Estimating Discrete Choice Models of Demand with Market Level Variation

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

June 1, 2012

Abstract

In this paper we develop and apply a new empirical approach to estimating a widely used class of models of demand for differentiated products. Our approach is applicable to the often encountered situation where the researcher has data on a sample of markets (i.e., the "many markets" setting). Our focus on the "many markets" setting stands in contrast to the "many products" setting that was famously studied by Berry, Levinsohn, and Pakes (1995). We first show that the key identification restriction underlying the Berry, Levinsohn, and Pakes (1995) approach to estimation does not have *any* identifying power in the many markets setting without further restrictions – the essential problem being the presence of sampling variability in market shares. Furthermore, their inference methods do not apply because of the presence of strategic dependence among products within a market. We show that if choice probabilities are bounded away from zero, the standard consumer model implies enough demand restrictions that we can use to form a system of moment inequalities which hold at the market level. We also construct a profiling approach for parameter inference with moment inequalities, which allows us to study models with a large number of parameters (as typically required in demand applications) by focusing attention on a generalized profile of the parameters, such as the price coefficient. We use our approach to study UPC level demand on scanner data from the Dominick's Fine Foods database, and find that even for the baseline logit model, demand elasticities nearly double when the "many markets" structure of the data is taken into account.

Keywords: Demand Estimation, Differentiated Products, Moment Inequality, Subvector Inference, Generalized Profile.

JEL: C01, C12, L10, L81.

[*] We are thankful to Jean-Pierre Dube, Philip Haile, Bruce Hansen, Ulrich Müller, Aviv Nevo, and Jack Porter 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 of "Junior Festival on New Developments in Microeconometrics", as well as seminar participants at Wisconsin-Madison, Wisconsin-Milwaukee, Cornell, Indiana, Princeton for their many helpful comments and questions.

1 Introduction

In this paper we develop and apply a theory for demand estimation for differentiated products that applies to the case where the researcher has data on demand from a sample of different markets (such as spatially and/or temporally separated markets). Markets level variation is the classic setting used to identify and estimate demand for homogeneous products and also underlies the standard textbook simultaneous equations supply and demand model. Beyond the textbook example of homogenous goods, variation in markets is also a standard feature of data from differentiated goods industries.[1] However the empirical problem of estimating discrete-choice demand models when the source of variation in the data is at the market level (which we shall call the "many markets" setting) remains an open issue, which is the key void the present paper aims to fill.

The fundamental difficulty for empirical work with many markets is that both the number of consumers and the number of products observed within a market are fixed attributes of the market, i.e., these attributes only vary *across* markets but not *within* a market. This stands in stark contrast to the "many products" case that was formally studied by Berry (1994), Berry, Levinsohn, and Pakes (1995), and Berry, Linton, and Pakes (2004) (we shall reference these papers collectively as BLP, and the last paper as BLintonP), where they assume a single large market in which both the number of products and consumers in the market grow with the sample size. Their focus on a single large market allows BLP to abstract away from two essential problems that we must directly confront in the many markets case.

First, a fundamentally new identification problem comes to light in the many markets case. Identification in BLP is based on a conditional mean restriction between a product's unobservable and a vector of instruments; applying this restriction to the data however requires that the choice probabilities predicted by the model be non-linearly inverted (in a parameter dependent way) so that each product's unobservable becomes a separable econometric error. Unfortunately, choice probabilities are never observed in applications. Instead, empirical market shares are observed, which by construction are noisy measures of choice probabilities. Thus the standard practice of inverting market shares introduces a source of nonlinear measurement error that undermines identification.[2] We show that this

[1] Variation in demand data across markets arises in the studies of cereal industry (e.g. Nevo (2000), Nevo (2001)), hospital demand (e.g. Capps, Dranove, and Satterthwaite (2003), Ho (2007)), yellow pages demand (e.g. Rysman (2004)), yogurt industry (e.g. Villas-Boas (2007)), newspaper industry (e.g. Fan (2008)), health insurance market(e.g. Lustig (2008)), airline market(e.g. Berry and Jia (2010)), etc.

[2] BLP implicitly avoid this problem in the "many products" environment by assuming (not without loss) that the number of consumers grows at a sufficiently rapid rate with the number of products in the market, and thus sampling variability vanishes at a controlled rate in the limit (where the limit is instead taken with respect to the number of products). Such assumptions cannot be easily made in the many markets case we consider.

problem is a severe one: in the many market setting, observing only empirical market shares rather than the choice probabilities themselves causes the conditional mean restriction to generically lose *all* of its identifying content for the demand parameters.

Our approach to solving this identification problem is to add information from the model to the observed market shares, which in turn restores the identifying power of the conditional mean restriction. We derive the additional information from the restrictions on market demand implied by the underlying consumer choice model. The only shape restriction on market demand that the BLP approach exploits is that demand is invertible. However discrete choice models of market demand impose shape restrictions beyond invertibility (Berry, Gandhi, and Haile (2011)). In particular, we show that the monotone structure of inverse demand allows us to form a system of conditional moment *inequalities* from the observed data. Because our approach does not require any new moment conditions or instrumental variable assumptions relative to the standard demand model already used in the literature, it is not surprising that we cannot generally achieve point identification of the demand parameters in the presence of non-linear measurement error. Nevertheless, the moment inequalities we do construct are *adaptive* to the information revealed by the empirical market shares (in a precise sense that we define), which makes our partial identifications approach especially useful for applications.

A second fundamental difference between the "many markets" and the "many products" setting is that with many markets, strategic interaction among the fixed and relatively small number of products within a market will generally cause the attributes of products *within* the market to be dependent in a non-standard way. In the many product setting of BLP, it is assumed the true underlying data generating process is such that the products have unobservable attributes that are *independent* of one another. Instead, we can allow for arbitrary strategic interaction and dependencies among product unobservables within a market by exploiting the variation across markets. In particular we show that our conditional moment inequality can be aggregated without information loss to the market level. Then using standard independence or weak dependence assumptions on the sampling of markets in the data, we can appeal to the literature on inference with moment inequalities to conduct inference with these inequalities.

Our third major contribution in the paper is that we provide a generalized profiling procedure for inference with moment inequalities that is applicable to the moment inequalities literature more generally, and particularly to our demand estimation context. The existing inference method for our moment inequality setup (e.g. Andrews and Shi (2009)) requires exhaustive grid search over the parameter space. The computational cost is particularly high for demand studies because at least a moderate number of control variables are needed to ensure validity of the instrument for price. Generalized profiling of moment inequalities allows us to circumvent this computational burden by performing inference directly on a generalize profile of the parameters, i.e., a function of the parameters that capture the policy relevant objects of interest, such as elasticity and welfare. The model parameters themselves are treated as nuisance parameters and profiled out when conducting the inference. The idea of profiling out nuisance parameters has received little attention in the partial-identification literature. Romano and Shaikh (2008) are to our knowledge the only one to suggest applying profiling to moment inequality models. They also show the validity of a subsampling procedure under high-level conditions, but the high-level conditions are not straightforward to verify. We fill in this void by showing the uniform validity of subsampling under low level conditions. More importantly, we design a bootstrap alternative to subsampling that is also uniformly valid and easy to implement.

We apply our inference strategy to the Dominick's Fine Foods (DFF) database, which is a publicly available and heavily studied source of data on consumer demand. An important feature of the data that has proven to be both a strength and weakness for empirical work is its high frequency, both in the product dimension and time dimension. Products are defined in the data at the bar-code or "UPC" level, which gives rise to a substantial number of available products even within narrowly defined product categories. Furthermore, the sales information is available at the weekly level, which is the time horizon over which the grocery store chain makes its pricing and promotion decision. These two aspects of the data combined to give rise to a phenomenon that draws particular attention to the problem of sampling variability in shares: many observations exhibit in the data zero sales. Because the standard BLP approach cannot explain zero sales in the data, this has required researchers to either ignore UPC's with zero demand (and put them into the definition of the outside good and thereby introduce a selection problem that we explain), or ignore UPC's altogether and instead aggregate UPC's into a composite products (such as a brand) that is no longer representative of the consumer problem. However zeroes are natural under our approach to the data: they are simply the outcome of sampling variability when the underlying choice probabilities are small. Applying our inference strategy to the data from the bath tissue category, which is a category that has been considered by several other papers, we find that demand becomes almost twice as elastic as compared to estimating demand using existing techniques that ignore sampling variability in shares.

2 Model

A market t consists of 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 a vector of observable demand shifters $x_t = (x_{1t}, \dots, x_{J_t t}) \in X$, where each $x_{it} \in \mathbb{R}^K$ for $i = 1, \dots, J_t$ is a vector of product attributes (typically including price) corresponding to the inside products. Let $\xi_t = (\xi_{1t}, \dots, \xi_{J_t t}) \in \mathbb{R}^{J_t}$ denote a vector of demand shocks, where each ξ_{it} for $i = 1, \dots, J_t$ is typically interpreted as the unobservable (to the econometrician) attribute of each inside product. Each market t also consists of a certain number of consumers n_t.[3]

The demand of each consumer $i = 1, \dots, n_t$ in market t is described by a random utility model. For simplicity, we use the standard linear in random coefficients random utility model employed by Berry (1994), but the ideas of this paper extend in a straightforward way to more general specifications. The utility to consumer i for product $j = 0, \dots, J_t$ in market t is

$$u_{ijt} = \delta_{jt} + \nu_{ijt}, \tag{2.1}$$

where

1. $\delta_{jt} = \beta_0 x_{jt} + \xi_{jt}$ 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 $\delta_t = (\delta_{1t}, \dots, \delta_{J_t t})$ denote the vector of mean utilities of the "inside" goods $j > 0$.

2. The vector $\nu_{it} = (\nu_{i0t}, \dots, \nu_{iJ_t t}) \sim F(\cdot \mid x_t; \lambda_0)$ is the random vector of tastes in market t. We will assume for simplicity that the random vector ν_{it} 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 ν_{ijt} is an idiosyncratic preference shock with full support (as in the mixed logit model or probit models), then full support of ν_{it} holds.[4]

3. The vector $\theta_0 = (\beta_0, \lambda_0) \in \Theta$ denotes the true value of the parameters, where Θ is a finite dimensional parameter space.

The random utility model can be aggregated to yield a system of choice probabilities

$$\pi_{jt} = \sigma_j(\delta_t, x_t; \lambda_0) \quad j = 0, 1, \dots, J_t, \tag{2.2}$$

and π_{jt} is the choice probability that a randomly sampled consumer from market t whose tastes are drawn from $F(\cdot \mid x_t; \lambda_0)$ would maximize utility by choosing good j. Let $\pi_t = (\pi_{0t}, \pi_{1t}, \dots, \pi_{J_t t})$ denote the vector of choice probabilities predicted by the random utility model.

[3] 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.

