Freyberger, Joachim

Working Paper

Asymptotic theory for differentiated products demand models with many markets

cemmap working paper, No. CWP19/12

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Freyberger, Joachim (2012) : Asymptotic theory for differentiated products demand models with many markets, cemmap working paper, No. CWP19/12, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2012.1912

This Version is available at:
https://hdl.handle.net/10419/64713

cemmap
centre for microdata methods and practice

Asymptotic theory for differentiated products demand models with many markets

Joachim Freyberger

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP19/12

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Asymptotic theory for differentiated products demand models with many markets*

Joachim Freyberger‡

This version: April 15, 2012†

Abstract

This paper develops asymptotic theory for estimated parameters in differentiated product demand systems with a fixed number of products, as the number of markets T increases, taking into account that the market shares are approximated by Monte Carlo integration. It is shown that the estimated parameters are $\sqrt{T}$ consistent and asymptotically normal as long as the number of simulations R grows fast enough relative to T. Monte Carlo integration induces both additional variance as well additional bias terms in the asymptotic expansion of the estimator. If R does not increase as fast as T, the leading bias term dominates the leading variance term and the asymptotic distribution might not be centered at 0. This paper suggests methods to eliminate the leading bias term from the asymptotic expansion. Furthermore, an adjustment to the asymptotic variance is proposed that takes the leading variance term into account. Monte Carlo results show that these adjustments, which are easy to compute, should be used in applications to avoid severe undercoverage caused by the simulation error.

Keywords: Demand estimation, differentiated products, many markets, asymptotic theory, simulation error, bias correction, adjusted standard errors.

*I am grateful to Ivan Canay, Joel Horowitz, and Elie Tamer for helpful discussions and suggestions. I have also received valuable feedback from Mike Abito, Steven Berry, Mark Chicu, Roland Eisenhuth, Amit Gandhi, Aviv Nevo, Henrique de Oliveira, and Ketan Patel.

‡Northwestern University, Department of Economics. Email: j.freyberger@u.northwestern.edu.

†First version: November 10, 2010.

1 Introduction

Discrete choice models have been widely used in the empirical industrial economics literature to estimate demand for differentiated products. In these models consumers in market t can typically choose one of J_t products or an outside option. The market share of good j in market t is calculated as the probability that a consumer chooses good j given prices and product characteristics. In models with heterogeneous consumers, such as the model of Berry, Levinsohn, and Pakes (1995) (often referred to as the BLP model), the market shares involve integrals over the distribution of random coefficients. When estimating the parameters of the model these integrals cannot be calculated analytically. Therefore, they are usually approximated by Monte Carlo integration with R random draws from the known distribution of the random coefficients.[1] The limiting distribution of the estimated parameters can be obtained by either letting the number of products, the number of markets, or both approach infinity. Since the asymptotic distribution of the estimator serves as an approximation of its unknown finite sample distribution, it depends on the particular data set which approximation is most suitable. While in some cases using an approximation where the number of products approaches infinity is appropriate (as in Berry, Levinsohn, and Pakes (1995)), in other cases the number of markets is a lot larger than the number of products (e.g. Nevo (2001)). As shown in this paper the asymptotic properties of the estimator differ a lot depending on which approximation is used. Therefore, it is important that both approximations are well understood.

Berry, Linton, and Pakes (2004) provide asymptotic theory for estimating the parameters of differentiated product demand systems for a *large number of products in one market*. They allow for three sources of errors: The sampling error in estimating the market shares, the simulation error in approximating the shares predicted by the model, and the underlying sampling error. In their paper all market shares go to 0 at the rate $1/J$ and a necessary condition for asymptotic normality is that J^2/R is bounded. In this case, without a sampling error in estimating the market shares, their estimator $\hat{\theta}$ of the parameter vector θ_0 satisfies

$$\sqrt{J}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, V_{GMM} + \lambda_1 V_{MC}\right)$$

where $\lambda_1 = \lim_{J,R\to\infty} J^2/R$. Here V_{GMM} denotes the variance of the estimator when the integral is calculated exactly and V_{MC} is an additional variance term due to the simulation

[1] Although the focus lies on the effect of using Monte Carlo integration to approximate integrals, non-stochastic approximations such as quadrature rules are also discussed.

error. Hence, if J^2/R is bounded away from 0, the asymptotic distribution is centered at 0 but the use of Monte Carlo integration leads to a larger variance.

This paper is concerned with the asymptotic theory for a *fixed number of products, J, in a growing number of markets T*. Since a market can be defined as a geographic region but also as different time periods, in many data sets used in applications the number of markets is a lot larger than the number of products (see for example Nevo (2001), Kim (2004), and Villas-Boas (2007)). In the study of Nevo (2001), for example, the number of markets is 1124 while the number of products is 24. For these cases, using an asymptotic approximation where the number of markets approaches infinity and the number of products is fixed is the natural choice. However, this setup has not been investigated in the literature so far. Furthermore, in a similar (but more general) class of models, Berry and Haile (2010) provide non-parametric identification results for a large number of markets and a fixed number of products. These identification results can serve as a basis for non-parametric or semi-parametric estimation of the model. Before such a flexible estimation procedure is developed, it is interesting to know how the commonly used fully parametric estimators behave in this setup. I prove consistency and asymptotic normality in a general setup for these cases where T approaches infinity and J is fixed. I use Nevo's widely used parameterization of the BLP model as an example throughout the paper. For this model, I also provide intuitive conditions for local identification and show under which circumstances point identification fails.

I find that the estimated parameters are $\sqrt{T}$ consistent and asymptotically normal as long as $\sqrt{T}/R$ is bounded. In this case, $\hat{\theta}$ satisfies

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(\lambda_2 \mu, V_{GMM}\right)$$

where $\lambda_2 = \lim_{T,R\to\infty} \sqrt{T}/R$, again V_{GMM} is the variance of the estimator without integration error, and $\mu \neq 0$. Hence, if $\sqrt{T}/R$ is bounded away from 0, Monte Carlo integration (as opposed to evaluating the integral) leads to an asymptotic normal distribution of the estimated parameters which is not centered at 0. The reason for this result is that Monte Carlo integration yields both additional bias as well as additional variance terms in the asymptotic expansion. If one uses different draws to evaluate the integral in different markets, the leading bias term may dominate the leading variance term. The leading additional variance term is of order $O_p(1/\sqrt{R})$ and does not enter the first order asymptotic distribution as long as $R \to \infty$. These results rely on using *different draws to approximate the integral in different markets*. If the same R draws are used in all markets one needs T/R to be bounded

to obtain $\sqrt{T}$ consistency which means that more draws are needed to approximate each integral relative to the number of markets.

In a similar way one can introduce sampling error in the observed markets shares. Similar to Berry, Linton, and Pakes (2004), one could assume that one does not observe the true markets shares but an approximation from n random consumers. Observing only approximated market shares leads to additional bias and variance terms in the asymptotic expansion. The rate at which n has grow to relative to T in order to obtain $\sqrt{T}$ consistency is identical to the rate requirement for R.

From these two asymptotic distributions it is apparent that there are important differences between letting the number of products or the number of markets approach infinity. With a large number of products, it is important to correct the variance for the use of Monte Carlo integration. With a large number of markets and different draws in each market, the asymptotic distribution might not be centered at 0 and hence, a bias corrected estimator is needed. In both cases if R is too small, confidence intervals based on the usual asymptotic GMM distribution have the wrong size even asymptotically. Also notice that with J approaching infinity one needs that J^2/R goes to 0 for an asymptotic distribution that is not affected by Monte Carlo integration. Contrary, if T goes to infinity one only needs that $\sqrt{T}/R$ goes to 0 to obtain this result. Hence, a lot fewer draws to evaluate each integral are needed relative to the sample size.

The finite sample properties of the estimator depend on the number of draws R due to both the additional bias and the additional variance induced by simulations. I suggest two different methods that allow eliminating the leading bias term from the asymptotic distribution. One method is an analytical bias adjustment. The other method is a jackknife bias adjustment. I also show how one can easily incorporate the leading variance term when calculating standard errors. These two corrections greatly improve finite sample results. In particular, Monte Carlo simulations demonstrate that using a small number of draws in comparison to the number of markets and using the usual GMM asymptotic distribution can yield distorted inference while the use of bias correction and adapted standard errors leads to a considerably better performance.

The focus of this paper is on understanding the asymptotic theory when Monte Carlo inte-

gration is employed because this is the method which has almost exclusively been used in applications. Furthermore, an advantage of Monte Carlo integration is that it can easily be used to integrate over complicated distributions such as the joint distribution of demographic characteristics as in Nevo (2001). For these distributions there is no closed form expression for the density function. An interesting alternative is to use non-stochastic approximations, such as quadrature rules recently advocated by Judd and Skrainka (2011). These approximations are shown to perform well in simulations when integrating over a normal distribution. However, it is not clear how a distribution of demographic characteristics, which does not have a closed form expression, can be handled with quadrature rules. Although the focus lies on Monte Carlo integration, the asymptotic expansions derived in this paper also provide insights into finite sample bias from non-stochastic approximations.

These results might suggest that practitioners can simply use a very large number of draws and ignore Monte Carlo integration issues. Although this might be feasible depending on the model and computing resources available, in applications this is often not possible for several reasons. First, one does not know in advance how many draws suffice to obtain satisfactory results. As discussed in Section 4, the number of draws needed depends, among others, on the sample size, the number of random coefficients as well as unknown parameters, such as the variance of the random coefficients. Second, taking a very large number of draws is computationally very demanding because one needs to solve a complicated nonlinear optimization problem to estimate the parameters. The Monte Carlo results of the random coefficients logit model presented in Section 4 are based on a small number of products ($J = 4$) and five random coefficients to make the problem tractable. However, in the same setup as in Section 4 but with a sample size of $J = 24$ and $T = 1,124$ (as in Nevo (2001)) it takes around 24 hours to minimize the objective function when $R = 2,000$ and the the starting values of the parameters are close to the true values.[2] Since we are dealing with a nonlinear optimization problem one needs to use several different starting values in applications. With an even larger number of draws or with a larger sample size estimating the model can take more than one week. Taking a smaller number of draws considerably speeds up calculations. Third, even when taking the same draws for each product and each random coefficient, the number of draws needed is $T \times R$. In the previous example this means that $2,248,000$ draws are used to calculate the shares and the draws have to be stored before

[2]Computational details are presented in the Monte Carlo section. The programs are run on the Northwestern Social Sciences Computing Cluster which use 260 AMD Opteron CPU cores, running at 2.8Ghz.

optimizing the function. As a consequence, more than 10 GB of memory is needed to run the program which is used to do the simulations in this paper. Finally, in case one wants to integrate over empirical distributions of demographic characteristics, R is constrained by the number of people in the database for a certain market.

The implication for empirical work that makes use of Monte Carlo integration is that in practice one should always use bias corrections and standard errors that correct for the simulation error. If the number of simulations is sufficiently large, the bias correction is close to 0 and the corrected standard errors will be very close to the usual GMM standard errors. If the number of simulations is small, the simulation error will affect the finite sample performance of the estimator, the usual GMM standard errors underestimate the true variance, and the estimates might be severely biased. In this case the proposed corrections, which can be computed easily, considerably improve the finite sample performance. Nevertheless, the number of draws should be as large as possible, subject to computational constraints and data availability, in order to improve the precision of the initial estimate which is used to calculate the bias correction.

The results in this paper are related to similar findings of Lee (1995) in simulated maximum likelihood estimation of discrete choice models. Lee (1995) also finds that the asymptotic distribution might not be centered at 0 if the number of draws is small relative to the sample size. Furthermore, similar bias corrections have been proposed by Arellano and Hahn (2007) in nonlinear fixed effects panel data models and by Kristensen and Salanie (2010) for a general class of approximate estimators. The results in the aforementioned papers do not directly apply to the setup presented here because there is no closed form analytic expression for the objective function which is constructed by solving a system of equations. The results are also in line with simulation results in a recent study by Judd and Skrainka (2011) who find among others finite sample bias and excessively tight standard errors when using Monte Carlo integration. Other recent contributions to literature on estimation of discrete choice demand models include Gandhi and Kim (2011) and Armstrong (2012). In both papers $J \to \infty$. Gandhi and Kim (2011) allow for interactions between the unobserved demand error and product characteristics which affects both the identification arguments and estimation method. Armstrong (2012) discusses the validity of commonly used instruments in models with a large number of products (i.e. $J \to \infty$) under conditions on economic primitives. He shows that in several models using product characteristics as instruments for price

yields inconsistent estimates and he shows how consistent estimates can be obtained.

