. Gupta (2009): "Simulated Maximum Likelihood Estimator for the Random Coefficient Logit Model Using Aggregate Data," *Journal of Marketing Research*, 46(4), 531–542.

Politis, D. N., and J. P. Romano (1994): "Large Sample Confidence Regions Based on Subsampling Under Minimal Assumptions," *The Annals of Statistics*, 22, 2031–2050.

Rockafellar, R. T. (1970): *Convex Analysis*. Princeton University Press.

Romano, J., and A. Shaikh (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models," *Journal of Statistical Planning and Inference*, (Special Issue in Honor of T. W. Anderson, Jr. on the Occasion of his 90th Birthday), 138, 2786–2807.

Romeo, C. J. (2005): "Estimating Discrete Joint Probability Distributions for Demographic Characteristics at the Store Level Given Store Level Marginal Distributions and a City-Wide Joint Distribution," *Quantitative Marketing and Economics*, 3, 71–93.

Santos, A. (2012): "Inference in Nonparametric Instrumental Variable with Partial Identification," *Econometrica*, 80, 213–275.

Serfling, R. J. (1980): *Approximation Theorems in Mathematical Statistics*. John Wiley and Sons, INC.

van der Vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer.

Werden, G. J., and L. M. Froeb (1994): "The Effects of Mergers in Differentiated Products Industries: Logit Demand and Merger Policy," *Journal of Law, Economics, & Organization*, 10, 407–426.

Wood, G. R. (1999): "Binomial Mixtures: Geometric Estimation of the Mixing Distribution," *The Annals of Statistics*, pp. 1706–1721.