[4] The main role of the full support assumption is computational convenience but is nevertheless a a useful assumption to maintain for the exposition. We could in principle proceed instead under the weaker "connected substitutes" structure of Berry, Gandhi, and Haile (2011).

The econometrician observes the aggregate demand of the n_t consumers in the market, which can be represented as a market share s_{jt} for $j = 0, 1, \ldots, J_t$ where

$$s_{jt} = \frac{\sum_{i=1}^{n_t} d_{ijt}}{n_t} \tag{2.3}$$

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 choice in the market has identical choice probabilities π_t. Thus the vector of empirical shares $s_t = (s_{0t}, s_{1t}, \ldots, s_{J_t t})$ is simply the sample analogue of the underlying population choice probabilities π_t. In particular, conditional on π_t and n_t, the vector $n_t s_t$ is exactly a multinomial random variable $MN(n_t, \pi_t)$.

The key identifying restriction proposed by BLP to identify the parameters $\theta_0 \in \Theta$ is the conditional mean restriction

$$E[\xi_{jt} \mid z_{jt}, J_t] = 0 \quad \forall j = 1, \ldots, J_t \ \forall t, \tag{2.4}$$

where z_{jt} is a vector of instruments. To understand how the CMR can potentially serve to identify θ_0, assume for the sake of argument that the underlying vector of choice probabilities π_t corresponding to each market can be observed and focus attention on a single product in each market, say $j = 1$. Then under general conditions (see Berry, Gandhi, and Haile (2011)) the demand system (2.2) can be inverted to recover the true vector of mean utilities utilities δ_t, which gives us the inverse relationship.

$$\sigma_j^{-1}(\pi_t, x_t; \lambda_0) = \beta_0 x_{jt} + \xi_{it} \tag{2.5}$$

Then the parameters θ are identified by the system of equations

$$E\left[\sigma_j^{-1}(\pi_t, x_t, \lambda_0) \mid z_{jt}, J_t\right] = \beta_0 E\left[x_{jt} \mid z_{jt}, J_t\right] \quad a.e. \ (z_{jt}, J_t). \tag{2.6}$$

Assuming that the exogenous variables (z_{jt}, J_t) can vary the conditional moments

$$E\left[\sigma_j^{-1}(\pi_t, x_t, \lambda_0) \mid z_{jt}, J_t\right] \quad and \quad E\left[x_{jt} \mid z_{jt}, J_t\right]$$

sufficiently, then the parameters $\theta_0 = (\beta_0, \lambda_0)$ are uniquely identified by the system of restrictions (2.6). We can now appreciate the two fundamental challenges for conducting inference of the parameters in the demand relationship (2.5) on the basis of the mean restriction (2.4).

The Identification Problem

What we now show is that, without further restrictions, the identifying content of the conditional mean restriction (2.4) completely breaks down, i.e., it loses *all* of its empirical content, when the underlying choice probabilities π_t are not observed (which is the actual empirical case) but rather only the empirical market shares s_t as defined by (2.3) are observed. This can be seen by focusing on the absolute simplest random utility of demand, i.e., a simple logit model with a single product and no covariates, i.e.

$$u_{it} = c + \xi_t + \epsilon_{it}$$

where $\epsilon_{it} \overset{iid}{\sim} EV$ where EV is the standardized (type 1) extreme value distribution. In this case it is straightforward to show that

$$\log\left(\frac{\pi_t}{1-\pi_t}\right) = c + \xi_t$$

and hence the constant c is identified by

$$E\left[\log\left(\frac{\pi_t}{1-\pi_t}\right)\right]. \qquad (2.7)$$

However suppose that instead of observing π_t we only observe the choices of n i.i.d. consumers in each market.[5] From realized choices, we observe the empirical share s_t where $ns_t \mid \pi_t \sim MN(n, \pi_t)$. Observe that

$$E[s_t] = E[E[s_t \mid \pi_t]] = E[\pi_t].$$

Hence the expectation of s_t identifies $E[\pi_t]$. However we now show that even the identification of the entire distribution of the random variable s_t from the data places *no* restrictions whatsoever on the the expectation of interest (2.7). This is because the distribution of the empirical shares s_t severely under-identifies the distribution of the choice probabilities 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 the observed distribution, which is the source of the problem.

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

To see this more precisely, first observe that the pmf of ns_t given π_t is

$$p_{ns_t|\pi_t}(l|\pi) = \left(\begin{array}{c} n \\ l \end{array} \right) \pi^l (1-\pi)^{n-l}, \; \forall l = 0, ..., n. \tag{2.8}$$

Suppose that the distribution of π_t is $F_\pi : [0,1] \to [0,1]$. Then the unconditional pmf of ns_t (which is identified in the data) under F_π is

$$p_{ns_t}(l; F_\pi) = \int \left(\begin{array}{c} n \\ l \end{array} \right) \pi^l (1-\pi)^{n-l} dF_\pi(\pi). \tag{2.9}$$

That is, the distribution of ns_t is a mixture of binomials with mixing probability F_π. Notice that because p_{ns_t} is a discrete distribution with $n+1$ support points, and the true distribution F_π is potentially continuous, there can be many different F_π that give rise to the same p_{ns_t}, i.e., the mixing distribution in a binomial mixture is not identified. However identification of the full distribution F_π is not necessary for us - we only seek to address the extent to which $\int \log(\pi) - \log(1-\pi) F_\pi$ can be identified from the distribution of the mixture ns_t. The answer we now show is that, generically, it cannot be identified at all: the information provided by the distribution of ns_t can only give trivial ($(-\infty, \infty)$) bound for $\int \log(\pi) - \log(1-\pi) F_\pi$ for a generic distribution of ns_t.

More precisely, the set of all possible mixture distributions p_{ns_t} can be geometrically constructed as follows. Let $\vec{p}_s(F_\pi) = (p_{ns_t}(0; F_\pi), ..., p_{ns_t}(n; F_\pi))'$, and let

$$P_s = \left\{ \vec{p}_s(F_\pi) : F_\pi(t) = 1\{t \geq x) \text{ for some } x \in [0,1] \right\}. \tag{2.10}$$

Let $\vec{p}_s^* = (p_s^*(0), ..., p_s^*(n))'$ where $p_s^*(l)$ 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 a generic point in the set because the boundary of $co(P_s)$ has measure zero.

Theorem 1. *The distribution $\vec{p}_s^*$ generically produces no informative restrictions on the expectation* (2.7)*, i.e., for any generic $\vec{p}_s^*$ and any $M > 0$ there exists distributions of π_t, F_π^M and F_π^{-M}, whose supports are strictly contained in* $(0,1)$ *and are consistent with $\vec{p}_s^*$, i.e., $\vec{p}_s^* = \vec{p}_s\left(F_\pi^M\right)$ and $\vec{p}_s^* = \vec{p}_s\left(F_\pi^{-M}\right)$, such that*

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

The proof is given in Appendix A and easily extends to the general demand model discussed above. The result has far reaching implications for empirical work with differentiated products. When only empirical shares rather than choice probabilities are observed, without further restrictions, the conditional mean restriction (2.4) has zero empirical content, i.e., empirical shares provide no restrictions on the expectation of interest

$$E\left[\sigma_j^{-1}\left(\pi_t, x_t, \lambda_0\right) \mid z_{jt}, J_t\right] \tag{2.11}$$

that identifies the parameters in (2.6).

The approach of BLP avoided this issue by appealing to the economics of "large" markets: as the number of products J in the cross section grow large, if the number of consumers N in the markets grows at a sufficiently faster rate, then in the limit the error in s_t as a measure of π_t can be ignored. This asymptotic experiment cannot be performed in our setting since the number of consumers n_t in each market t is a fixed attribute of each market and does not change in the limit as the number of *markets* as opposed to products grows large. This problem can be seen manifest in many data sets involving a large number of markets relative to products, which can exhibit a large fraction zero market shares for products in the data and directly rejects the model $\pi_t \in (0, 1)$. Thus a fundamental problem we face is how to rescue the identifying content of (2.4) with data on empirical shares.

The Inference Problem

The second fundamental problem that arises in many markets data is that the product/market level (j, t) observations are *not* independent realizations. In particular, the data generating process is such that only the exogenous attributes of products across different markets $\{(z_{jt}, \xi_{jt})\}_{j \in J_t}$ and $\left\{(z_{jt'}, \xi_{jt'})\right\}_{j \in J_{t'}}$ are independent or weakly dependent. However the attributes of products within a market $\{(z_{jt}, \xi_{jt})\}_{j \in J_t}$ can be dependent in complicated ways due to the strategic interaction among firms in their "product location" decisions. BLintonP sidestep this problem by appealing to the economics of "large" markets - that is they consider a large cross section of J products in a single market, and assume that the unobservable attributes $\{\xi_{jt}\}_{j \in J_t}$ are independent. Their asymptotic theory appeals to this independence and a few high level conditions on the sample averages of the observables. Though arguments may exist to justify the assumption of independent unobservables in a large J context, they are likely to be much less convincing for markets with a small J. In a small market with relatively few products, the firms are inevitably strategically linked in dimensions both observable and unobservable to the econometrician. Our second key problem is thus to allow for general stochastic dependence in the product locations of firms within a market and only maintain standard assumptions on the way markets rather than products are sampled.

3 Identification using Moment Inequalities

3.1 Constructing Product Level Moment Inequalities

The negative conclusion of Theorem 1 is driven by the fact that distribution of empirical shares s_{jt} can be rationalized by a distribution of choice probabilities π_{jt} that contains support points arbitrarily close to 0 or 1. Thus to avoid this conclusion it is necessary to restrict the support of π_{jt} ex-ante so that it is bounded away from zero, which we now introduce as a formal assumption.

Assumption 1. $\pi_t \in \Delta^{\varepsilon}_{J_t}$ *where* $\Delta^{\varepsilon}_{J_t} = \{\pi_t = (\pi_{0t}, \pi_{1t}, ..., \pi_{J_t t}) \in [\varepsilon_t, 1]^{J_t} : \sum_{j=0}^{J_t} \pi_{jt} = 1\}$

That is, there exists a lower bound ε_t, which is the lowest value that any choice probability π_{jt} can possibly take in market t, which is a necessary assumption in light of Theorem 1. The same assumption is also imposed in BLintonP,[6], but in our many markets context we can formally see clearly why it is necessary for identification. In practice, one can set ε, for example, to the smallest positive machine number, or according to ones prior on the minimum choice probability in the market to sustain the fixed costs of the product being available in the market, etc.