The remainder of this paper is organized as follows. The next section introduces the models of Berry, Levinsohn, and Pakes (1995) and Nevo (2001) in detail. I provide intuitive conditions for local identification and show under which circumstances point identification fails. Then I prove consistency and asymptotic normality in a general setup. I argue that the assumptions are satisfied in the model of Nevo (2001). Finally, I discuss Monte Carlo results to demonstrate that the proposed corrections considerably improve the finite sample performance.

2 A motivating example

The following widely used model of Berry, Levinsohn, and Pakes (1995), and in particular the parameterization of Nevo (2001), is used as a motivating example throughout the paper. The asymptotic theory provided in the section holds more generally. Since identification is one of the main assumption in the next sections, I provide intuitive conditions for local identification for this model and show under which circumstances point identification fails.

2.1 Model

In this model, the utility of consumer i from product j in market t is assumed to be

$$u_{i,j,t} = x_{1,j,t}\beta_{1,i,t} + x_{2,j}\beta_{2,i,t} - \alpha_{i,t}p_{j,t} + \xi_j + \Delta\xi_{j,t} + \epsilon_{i,j,t}, \quad i = 1,\dots,I,\ j = 1,\dots,J,\ t = 1,\dots,T.$$

Here $x_{1,j,t}$ is a K_1 dimensional vector of product characteristics that differ across markets and $x_{2,j}$ is a K_2 dimensional vector of product characteristics that are identical in all markets. The price of product j in market t is denoted by $p_{j,t}$. Unobserved product characteristics which are identical in each market for product j are called ξ_j, while $\Delta\xi_{j,t}$ are the unobserved characteristics that vary across markets (deviations from the mean ξ_j). Finally, $\epsilon_{i,j,t}$ is a mean zero stochastic term. The utility of an outside option, $j = 0$, is standardized to 0. Nevo (2001) assumes that

$$\begin{pmatrix} \alpha_{i,t} \\ \beta_{1,i,t} \\ \beta_{2,i,t} \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} + \Pi D_{i,t} + \Sigma v_i, \qquad v_i \sim N(0, I_{K+1})$$

where $K = K_1 + K_2$. Here, $D_{i,t}$ is a draw from the known distribution of demographics, with dimension $B \times 1$ which may vary across markets. The matrix Σ is assumed to be

diagonal. Next define $\beta_i = (\beta'_{1,i} \ \beta'_{2,i})'$ and $\theta = (\theta'_1, \theta'_2)'$ where $\theta_1 = (\alpha, \beta', \xi_1, \ldots, \xi_J)'$ and $\theta_2 = (vec(\Pi)', vec(\Sigma)')'$. Using the previous specifications the utility can be rewritten as

$$u_{i,j,t} = \delta(x_{1,j,t}, x_{2,j}, p_{j,t}, \xi_j, \Delta\xi_{j,t}; \theta_1) + \mu(x_{1,j,t}, x_{2,j}, p_{j,t}, v_i, D_i; \theta_2) + \epsilon_{i,j,t},$$

where

$$\begin{aligned} \delta(x_{1,j,t}, x_{2,j}, p_{j,t}, \xi_j, \Delta\xi_{j,t}; \theta_1) &\equiv x_{1,j,t}\beta_1 + x_{2,j}\beta_2 - \alpha p_{j,t} + \xi_j + \Delta\xi_{j,t}, \\ \mu(x_{1,j,t}, x_{2,j}, p_{j,t}, v_i, D_i; \theta_2) &\equiv [p_{j,t}, x_{1,j,t}, x_{2,j}]'(\Pi D_{i,t} + \Sigma v_i). \end{aligned}$$

Also define

$$\gamma(x_{1,j,t}, x_{2,j}, p_{j,t}, \xi_j, \Delta\xi_{j,t}, v_i, D_i; \theta) \equiv \delta(x_{1,j,t}, x_{2,j}, p_{j,t}, \xi_j, \Delta\xi_{j,t}; \theta_1) + \mu(x_{1,j,t}, x_{2,j}, p_{j,t}, v_i, D_i; \theta_2).$$

Furthermore, there are instruments $z_t \in \mathbb{R}^{J \times L}$ such that

$$E(z'_t \Delta\xi_t) = 0.$$

Assuming that $\epsilon_{i,j,t}$ is independent extreme value distributed, it can be shown that the market share of product j in market t is given by

$$s_{j,t} = \int \frac{\exp(\gamma(x_{1,j,t}, x_{2,j}, p_{j,t}, \xi_j, \Delta\xi_{j,t}, v, D; \theta))}{1 + \sum_{m=1}^{J} \exp(\gamma(x_{1,m,t}, x_{2,m}, p_{m,t}, \xi_m, \Delta\xi_{m,t}, v, D; \theta))} dQ_t(D) dF(v).$$

The distribution functions of demographics and v in market t are denoted by Q_t and F, respectively, where the latter is market invariant.

For each market t, given data on market shares, prices, and product characteristics, as well as a value of θ_2, Berry (1994) showed that there is a unique value of $(\delta_{1,t}, \ldots, \delta_{J,t})$ that sets the observed market shares equal to the market shares generated by the model. That is for each market t,

$$s_{j,t} = \int \frac{\exp(\delta_{j,t} + \mu(x_{1,j,t}, x_{2,j}, p_{j,t}, v, D; \theta_2))}{1 + \sum_{m=1}^{J} \exp(\delta_{m,t} + \mu(x_{1,m,t}, x_{2,m}, p_{m,t}, v, D; \theta_2))} dQ_t(D) dF(v) \qquad (1)$$

with $j = 1, \ldots, J$ constitutes a system of J equations and J unknowns $(\delta_{1,t}, \ldots, \delta_{J,t})$ which can be solved uniquely. Nevo (2001) denotes this solution by $\delta_{j,t}(s_t, x_{1,t}, x_2, p_t; \theta_2)$ for $j = 1, \ldots, J$. The model is commonly estimated by taking R random draws from the known distributions of v and D, approximating the integral by an average, and solving the above system of equations numerically for $\delta_{j,t}(s_t, x_{1,t}, x_2, p_t; \theta_2)$. Now define

$$\omega_{j,t}(\theta) = \delta_{j,t}(s_t, x_{1,t}, x_2, p_t; \theta_2) - x_{2,j}\beta_2 - \xi_j - x_{1,j,t}\beta_1 + \alpha p_{j,t}$$

which can then be used to estimate the parameters by a nonlinear instrumental variables estimator. The identification condition is that there is a unique value θ_0 such that

$$E(z_t'\Delta\xi_t) = E(z_t'\omega_t(\theta_0)) = 0.$$

This identification condition cannot hold if $\xi_j \neq 0$ and $K_2 > 0$ due to collinearity of the regressors. Thus, Nevo defines product dummies $d_j = x_{2,j}\beta_2 + \xi_j$ and uses instead

$$\omega_{j,t}(\theta) = \delta_{j,t}(s_t, x_{1,t}, x_2, p_t; \theta_2) - d_j - x_{1,j,t}\beta_1 + \alpha p_{j,t}$$

where now $\theta_1 = (\alpha, \beta_1, d_1, \ldots, d_J)$. Notice that one can assume without loss of generality that $E(\Delta\xi_{j,t}) = 0$ because ξ_j captures the mean of each product, but not that $E(\xi_j + \Delta\xi_{j,t}) = 0$. Using this parameterization, all price elasticities can be identified. I discuss identification in more detail below.

The model is estimated using instrumental variables and not simply by a minimum distance procedure because it is usually assumed that prices are endogenous in the sense that they are correlated with $\Delta\xi_t$. The reason is that firms take all product characteristics into account when setting prices. The analysis remains the same if some of the observed product characteristics, $x_{1,t}$, are treated as endogenous as well. If $x_{1,t}$ is not correlated with $\Delta\xi_t$, then z_t includes $x_{1,t}$.

2.2 Identification

The consistency arguments below require that the model is point identified. Nonparametric identification in a setup that nests the one in this paper is shown in Berry and Haile (2010). This, however, does not imply that any parameterization of the model yields point identification. In order to provide insight on identification of the model of Nevo (2001), I mainly focus on local identification conditions and discuss under which circumstances point identification fails.

First define $\varsigma = diag(\Sigma)$. As already mentioned, β_2 cannot be identified. Furthermore, if there are more product characteristics that do not change over markets than products, the model is not identified. To see this, notice that $\mu(x_{1,j,t}, x_{2,j}, p_{j,t}, v, D; \theta_2)$ contains

$$\sum_{k=K_1+1}^{K} x_{2,j}^k \varsigma_k v^k \sim N\left(0, \sum_{k=K_1+1}^{K} \left(x_{2,j}^k \varsigma_k\right)^2\right).$$

Now define

$$\varphi_j^2 = \sum_{k=K_1+1}^{K} \left(x_{2,j}^k \varsigma_k^0\right)^2, \qquad j = 1, \ldots, J$$

where ς^0 is the true value of ς. But if $K_2 > J$, then

$$\varphi_j^2 = \sum_{k=K_1+1}^{K} \left(x_{2,j}^k \varsigma_k\right)^2, \qquad j = 1, \ldots, J$$

is a system of more unknowns (K_2) than equations (J). Hence, ς^0 is not the unique solution. It follows that the model is not identified without normalizations because several parameter values predict the same market shares given the product characteristics.[3] Notice that this argument relies on the facts that the normal distribution is just determined by the mean and variance and that a sum of normals is normal. So if v had a different distribution, one might get identification from higher order moments. In case $K_2 > J$ one can normalize $\varsigma_{K_1+J+1} = \ldots = \varsigma_K = 0$.

Similarly observe that $\mu(x_{1,j,t}, x_{2,j}, p_{j,t}, v, D; \theta_2)$ contains

$$\sum_{r=1}^{B} D_b \sum_{k=K_1+1}^{K} x_{2,j}^k \pi_{kb} \qquad j = 1, \ldots, J.$$

Again, if $K_2 > J$, for each $b = 1, \ldots, B$ different values of π_{kb} yield the same value

$$\sum_{k=K_1+1}^{K} x_{2,j}^k \pi_{kb} \qquad j = 1, \ldots, J.$$

Thus, one has to normalize similar as before.

Sufficient conditions for local identification in GMM estimation include that the matrix

$$\frac{\partial}{\partial \theta} E(z_t' \omega_t(\theta))\Big|_{\theta=\theta_0}$$

has full rank (see Rothenberg (1971)). This is convenient here because there is a closed form expression for the derivative of the moment condition although there is no closed form expression for the objective function. For simplicity assume that there is only one inside and one outside good. Then the index j can be dropped. Moreover, assume that there are

[3] This result is similar to non-identification in a linear IV model if the matrix of regressors does not have full rank. For any vector of parameter values there is a different parameter vector that predicts the same distribution of the dependent variable given the regressors.

two product characteristics with random coefficients, one that changes over markets and one that does not. Furthermore, assume that $B = 1$. Then

$$\omega_t(\theta) = \delta_t(s_t, x_{1,t}, x_2, p_t; \varsigma_1, \varsigma_2, \pi_1, \pi_2) - d - x_{1,t}\beta_1 + \alpha p_t,$$

where $\delta_t = \delta_t(s_t, x_{1,t}, x_2, p_t; \varsigma_1, \varsigma_2, \pi_1, \pi_2)$ solves

$$s_t = \int \frac{\exp(\delta_t + x_{1,t}\varsigma_1 v^1 + x_{1,t}\pi_1 D + x_2\varsigma_2 v^2 + x_2\pi_2 D)}{1 + \exp(\delta_t + x_{1,t}\varsigma_1 v^1 + x_{1,t}\pi_1 D + x_2\varsigma_2 v^2 + x_2\pi_2 D)} dQ_t(D)dF(v).$$

Here s_t is the market share of the inside good in market t. By the Implicit Function Theorem

$$\begin{aligned}
\frac{\partial \delta_t}{\partial \varsigma_k} &= -\frac{\partial s_t}{\partial \varsigma_k}\left(\frac{\partial s_t}{\partial \delta_t}\right)^{-1} \\
\frac{\partial \delta_t}{\partial \pi_k} &= -\frac{\partial s_t}{\partial \pi_k}\left(\frac{\partial s_t}{\partial \delta_t}\right)^{-1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial}{\partial d}E(z_t'\omega_t(\theta))\Big|_{\theta=\theta_0} &= -E\left(z_t'\right) \\
\frac{\partial}{\partial \beta_1}E(z_t'\omega_t(\theta))\Big|_{\theta=\theta_0} &= -E\left(z_t'x_{1,t}\right) \\
\frac{\partial}{\partial \alpha}E(z_t'\omega_t(\theta))\Big|_{\theta=\theta_0} &= E\left(z_t'p_t\right) \\
\frac{\partial}{\partial \varsigma_1}E(z_t'\omega_t(\theta))\Big|_{\theta=\theta_0} &= -E\left(z_t'x_{1,t}\int f(v, x_2, \delta_t; \varsigma_0)v^1 dQ_t(D)dF(v)\right) \\
\frac{\partial}{\partial \varsigma_2}E(z_t'\omega_t(\theta))\Big|_{\theta=\theta_0} &= -E\left(z_t'x_2\int f(v, x_2, \delta_t; \varsigma_0)v^2 dQ_t(D)dF(v)\right) \\
\frac{\partial}{\partial \pi_1}E(z_t'\omega_t(\theta))\Big|_{\theta=\theta_0} &= -E\left(z_t'x_{1,t}\int f(v, x_2, \delta_t; \varsigma_0)D dQ_t(D)dF(v)\right) \\
\frac{\partial}{\partial \pi_2}E(z_t'\omega_t(\theta))\Big|_{\theta=\theta_0} &= -E\left(z_t'x_2\int f(v, x_2, \delta_t; \varsigma_0)D dQ_t(D)dF(v)\right)
\end{aligned}$$

where

$$f(v, D, x_2, x_{1,t}, x_2, p_t, \theta) = \frac{\frac{\exp(\delta_t + x_{1,t}\varsigma_1 v^1 + x_{1,t}\pi_1 D + x_2\varsigma_2 v^2 + x_2\pi_2 D)}{(1+\exp(\delta_t + x_{1,t}\varsigma_1 v^1 + x_{1,t}\pi_1 D + x_2\varsigma_2 v^2 + x_2\pi_2 D))^2}}{\int \frac{\exp(\delta_t + x_{1,t}\varsigma_1 v^1 + x_{1,t}\pi_1 D + x_2\varsigma_2 v^2 + x_2\pi_2 D)}{(1+\exp(\delta_t + x_{1,t}\varsigma_1 v^1 + x_{1,t}\pi_1 D + x_2\varsigma_2 v^2 + x_2\pi_2 D))^2} dQ_t(D)dF(v)}$$

can be seen as a weighting function.

If there are L instrument, the gradient is the $L \times 7$ matrix containing the derivatives. Since for identification the gradient needs to have rank 7, one needs at least 7 instruments (including a constant term or x_2 and $x_{1,t}$). In applications it is often mentioned that one needs

an instrument for the price because the price is endogenous. However, z_t also has to be correlated with the interactions of product characteristics (including price) and weighted averages of distributions. In case one assumes that z_t is mean independent of $\Delta\xi_t$ and that z_t is chosen based on its correlation with the product characteristics, it makes sense to use nonlinear functions of z_t as instruments as well. Similarly, it also may make sense to take the mean of the distribution of demographics in each market as an instrument if the mean varies over markets.[4]

Full rank means that the rows are linearly independent. This fails for example if $\pi_1 = \pi_2 = \alpha = \beta_1 = \varsigma_1 = 0$ (or $\alpha = 0$, no product characteristics change over markets, and the distribution of demographics is the same in all markets). Then $f(v, D, x_2, x_{1,t}, x_2, p_t, \theta)$ only varies over markets because of $\Delta\xi_t$ which usually is assumed to be mean independent of z_t. However, one would of course assume that $\alpha > 0$. Nevertheless, it may happen that no product characteristic changes over markets (as in Nevo (2001)) and that the distributions of the random coefficients are the same in all markets. If in such cases α is close to 0, one is in the setup of Andrews and Cheng (2010) and the normal approximation is poor in finite samples. With more products, the expressions become more complicated, but the previous intuition remains the same.

3 Asymptotic theory

The setup in this section is more general than the motivating example presented in the last section.

3.1 Notation and definitions

I denote the norm of any $m \times n$ matrix A by $||A|| = tr(A'A)^{1/2}$. For any two vectors $a, b \in \mathbb{R}^n$, I define

$$\rho(a, b) \equiv n^{-1}||a - b||^2 = n^{-1}\sum_{i=1}^{n}(a_i - b_i)^2.$$

Furthermore, I define the neighborhoods

$$\mathcal{N}_{\Delta\xi_0}(\theta; \varepsilon) \equiv \{\Delta\xi \in \mathbb{R}^J : \rho(\Delta\xi, \Delta\xi(\theta, P_0)) \leq \varepsilon\},$$

[4]Using mean demographics as instruments has been suggested by Romeo (2010) to improve the numerical performance in the BLP model. See also Abito (2011).

$$\mathcal{N}_{\theta_0}(\varepsilon) \equiv \{\theta \in R^q : \rho(\theta, \theta_0) \leq \varepsilon\},$$

and

$$\mathcal{N}_{P_0}(\varepsilon) \equiv \left\{ P \in \mathcal{P} : \sup_{v \in \mathbb{R}^d} |P_0(v) - P(v)| \leq \varepsilon \right\}.$$

I follow the approach of Berry, Linton, and Pakes (2004) but I neglect the sampling error in the observed market shares. I comment below on how this additional source of error can be incorporated without affecting the main results. I prove consistency and asymptotic normality in a general setup. Let $x_t \equiv (x_{1,t}, x_2) \in \mathbb{R}^{J \times K}$ be the observed product characteristics, $\Delta\xi_t \in \mathbb{R}^J$ the unobserved product characteristics that differ across markets, and $p_t \in \mathbb{R}^J$ the price vector in market t. Define the map $\vartheta : \mathbb{R}^p \times \Theta \times \mathbb{R}^d \to \mathbb{R}^J$ where p is the dimension of the vector of stacked elements of $(x_t, p_t, \Delta\xi_t)$, the parameter space Θ is a compact subset of $\mathbb{R}^q$ where q is the dimension of the parameter vector of interest θ, and d is the dimension of a random variable in market t with known distribution P_t. I assume that $P_t \in \mathcal{P}$ where $\mathcal{P}$ is a space of probability distributions which is restricted in the assumptions that follow. The $J \times 1$ vector of market shares generated by the model is assumed to have the form

$$\sigma(x_t, p_t, \Delta\xi_t, \theta, P_t) \equiv \int \vartheta(x_t, p_t, \Delta\xi_t, \theta, v) dP_t(v), \qquad \forall t = 1, \ldots, T$$

where $\sigma : \mathbb{R}^p \times \Theta \times \mathcal{P} \to \mathbb{R}^J$. For notational convenience, I refer to $\vartheta(x_t, p_t, \Delta\xi_t, \theta, v)$ as $\vartheta_t(\Delta\xi_t, \theta, v)$ and to $\sigma(x_t, p_t, \Delta\xi_t, \theta, P_t)$ as $\sigma_t(\Delta\xi_t, \theta, P_t)$. The jth element of $\sigma_t(\Delta\xi_t, \theta, P_t)$ is denoted by $\sigma_{j,t}(\Delta\xi_t, \theta, P_t)$. The $J \times 1$ vector of observed market shares in market t is denoted by s_t. Hence, I assume that

$$s_t = \int \vartheta(x_t, p_t, \Delta\xi_t, \theta, v) dP_t(v), \qquad \forall t = 1, \ldots, T$$

for some θ which is analogous to the discussion in the previous section but without specific functional form or distributional assumptions. Notice that $\Delta\xi_t \in \mathbb{R}^J$ and that the distribution function P_t can differ in each market.

Given the assumptions made in this paper, Berry, Levinsohn, and Pakes (1995) proved that for any pair (P, θ), there is a unique solution $\Delta\xi$ to

$$s_t - \sigma_t(\Delta\xi, \theta, P) = 0$$

where s_t are the true observed market shares. This solution is denoted by $\Delta\xi(\theta, P, s_t, p_t, x_t)$ and usually abbreviated by $\Delta\xi_t(\theta, P)$.

Now define the function

$$G(\theta, P_t) \equiv E(z_t' \Delta\xi(\theta, P_t, s_t, p_t, x_t))$$

where $s_t, p_t, \Delta\xi_t \in \mathbb{R}^J$ and $z_t \in \mathbb{R}^{J \times L}$. The moment condition of the model is

$$E(z_t' \Delta\xi(\theta_0, P_{0,t}, s_t, p_t, x_t)) = 0$$

where θ_0 and $P_{0,t}$ denote the true value of θ and the true probability distribution, respectively.

The market shares generated by the model can be approximated by using the empirical probability measure $P_{R,t}$ of an i.i.d. sample $v_{1,t}, \ldots, v_{R,t}$ from P_t. These estimated shares are given by

$$\sigma_t(\Delta\xi_t, \theta, P_{R,t}) = \int \vartheta_t(\Delta\xi_t, \theta, v) dP_{R,t}(v) = \frac{1}{R} \sum_{r=1}^{R} \vartheta_t(\Delta\xi_t, \theta, v_{r,t}).$$

Throughout this paper I assume that the number of draws, R, is a function of T and all limits are taken as $T \to \infty$. For any function $h(v, x)$ denote by $E_t^*(h(v, x))$, the expectation with respect to $P_{0,t}$ given the data. Next define the sample moment

$$G_T(\theta, P_R) \equiv \frac{1}{T} \sum_{t=1}^{T} z_t' \Delta\xi(\theta, P_{R,t}, s_t, p_t, x_t).$$

Finally, define the estimator

$$\hat{\theta} \equiv \arg\min_{\theta \in \Theta} \left|\left| W_T^{1/2} G_T(\theta, P_R) \right|\right|$$

where W_T is a $L \times L$ symmetric positive definite weighting matrix such that $W_T \xrightarrow{p} W$ for some positive definite nonstochastic matrix $W \in \mathbb{R}^{L \times L}$. The estimator $\hat{\theta}$ is the one used in practice and its asymptotic properties are analyzed in this paper.

3.2 Consistency

I first make high level assumptions which provide sufficient conditions for consistency. After stating the consistency theorem, I discuss the role of each assumption. I also provide sufficient primitive conditions for the most abstract assumptions.

Assumption A1. (i) For any fixed $(\Delta\xi, \theta)$ and $\forall t = 1, \ldots, T$,

$$\sigma_t(\Delta\xi, \theta, P_{R,t}) - \sigma_t(\Delta\xi, \theta, P_{0,t}) = \frac{1}{R} \sum_{r=1}^{R} \varepsilon_{r,t}(\Delta\xi, \theta)$$

where, conditional on the data, $\varepsilon_{r,t}(\Delta\xi, \theta)$ is independent across r and t and has mean 0. Each element of the vector $\varepsilon_{r,t}(\Delta\xi, \theta)$ is bounded, continuous, and differentiable in $\Delta\xi$ and θ.

(ii) For all $j = 1, \ldots, J$,

$$\sup_{\theta\in\Theta} \max_{1\leq t\leq T} |\sigma^R_{j,t}(\theta) - \sigma_{j,t}(\theta)| \xrightarrow{p} 0, \qquad \text{as } T \to \infty$$

where $\sigma^R_{j,t}(\theta) = \sigma_{j,t}(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{R,t})$ and $\sigma_{j,t}(\theta) = \sigma_{j,t}(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t})$.

Assumption A2. (i) For every $j = 1, \ldots, J$ and $t = 1, \ldots, T$, for all $\theta \in \Theta$, and for all P_t in a neighborhood of $P_{0,t}$, $\frac{\partial}{\partial\Delta\xi_{k,t}}\sigma_{j,t}(\Delta\xi_t, \theta, P_t)$ exists, and is continuously differentiable in both $\Delta\xi_t$ and θ, with

$$\frac{\partial}{\partial\Delta\xi_{j,t}}\sigma_{j,t}(\Delta\xi_t, \theta, P_t) > 0,$$

and

$$\frac{\partial}{\partial\Delta\xi_{k,t}}\sigma_{j,t}(\Delta\xi_t, \theta, P_t) \leq 0$$

for all $k \neq j$ where $k, j = 1, \ldots, J$.