Our approach to the identification problem is to exploit Assumption 1 to construct a new inversion mapping. In particular we will use Assumption 1 along with the structure of the discrete choice model to construct mappings $\sigma^{-1}_{j,l}(s_t, x_t, \lambda_0)$ and $\sigma^{-1}_{j,u}(s_t, x_t, \lambda_0)$ such that the expectation of interest (2.11) can be bounded, i.e.,

$$E\left[\sigma^{-1}_{j,l}(s_t, x_t, \lambda_0) \mid z_{jt}, J_t\right] \le E\left[\sigma^{-1}_{j}(\pi_t, x_t, \lambda_0) \mid z_{jt}, J_t\right] \le E\left[\sigma^{-1}_{j,u}(s_t, x_t, \lambda_0) \mid z_{jt}, J_t\right] \tag{3.1}$$

This enables to express the identification condition as a system of moment inequalities, which we will explain below have the property of being *adaptive* to the observed information.

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

$$E\left[\sigma^{-1}_j(s_t, x_t, \lambda_0) \mid z_{jt}, J_t\right] \tag{3.2}$$

to no longer equal the expectation of interest (2.11) because the inverse function σ^{-1} is nonlinear. The situation is actually even more complicated because the expectation (3.2) does not even exist. This is because there is always some positive probability that the

[6] Condition S of BLintonP.

empirical shares s_{jt} can be zero even if $\pi_{jt} \in (0,1)$, 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).

1. Our first step towards constructing bounds on expectation of interest (2.11) is to transform empirical shares so they move strictly to the interior of the simplex. We do so using a natural transformation: 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 estimator can be interpreted as the Bayesian posterior of π_t under a uniform prior on the J_t-dimensional unit simplex (which is also useful then for counterfactual purposes).[7] Using $\tilde{s}_t$ in $\sigma_j^{-1}(\cdot, x_t; \lambda_0)$ in place of s_t, solves the problem of the expectation (3.2) existing. However, we still have that $E\left[\sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda_0\right) \mid n_t, , z_t, J_t\right] \neq E\left[\sigma_j^{-1}\left(\pi_t, x_t; \lambda_0\right) \mid z_t, J_t\right]$.

2. To circumvent this problem, we will now exploit a monotonicity feature of demand that has thus far not been recognized nor applied to empirical work. In particular, for any market t and for each product $j = 1, ..., J_t$, and for any λ, there exists a unique real valued function $\eta_j(n_t, \pi_t, x_t; \lambda)$ defined implicitly by the unique solution of η in :
$$E\left[\sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right) | n_t, \pi_t, x_t\right] = E\left[\sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right) | n_t, \pi_t, x_t\right] = \sigma_j^{-1}(\pi_t, x_t; \lambda), \tag{3.3}$$
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 that the structure of the underlying discrete choice model ensures the existence of a unique such implicit function.

3. Let
$$\eta_{jt}^u(\lambda) := \eta_j^u(n_t, x_t; \lambda) = \sup_{\pi_t \in \Delta_{J_t}^{\varepsilon}} \eta_j(n_t, \pi_t, x_t; \lambda). \tag{3.4}$$
As the proof of Lemma B.1 makes clear, the function σ_j^{-1} is monotone in the jth share, which gives us the inequality
$$E\left[\sigma_j^{-1}\left(\tilde{s}_t + \eta_{jt}^u \cdot e_j, x_t; \lambda\right) | n_t, \pi_t, x_t\right] \geq \sigma_j^{-1}(\pi_t, x_t; \lambda), \tag{3.5}$$
for all $\pi_t \in \Delta_{J_t}^{\varepsilon}$. Hence taking expectations of both sides of this inequality conditional

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

on the instruments (z_{jt}, J_t) we have

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

Similarly, we let

$$\eta_{jt}^l(\lambda) := \eta_j^l(n_t, x_t; \lambda) = \inf_{\pi_t \in \Delta_{J_t}^{\varepsilon}} \eta_j(n_t, x_t; \lambda). \tag{3.6}$$

and a valid lower bound can be defined as:

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

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

As can be seen, letting $\sigma_{j,l}^{-1}(s_t, x_t, \lambda_0) := \sigma_j^{-1}\left(\tilde{s}_t + \eta_{jt}^l \cdot e_j, x_t; \lambda\right)$ and $\sigma_{j,u}^{-1}(s_t, x_t, \lambda_0) := \sigma_j^{-1}\left(\tilde{s}_t + \eta_{jt}^u \cdot e_j, x_t; \lambda\right)$, we have thus constructed new inversion mappings that satisfy the bounds (3.1) on the expectation of interest. We are now able use the conditional mean restriction $E[\xi_{jt} \mid z_{jt}, J_t] = 0$ to express the empirical content of the model as a system of moment inequalities

$$\begin{aligned} E\left[\sigma_{j,u}^{-1}(s_t, x_t; \lambda) - \alpha_0 p_{jt} - \beta_0 x_{jt} \mid z_{jt}, J_t\right] &\geq 0 \\ E\left[\alpha_0 p_{jt} + \beta_0 x_{jt} - \sigma_{j,l}^{-1}(s_t, x_t; \lambda) \mid z_{jt}, J_t\right] &\geq 0 \end{aligned} \tag{3.7}$$

for all $j = 1, \ldots, J_t$. Letting $w_t = (s_t, n_t, x_t, z_t, J_t)$, we can express this system more succinctly as

$$E\left[m_j(w_t; \theta_0) \mid z_{jt}, J_t\right] \geq 0 \quad j = 1, \ldots, J_t,\ t = 1, ..., T, \tag{3.8}$$

where $m_j(w_t; \theta_0)$ is a stacked vector of the two moments in (3.7).

The true parameter value θ_0 is not necessarily point-identified by the conditional moment inequality restrictions (3.8). Let Θ_0 be the collection of all $\theta \in \Theta$ that satisfy (3.8):

$$\Theta_0 = \left\{\theta \in \Theta : E\left[m_j(w_t; \theta) \mid z_{jt}, J_t\right] \geq 0 \quad j = 1, \ldots, J_t,\ t = 1, ..., T\right\}. \tag{3.9}$$

The set Θ_0 usually is called the identified set of θ_0.

Remark 1. Our bounds approach is "adaptive" in a way that makes them useful for applied work. To understand this adaptive property, let us consider how one could proceed under Assumption 1 without exploiting the structure of the underlying discrete choice model, which we shall call the "naive bounding" approach. Suppose we define

$$\begin{aligned}\mu^l_{j,naive}(x_t,n_t;\lambda) &= \sup_{\pi_t\in\Delta^{\varepsilon}_{J_t}}\{E[\sigma_j^{-1}(\pi_t,x_t;\lambda)-\sigma_j^{-1}(\tilde{s}_t,x_t;\lambda)\,|n_t,\pi_t,x_t]\}\\ \mu^u_{j,naive}(x_t,n_t;\lambda) &= \inf_{\pi_t\in\Delta^{\varepsilon}_{J_t}}\{E[\sigma_j^{-1}(\pi_t,x_t;\lambda)-\sigma_j^{-1}(\tilde{s}_t,x_t;\lambda)\,|n_t,\pi_t,x_t]\}.\end{aligned} \tag{3.10}$$

The functions $\sigma_j^{-1}(\tilde{s}_t,x_t;\lambda)+\mu^l_{j,naive}(x_t,n_t;\lambda)$ and $\sigma_j^{-1}(\tilde{s}_t,x_t;\lambda)+\mu^u_{j,naive}(x_t,n_t;\lambda)$ could thus also be used to bound the expectation of interest (2.11). However, because these "naive" correction factors $\mu^l_{j,naive}(x_t,n_t;\lambda)$ and $\mu^u_{j,naive}(x_t,n_t;\lambda)$ do not depend upon s_t, they typically are large because the bias in $\sigma_j^{-1}(\tilde{s}_t,x_t;\lambda)$ is large for π_t close to the boundary of $\Delta^{\varepsilon}_{J_t}$. The large bias correction is applied indiscriminately to all realizations of $\sigma_j^{-1}(\tilde{s}_t,x_t;\lambda)$ even if s_t is far from the boundary.

Consider on the other hand the correction factors implicitly defined by our bounds construction, i.e.

$$\begin{aligned}\sigma_j^{-1}\left(\tilde{s}_t+\eta^l_{jt}\cdot e_j,x_t;\lambda\right) &= \sigma_j^{-1}(\tilde{s}_t,x_t;\lambda)+\mu^l_j(s_t,x_t,n_t;\lambda)\\ \sigma_j^{-1}\left(\tilde{s}_t+\eta^u_{jt}\cdot e_j,x_t;\lambda\right) &= \sigma_j^{-1}(\tilde{s}_t,x_t;\lambda)+\mu^u_j(s_t,x_t,n_t;\lambda)\end{aligned}$$

Thus the correction factors $\mu^u_j(n_t,x_t,\pi_t)$ and $\mu^l_j(n_t,x_t,\pi_t)$ are adaptive, i.e., they change with the realization of s_t. They are larger (in absolute value) when noise in $\tilde{s}_t$ affects the expectation of $\sigma_j^{-1}(\tilde{s}_t,x_t;\lambda)$ more, and vice versa. The adaptiveness comes from the fact that our correction η^u_{jt} and η^l_{jt} enter in the same way as the noise $(\tilde{s}_t-\pi_t)$. Because of the way it enters, when the noise affects the expectation more, the adjustments η^u_{jt} and η^l_{jt} will also affect the expectation more making μ^u_j more negative and μ^l_j more positive, and vice versa.

Because our correction factors utilize more information, we will have that $\mu^u_{jt}\le\mu^u_{jt,naive}$ and $\mu^l_{jt}\ge\mu^l_{jt,naive}$ and thus

$$\begin{aligned}E\left[\mu^u_{jt}\mid z_{jt},J_t\right] &\le E\left[\mu^u_{jt,naive}\mid z_{jt},J_t\right]\\ E\left[\mu^l_{jt}\mid z_{jt},J_t\right] &\ge E\left[\mu^l_{jt,naive}\mid z_{jt},J_t\right],\end{aligned}$$

That is, our approach will more tightly bound the expectation of interest than the "naive bounds" because of the adaptiveness, which makes our approach useful in practice as our empirical illustration shall illustrate.

3.2 Aggregating Moment Inequalities to the Market Level without Information Loss

In Section 3.1, it is shown that the aggregate demand model can be written as

$$E\left[m_j\left(w_t;\theta_0\right) \mid z_{jt}, J_t\right] \geq 0 \quad j = 1,\dots,J_t,\ t = 1,...,T. \tag{3.11}$$

The model (3.11) appears almost the same as the conditional moment inequality model discussed extensively in, e.g., Andrews and Shi (2009) 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. One immediate challenge that the moment inequalities (3.11) presents is that observations $\{x_{jt}, z_{jt}, \xi_{jt}\}$ will tend to be correlated across observations j *within* the same market t in non-standard ways due to the strategic interaction between products in a market. [8]

Instead of treating each (jt) 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. Assuming that there is no strategic linkage in the realization of the unobservables across markets, then the market level data can more naturally be assumed to satisfy either the independence or weak dependence in a conventional sense.

The aggregation we seek needs to be done properly to preserve all the identification information there is in (3.11) under acceptable assumptions on the data generating process. The first step is to transform (3.11) into moments not conditioning on product level variables – moments that can be aggregated. Let $g(z_{jt})$ be a real-valued function that lies in the collection $\mathcal{G}$. The collections are collections of indicator functions:

$$\mathcal{G} = \{1(z \in C) : C \in \mathcal{C}\}, \tag{3.12}$$

where $\mathcal{C}$ is a collection of subsets of $\mathcal{Z}$, the support of z_{jt}. The following Lemma shows the equivalent form of (3.11). The proof is the same as that of Lemma 3 in Andrews and Shi (2009) 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 $\mathcal{C} \cup \{\emptyset\}$*

[8] The dependence of $m_j(w_t, \theta_0)$ on other products' characteristics cannot be captured by a market level fixed effect. It is not helpful 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 either because J_t varies across markets.

The sample $\{x_{jt}, z_{jt}, s_{jt}\}$ can be consider to be a cluster sample with each market being a cluster. Unfortunately, empirical process theory for cluster samples is not readily available but is needed for the asymptotic justification of our inference procedure.

equals $\mathcal{B}(\mathcal{Z})$ – the Borel sigma field on $\mathcal{Z} \subseteq R^{d_z}$.[9]

Then, (3.11) *holds if and only if*

$$E[m_j(w_t, \theta_0) g(z_{jt}) | J_t] \geq 0 \quad j = 1, ..., J_t, \; t = 1, ..., T, \; \forall g \in \mathcal{G}. \tag{3.13}$$

The second step is to aggregate up the moments in (3.13) to market level:

$$E\left[\sum_{j=1}^{J_t} m_j(w_t, \theta_0) g(z_{jt}) \middle| J_t \right] \geq 0, \quad t = 1, ..., T, \; \forall g \in \mathcal{G}. \tag{3.14}$$

The aggregated moment condition contains exactly the same information as (3.13) if products within a market satisfy a standard exchangeability requirement. The exchangeability requirement in this context is that the econometrician is agnostic about how products from different markets are linked to each other, i.e., there is no ex-ante information in the index j we use to index a product.Then we have for all $j^{'} = 1, 2, ..., J_t$,

$$E\left[\sum_{j=1}^{J_t} m_j(w_t, \theta_0) g(z_{jt}) \middle| J_t \right] = J_t E\left[m_{j'}(w_t, \theta_0) g(z_{j't}) \middle| J_t \right]. \tag{3.15}$$

It is then immediate that the market level moment condition (3.14) holds if and only if (3.13) does.

We assume that the number of products in a market is bounded by $\bar{J}$. Let $\mathcal{C}^J$ be a semi-ring of subsets of $\{1, ..., \bar{J}\}$ and $\mathcal{G}^J = \{g^J(y) = 1\{y \in C^J\} : C^J \in \mathcal{C}^J\}$. Let

$$\rho(w_t, \theta, g, g^J) = \sum_{j=1}^{J_t} m_j(w_t, \theta) g(z_{jt}) g^J(J_t). \tag{3.16}$$

Suppose that $\{1, ..., \bar{J}\}$ can be written as the union of countable disjoint sets in $\mathcal{C}^J$ and the sigma field generated by $\mathcal{C}^J$ is the power set of $\{1, ..., \bar{J}\}$. Then similar to Lemma 1, we can show that

$$E[\rho(w_t, \theta_0, g, g^J)] \geq 0, \; \forall g \in \mathcal{G}, g^J \in \mathcal{G}^J \tag{3.17}$$

if and only if (3.14) holds. Thus, we have aggregated up the individual product level moments into market level without loss of information. The following lemma collects all the assumptions and state the aggregation formally. The proof is omitted because its

[9] 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 $\mathcal{C}_c = \{[a, b) : a, b \in [0, 1]\} \cup \{\{b\}\}$.

supporting arguments are already given above.

Lemma 2. *Suppose that* (i) $J_t \in \{1, ..., \bar{J}\}$ *for some* $\bar{J} < \infty$, (ii) $\mathcal{C} \cup \{\emptyset\}$ *and* $\mathcal{C}^J \cup \{\emptyset\}$ *are semi-rings of subsets of* $\mathcal{Z}$ *and* $\{1, ..., \bar{J}\}$ *respectively,* (iii) $\mathcal{Z}$ *and* $\{1, ..., \bar{J}\}$ *can be written as the union of countable disjoint sets in* $\mathcal{C}$ *and in* $\{1, ..., \bar{J}\}$ *respectively and* (iv) *the sigma field generated by* $\mathcal{C}$ *and* $\mathcal{C}^J$ *are* $\mathcal{B}(\mathcal{Z})$ *and* $2^{\{1,...,\bar{J}\}}$*, respectively. Then*

$$\Theta_0 = \{\theta \in \Theta : E[\rho(w_t, \theta, g, g^J)] \geq 0, \ \forall g \in \mathcal{G}, g^J \in \mathcal{G}^J\}.$$

The next section takes the model in (3.17) as the starting point and develop a generalized profiling method for the inference of any parameter that is identified through a (possibly set valued) function of θ_0.

4 Estimation and Inference

4.1 Inference Using Generalized Profiling

The model (3.17) is a moment inequality model with many moment conditions. One could use the method developed in Andrews and Shi (2009) to construct a confidence set for θ_0. However, Andrews and Shi (2009)'s confidence set is constructed by inverting an Anderson-Rubin test: $CS = \{T(\theta) \leq c(\theta)\}$ for some test statistic $T(\theta)$ and critical value $c(\theta)$. Computing the set amounts to constructing 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 $E(\xi_{jt}|z_{jt}, J_t) = 0$ 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 *generalized profiling* method to profile out the nuisance parameters and only construct confidence sets for a parameter of interest.

The generalized profiling approach applies to general moment inequality models with many moment inequalities. Thus from this point on, we treat $\rho(w_t, \theta, g, g^J)$ 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:

$$\gamma_0 \in \Gamma(\theta_0) \subseteq R^{d_\gamma}, \tag{4.1}$$

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. $\Gamma(\theta) = \{e_j(p, \pi, \theta, x) = (\alpha p_j)/(\pi_j \partial \sigma_j^{-1}(\pi, x, \sigma_0)/\partial \pi_j)\}$: γ_0 is the own-price demand elasticity of product j at a given value of the price vector p, the market share 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 a market share vector that is known to lie between π^l and π^u. This example is particularly useful when the elasticity depends on the market share but the market share is not precisely observed. The interval $[\pi^l, \pi^u]$ can be a confidence interval of the market share.

The generalized profiling approach constructs a confidence set for γ_0 by inverting a test of the hypothesis:

$$H_0 : \gamma_0 \in \Gamma_0, \tag{4.2}$$

where Γ_0 is the identified set of γ_0: $\Gamma_0 = \{\gamma \in R^{d_\gamma} : \exists \theta \in \Theta_0 \ s.t. \ \Gamma(\theta) \ni \gamma\}$. Let $\Gamma^{-1}(\gamma) = \{\theta \in \Theta : \Gamma(\theta) \ni \gamma\}$.

The test to be inverted uses the *profiled* test statistic:

$$\hat{T}_T(\gamma) = T \times \min_{\theta \in \Gamma^{-1}(\gamma)} \hat{Q}_T(\theta), \tag{4.3}$$

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 $c_T(\gamma, p)$:

$$CS_T = \{\gamma \in R^{d_\gamma} : \hat{T}_T(\gamma) \leq c_T(\gamma, p)\}. \tag{4.4}$$

Notice that the new confidence set only involves computing a d_γ-dimensional level set, where d_γ is often 1. The generalized profiling transfers the burden of searching (for minimum) 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

$$\lim_{T \to \infty} \inf_{(\gamma, F) \in \mathcal{H}_0} \Pr(\hat{T}_T(\gamma) > c_T(\gamma, p)) \leq 1 - p, \tag{4.5}$$

where F is the distribution on $(w_t)_{t=1}^T$ and $\mathcal{H}_0$ is the null parameter space of (γ, F).[10] As a result, the confidence set asymptotically has the correct minimum coverage probability:

$$\liminf_{T\to\infty} \inf_{(\gamma,F)\in\mathcal{H}_0} \Pr_F(\gamma \in CS_T) \geq p. \tag{4.6}$$

The left hand side is called the "asymptotic size" of the confidence set in Andrews and Shi (2009). 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 details and show that (4.6) holds.

4.2 Test Statistic

Let $\mathcal{G} = \mathcal{G}^z \times \mathcal{G}^J$ and $g(z_{jt}, J_t) = g^z(z_{jt}) \times g^J(J_t)$. The test statistic $\hat{T}_T(\gamma) = T \times \min_{\theta\in\Gamma^{-1}(\gamma)} \hat{Q}_T(\theta)$ with

$$\hat{Q}_T(\theta) = \int_{\mathcal{G}_T} S(\bar{\rho}_T(\theta, g), \hat{\Sigma}_T^\iota(\theta, g))d\mu(g), \tag{4.7}$$

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

$$\begin{aligned}
\bar{\rho}_T(\theta, g) &= T^{-1}\sum_{t=1}^T \rho(w_t, \theta, g), \\
\hat{\Sigma}_T^\iota(\theta, g) &= \hat{\Sigma}_T(\theta, g) + \iota \times \hat{\Sigma}_T(\theta, 1) \\
\hat{\Sigma}_T(\theta, g) &= T^{-1}\sum_{t=1}^T \rho(w_t, \theta, g)\rho(w_t, \theta, g)' - \bar{\rho}_T(\theta, g)\bar{\rho}_T(\theta, g)'.
\end{aligned} \tag{4.8}$$

In the above definition, ι is a small positive number which is used because in some form of S defined below, 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 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

[10] The definition of $\mathcal{H}_0$ along with other technical assumptions are given in Appendix C.

$\min_{\theta\in\Gamma^{-1}(\gamma)} Q_F(\theta)$ where

$$Q_F(\theta) = \int_{\mathcal{G}} S(\rho_F(\theta,g), \Sigma_F^{\iota}(\theta,g))d\mu(g), \tag{4.9}$$

with

$$\begin{aligned} \rho_F(\theta,g) &= E_F(\rho(w_t,\theta,g)) \\ \Sigma_F(\theta,g) &= Cov_F(\rho(w_t,\theta,g)) \text{ and} \\ \Sigma_F^{\iota}(\theta,g) &= \Sigma_F(\theta,g) + \iota\Sigma_F(\theta,1). \end{aligned} \tag{4.10}$$

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 $\gamma \in \Gamma_{0,F}$. These two results imply 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 3 summarizes those results. The parameter space $\mathcal{H}$ of (γ, F) appearing in the lemma is defined in Assumption C.2 in the appendix.