(ii) The matrix $\frac{\partial}{\partial\Delta\xi_t}\sigma_t(\Delta\xi_t, \theta, P_t)$ is invertible for all P_t in a neighborhood of $P_{0,t}$, for all $\theta \in \Theta$, and for all $t = 1, \ldots, T$.

(iii) There exists an $\varepsilon > 0$ such that for every $j = 0, \ldots, J$ and $t = 1, \ldots, T$,

$$\varepsilon \leq s_{j,t} \leq 1 - \varepsilon.$$

Assumption A3. Define the $JT \times L$ matrix of instruments Z. The instruments are such that the matrix $Z'Z/T$ has full rank and is stochastically bounded, i.e. for all $\varepsilon > 0$ there exists an M_ε such that $\Pr(||Z'Z/T|| > M_\varepsilon) < \varepsilon$.

Assumption A4. For all $\delta > 0$, there exists $C(\delta) > 0$ such that for all $t = 1, \ldots T$,

$$\lim_{T\to\infty} \Pr\left(\inf_{\theta\in\Theta}\left\{\inf_{\Delta\xi\notin\mathcal{N}_{\Delta\xi_t(\theta,P_{R,t})}(\theta;\delta)}\left\{\rho\left(\sigma_t(\Delta\xi, \theta, P_{R,t}), \sigma_t(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t})\right)\right\}\right\} > C(\delta)\right) = 1.$$

Assumption A5. For all $\delta > 0$, there exists $C(\delta) > 0$ such that

$$\lim_{T\to\infty} \Pr\left(\inf_{\theta\notin\mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0) - G_T(\theta_0, P_0)|| \geq C(\delta)\right) = 1.$$

I now provide the main theorem of this section.

Theorem 1. *Suppose that Assumptions A1-A5 hold. Then* $\hat{\theta} \xrightarrow{p} \theta$ *as* $T \to \infty$.

The proof can be found in the appendix.

Assumptions A1 and A4 are needed because in the proof its is required that

$$\sup_{\theta \in \Theta} ||G_T(\theta, P_R) - G_T(\theta, P_0)|| = \sup_{\theta \in \Theta} \left|\left| \frac{1}{T} \sum_{t=1}^{T} z_t' \left(\Delta\xi_t(\theta, P_{R,t}) - \Delta\xi_t(\theta, P_{0,t}) \right) \right|\right| = o_p(1).$$

A sufficient condition for this to hold is that $\Delta\xi_t(\theta, P_{R,t}) - \Delta\xi_t(\theta, P_{0,t})$ converges to 0 in probability uniformly over θ and t. Since there is no expression for $\Delta\xi_t$, I assume instead that the market shares generated by the model, with the true and the approximated distribution, are uniformly close (Assumption A1) and that this would be violated if $\Delta\xi_t(\theta, P_{0,t})$ was not close to $\Delta\xi_t(\theta, P_{R,t})$ (Assumption A4). For the fourth assumption notice that $\sigma_t(\Delta\xi_t, \theta, P_{R,t}) = s_t$ and that Berry, Levinsohn, and Pakes (1995) prove that in their model for each θ and each t, the vector $\Delta\xi_t(\theta, P_{R,t})$ is the unique solution to $\sigma_t(\Delta\xi_t, \theta, P_{R,t}) = s_t$. So if $\Delta\xi_t \notin \mathcal{N}_{\Delta\xi_t(\theta, P_{R,t})}(\theta; \delta)$, then

$$\rho(\sigma_t(\Delta\xi_t, \theta, P_{R,t}), \sigma_t(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t})) > C(\delta)$$

for some $C(\delta) > 0$. The assumption says that $C(\delta)$ does not depend on t or T, but the statement only has to hold in the limit with probability 1. The assumption ensures that, at least asymptotically, the $\Delta\xi$ that sets the models predictions for shares equal to the actual shares can be distinguished from other values of $\Delta\xi$.

Assumption A1 implies that $R(T) \to \infty$ as $T \to \infty$. In this case for each $t = 1, \ldots, T$,

$$\sup_{\theta \in \Theta} |\sigma_{j,t}^R(\theta) - \sigma_{j,t}(\theta)| \xrightarrow{p} 0$$

by a standard uniform law of large numbers (see for example Amemiya (1985)). Assumption 1 is more difficult to verify due to the maximum over all markets. The following lemma provides sufficient conditions for Assumption 1 to hold in the model discussed in Section 2 without the normality assumption on the distribution of random coefficients.

Lemma 1. Let $v_t \sim P_t$ and let $\tilde{\mathcal{P}} = \{P^1, \ldots, P^m\}$ be finite set of distributions. Assumption A1 holds in the random coefficients logit model if

1. $\ln(T)/R \to 0$,
2. $m_t = (p_t, x_t) \in \mathcal{M}$ where $\mathcal{M}$ is a compact set,

3. the lth element of v_t, namely $v_{l,t}$, satisfies (a) $v_{l,t} = g(a_{l,t}, w_{l,t})$ where $w_{l,t} \sim \tilde{P}_{l,t}$ and $\tilde{P}_{l,t} \in \tilde{\mathcal{P}}$, g is continuously differentiable in both arguments, and $a_t \in \Upsilon \subset \mathbb{R}^A$ where Υ is a compact set or (b) the support of $v_{l,t}$ is compact.

The proof is in the appendix. It follows from similar arguments as the proof of Jennrich's uniform law of large numbers (see Jennrich (1969)). Other sufficient conditions that do not require boundedness of the data are, for example, that $\ln(T)/R \to 0$ and that $|\sigma^R_{j,t}(\theta) - \sigma_{j,t}(\theta)|$ converge to 0 in probability uniformly in θ at an exponential rate. This holds under continuity conditions given in Xu (2010) that might be hard to verify in practice. All of these conditions require that $R(T) \to \infty$ because as opposed to the setup of Pakes and Pollard (1989), the integral over $P_{R,t}$ enters the objective function non-linearly. Assuming compactness is not a very desirable assumption but this assumption is basically implied by assuming that all market shares are bounded away from 0 or by Condition S in Berry, Linton, and Pakes (2004). The last condition holds if the random coefficients have the same distribution in all markets but the distribution is also allowed to change across markets. For example one could have a log-normal distribution with a different mean and a different variance in each market. The assumption does not allow for an arbitrary distribution in each market. Similar to the assumptions on the data, the family of distribution has to be restricted in some way to obtain consistency. Assumption A1 holds in other models under the same conditions as in the previous lemma as long as $\Delta\xi_t(\theta, P_{0,t})$ is an element from a compact set. I conjecture that Assumption A4 is also satisfied in the models of Berry, Levinsohn, and Pakes (1995) and Nevo (2001) under the conditions of Lemma 1 but the verification of this conjecture is beyond the scope of this paper.

Assumption A2(i) is easily verifiable in practice and holds in the models of Berry, Levinsohn, and Pakes (1995) and Nevo (2001) in particular due to the parametric assumptions. Assumption A2(ii) also holds in these models (see for example Dubé, Fox, and Su (2009)). Assumption A2(iii) is similar to Condition S in Berry, Linton, and Pakes (2004). Assumption A2 guarantees among others that the demand system is invertible. Assumption A3 is mild and depends on the data. Assumption A5 is an identification condition and depends on the particular model under consideration.

3.3 Asymptotic normality

Next I present additional assumptions which are sufficient for asymptotic normality. Then I provide the main theorem which states the asymptotic expansion of the estimator. A

corollary to this theorem establishes asymptotic normality. In order to provide some intuition on the results I outline the proof for the case where $J = 1$. The details for the more general case are in the appendix.

Assumption B1. Assume that θ_0 is an interior point of Θ.

Assumption B2. For all $P \in \mathcal{P}$, the function $G_T(\theta, P)$ is differentiable at θ_0. Define the derivative matrix $\Gamma_t \equiv \frac{\partial G(\theta_0, P_{0,t})}{\partial \theta}$ and assume that $\frac{1}{T}\sum_{t=1}^{T} \Gamma_t$ converges to a matrix, Γ, of full rank.

Assumption B3. (i) For any fixed $(\Delta\xi, \theta)$ and $\forall t = 1, \ldots, T$ and $\forall j = 1, \ldots, J$,

$$\frac{\partial \sigma_t(\Delta\xi, \theta, P_{R,t})}{\Delta\xi_j} - \frac{\partial \sigma_t(\Delta\xi, \theta, P_{0,t})}{\partial \Delta\xi_j} = \frac{1}{R}\sum_{r=1}^{R} d\varepsilon_{r,j,t}(\Delta\xi, \theta)$$

where, conditional on the data, $d\varepsilon_{r,j,t}(\Delta\xi, \theta)$ is independent across r and t and has mean 0. Each element of the vector $d\varepsilon_{r,j,t}(\Delta\xi, \theta)$ is bounded, continuous, and differentiable in $\Delta\xi$ and θ.

(ii) For any fixed $(\Delta\xi, \theta)$ and $\forall t = 1, \ldots, T$ and $\forall j, k = 1, \ldots, J$,

$$\frac{\partial^2 \sigma_t(\Delta\xi, \theta, P_{R,t})}{\partial \Delta\xi_j \partial \Delta\xi_k} - \frac{\partial^2 \sigma_t(\Delta\xi, \theta, P_{0,t})}{\partial \Delta\xi_j \partial \Delta\xi_k} = \frac{1}{R}\sum_{r=1}^{R} d^2\varepsilon_{r,j,k,t}(\Delta\xi, \theta)$$

where, conditional on the data, $d^2\varepsilon_{r,j,k,t}(\Delta\xi, \theta)$ is independent across r and t and has mean 0. Each element of the vector $d^2\varepsilon_{r,j,k,t}(\Delta\xi, \theta)$ is bounded, continuous, and differentiable in $\Delta\xi$ and θ.

Assumption B4. Let $v_t \sim P_t$ and let $\tilde{\mathcal{P}} = \{P^1, \ldots, P^m\}$ be finite set of distributions. Assume that

(i) $\ln(T)/R(T) \to 0$ as $T \to \infty$,

(ii) $v_{r,t}$ is i.i.d. across r and independent of x_t and z_t, and

(iii) the lth element of v_t, namely $v_{l,t}$, satisfies (a) $v_{l,t} = g(a_{l,t}, w_{l,t})$ where $w_{l,t} \sim \tilde{P}_{l,t}$ and $\tilde{P}_{l,t} \in \tilde{\mathcal{P}}$, g is continuously differentiable in both arguments, and $a_t \in \Upsilon \subset \mathbb{R}^A$ where Υ is a compact set or (b) the support of $v_{l,t}$ is compact.

Assumption B5. The random variables x_t, and p_t have bounded support and for all $l = 1, \ldots, L$, $t = 1, \ldots, T$ and $j = 1, \ldots, J$

$$E(|z_{l,j,t}|^4) \leq M$$

for some constant M.

Assumption B6. The absolute value of each element of

$$\left(\frac{\partial \sigma_t(\Delta\xi(\theta, P_{0,t}), \theta, P_{0,t})}{\partial \Delta\xi}\right)^{-1}$$

is bounded above by some constant M in a neighborhood of θ_0 for all $t = 1, \ldots, T$.

Assumption B7. All partial derivatives up to order 3 of the function $\sigma_{j,t}(\Delta\xi, \theta_0, P_{0,t})$ with respect to $\Delta\xi$ are bounded in absolute value by some constant M for all $t = 1, \ldots, T$ and $j = 1, \ldots, J$.

Assumption B8. There exists a $J \times K$ matrix $H(v)$ such that each element has 4 bounded absolute moments with respect to $P_{0,t}$ and

$$\left|\frac{\partial \vartheta_t(\Delta\xi_t, \theta, v)}{\partial \theta}\right| \leq H(v)$$

and for all $j = 1, \ldots, J$

$$\left|\frac{\partial^2 \vartheta_t(\Delta\xi_t, \theta, v)}{\partial \theta \partial \Delta\xi_{j,t}}\right| \leq H(v)$$

where the inequalities are understood element by element.