Lemma 3. *Suppose that the conditions in Lemma* 2 *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}, J_t) into discrete instruments, $z_{d,jt}$, and continuous instruments $z_{c,jt}$.[11] 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, $\hat{\Sigma}_{z_c}$ is the sample covariance matrix of $z_{c,jt}$. The set $\mathcal{G}$ is defined as

$$\begin{aligned} \mathcal{G} = \{g_{a,r,\zeta}(z_d, z_c) = 1(\tilde{z}_c \in C_{a,r}, z_d = \zeta) : C_{a,r} \in \mathcal{C}_{cc}, \zeta \in \mathcal{Z}_d\}, \text{ where} \\ \mathcal{C}_{cc} = \{\times_{u=1}^{d_{z_c}}((a_u - 1)/(2r), a_u/(2r)] : \ a_u \in \{1, 2, ..., 2r\}, \text{ for } u = 1, ..., d_{z_c} \\ \text{and } r = r_0, r_0 + 1, ...\} \end{aligned} \tag{4.11}$$

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 runs from r_0 to $\bar{r}_T$ where $\bar{r}_T \to \infty$ as $T \to \infty$.

[11] J_t naturally belongs to the $z_{c,jt}$ part.

For S , we use

$$S(m, \Sigma) = \sum_{j=1}^{d_m} [m_j]_-^2, \tag{4.12}$$

where m_j is the jth coordinate of m and $[x]_- = |\min\{x, 0\}|$. There may be efficiency loss from not using the information in the variance matrix, but this S function brings great computational convenience because it makes the minimization problem in (4.7) a convex one. For $\mu(\cdot)$, we use

$$\mu(\{g_{a,r,\zeta}\}) \propto (100 + r)^{-2}(2r)^{-d_{zc}} K_d^{-1} \text{ for } g \in \mathcal{G}_{d,cc}, \tag{4.13}$$

where K_d is the number of elements in $\mathcal{Z}_d$.

4.3 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.[12] 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 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

$$c_T^{sub}(\gamma, p) = c^*_{sub}(\gamma, p + \eta^*) + \eta^*, \tag{4.14}$$

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 (2009).

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;

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

denote the bootstrap sample by $t_1, ..., t_T$, [2] let the bootstrap statistic be

$$T_T^*(\gamma) = \min_{\theta\in\Theta:\gamma\in\Gamma(\theta)} \int_{\mathcal{G}} S(\hat{\nu}_T^*(\theta, g) + \kappa_T^{1/2}\bar{\rho}_T(\theta, g), \hat{\Sigma}_T^{\iota}(\theta, g))d\mu(\mathcal{G}), , \tag{4.15}$$

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

$$c_T^{bt}(\gamma, p) = c_{bt}^*(\gamma, p + \eta^*) + \eta^*, \tag{4.16}$$

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

4.4 Coverage Probability

We show that the confidence sets defined in (4.4) 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 2 and Assumptions* C.1-C.3 *and* C.5-C.7 *hold, then*
(a) (4.6) *holds with* $c_T(\gamma, p) = c_T^{sub}(\gamma, p)$*, and*
(b) (4.6) *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 why it is lengthy and how it works. 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

$$\hat{\nu}_T(\theta, g) = \sqrt{T}(\bar{\rho}_T(\theta, g) - \rho_F(\theta, g)). \tag{4.17}$$

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 emphasize that the issue is more complicated here than in Andrews and Shi (2009) 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 nonincreasing 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. 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) with a certain choice of their moment selection function. The essence of both our subsampling and bootstrap procedures is 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 $\theta \neq \Theta_0$, $\sqrt{\kappa_T}\rho_{F,j}(\theta, g)$ might be bigger than $\sqrt{T}\rho_{F,j}(\theta, g)$ making $\int_{\mathcal{G}_T} S(\sqrt{T}\bar{\rho}_T(\theta, g), \hat{\Sigma}_T^{\iota}(\theta, g))d\mu(g)$ 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 integrals $\int_{\mathcal{G}_T} S(\sqrt{T}\bar{\rho}_T(\theta, g), \hat{\Sigma}_T^{\iota}(\theta, g))d\mu(g)$ 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 and put off to the appendix.

5 Empirical Application

5.1 Data Description

We now apply our inference strategy to a scanner data on demand for consumer goods, which is both a demand setting that has wide policy relevance and one where sampling variability in shares appears to be a prominent problem. 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.[13]

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

The data comprises all Dominick's Fine 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 the finest level of product description, i.e., the bar code that identifies a product. The set of UPC's that a store places on its shelves exactly corresponds to the choice set consumers who enter the store face. All the relevant marketing decisions made by the store, i.e., which UPC's to offer on its shelves, where to place on shelves, how much to price and discount, etc, are decided on a weekly basis. Thus different markets (i.e., a period of time over which the choice set is stable) are naturally defined by different store/week pairs. The data in principle provide about 40,000 such store/week pairs.

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 product simply was not stocked that week (this can be difficult to disentangle in other scanner data sets that only records data on a upc if it sells in a given store/week). We summarize the percents of zero sales in the entire Dominick's database at the UPC level for all the categories in Table 1. We can see that the fraction of observations with zero sales can even exceed 60% for some categories.

Table 1: Percent of Zero Sales in Dominick's Database

Category	Zeros(%)	Category	Zeros(%)	Category	Zeros(%)
Analgesics	58.02	Dish Detergent	42.39	Refrigerated Juices	27.83
Bath Soap	74.51	Front-end-candies	32.37	Soft Drinks	38.54
Beer	50.45	Frozen Dinners	38.32	Shampoos	69.23
Bottled Juices	29.87	Frozen Entrees	37.30	Snack Crackers	34.53
Cereals	27.14	Frozen Juices	23.54	Soaps	44.39
Cheeses	27.01	Fabric Softeners	43.74	Toothbrushes	58.63
Cigarettes	66.21	Grooming Products	62.11	Canned Tuna	35.34
Cookies	42.57	Laundry Detergents	50.46	Toothpastes	51.93
Crackers	37.33	Oatmeal	26.15	Bathroom Tissues	28.14
Canned Soup	19.80	Paper Towels	48.27		

We choose 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). Second, this category has a smaller fraction of zeroes as compared to some other product categories, and thus is far from a "worst case" scenario for the selection problem caused by zero sales for BLP, an issue we explore below.

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 demand because there is rich demographic variation associated with the zip codes of different stores and demand parameters could differ in arbitrary ways across different stores due to the different demographic surroundings of the stores. (see e.g., Hoch, Kim, Montgomery, and Rossi (1995)). We respect this concern by focusing attention on a single store.[14] 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.

The market share for each UPC is constructed by dividing the weekly sales at the store for each UPC by the "Customer Count" variable.[15] Also, we invert DFF's data on gross margin to calculate the chain's wholesale costs,[16] which are used as the instruments for the retail prices and is a standard choice of price instrument in the literature that looks at the DFF data. The total number of observations (UPC/week) of our sample is 4438, which consists of 104 weeks with an average number of UPC's in each week being 43.

5.2 Utility Specification

The indirect random utility specification is given in (2) and rewritten here for easy reference:

$$u_{ijt} = \delta_{jt} + \epsilon_{ijt} \equiv x_{jt}^{'}\beta - \alpha p_{jt} + \xi_{jt} + \epsilon_{ijt}, \quad j = 1, \ldots, J_t, \tag{5.1}$$

where p_{jt} is the retail price, x_{jt} includes 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 key source of the price variation in the data is the decision by the store to put a product on sale. In Figure 1 we show the time series of price for an arbitrary UPC,

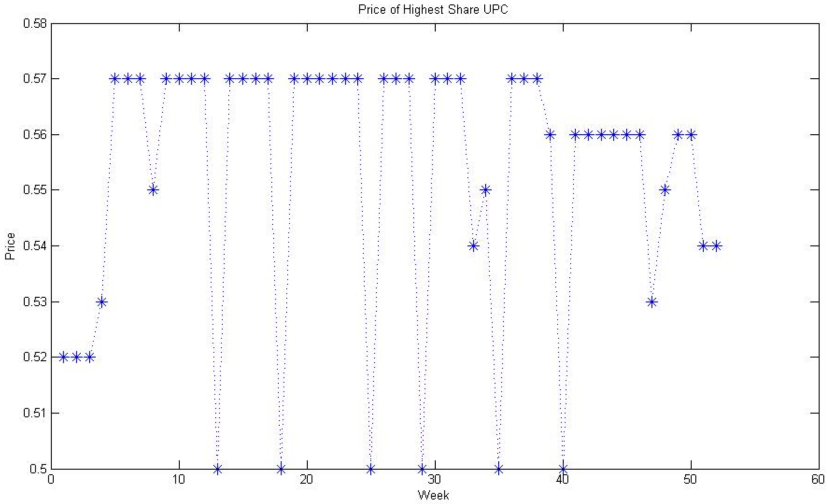
[14] We select as the only store in a pricing zone (zone number is 13 in the dataset), which belongs to "medium" price tier and is one of the 16 DFF's reported zones. The store is numbered as 134 in the dataset, which locates at the city of West Chicago. Our results are in no way dependent upon the selection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

[15] This is the number of customers visiting the store during a week and purchasing something.

[16] The gross margin is defined as (retail price - wholesale cost)/retail price, so we get wholesale cost using retail price×(1 - gross margin).

and as can be seen, the price variation largely takes the form of the product going on a temporary sale and then reverting back to an "everyday" price. The sales aspect of the price variation 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.[17]

Figure 1: Price Variation of a UPC



We will model the random utility terms ϵ_{ijt} as i.i.d. across i, j and t with the standard type-I "extreme value" distribution function $\exp(-\exp(-\epsilon))$, i.e., the "logit model".[18] In the logic 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.

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

[17] 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 instablity and sensitivity stemming from colinearity among the upc's and other covariates. Our current strategy appears to control for much of what a UPC fixed effect strategy seems to empirically offer.

[18] The outside good has utility $u_{i0t} = \epsilon_{i0t}$ where ϵ_{i0t} is also type-I extreme value.

motivate potentially richer specifications. Our strategy in isolating the logit is intentional: we wish to demonstrate that even for this widely recognized and seemingly well understood model, the problem of the demand inference with many markets still poses serious empirical problems and that our inference strategy can actually reveal new insights in this context.[19]

5.3 The BLP Approach on the Bath Tissue Data

Let us recall that the BLP approach is designed for a "many products" environment in which it is assumed that

1. the number of consumers increases at a sufficiently fast rate with the number of products and

2. product unobservables are presumed to be independent,

and is based on inverting empirical market shares to give rise to the estimating equation (2.5). In the logit case, the inverse market share function σ^{-1} takes an analytically closed form, and the estimating equation (2.5) can be expressed as

$$\log\left(\frac{\pi_{jt}}{\pi_{0t}}\right) = x_{jt}^{'}\beta - \alpha p_{jt} + \xi_{jt}. \tag{5.2}$$

Under their assumption that the empirical market share s_t approaches the true market share π_t sufficiently fast, one can replace π_t by s_t and estimate α, β based on (5.2) using an instrumental variable regression.

In the "many markets" case, which our empirical setting clearly represents, we have already argued that BLP approach is no longer appropriate. The critical identification problem that we brought to light in our earlier analysis was that s_{jt} will no longer approximates π_{jt} well enough because the consumers in a market does not grow asymptotically with the number of markets, and thus the finite sample error in s_{jt} remains present in the limit. In the context of the logit model (5.2), this problem can be seen in a fairly concrete way. First, observe that the s_{jt} and s_{0t} enters the regression nonlinearly, so that there is a nonlinear measurement error problem if sampling error in shares is used instead of the true shares π_{jt} and π_{0t}. As the literature has shown (e.g. Abrevaya and Hausman (2004)), a non-linear measurement error problem can cause the direction of asymptotic bias in the parameter estimate of (α, β) to go in various direction. Second, sampling error can lead to $s_{jt} = 0$, i.e., a zero observed market share, which is a prominent phenomenon in our data as shown in Table 1. Under the BLP approach, such zeroes would actually rejects the model as the inverse share mapping σ^{-1} does not exist at the boundary of the simplex (i.e.,

[19] There is nothing that would prevent us from adding random coefficients or nests among products.

$\log\left(\frac{s_{jt}}{s_{0t}}\right)$ doesn't exist when $s_{jt} = 0$). This forces the researcher using the BLP approach to exclude these observations from the regression, which gives rise to a sample selection problem as observations are selected based on the outcome of the dependent variable. It is fairly easily seen that the selection bias tends to produce an attenuation bias on the price coefficient α, that is, to produce demand estimates that are too inelastic.[20] The combined effect of the nonlinear measurement error and the selection may vary from one empirical setting to another.

If we impose the BLP assumptions on the data, i.e., systematically exclude the UPC's with zero shares, ignore any remaining sampling error in shares, then the result is the following parameter estimates for the price coefficient – the main parameter of interest – and the average own-price elasticity which is computed from the price coefficient.

Table 2: Results of the Logit Models

	Price Coefficient	Average Own Price Elasticity	No. of Observations
IV Logit	-1.50 [-1.90, -1.11]	-2.40 [-3.04, -1.78]	3565
OLS Logit	-2.17 [-2.37, -1.98]	-3.472 [-3.79, -3.17]	3565

Note: 95% Confidence Intervals are in [$\cdot$] (cluster at market level)

As can be seen, the IV changes the OLS estimates in the expected direction as we already anticipated based on the likely negative relationship between price and the unobservable as discussed above. However all the concerns raised above should make us wary of these results. In order to get a better sense of the general nature of the bias of these estimates and to help provide guidance on the choice of tuning parameters for our inference strategy, we now turn to a Monte Carlo analysis that mimics the structure of our data.

[20] To see this selection bias, consider for simplicity a binary choice model where

$$\delta_t = \alpha p_t + \xi_t,$$

and ξ_t is independent of x_t. The choice probability of the product is increasing in the mean utility δ_j. Because $\alpha < 0$, this choice probability $\pi_t = \sigma(\alpha p_t + \xi_t)$ is increasing in ξ_t and decreasing in p_t. If the sample share is from finite number of individual consumers, then the probability of the sampled share s_t being non-zero is an increasing function of π_t. Let us consider a simplified version of this selection mechanism so as to make the point transparent: suppose that we observe market t in the sample iff $\sigma(\alpha p_t + \xi_t) \geq \underline{\pi}$. Then we have that $d = 1$ iff $\xi_t \geq \sigma^{-1}(\underline{\pi}) - \alpha p_t$ which is *increasing* in p_t. Thus the selection mechanism is such that $E[\xi_t \mid p_t, d_t = 1]$ is an increasing function of p_t. If $E[\xi_t \mid p_t, d_t = 1] = \gamma p_t$ where $\gamma > 0$, then the regression

$$E[\delta_t \mid p_t, d_t = 1] = (\alpha + \gamma) p_t,$$

will tend to bias the slope coefficient towards zero. Said another way, it will bias the price elasticity towards being too inelastic.

5.4 A Monte Carlo Analysis

We now consider a Monte Carlo study that is designed to serve two main purposes: 1) to gain some further intuition on the possible nature of the bias of the above BLP estimates and 2) to provide guidance on the performance of different choices of the tuning parameters κ_T underlying the bootstrap procedure for the generalized profiling of the moment inequalities that forms the basis of our empirical strategy.

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 our data. We generate the data from a logit model with a single observable product characteristic x whose distribution we wish to bear some similarity to the distribution of our main covariate of interest in our data, namely price. We thus simulate x as uniform $[0, 10]$ (this being the approximate range of prices in the data). The x characteristic is independent across products and markets, and each consumer $i = 1, \ldots, n_t$ has utility for a product given by

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

and $u_{i0t} = \epsilon_{i0t}$, where ϵ_{ijt} are i.i.d. type-I extreme value and $(\alpha_0, \beta_0) = (1, 1)$. The unobservable ξ_{jt} is simulated in such a way to satisfy the conditional mean restriction $E\left[\xi_{jt} \mid x_{jt}\right] = 0$, and to exhibit a simple form of heteroskedasticity. In particular we take $\xi_t \sim (x_t \geq 5)\, Unif\, [-.05, .05] + (x_t < 5)\, Unif\, \left[-.5\bar{\xi}, .5\bar{\xi}\right]$, which introduces a simple pattern of heteroskedasticity in which there is larger variance in the unobservable for higher priced markets (lower x) products. We focus on a range of $\bar{\xi}$ that makes the fraction of products with zero shares lie within the same general range as found in our data, i.e., roughly 20-30 percent of product level observations.

We implement both the "naive" logit BLP estimator described in Section 5.3 above, and our generalized profiling procedure for the moment inequalities we constructed in Section 3.2 for the many markets environment. In particular we profile out the nuisance parameter α and employ the bootstrapping procedure described in Section 4.3 to obtain confidence sets for β_0. One implementation issue with the bootstrapping procedure is 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 asymptotic theory to provide 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.[21] 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 (2009). We choose the constant c through a series of Monte Carlo simulation of the

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

coverage probability (of the true value $\beta_0 = 1$, CP) and the false coverage probability (of a point outside the identified set of β, FCP). We find that the CP's are always 1, showing that our confidence set does not under cover. The FCP's are shown in Table 3 below. As the table shows, at $c = 0.5 - 0.6$ our confidence set has decent FCP's. [22]

Table 3: False Coverage Probabilities (FCP) of the 95% Confidence Interval

$c \backslash \bar{\xi}$	$\bar{\xi} = 11$	13	15	17
0.1	1	1	1	1
0.3	0.577	0.960	0.768	1
0.4	0.510	0.288	0.285	0.644
0.5	0.636	0.256	0.244	0.455
0.6	0.894	0.344	0.305	0.477
0.7	0.994	0.591	0.483	0.595
0.9	1	0.995	0.961	0.936

Note: The FCPs are computed at 0.95, 0.94, 0.93, 0.91 for $\bar{\xi} = 11, 13, 15, 17$, respectively. These numbers are chosen to yield nontrivial FCP's.

The results of both the naive BLP and our approaches are shown in Table 4, which reports the BLP estimates along with our 50% and 95% confidence intervals (CS). As can be seen, the selection bias with the "naive" logit goes in the anticipated direction of attenuating the coefficient on the variable of interest towards zero, sometimes severely so. On the other, our confidence intervals based on the moment inequalities, which were designed specifically for the many markets environment, always contain the true value and for the whole range of $\bar{\xi}$ and exclude the biased BLP-logit point estimates. Moreover, our confidence intervals are fairly informative even when the degree of heteroskedasticity and hence selection in the data as determined by $\bar{\xi}$ is large.

[22] Another implementation details for our inference strategy are the set of g functions, We follow the suggestions in Section 4.2 and let $\bar{r}_T = 50$, which yields on average 50 product/markets in each of the smallest hyperboxes. The final implementation detail is the lower bound for the true share: ϵ_t. We set ϵ_t to be machine precision 10^{-16}.

Table 4: Monte Carlo Results: Point and Bound Estimates

$\bar{\xi}$	Logit Point Estimate and 95% CS	50% CS	95% CS	Percent of Positive Shares
1	0.94 [0.93, 0.94]	[0.98, 1.02]	[0.97, 1.04]	82.4%
3	0.90 [0.89, 0.91]	[0.97, 1.03]	[0.96, 1.06]	82.5%
5	0.85 [0.84, 0.86]	[0.97, 1.04]	[0.95, 1.12]	82.3%
7	0.77 [0.76, 0.78]	[0.97, 1.03]	[0.95, 1.15]	80.2%
9	0.69 [0.68, 0.71]	[0.97, 1.03]	[0.94, 1.06]	78.9%
11	0.61 [0.60, 0.63]	[0.97, 1.02]	[0.94, 1.05]	77.9%
13	0.53 [0.51, 0.55]	[0.97, 1.02]	[0.94, 1.04]	76.7%
15	0.44 [0.42, 0.46]	[0.96, 1.03]	[0.94, 1.05]	75.6%
17	0.34 [0.32, 0.36]	[0.94, 1.04]	[0.92, 1.06]	74.1%
19	0.24 [0.21, 0.26]	[0.90, 1.10]	[0.84, 1.26]	72.7%
20	0.18 [0.16, 0.21]	[0.86, 1.16]	[0.75, 1.47]	71.8%

Note: True value = 1, $T = 100$, $J = 50$, $\kappa_T = T/(0.5 \cdot \log(T))$

5.5 The Many Markets Approach on the Bath Tissue Data

We now perform inference on the price coefficient and the resulting elasticities in the logit model using our generalized profiling strategy applied to the moment inequalities that we derived for the many markets model. The Monte Carlo analysis above suggests that the BLP estimates shown in Table 2 are biased towards zero, and thus generate price elasticities that are too small. Relative to the Monte Carlo, we now must profile out many more coefficients besides the constant term because the specification in Section 5.2 includes many more control variables. [23] [24]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 generalized profiling procedure allows.