Assumption B9. Define

$$H_{0,t} \equiv \frac{\partial \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial \Delta\xi}.$$

Assume that

$$\lim_{T\to\infty} E\left(\frac{1}{T}\sum_{t=1}^{T} z_t' \Delta\xi(\theta_0, P_{0,t}, s_t, p_t, x_t)\Delta\xi(\theta_0, P_{0,t}, s_t, p_t, x_t)' z_t\right) = \Phi_1$$

and that

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} Var\left(z_t' H_{0,t}^{-1} \varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0)\right) = \Phi_2$$

for positive definite matrices Φ_1 and Φ_2. Furthermore, assume that $\Delta\xi(\theta_0, P_{0,t}, s_t, p_t, x_t)$ is uncorrelated across t conditional on z_t and that all conditions of the Lindeberg-Feller central limit theorem hold for the random variables $z_t'\Delta\xi(\theta_0, P_{0,t}, s_t, p_t, x_t)$ and $z_t' H_{0,t}^{-1} \varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0)$.[5]

[5]These conditions limit the dependence structure of the data as well the degree to which the distribution can differ over markets. These conditions are probably implied by more primitive conditions on the data which are simply assumed here.

Assumption B10. Assume that the following limit exists

$$\bar{\mu} \equiv \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J} E\left(z'_{j,t}\left(e'_j H_{0,t}^{-1} E_t^*\left(d\varepsilon_{r,t}(\Delta\xi_t(\theta_0,P_{0,t}),\theta_0)H_{0,t}^{-1}\varepsilon_{r,t}(\Delta\xi_t(\theta_0,P_{0,t}),\theta_0)\right)\right.\right.$$
$$\left.\left. -\frac{1}{2}E_t^*\left(\varepsilon_{r,t}(\Delta\xi_t(\theta_0,P_{0,t}),\theta_0)' I_{0,t,j}\varepsilon_{r,t}(\Delta\xi_t(\theta_0,P_{0,t}),\theta_0)\right)\right)\right)$$

where

$$I_{0,t,j} \equiv \sum_{k=1}^{J} H_{0,t}^{-1} K_{0,t,k} H_{0,t}^{-1} e_j e'_k H_{0,t}^{-1}$$

and

$$K_{0,t,k} \equiv \frac{\partial^2 \sigma_t\left(\Delta\xi_t(\theta_0,P_{0,t}),\theta_0,P_{0,t}\right)}{\partial\Delta\xi\partial\Delta\xi_k}.$$

The vector e_j denotes the jth column of the $J\times J$ identity matrix.

Notice that $\varepsilon_{r,t}(\Delta\xi_t(\theta_0,P_{0,t})$ is $J\times 1$ and $d\varepsilon_{r,t}(\Delta\xi_t(\theta_0,P_{0,t})$ as well as $H_{0,t}$ and $I_{0,t,j}$ are $J\times J$ matrices.

Assumptions B1 and B2 are very common. Differentiability holds in the BLP model in particular. Assumption B4 places additional restrictions on the distribution of random coefficients. Assumption B5 is common in the simulation based estimation literature (for example McFadden (1989), Lee (1995), and Berry, Linton, and Pakes (2004)). Assumptions B3, B7, and B8 hold in the BLP model in particular. For example it is easy to verify that each element of $\frac{\partial^2\vartheta_t(\Delta\xi_t,\theta,v)}{\partial\theta\partial\Delta\xi_{j,t}}$ and $\frac{\partial\vartheta_t(\Delta\xi_t,\theta,v)}{\partial\theta}$ is dominated by $C\sum_{d=1}^{q}|v_d|$ where q is the dimension of v and C is some constant. The remaining assumptions place restrictions on the data generating process and are immediate if $P_{0,t}$ is the same in all markets and the data is i.i.d. across markets.

These assumptions lead to the main theorem of this section.

Theorem 2. *Assume that Assumptions A1-A5 and B1-B10 hold. Then*

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right) = \left(\left(\Gamma' W\Gamma\right)^{-1}\Gamma' W + o_p(1)\right)\left(Q_{1,T} + \frac{1}{\sqrt{R}}Q_{2,T,R} + \frac{\sqrt{T}}{R}Q_{3,T,R} + o_p\left(\frac{\sqrt{T}}{R}\right)\right)$$

where

$$Q_{1,T} \xrightarrow{d} N\left(0,\Phi_1\right),$$

$$Q_{2,T,R} \xrightarrow{d} N\left(0,\Phi_2\right),$$

and

$$Q_{3,T,R} \xrightarrow{p} \bar{\mu}.$$

Furthermore, $Q_{1,T}$ and $Q_{2,T,R}$ are asymptotically independent.

The proof is in the appendix. An immediate consequence of Theorem 2 is the following result.

Corollary 1. *Assume that Assumptions A1-A5 and B1-B10 hold. If $\lambda = \lim_{T\to\infty} \frac{\sqrt{T}}{R} < \infty$, then*

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(\lambda \left(\Gamma' W \Gamma\right)^{-1} \Gamma' W \bar{\mu}, V_1\right)$$

where

$$V_1 = (\Gamma' W \Gamma)^{-1} \Gamma' W \Phi_1 W \Gamma (\Gamma' W \Gamma)^{-1}.$$

Theorem 2 shows that the use of Monte Carlo integration (as opposed to evaluating the integral exactly) leads to additional variance and additional bias terms. The leading variance term is of order $1/\sqrt{R}$ while the leading bias term is of order $\sqrt{T}/R$. Hence, if R grows slower than T, the leading bias term dominates the leading variance term which may lead to an asymptotic distribution that is not centered at 0. As a consequence, if $\lambda > 0$, confidence intervals based on the usual GMM asymptotic distribution have the wrong size asymptotically. If R grows faster than T, the leading variance term becomes dominating but the first order asymptotic distribution is not affected by Monte Carlo integration. It can also be shown that, under slightly different assumptions, if the distribution of random coefficients is the same in all markets and *if one uses the same draws from $P_{0,t}$ in all markets* then

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) = \left(\left(\Gamma' W \Gamma\right)^{-1} \Gamma' W + o_p(1)\right)\left(Q_{1,T} + \frac{\sqrt{T}}{\sqrt{R}} Q^*_{2,T,R} + \frac{\sqrt{T}}{R} Q_{3,T,R} + o_p\left(\frac{\sqrt{T}}{R}\right)\right)$$

where $Q^*_{2,T,R}$ converges to a normally distributed random variable with mean 0 as well. In this case one needs T/R to be bounded to obtain asymptotic normality (which is a stronger condition than the rate in Corollary 1) and the additional variance term dominates the additional bias term. This means that if $\tilde{\lambda} = \lim_{T\to\infty} \frac{\sqrt{T}}{\sqrt{R}} < \infty$, then

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, V_1 + \tilde{\lambda}^2 V_2\right)$$

where

$$V_1 = (\Gamma' W \Gamma)^{-1} \Gamma' W \Phi_1 W \Gamma (\Gamma' W \Gamma)^{-1}$$

and

$$V_2 = (\Gamma' W \Gamma)^{-1} \Gamma' W \Phi_2 W \Gamma (\Gamma' W \Gamma)^{-1}.$$

This result is very similar to the one of Berry, Linton, and Pakes (2004).

The expansions are very similar if one uses non-stochastic approximations such as quadrature rules. However, the nodes are not random in these cases and hence, one cannot use laws of large numbers or a central limit theorems to deal with the terms $Q_{2,T,R}$ and $Q_{3,T,R}$. Since none of these terms has a mean of 0 with non-stochastic approximations, quadrature rules lead to an additional bias only and one obtains asymptotic normality if R grows fast enough relative to T.

To get a better sense of where the terms $Q_{2,T,R}$ and $Q_{3,T,R}$ come from I now present an intuitive outline of the asymptotic normality proof with $J = 1$. The intuition for the general case is very similar. The objective is to minimize $G_T(\theta, P_R)' W_T G_T(\theta, P_R)$ where

$$G_T(\theta, P_R) \equiv \frac{1}{T} \sum_{t=1}^{T} z_t' \Delta \xi_t(\theta, P_{R,t}).$$

The first order condition is

$$\left(\frac{\partial}{\partial \theta} G_T(\hat{\theta}, P_R)' \right) W_T G_T(\hat{\theta}, P_R) = 0.$$

Now define

$$D_T(\hat{\theta}, P_R) = \frac{\partial}{\partial \theta} G_T(\hat{\theta}, P_R).$$

Using a first order expansion of $G_T(\hat{\theta}, P_R)$ around $\theta = \theta_0$ yields

$$D_T(\hat{\theta}, P_R)' W_T \left(G_T(\theta_0, P_R) + D_T(\tilde{\theta}, P_R) \left(\hat{\theta} - \theta_0 \right) \right) = 0$$

where $\tilde{\theta}$ is between θ_0 and $\hat{\theta}$. Thus

$$\sqrt{T} \left(\hat{\theta} - \theta_0 \right) = \left(D_T(\hat{\theta}, P_R)' W_T D_T(\tilde{\theta}, P_R) \right)^{-1} D_T(\hat{\theta}, P_R)' W_T \sqrt{T} G_T(\theta_0, P_R).$$

It is shown in the appendix

$$\left(D_T(\hat{\theta}, P_R)' W_T D_T(\tilde{\theta}, P_R) \right)^{-1} D_T(\hat{\theta}, P_R)' W_T \xrightarrow{p} (\Gamma' W \Gamma)^{-1} \Gamma' W.$$

Now consider $\sqrt{T} G_T(\theta_0, P_R)$, write

$$G_T(\theta_0, P_R) = G_T(\theta_0, P_0) + G_T(\theta_0, P_R) - G_T(\theta_0, P_0),$$

and notice that $\sqrt{T}G_T(\theta_0, P_0)$ converges to a normally distributed random variable by Assumption B9. Next write

$$G_T(\theta_0, P_R) - G_T(\theta_0, P_0) = \frac{1}{T}\sum_{t=1}^{T} z_t' \left(\Delta\xi_t(\theta_0, P_{R,t}) - \Delta\xi_t(\theta_0, P_{0,t})\right).$$

Also notice that for all t, $\Delta\xi_t(\theta, P)$ solves

$$s_t = \sigma_t\left(\Delta\xi, \theta, P\right)$$

where s_t are the observed market shares. This implies that we can write

$$\Delta\xi_t(\theta_0, P) = \sigma_t^{-1}\left(s_t, \theta_0, P\right).$$

It follows that

$$s_t = \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right) = \sigma_t\left(\Delta\xi_t(\theta_0, P_{R,t}), \theta_0, P_{R,t}\right)$$

and

$$\sigma_t^{-1}\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right), \theta_0, P_{R,t}\right) = \Delta\xi_t(\theta_0, P_{0,t}).$$

Thus,

$$G_T(\theta_0, P_R) - G_T(\theta_0, P_0) = \frac{1}{T}\sum_{t=1}^{T} z_t' \left(\sigma_t^{-1}\left(s_t, \theta_0, P_{R,t}\right) - \sigma_t^{-1}\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right), \theta_0, P_{R,t}\right)\right).$$

Next consider only

$$\sigma_t^{-1}\left(s_t, \theta_0, P_{R,t}\right) - \sigma_t^{-1}\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right), \theta_0, P_{R,t}\right)$$

and the function $f_t : \mathbb{R} \to \mathbb{R}$ where

$$f_t(s) = \sigma_t^{-1}\left(s, \theta_0, P_{R,t}\right).$$

By a third order Taylor expansion

$$\begin{aligned} f_t(s_t) \;=\; & f_t\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right) \\ & + f_t'\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right)\left(s_t - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right) \\ & + \frac{1}{2} f_t''\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right)\left(s_t - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right)^2 \\ & + \frac{1}{6} f_t'''\left(\tilde{s}_t\right)\left(s_t - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right)^3 \end{aligned}$$

where $\tilde{s}_t$ is between s_t and $\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)$. For any s_0,

$$\left.\frac{\partial f_t'(s)}{\partial s}\right|_{s=s_0} = \left.\frac{\partial \sigma_t^{-1}\left(s, \theta_0, P_{R,t}\right)}{\partial s}\right|_{s=s_0} = \left(\left.\frac{\partial \sigma_t\left(\Delta\xi, \theta_0, P_{R,t}\right)}{\partial \Delta\xi}\right|_{\Delta\xi=\sigma_t^{-1}\left(s_0,\theta_0,P_{R,t}\right)}\right)^{-1}$$

and

$$\begin{aligned}\left.\frac{\partial f_t''(s)}{\partial s}\right|_{s=s_0} &= \left.\frac{\partial^2 \sigma_t^{-1}\left(s, \theta_0, P_{R,t}\right)}{\partial^2 s}\right|_{s=s_0} \\ &= \left(-\left.\frac{\partial^2 \sigma_t\left(\Delta\xi, \theta_0, P_{R,t}\right)}{\partial^2 \Delta\xi}\right|_{\Delta\xi=\sigma_t^{-1}\left(s_0,\theta_0,P_{R,t}\right)}\right)\left(\left.\frac{\partial \sigma_t\left(\Delta\xi, \theta_0, P_{R,t}\right)}{\partial \Delta\xi}\right|_{\Delta\xi=\sigma_t^{-1}\left(s_0,\theta_0,P_{R,t}\right)}\right)^{-3}.\end{aligned}$$