The results of our inference is shown in Table 5. As can be seen our inference strategy produces substantially larger price coefficients in magnitude, which is consistent with the results of the Monte Carlo analysis. We have shown the results for both the choice of the tuning parameter constants $c = .6$ and $c = .5$, which emerged from the Monte Carlo as having the desirable size and power characteristics.[25]

[23] This expanded set of covariates thus requires thus more g functions than the Monte Carlo. We construct the set $\mathcal{G}_T$ as described in (4.11). 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 (4.13), the weight for a g function indexed by r is $(100+r)^{-2}(2r \times 11 \times 9 \times 2 \times 2 \times 2)^{-1}$. To be consistent with the literature, we do not use J_t as an instrument even though our theory suggests such a possibility. We select the minimum possible choice probability ϵ_t (to be the same across t) by taking the smallest share in the data and dividing it by 100,000. The divider $100,000$ is the the largest number of the form 10^x for $x \in \mathbb{N}$ that guarantees nonempty confidence sets.

[24] Determine 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.

[25] One question that arises is how to compute an elasticity when the underlying choice probability π_{jt} is not known. Here we show that the noise in s_t does not affect the estimation of the elasticities that we choose to focus on – the average elasticity across markets:

$$\epsilon_I = \alpha E\left[p_{I_t t}\left(1 - \pi_{I_t t}\right)\right]$$

where the expectation is taken across markets t and I_t is some set of products. To estimate ϵ_I, we must estimate $E\left[p_{I_t t}\left(1 - \pi_{I_t t}\right)\right]$. But notice that where the last equality follows because $E\left[s_{I_t t} \mid \pi_{I_t t}, p_{I_t t}\right] = \pi_{I_t t}$ due to the way that $s_{I_t t}$ is generated. We can estimate $E\left[p_{I_t t}\left(1 - s_{I_t t}\right)\right]$ consistently using the sample analogue

$$T^{-1} \sum_{t=1}^{T} p_{I_t t}\left(1 - s_{I_t t}\right).$$

Thus, $E[p_{I_t t}(1 - \pi_{I_t t})]$ is consistently estimated by $T^{-1} \sum_{t=1}^{T} p_{I_t t}\left(1 - s_{I_t t}\right)$. This strategy can be applied to any subset of products, including a single UPC (in which case the price is simply the price of the UPC p_{jt}).

Table 5: 95% Confidence Intervals of Price Coefficient and Average Own Price Elasticity

c	Price Coefficient	Average Own Price Elasticity
0.6	[-4.17, -3.11]	[-7.73, -5.77]
0.5	[-4.55, -3.06]	[-8.44, -5.67]

Note: $\kappa_T = T/(c \cdot \log(T))$ and $\epsilon = \min_{j,t}\left[\hat{s}_{jt} 1\left(\hat{s}_{jt} > 0\right)\right]/100,000$.

To better understand our estimates in comparison to the naive BLP logit, we translate our price coefficient from the UPC level demand system into an average brand level elasticities for all weeks in the data, which is given in Table 6. 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 purchases with a representative agent model of aggregate demand, the elasticities we derive should be at least as large as the HL estimates as our data reflects store level purchases (hence there can be stockpiling effects as well as substitution to other stores). Observe that the brand elasticities derived under the BLP logit are all considerably less elastic than the HL estimates, which is contrary to the standard intuition (we note that these BLP logit-type elasticities 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)).

Our estimates on the other hand show elasticities that are at least as elastic, and for all but one brand contain the HL point elasticities. The only brand where we see a lack of intersection between our estimates and the HL estimates is Charmin, and this can be explained by the fact that we restricted attention to the data before 1993 to avoid the product introduction of Charmin's ultra line of products, whereas HL use data from 1992 to 1995 . The Charmin "ultra" line and its popularity undoubtedly made Charmin a less elastic overall brand. We also note that the naive 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). Our finding of more elastic demand when the many markets features of the data are taken seriously, which our approach does, 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 potentially points to one possible source of this general problem and its solution.

Table 6: Own Price Elasticity Comparison

Brand	95% CI 1 ($c = 0.5$)	95% CI 2 ($c = 0.6$)	IV Logit 95% CI	Hausman and Leonard	Brand-Level IV Logit 95% CI
Angel Soft	[-5.33, -3.58]	[-4.88, -3.64]	[-2.23, -1.30]	-4.07	[-1.89, -1.48]
Charmin	[-8.66, -5.82]	[-7.93, -5.92]	[-3.61, -2.11]	-2.29	[-3.21, -2.52]
Cottonelle	[-7.48, -5.03]	[-6.86, -5.11]	[-3.12, -1.83]	-3.29	[-2.65, -2.08]
Kleenex	[-5.09, -3.42]	[-4.66, -3.48]	[-2.12, -1.24]	-3.29	[-1.80, -1.42]
Quilted Northern	[-6.53, -4.39]	[-5.98, -4.46]	[-2.73, -1.59]	-3.08	[-2.31, -1.82]
Scott	[-2.97, -1.99]	[-2.72, -2.03]	[-1.24, -0.72]	-1.80	[-1.05, -0.83]

6 Conclusion

We have shown that in the many markets setting, i.e., the researcher has a sample of separate markets for a differentiated good, the sampling error in market shares and the strategic dependence among products within a market give rise to fundamentally new identification and inference problems for demand estimation. In particular we show that 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 "many markets" structure of the data into account has economically important implications for price elasticities.

A key message from our analysis is that it is critical to not ignore the sampling variability in shares when working with "many markets" 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 relative to the number of products can be small. This can manifest itself in a particularly problematic fashion for demand estimation - the data can exhibit zero market shares for some products. Standard discrete choice models always predicts positive market level demand, and hence the mere observation of a zero share in the data rejects the model. On the other hand, if the researcher systematically leaves these products out of the estimation (a common strategy in practice, seemingly justified by associating these products with the "outside good"), this causes a selection problem that biases elasticities in the direction of being too inelastic. However a zero share in the data can be seen as an entirely natural outcome when sampling variability is taken into account - zeroes are merely the outcome of sampling error (i.e., not enough consumer draws) when the underlying choice probabilities of products in the market are small relative to the outside option (which is the norm in applications). Thus our empirical strategy accommodates the presence of zero shares as a natural consequence of sampling variability in market shares, and our application to the DFF scanner data indeed shows that the proper treatment of the zeroes in the data can have a major impact on the measurement of price elasticities.

A potentially fruitful area for future applications of our approach is to individual level choice data where aggregation is necessary to control for price endogeneity, such as described by Berry, Levinsohn, and Pakes (2004) and Goolsbee and Petrin (2004). If the demand data is a consumer survey or household panel, sampling variability in market shares when we aggregate the data is a clear problem, and the approach we describe offers a novel solution for using micro data sets to address the joint problem of endogenous prices and consumer heterogeneity.

A Proof of Theorem 1

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: $P_s^\zeta = \{\vec{p}_s(1\{\cdot \geq x\}) : x \in (\zeta, 1-\zeta)\}$ 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 $\vec{p}_s^\dagger \in int(co(P_s^{1/m^*}))$. 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 non-zero points in $(\zeta, 1-\zeta)$. For $j = 1, ..., n$, define the vector $\vec{p}_s^j$ be

$$p_s^j(k) = \begin{pmatrix} n \\ k \end{pmatrix} (x_j)^k (1-x_j)^{n-k}. \tag{A.5}$$

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

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

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, $\vec{p}_s \in cl(co(P_s))$. 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 generalized Carathéodory's theorem (Steinitz (1914)) for convex hull, 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 $\vec{p}_s \in \cup_{m=1}^{\infty} co(P_s^{1/m^*})$. 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 $f(\eta) := E\left[\sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right)|n_t, \pi_t, x_t\right]$ 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 a given realization of $\tilde{s}_t$ and observe that $\tilde{s}_t + \eta e_j \geq \tilde{s}_t$ for $\eta > 0$. Thus using the fact σ^{-1} is an inverse isotone mapping as shown by Theorem 1 in Berry, Gandhi, and Haile (2011), we have that $\sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right) \geq \sigma_j^{-1}\left(\tilde{s}_t, x_t, \lambda\right)$. Strict monotoncity follows from the fact that $\sigma^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right) \neq \sigma^{-1}\left(\tilde{s}_t, x_t, \lambda\right)$ (because inverse isotone implies σ is invertible) and the fact the connected substitutes structure in Berry, Gandhi, and Haile (2011) (which is satisfied by our model) ensures that $\sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right) \neq \sigma_j^{-1}\left(\tilde{s}_t, x_t; \lambda\right)$. Because this holds for all realizations of $\tilde{s}_t$, strict monotonicity also hold for the expectation taken with respect to realizations of $\tilde{s}_t$. Observe finally that as $\eta \to -1/(n_t + J_t + 1)$ then the share of good j in the vector $\tilde{s}_t + \eta \cdot e_j$ is approaching 0 for the realization $\tilde{s}_t = 0$, and thus $\sigma_j^{-1}\left(\tilde{s}_t + \eta \cdot e_j, x_t; \lambda\right)$ must approach $-\infty$ for the realization $\tilde{s}_t = 0$ as a consequence of the full support assumption on ν_{ijt}. However, because all other realizations of $\tilde{s}_t$ 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 $\tilde{s}_t$ 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 $\tilde{s}_{jt} = 1$. □

Example: Logit Demand

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

$$\max_{\pi_t \in \Delta_{J_t}^{\varepsilon}, \eta \in [-1/(n_t + J_t + 1), 1/(n_t + J_t + 1)]} \eta$$

$$s.t.\ E\left[\sigma_j^{-1}\left(\frac{n_t s_t + 1}{n_t + J_t + 1} + \eta \cdot e_j, x_t; \lambda_0\right)|n_t, \pi_t, x_t, z_t, J_t\right] = \sigma_j^{-1}\left(\pi_t, x_t; \lambda_0\right), \tag{B.1}$$

where $n_t s_t | n_t, \pi_t, x_t, z_t, J_t \sim MN(n_t, \pi_t)$. 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 $\pi_{0t} = 1 - \sum_{j=1}^{J_t} \pi_{jt}$. Then, the constraint in the above problem is simplified to

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

where $n_t(s_{jt}, s_{0t}, 1-s_{jt}-s_{0t})|n_t, \pi_t \sim MN(n_t, (\pi_{jt}, \pi_{0t}, 1-\pi_{jt}-\pi_{0t}))$. 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) $\eta^u_{jt} = -\eta^l_{jt}$ 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 generalized 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 (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.

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 the 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 $\bar{\mathcal{G}} = \mathcal{G} \cap \{1\}$. 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 $\{\rho(w_t, \theta^{(1)}, g^{(1)})\rho(w_t, \theta^{(2)}, g^{(2)}) : (\theta^{(1)}, g^{(1)}), (\theta^{(2)}, g^{(2)}) \in \Gamma^{-1}(\gamma)\times \bar{\mathcal{G}}\}$ 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 $(\theta^{(1)}, \theta^{(2)})$, $\sup_{(\gamma,F)\in\mathcal{H},g\in\mathcal{G}} ||vec(G_F(\theta^{(1)}, g)) - vec(G_F(\theta^{(2)}, g))|| \leq C \times ||\theta^{(1)} - \theta^{(2)}||^{\delta_1}$, 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 set of $k \times k$ positive semi-definite matrices, and $\{vech(\Sigma_F(\cdot, g)) : \Gamma^{-1}(\gamma) \to R^{(d_m^2+d_m)/2} : (\gamma, F) \in \mathcal{H}, g \in \bar{\mathcal{G}}\}$ 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 Liptschiz coefficient being integrable and (2) the set $\mathcal{C}^z$ in the definition of $\mathcal{G}^z$ forms a Vapnik-Chervonenkis set. The Liptschitz 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 $(\gamma, F) \in \mathcal{H}$ and $\theta \in \Gamma^{-1}(\gamma)$.*

Remark. Part (b) is a 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 3 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)$\limsup_{T\to\infty} \sup_{(\gamma,F)\in\mathcal{H}_0} \sup_{\theta\in\Gamma^{-1}(\gamma)} \int_{\mathcal{G}/\mathcal{G}_T} S(\sqrt{T}\rho_F(\theta, g), \Sigma_F(\theta, g))d\mu(g) = 0$.

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

Assumption C.6. (a) *$S(m, \Sigma)$ is continuous in (m, Σ) on $(-\infty, \infty]^{d_m} \times \Psi^\xi$ for some $\xi > 0$.*

(b) *There exists a constant $C > 0$ and $\xi > 0$ such that for any $m_1, m_2 \in R^{d_m}$ 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 nonincreasing in m.*

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

(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 (4.12) as well as the SUM and MAX functions in Andrews and Shi (2009):

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

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 (2009): $S(m, \Sigma) = \min_{t\geq0}(m - t)'\Sigma^{-1}(m - t)$. The QLR functions are computationally much more cumbersome than the other choices, as discussed in Andrews and Shi (2009), and thus much less appealing in practice. On the other hand, imposing these two assumptions makes our proof techniques (using uniform asymptotic approximation) possible.

Lemma C.1. (a) *Assumption* C.6 *is satisfied by the S function in* (4.12) *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 3 and C.1

Proof of Lemma 3. (a) Assumptions C.2(c)-(d) imply that under F,

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

The second convergence implies that

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

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 $[-M^*,M^*]^{d_m} \times \Psi$. By Assumption C.2(f) and Equations (D.1) and (D.2), we have $\{(\bar{\rho}_T(\theta,g),\hat{\Sigma}_T^\iota(\theta,g)) : (\theta,g) \in \Gamma^{-1}(\gamma)\times\mathcal{G}\} \subseteq [-M^*-\xi, M^*+\xi]^{d_m} \times \Psi^\xi$ with probability approaching one for any $\xi > 0$. By Assumption C.6(a), $S(m,\Sigma)$ is uniformly continuous on $[-M^*,M^*]^{d_m} \times \Psi$. 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

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

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), $\min_{\theta\in\Gamma^{-1}(\gamma)} Q_F(\theta)\ge 0$, is implied by Assumption C.6(d).

Suppose $\gamma\in\Gamma_{0,F}$. Then there exists a $\theta^*\in\Gamma^{-1}(\gamma)$ such that $\rho_F(\theta^*,g)\ge 0$ for all $g\in\mathcal{G}$ by Lemma 2. 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 $\min_{\theta\in\Gamma^{-1}(\gamma)} Q_F(\theta)\le Q_F(\theta^*)=0$, 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 $j\le d_m$ by Lemma 2. 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

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

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

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

where the first inequality holds by the inequality $(a+b)^2 \le 2(a^2+b^2)$ and the second inequality holds because $(\sigma_{2,j}-\sigma_{1,j})^2 \le |\sigma_{2,j}^2 - \sigma_{1,j}^2| \le ||vech(\Sigma_1-\Sigma_2)||$ and because $\sigma_{1,j}^2, \sigma_{2,j}^2 \ge (\xi_2-\xi)^{1/2}$. Plug (D.6) in (D.5), 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 $\max_{j=1,\dots,k}$. The second change is to replace the Cauchy-Schwartz inequality used in (D.5) by the inequality $|\max_j a_j b_j| \le (\max_j a_j^2 \times \max_j b_j^2)^{1/2}$. 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. Let $\Lambda_T(\theta,\gamma) = \{\lambda :$

$\theta+\lambda/\sqrt{T}\in\Gamma^{-1}(\gamma),\ d(\theta+\lambda/\sqrt{T},\Theta_{0,F}(\gamma))=||\lambda||/\sqrt{T}\}$. The approximation is of the form:

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

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

Theorem E.1. *Suppose that the conditions in Lemma 2 and Assumptions* C.1-C.3 and C.5-C.6 *hold. Then for any real sequence* $\{x_T\}$ *and scalar* $\eta>0$ *,*

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

Theorem E.1 is a key step in the proof of Theorem 2 and is proved in the next 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 Liptschiz norm bounded by one. Let E^* and E_* denote outer and inner expectations respectively and $\Pr^*$ and $\Pr_*$ denote outer and inner probabilities.

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*

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

(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 Liptschiz 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 Liptschitz 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

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

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 $\{x_T\}$,

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

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

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

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

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

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 $(m_2, \Sigma_2) \in R^k \times \Psi^\xi$,

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

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

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

Solve the quadratic inequality for Δ_S, we have

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

This shows (E.9).

Observe that

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

where $c(x) = x + \sqrt{x^2 + 8x}$, 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

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

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

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

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

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

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 $\lambda \in \Lambda_T^{B_2}(\theta, \gamma)$,

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

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:

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

The second result holds by Assumption C.2(e).

Define an intermediate statistic

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

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 $\lambda \in \Lambda_T^{B_2}(\theta,\gamma)$,

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

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

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

where $\bar{C}$ is any constant such that $\bar{C} > 2C \times (6\sup_{\Sigma\in\Psi} S(-B_1 1_k,\Sigma) + 6B_2)$, 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) and the uniform Donsker theorem, 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

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

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

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

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

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

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

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

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

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

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

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 following 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 $\hat{\nu}_T(\theta,g) : \Theta\times\mathcal{G} \to R^{d_m}$. The uniform equicontinuity is implied by Assumptions C.2(b) and (c). 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) is implied by: for any $\theta \in \Theta_{0,F}(\gamma)$ and $\lambda \in \Lambda_T^{B_2}(\theta,\gamma)$,

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

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(b) and (c) imply that for any $\epsilon > 0$, there exists $B_{1,\epsilon}$ large enough such that

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

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:

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

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 (b), (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

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

But

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

where $c(x)=(x+\sqrt{x^2+8x})$ 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

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

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

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

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

STEP 4. In order to show (E.7), 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

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

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

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

where the rhs 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

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

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 there exists a λ^* such that

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

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), $I(0) \leq \sup_{\theta\in\Theta_{0,F}(\gamma)} \int_{\mathcal{G}} S(\nu_F(\theta,g), \Sigma^\iota_F(\theta,g))d\mu(g)$.. Thus, $I(a^*\lambda^{**}) < \sup_{\Sigma\in\Psi^\xi} S(-B_{1,\epsilon}1_k, \Sigma)$. 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

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

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

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

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

$$\begin{aligned}
\Delta_T = &||G_F(\theta^*,g)\lambda^* + \sqrt{T}\rho_F(\theta^*,g) - \sqrt{T}\rho_F(\theta^* + \lambda^*/\sqrt{T}, g)||^2\\
&+ ||vech(\Sigma^\iota_F(\theta^* + \lambda^*/\sqrt{T}, g) - \Sigma^\iota_F(\theta^*,g))||. \qquad \text{(E.46)}
\end{aligned}$$

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,

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

Thus

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

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,

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

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

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

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

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*

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

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

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

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 $\{\rho(w_t, \theta, g) : (\theta, g) \in \Theta \times \mathcal{G}\}$ 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:

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

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

(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 $X_T^{(1)} = \hat{\nu}_T^*$, $X_T^{(2)} = \nu_F$, $G_1 = \{\nu : \sup_{||\theta^{(1)}-\theta^{(2)}|| \leq \Delta_\varepsilon, g} ||\nu(\theta^{(1)}, g) - \nu(\theta^{(2)}, g)|| \geq \varepsilon\}$, and $G_0 = \{\nu : \sup_{||\theta^{(1)}-\theta^{(2)}|| \leq \Delta_\varepsilon, g} ||\nu(\theta^{(1)}, g) - \nu(\theta^{(2)}, g)|| > \varepsilon/2\}$, where Δ_ε satisfies:

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

Such a Δ_ε exists because $\{\rho(w_t, \theta, g) : (\theta, g) \in \Theta \times \mathcal{G}\}$ 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,

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

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:

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

Define an intermediate quantity first:

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

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 $\epsilon > 0$,

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

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

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

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 $\{x_T\}$,

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

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

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

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

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

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

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

which implies (E.56).

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

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

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

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

where the first inequality holds by (E.67), the second inequality holds by Theorem E.1 and the third inequality holds because $T_{\kappa_T}^{appr}(\gamma) \geq T_T^{appr}(\gamma)$ 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 $T_{\kappa_T}^{appr}(\gamma)$ 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, $F_{S_n}(x,\gamma)$ 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 $\{x_T\}$,

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

When following the arguments for Theorem E.1, we simply need to observe the resemblence 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 $\{x_n\}$,

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

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

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

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

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.

Andrews, D. W. K., and X. Shi (2009): "Inference Based on Conditional Moment Inequality Models," unpublished manuscript, Department of Economics, Yale University.

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.

——— (2004): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 112(1), 68–105.

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.

Capps, C., D. Dranove, and M. Satterthwaite (2003): "Competition and Market Power in Option Demand Markets," *RAND Journal of Economics*, 34, 737–763.

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.

Fan, Y. (2008): "Market Structure and Product Quality in the U.S. Daily Newspaper Market," Discussion paper, Yale University.

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

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

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.

Ho, K. (2007): "Insurer-Provider Networks in the Medical Care Market," Discussion paper, Columbia University.

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.

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

Lustig, J. (2008): "The Welfare Effects of Adverse Selection in Privatized Medicare," Discussion paper, Boston University.

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

Nevo, A. (2000): "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry," *RAND Journal of Economics*, 31, 395–421.

——— (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69, 307–342.

Romano, J., and A. Shaikh (2008): *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.

Rysman, M. (2004): "Competition Between Networks: A Study of the Market for Yellow Pages," *Review of Economic Studies*, 71, 483–512.

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

Steinitz, E. (1914): "Bedingt konvergente Reihen und konvexe Systeme," *J. Reine Angew. Math.*, 143, 128–175.

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

Villas-Boas, S. (2007): "Vertical Relationships between Manufacturers and Retailers: Inference with Limited Data," *Review of Economic Studies*, 74, 625–652.

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.