In the expansion above, the derivatives are evaluated at $s_0 = \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)$ which implies that

$$\sigma_t^{-1}\left(s_0, \theta_0, P_{R,t}\right) = \sigma_t^{-1}\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right), \theta_0, P_{R,t}\right) = \Delta\xi_t(\theta_0, P_{0,t}).$$

Combining these results yields

$$\begin{aligned}\Delta\xi_t(\theta_0, P_{R,t}) &= \Delta\xi_t(\theta_0, P_{0,t}) \\ &\quad + \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial \Delta\xi}\right)^{-1} \\ &\qquad \times \left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right) - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right) \\ &\quad + \frac{1}{2}\left(-\frac{\partial^2 \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial^2 \Delta\xi}\right)\left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial \Delta\xi}\right)^{-3} \\ &\qquad \times \left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right) - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right)^2 \\ &\quad + \frac{1}{6} f_t'''\left(\tilde{s}_t\right)\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right) - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right)^3.\end{aligned}$$

Now define

$$\begin{aligned}H_{R,t} &\equiv \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial \Delta\xi}\right) \\ H_{0,t} &\equiv \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial \Delta\xi}\right) \\ I_{R,t} &\equiv \left(\frac{\partial^2 \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial^2 \Delta\xi}\right)\left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial \Delta\xi}\right)^{-3} \\ I_{0,t} &\equiv \left(\frac{\partial^2 \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial^2 \Delta\xi}\right)\left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial \Delta\xi}\right)^{-3}.\end{aligned}$$

Notice that $H_{R,t}$ and $I_{R,t}$ are smooth functions of the sample averages

$$\frac{1}{R}\sum_{r=1}^{R} d\varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0) \qquad \text{and} \qquad \frac{1}{R}\sum_{r=1}^{R} d^2\varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0)$$

while $H_{0,t}$ and $I_{0,t}$ are smooth functions of the corresponding conditional expectations. It now follows that

$$\begin{aligned}
\Delta\xi_t(\theta_0, P_{R,t}) &= \Delta\xi_t(\theta_0, P_{0,t}) \\
&\quad + H_{0,t}^{-1}\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right) - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right) \\
&\quad + \left(H_{R,t}^{-1} - H_{0,t}^{-1}\right)\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right) - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right) \\
&\quad - \frac{1}{2} I_{0,t}\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right) - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right)^2 \\
&\quad + \frac{1}{2}\left(I_{0,t} - I_{R,t}\right)\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right) - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right)^2 \\
&\quad + \frac{1}{6} f_t'''\left(\tilde{s}_t\right)\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right) - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right)^3.
\end{aligned}$$

We can also write

$$\begin{aligned}
H_{R,t}^{-1} - H_{0,t}^{-1} &= -\left(\frac{\partial\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial\Delta\xi}\right)^{-2} \\
&\qquad \times \left(\frac{\partial\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial\Delta\xi} - \frac{\partial\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial\Delta\xi}\right) \\
&\quad + \left(\tilde{\sigma}_t\right)^{-3}\left(\frac{\partial\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial\Delta\xi} - \frac{\partial\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial\Delta\xi}\right)^2
\end{aligned}$$

where $\tilde{\sigma}_t$ is between $\frac{\partial\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial\Delta\xi}$ and $\frac{\partial\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial\Delta\xi}$. Now define

$$e_{R,t} \equiv \left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right) - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)\right) = \frac{1}{R}\sum_{r=1}^{R} \varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0)$$

and

$$\frac{\partial e_{R,t}}{\partial\Delta\xi} \equiv \left(\frac{\partial\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial\Delta\xi} - \frac{\partial\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial\Delta\xi}\right) = \frac{1}{R}\sum_{r=1}^{R} d\varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0).$$

Moreover, define $\varepsilon_{r,0,t} \equiv \varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0)$ and $d\varepsilon_{r,0,t} \equiv d\varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0)$ Both $e_{R,t}$ and $\frac{\partial e_{R,t}}{\partial\Delta\xi}$ are averages of R terms with conditional expectation of 0. It follows that

$$\begin{aligned}
\Delta\xi_t(\theta_0, P_{R,t}) &= \Delta\xi_t(\theta_0, P_{0,t}) - H_{0,t}^{-1} e_{R,t} \\
&\quad - \frac{1}{2} I_{0,t}\left(e_{R,t}\right)^2 + H_{0,t}^{-2}\left(\frac{\partial e_{R,t}}{\partial\Delta\xi}\right) e_{R,t} \\
&\quad + error_{R,t}
\end{aligned}$$

where

$$\begin{aligned}
error_{R,t} &\equiv \frac{1}{6} f_t'''\left(\tilde{s}_t\right)\left(\sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right) - \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)\right)^3 \\
&\quad + \left(I_{R,t} - I_{0,t}\right)\left(e_{R,t}\right)^2 + \left(\tilde{\sigma}_t\right)^{-3}\left(\frac{\partial e_{R,t}}{\partial\Delta\xi}\right)^2 e_{R,t}.
\end{aligned}$$

Plugging this expansion back into the objective function gives

$$\begin{aligned}\sqrt{T}G_T(\theta_0, P_R) &= \sqrt{T}G_T(\theta_0, P_0) + \sqrt{T}\left(G_T(\theta_0, P_R) - G_T(\theta_0, P_0)\right) \\ &= \sqrt{T}G_T(\theta_0, P_0) - \frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t' H_{0,t}^{-1} e_{R,t} \\ &\quad + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t' \left(H_{0,t}^{-2}\left(\frac{\partial e_{R,t}}{\partial \Delta\xi}\right) e_{R,t} - \frac{1}{2} I_{0,t}\left(e_{R,t}\right)^2\right) \\ &\quad + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t' error_{R,t}.\end{aligned}$$

We now have four terms. The first term, $\sqrt{T}G_T(\theta_0, P_0)$, is $O_p(1)$ and belongs to the GMM objective function without simulation error. Therefore, by Assumption B9 it converges to a normally distributed random variable. The second term, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t' H_{0,t}^{-1} e_{R,t}$, is $O_p\left(\frac{1}{\sqrt{R}}\right)$ and converges to a normally distributed random variable as well when multiplied by $\sqrt{R}$ by Assumption B9. These two normal terms are asymptotically independent.

The third term does not have a mean of zero because

$$\begin{aligned}&E\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t' \left(H_{0,t}^{-2}\left(\frac{\partial e_{R,t}}{\partial \Delta\xi}\right) e_{R,t} - \frac{1}{2} I_{0,t}\left(e_{R,t}\right)^2\right)\right) \\ &\quad = \frac{\sqrt{T}}{R}\frac{1}{T}\sum_{t=1}^{T} E\left(z_t'\left(H_{0,t}^{-2} cov_t^*\left(\varepsilon_{r,0,t}, d\varepsilon_{r,0,t}\right) - \frac{1}{2} I_{0,t} E_t^*\left(\varepsilon_{r,0,t}^2\right)\right)\right).\end{aligned}$$

By Assumption B10

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} E\left(z_t'\left(H_{0,t}^{-2} cov_t^*\left(\varepsilon_{r,0,t}, d\varepsilon_{r,0,t}\right) - \frac{1}{2} I_{0,t} E_t^*\left(\varepsilon_{r,0,t}^2\right)\right)\right) = \bar{\mu}.$$

Furthermore, by the weak law of large numbers

$$\frac{R}{\sqrt{T}}\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t'\left(H_{0,t}^{-2}\left(\frac{\partial e_{R,t}}{\partial \Delta\xi}\right) e_{R,t} - \frac{1}{2} I_{0,t}\left(e_{R,t}\right)^2\right) \xrightarrow{p} \bar{\mu}$$

which implies that the third term is $O_p\left(\frac{\sqrt{T}}{R}\right)$ and converges in probability to a constant when multiplied by $\frac{R}{\sqrt{T}}$.

It follows from the proof of Theorem 2 that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t' error_{R,t} = o_p\left(\frac{\sqrt{T}}{R}\right).$$

and therefore

$$\sqrt{T}G_T(\theta_0, P_R) \;=\; O_p(1) + O_p\left(\frac{1}{\sqrt{R}}\right) + O_p\left(\frac{\sqrt{T}}{R}\right) + o_p\left(\frac{\sqrt{T}}{R}\right).$$

Furthermore, under the assumptions of Theorem 2 it also holds that

$$\left(D_T(\hat{\theta}, P_R)' W_T D_T(\tilde{\theta}, P_R)\right)^{-1} D_T(\hat{\theta}, P_R)' W_T \xrightarrow{p} \left(\Gamma' W \Gamma\right)^{-1} \Gamma' W.$$

Hence, $\sqrt{T}$ consistency is only achieved if $\frac{\sqrt{T}}{R}$ does not diverge. If $\frac{\sqrt{T}}{R} \to \lambda$, then

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(\lambda \left(\Gamma' W \Gamma\right)^{-1} \Gamma' W \bar{\mu}, V_1\right)$$

where V_1 is the usual GMM variance given by

$$V_1 \equiv (\Gamma' W \Gamma)^{-1} \Gamma' W \Phi_1 W \Gamma (\Gamma' W \Gamma)^{-1}.$$

This implies that if R converges to 0 at a smaller rate than T, the additional bias (second order term in the Taylor expansion) dominates the additional variance (first order). Furthermore, the additional bias term yields a rate restriction on R relative to T whereas the first order term converges to 0 as $R \to \infty$ at any rate. In the following section it will be shown how the leading bias term can be removed and how and the additional variance can be taken into account when calculating standard errors.

3.4 Bias and variance correction

This section shows how the leading bias term, namely the $O_p\left(\frac{\sqrt{T}}{R}\right)$ term in the expansion above, can be eliminated by using either an analytic bias correction or a jackknife method. Similar methods have been suggested by Lee (1995), Arellano and Hahn (2007) and Kristensen and Salanie (2010) in related setups. Furthermore, the leading additional variance term, in particular the $O_p\left(\frac{1}{R}\right)$ term can easily be taken into account when calculating standard errors.

Let $\tilde{R}$ be large relative to R. Estimators for the bias and the variance can be obtained by replacing $P_{0,t}$ with $P_{\tilde{R},t}$ and θ_0 with $\hat{\theta}$. Moreover, moments are replaced by the corresponding sample analogs. I use $P_{\tilde{R},t}$ instead of $P_{R,t}$ in order to obtain a better estimate $P_{0,t}$ and to make the corrections less dependent on the number of draws. Notice that the computational costs are quite low because one has to solve for $\Delta\xi\left(\hat{\theta}, P_{\tilde{R},t}, s_t, p_t, x_t\right)$ only once and not repeatedly as during the optimization procedure.[6]

[6]For the the Monte Carlo simulations in this paper, I use R between 50 and 800 and set $\tilde{R} = 20,000$.

3.4.1 Analytic bias correction

Define the bias adjusted estimator as

$$\hat{\theta}_A \equiv \hat{\theta} - \frac{1}{R}\left(\hat{\Gamma}' W_T \hat{\Gamma}\right)^{-1} \hat{\Gamma}' W_T \hat{\bar{\mu}}$$

where

$$\hat{\Gamma} = \frac{\partial}{\partial \theta} G_T\left(\hat{\theta}, P_{\tilde{R}}\right) = \frac{\partial}{\partial \theta} \frac{1}{T} \sum_{t=1}^{T} z_t' \Delta\xi\left(\hat{\theta}, P_{\tilde{R},t}, s_t, p_t, x_t\right)$$

and

$$\hat{\bar{\mu}} = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J} z_{j,t}' \left(e_j' \hat{C}_{R,t} - \frac{1}{2} \hat{S}_{R,j,t}\right)$$

where

$$\begin{aligned} \hat{S}_{R,j,t} &= \frac{1}{R} \sum_{r=1}^{R} \left(\vartheta_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, v_{r,t}\right) - \sigma_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, P_{\tilde{R},t}\right)\right)' \hat{I}_{\tilde{R},t,j} \\ &\quad * \left(\vartheta_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, v_{r,t}\right) - \sigma_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, P_{\tilde{R},t}\right)\right) \end{aligned}$$

and

$$\begin{aligned} \hat{C}_{R,t} &= \frac{1}{R} \sum_{r=1}^{R} \hat{H}_{\tilde{R},t}^{-1} \left(\frac{\partial \vartheta_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, v_{r,t}\right)}{\partial \Delta\xi} - \frac{\partial \sigma_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, P_{\tilde{R},t}\right)}{\partial \Delta\xi}\right) \hat{H}_{\tilde{R},t}^{-1} \\ &\quad * \left(\vartheta_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, v_{r,t}\right) - \sigma_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, P_{\tilde{R},t}\right)\right) \end{aligned}$$

with

$$\hat{H}_{\tilde{R},t} = \left(\frac{\partial \sigma_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, P_{\tilde{R},t}\right)}{\partial \Delta\xi}\right)$$

and

$$\hat{I}_{\tilde{R},t,j} = \sum_{k=1}^{J} \hat{H}_{\tilde{R},t}^{-1} \left(\frac{\partial^2 \sigma_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, P_{\tilde{R},t}\right)}{\partial \Delta\xi \partial \Delta\xi_k}\right) \hat{H}_{\tilde{R},t}^{-1} e_j e_k' \hat{H}_{\tilde{R},t}^{-1}.$$

Again e_j denotes the jth column of the $J \times J$ identity matrix.

Subtracting an estimate of $\frac{1}{R}\left(\Gamma' W \Gamma\right)^{-1} \Gamma' W \bar{\mu}$ from $\hat{\theta}$ eliminates the leading bias term from the asymptotic expansion which is established by the following theorem.

Using $\tilde{R} = R$ performs worse in terms of nominal coverage rates, especially if R is small, but the this would not change the result in Theorem 3.

Theorem 3. *Assume that Assumptions A1-A5 and B1-B10 hold. Then*

$$\sqrt{T}\left(\hat{\theta}_A - \theta_0\right) = \left(\left(\Gamma' W \Gamma\right)^{-1} \Gamma' W + o_p(1)\right)\left(Q_{1,T} + \frac{1}{\sqrt{R}} Q_{2,T,R} + o_p\left(\frac{\sqrt{T}}{R}\right)\right).$$

As opposed to the results of Theorem 2, it follows that as long as $\frac{\sqrt{T}}{R}$ is bounded, it holds that $\sqrt{T}\left(\hat{\theta}_A - \theta_0\right) \xrightarrow{d} N(0, V_1)$.

3.4.2 Jackknife bias correction

A second possibility to eliminate the leading term of the bias is to use a Jackknife style bias correction. Let $\hat{\theta}_{R/2,n}$, $n = 1, \ldots, N$ be estimators of θ using $R/2$ independent draws to approximate the integral. Define

$$\hat{\theta}_{JK} = 2\hat{\theta} - \frac{1}{N}\sum_{n=1}^{N} \hat{\theta}_{R/2,n}.$$

It is easily verified that this procedure eliminates the leading term of the bias as well. However, this estimator is computationally costly. If $N = 2$, the optimization problem has to be solved two additional times. Furthermore, this procedure increases the additional variance due to the simulations by a factor of $4 + 2/N$. Therefore, in the Monte Carlo study below only the analytic bias correction is pursued. Similarly, the panel jackknife bias correction suggested by Hahn and Newey (2004) is not very appealing in this setting because it requires to estimate the parameter vector $R + 1$ times which is computationally too demanding.

3.4.3 Variance correction

The variance of the estimator can be estimated by

$$\hat{V} = \left(\hat{\Gamma}' \hat{W} \hat{\Gamma}\right)^{-1} \hat{\Gamma}' \hat{W} \left(\hat{\Phi}_1 + \frac{1}{R}\hat{\Phi}_2\right) \hat{W} \hat{\Gamma} \left(\hat{\Gamma}' \hat{W} \hat{\Gamma}\right)^{-1}$$

where

$$\hat{\Phi}_1 = \frac{1}{T}\sum_{t=1}^{T} z_t' \Delta\xi\left(\hat{\theta}, P_{\tilde{R},t}, s_t, p_t, x_t\right) \Delta\xi\left(\hat{\theta}, P_{\tilde{R},t}, s_t, p_t, x_t\right)' z_t$$

and

$$\hat{\Phi}_2 = \frac{1}{R}\frac{1}{T}\sum_{r=1}^{R}\sum_{t=1}^{T} z_t' \hat{H}_{\tilde{R},t}^{-1} \hat{\varepsilon}_{r,t} \hat{\varepsilon}_{r,t}' \hat{H}_{\tilde{R},t}^{-1'} z_t$$

and

$$\hat{\varepsilon}_{r,t} = \vartheta_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, v_{r,t}\right) - \sigma_t\left(\Delta\xi_t\left(\hat{\theta}, P_{\tilde{R},t}\right), \hat{\theta}, P_{\tilde{R},t}\right).$$

In this way the variance of the $O_p\left(\frac{1}{R}\right)$ term is taken into account as well. In case a bias adjustment is used, $\hat{\theta}$ can be replaced by $\hat{\theta}_A$ or $\hat{\theta}_{JK}$.

4 Monte Carlo simulation

In this section, I illustrate that the simulation error will affect the finite sample performance of the estimator because the usual GMM standard errors underestimate the true variance and the estimates are biased. I use the model described in Section 2. The setup for the Monte Carlo simulation is adapted from Dubé, Fox, and Su (2009) with very few changes to accommodate the asymptotics in the number of markets. This setup is also used by Judd and Skrainka (2011). The number of products is set to 4 and I vary the number of markets, T, and draws, R. I use a constant term and three product characteristics next to the price. Two of these three product characteristics vary across markets and one product characteristic does not. The product characteristics are distributed as

$$\begin{pmatrix} x_{1,j} \\ x_{2,j,t} \\ x_{3,j,t} \end{pmatrix} \sim TN\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.8 & 0.3 \\ -0.8 & 1 & 0.3 \\ 0.3 & 0.3 & 1 \end{pmatrix} \right), \qquad t = 1, \ldots, T$$

where TN denotes the standard normal distribution truncated at -4 and 4. There is also a constant term, $x_{0,j} = 1$ for all j. The unobserved product characteristics are $\xi_{j,t} = \frac{1}{2}(\xi_j + \Delta\xi_{j,t})$ where

$$\xi_j \sim TN(0,1), \qquad j = 1, \ldots, 4$$

and

$$\Delta\xi_{j,t} \sim TN(0,1), \qquad j = 1, \ldots, 4, \qquad t = 1, \ldots, T.$$

The price is generated by

$$p_{j,t} = \frac{1}{2}\left|0.5\xi_{j,t} + e_{j,t} + 1.1\left(x_{1,j} + x_{2,j,t} + x_{3,j,t}\right)\right|$$

where $e_{j,t} \sim TN(0,1)$. There is a random coefficient on all product characteristics including price and the constant term. The random coefficient are distributed as follows

$$\begin{pmatrix} \beta_i^0 \\ \beta_i^1 \\ \beta_i^2 \\ \beta_i^3 \\ \alpha_i \end{pmatrix} \sim N\left(\begin{pmatrix} -1 \\ 1.5 \\ 1.5 \\ 0.5 \\ 3 \end{pmatrix}, \begin{pmatrix} 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.2 \end{pmatrix} \right), \qquad t = 1, \ldots, T.$$

For each product j in market t, I generate 6 instruments ($b = 1, \ldots, 6$) as

$$z_{b,j,t} = a_{b,j,t} + 0.25 \left(e_{j,t} + 1.1 \left(x_{1,j} + x_{2,j,t} + x_{3,j,t} \right) \right), \quad j = 1, \ldots, 4, \ t = 1, \ldots, T.$$

Here $a_{b,j,t} \sim U(0,1)$. For reasons explained in Section 2.2, next to $z_{b,j,t}$, $x_{2,j,t}$ and $x_{3,j,t}$, I also use product dummies, $z^2_{b,j,t}$ and $z^3_{b,j,t}$ for all b, $x^2_{2,j,t}$, $x^2_{3,j,t}$, $x^3_{2,j,t}$, $x^3_{3,j,t}$ as well as $\prod_{b=1}^{6} z_{b,j,t}$, $x_{2,j,t}x_{3,j,t}$, $x_{2,j,t}z_{b,j,t}$ and $x_{3,j,t}z_{b,j,t}$ as instruments.

I make use of a nested fixed point approach. The code is written in Matlab and C++. The latter program is used to calculate the market shares and their derivative. Using C++ for these calculations yields a large time improvement compared to Matlab which reduces the time advantage of the MPEC approach of Dubé, Fox, and Su (2009).[7] As mentioned in the introduction estimating the model is computationally demanding, especially if R is large. The nested fixed point approach solves the non-linear system of equations given in (1) using a contraction mapping when evaluating the objective function for a certain parameter value. In each step of the contraction mapping the integrals have to be calculated. With the MPEC approach the moments are treated as additional parameters which adds $J \times T$ parameters. The system of equations in (1) is then enforced as constraints and for each of these constraints the integrals have to be calculated. For further computational details see Nevo (2001) and Dubé, Fox, and Su (2009).

Below I investigate the actual coverage rate of a 95% nominal confidence interval for $\alpha = E(\alpha_i) = 3$ using bias correction methods as well as standard errors with and without correcting for the simulation error. The usual GMM standard errors are estimated using

$$\hat{V}_1 = \left(\hat{\Gamma}' \hat{W} \hat{\Gamma} \right)^{-1} \hat{\Gamma}' \hat{W} \hat{\Phi}_1 \hat{W} \hat{\Gamma} \left(\hat{\Gamma}' \hat{W} \hat{\Gamma} \right)^{-1}$$

and the adjusted standard errors are calculated using

$$\hat{V} = \left(\hat{\Gamma}' \hat{W} \hat{\Gamma} \right)^{-1} \hat{\Gamma}' \hat{W} \left(\hat{\Phi}_1 + \frac{1}{R} \hat{\Phi}_2 \right) \hat{W} \hat{\Gamma} \left(\hat{\Gamma}' \hat{W} \hat{\Gamma} \right)^{-1}.$$

In both cases, as well as for estimating the bias, $\tilde{R} = 20,000$. I also compare the median length of the confidence intervals obtained from different simulations. I make use of $50 - 800$ draws from the normal distribution as well as $100 - 800$ markets. I use different draws to

[7] I am very grateful to Ketan Patel for sharing his code. The CPU time using the NFP approach is in fact lower than the one reported by Judd and Skrainka (2011) in their Tables 4 and 5 with the MPEC approach when I use the same setup.

approximate the integral in different markets. Hence, with 100 markets and 800 draws I sample in total $80,000$ times from the normal distribution. The computational costs only depend on the number draws that are used to evaluate each integral which is 800 in this case. All coverage rates are based on 1000 Monte Carlo iterations.

Table 1 shows that the number of simulations affects the actual coverage rate of a nominal 95% confidence interval if the usual GMM asymptotic distribution is employed. For example with 800 markets the actual coverage rate is only 68.8% with 50 draws while it increases to 90.8% with 800 draws. It is also striking that the coverage rate tends to decrease if R is fixed but T increases which is intuitive because the relation between T and R matters for the asymptotic results. Using the bias corrected estimator leads to an improvement in these coverage rates. However, in general there is still a large difference between using a small number and large number of draws. For instance, with 800 marktes and 50 draws one obtains a coverage rate of 85.6% while 800 draws yield a coverage rate of 91.3%. The same holds when simulation adjusted standard errors but no bias adjustment is used. In case one uses both the analytical bias adjustement and the standard error adjustement, the coverage rate is very close to 95% even with a small number of draws. For example with 800 markets and 50 draws one obtains a coverage rate of 93.4%.

The cost of the improved coverage rate is a wider confidence interval. Table 2 shows that with the usual GMM asymptotic distribution, using a large number of draws has almost no effect on the median length of the confidence intervals. This is not the case for the adjusted standard errors. For example with 100 markets, the median length is 1.059 with 50 draws and 0.748 with 800 draws. It is also striking that the lengths are very similar with 800 draws for both types of standard errors. Thus, corrected standard errors only affect the coverage rate and the median length when the number of simulations is small compared to the number of markets. This was expected since the variance of the second term of the asymptotic expansion decreases with the number of draws.

Table 3 shows the finite sample bias with and without bias correction. It can be seen that the finite sample bias decreases as R increases. Nevertheless, especially if T is small, the finite sample bias is still substantial even if R is large. The reason is that we are dealing with a nonlinear estimator which is biased in finite samples even if $R = \infty$. The bias correction reduces the finite sample bias. The bias correction works particularly well if T is large. The

Table 1: Coverage rates of 95% confidence intervals for α

	50 draws	100 draws	200 draws	400 draws	800 draws
	GMM point estimates and GMM standard errors				
100 markets	0.788	0.863	0.902	0.910	0.911
200 markets	0.720	0.790	0.848	0.887	0.899
400 markets	0.638	0.755	0.840	0.896	0.914
800 markets	0.688	0.802	0.846	0.899	0.886
	Bias corrected point estimates and GMM standard errors				
100 markets	0.866	0.912	0.920	0.936	0.932
200 markets	0.826	0.854	0.894	0.912	0.917
400 markets	0.809	0.827	0.894	0.921	0.931
800 markets	0.858	0.877	0.872	0.903	0.898
	GMM point estimates and adjusted standard errors				
100 markets	0.872	0.901	0.919	0.927	0.923
200 markets	0.842	0.863	0.884	0.906	0.910
400 markets	0.801	0.846	0.902	0.919	0.931
800 markets	0.846	0.885	0.890	0.938	0.906
	Bias corrected point estimates and adjusted standard errors				
100 markets	0.920	0.940	0.941	0.946	0.942
200 markets	0.907	0.917	0.922	0.931	0.925
400 markets	0.902	0.898	0.930	0.936	0.942
800 markets	0.935	0.933	0.917	0.939	0.920

The nominal coverage rate is 0.95 and the number of Monte Carlo simulations is 1000. The true value of α is 3. If the actual coverage rate is 95%, the standard error with 1000 simulations is around 0.0069. If the actual coverage rate is 80%, the standard error increases to 0.0126.

reason is that estimation of the bias relies on an initial estimate of θ_0 which is more precise if T is large. With 800 markets the finite sample bias of the bias adjusted estimator is very close to 0 even if only a small number of draws is used.

Many parameter choices drive the results in this Monte Carlo study. First, there is the ratio of the variance of the error term and the variance of the products characteristics. Second, there is the strengths of the instruments. A low variance ratio or strong instruments imply that one gets more precise estimates for a given number of markets. Furthermore, the effect

Table 2: Median length of confidence intervals for α

	50 draws	100 draws	200 draws	400 draws	800 draws
	GMM point estimates and GMM standard errors				
100 markets	0.772	0.769	0.757	0.758	0.748
200 markets	0.603	0.602	0.603	0.596	0.604
400 markets	0.478	0.475	0.453	0.474	0.480
800 markets	0.375	0.369	0.368	0.365	0.371
	Bias corrected point estimates and adjusted standard errors				
100 markets	1.059	0.945	0.859	0.818	0.786
200 markets	0.868	0.772	0.701	0.667	0.642
400 markets	0.796	0.648	0.573	0.531	0.512
800 markets	0.732	0.541	0.464	0.420	0.399

Table 3: Finite sample bias of α

	50 draws	100 draws	200 draws	400 draws	800 draws
	GMM point estimates				
100 markets	-0.229	-0.185	-0.153	-0.128	-0.115
200 markets	-0.188	-0.159	-0.130	-0.108	-0.091
400 markets	-0.158	-0.125	-0.092	-0.071	-0.064
800 markets	-0.097	-0.065	-0.054	-0.041	-0.041
	Bias corrected point estimates				
100 markets	-0.198	-0.160	-0.130	-0.112	-0.108
200 markets	-0.161	-0.134	-0.104	-0.080	-0.081
400 markets	-0.087	-0.080	-0.056	-0.046	-0.053
800 markets	-0.004	0.006	0.001	-0.008	-0.017

of the number of draws depends on the variance of the random coefficients relative to the variance of the product characteristics. A high variance of the random coefficient implies that a lot of draws are needed to eliminate the effect of the second term of the asymptotic expansion.

This highlights that one cannot give a general guideline of how many draws (or how many markets) suffice to obtain satisfactory results. These Monte Carlo results, however, demonstrate that one should always use bias corrections and standard errors that correct for the

simulation error when making use of Monte Carlo integration in this setup. If the number of simulations is sufficiently large, the corrected estimates and standard errors will be very close to the GMM standard errors. If the number of simulations is small, the simulation error will affect the finite sample performance of the estimator and using the usual GMM asymptotic distributions yields biased estimates and underestimation of the true variance.[8] As mentioned before, there might be computational constraints that do not allow taking a very large number of draws with a larger sample size or with a larger number of random coefficients. In other cases, i.e. when using empirical distributions of demographic characteristics, R might be fixed. Nevertheless, the number of draws should be as large as possible, subject to computational constraints and data availability. The reason is that a large number of draws improves the precision of the initial estimator which is in turn used to calculate the bias correction.

5 Conclusion

One could easily introduce errors in the observed markets shares. Similar to Berry, Linton, and Pakes (2004), one could assume in this setup that one does not observe the true markets shares but an approximation from n random consumers. Also in this case, observing only approximated market shares lead to additional bias and variance terms in the asymptotic expansion. The rate at which n has to go to infinity relative to T in order to obtain $\sqrt{T}$ consistency is identical to the rate requirement for R. This is simple to incorporate because the error in the market shares can be assumed to be independent of the other errors. Moreover, this error does not even depend on the parameter θ. Hence, uniform convergence in any of the additional assumptions employed is not required. It is not treated in this paper because in applications n is usually a lot larger than T in which case this additional error is negligible. Furthermore, the results in this paper are also likely to hold if one observes an unbalanced panel where, asymptotically, all products are observed in infinitely many markets and the total number of products is bounded.

[8] I obtain the same conclusions in various simulation setups.

A Useful lemmas

Lemma A1. Let $f(x,\theta,v) : \mathcal{X} \times \Theta \times \mathbb{R}^q \to [-M_1, M_2]$ be a continuously differentiable function in all arguments where $\mathcal{X}$ is a compact subset of $\mathbb{R}^p$ and Θ is a compact subset of $\mathbb{R}^d$. Let $v_{1,t}, \ldots, v_{R,t}$ be i.i.d. draws from $P_t \in \mathcal{P}$. Let $x_t \in \mathcal{X}$ denotes the (random) data. Assume that

(i) $\ln(T)/R(T) \to 0$ as $T \to \infty$,

(ii) $v_{r,t}$ and x_t are independent, and

(iii) The lth element of $v_{r,t}$, namely $v_{r,t,l}$, satisfies (a) $v_{r,t,l} = g(\gamma_{t,l}, w_{r,t,l})$ where $w_{r,t,l} \sim \tilde{P}_t \in \{P^1, \ldots, P^m\}$ with m finite, g is continuously differentiable in both arguments, and $\gamma_{t,l} \in \Gamma$ where Γ is a compact subset of $\mathbb{R}^k$ or (b) the support of $v_{r,t,l}$ is compact.

Then

$$\sup_{\theta\in\Theta} \max_{1\le t\le T} \left| \frac{1}{R}\sum_{r=1}^{R} f(x_t,\theta,v_{r,t}) - \int f(x_t,\theta,v) dP_t(v) \right| \xrightarrow{p} 0 \qquad \text{as } T \to \infty.$$

Proof. Denote the expectation with respect to the distribution $v_{r,t}$ and conditional on x_t as E_t^*. Then we have to show that

$$\sup_{\theta\in\Theta} \max_{1\le t\le T} \left| \frac{1}{R}\sum_{r=1}^{R} \left(f(x_t,\theta,v_{r,t}) - E_t^*(f(x_t,\theta,v_{r,t}))\right) \right| \xrightarrow{p} 0$$

or that for any $\varepsilon > 0$,

$$E_x\left(\Pr_t^*\left(\max_{1\le t\le T} \sup_{\theta\in\Theta} \left| \frac{1}{R}\sum_{r=1}^{R} \left(f(x_t,\theta,v_{r,t}) - E_t^*(f(x_t,\theta,v_{r,t}))\right) \right| > \varepsilon \right)\right) \to 0 \text{ as } T \to \infty$$

where $\Pr_t^*$ denotes the probability with respect to the distribution v and conditional on x.

Notice that

$$\sup_{\theta\in\Theta} \left| \frac{1}{R}\sum_{r=1}^{R} \left(f(x_t,\theta,v_{r,t}) - E_t^*(f(x_t,\theta,v_{r,t}))\right) \right| \xrightarrow{p} 0$$

simply follows from Jennrich's uniform law of large numbers. The proof now follows from arguments similar to the proof of the uniform law of large numbers of Jennrich. First define $\lambda = (x,\theta)$ and $\Lambda = \mathcal{X} \times \Theta$. Furthermore, denote $f(\lambda, v) = f(x,\theta,v)$. Now partition Λ in $\Lambda_1^n, \ldots, \Lambda_n^n$ such that the difference between any two elements in Λ_i^n goes to 0 as $n \to \infty$ for all i. Let λ_i^n be an arbitrary

element from Λ_i^n for all i. Then

$$\begin{aligned}
&\Pr_t^*\left(\sup_{\theta\in\Theta}\max_{1\le t\le T}\left|\frac{1}{R}\sum_{r=1}^{R}\left(f(x_t,\theta,v_{r,t})-E_t^*(f(x_t,\theta,v_{r,t}))\right)\right|>\varepsilon\right)\\
&\le\sum_{t=1}^{T}\Pr_t^*\left(\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}\left|\frac{1}{R}\sum_{r=1}^{R}\left(f(x,\theta,v_{r,t})-E_t^*(f(x,\theta,v_{r,t}))\right)\right|>\varepsilon\right)\\
&=\sum_{t=1}^{T}\Pr_t^*\left(\sup_{\lambda\in\Lambda}\left|\frac{1}{R}\sum_{r=1}^{R}\left(f(\lambda,v_{r,t})-E_t^*(f(\lambda,v_{r,t}))\right)\right|>\varepsilon\right)\\
&\le\sum_{t=1}^{T}\Pr_t^*\left(\bigcup_{i=1}^{n}\sup_{\lambda\in\Lambda_i^n}\left|\frac{1}{R}\sum_{r=1}^{R}\left(f(\lambda,v_{r,t})-E_t^*(f(\lambda,v_{r,t}))\right)\right|>\varepsilon\right)\\
&\le\sum_{t=1}^{T}\sum_{i=1}^{n}\Pr_t^*\left(\sup_{\lambda\in\Lambda_i^n}\left|\frac{1}{R}\sum_{r=1}^{R}\left(f(\lambda,v_{r,t})-E_t^*(f(\lambda,v_{r,t}))\right)\right|>\varepsilon\right)\\
&\le\sum_{t=1}^{T}\sum_{i=1}^{n}\Pr_t^*\left(\left|\frac{1}{R}\sum_{r=1}^{R}\left(f(\lambda_i^n,v_{r,t})-E_t^*(f(\lambda_i^n,v_{r,t}))\right)\right|>\varepsilon/2\right)\\
&\quad+\sum_{t=1}^{T}\sum_{i=1}^{n}\Pr_t^*\left(\frac{1}{R}\sum_{r=1}^{R}\sup_{\lambda\in\Lambda_i^n}\Big|f(\lambda,v_{r,t})-f(\lambda_i^n,v_{r,t})\right.\\
&\qquad\left.+E_t^*(f(\lambda_i^n,v_{r,t}))-E_t^*(f(\lambda,v_{r,t}))\Big|>\varepsilon/2\right)
\end{aligned}$$

The first term converges to 0 if $\frac{\ln(T)}{R}\to 0$ because by the Bernstein inequality for bounded random variables, there exists a constant C such that for each fixed λ and t

$$\Pr_t^*\left(\left|\sum_{r=1}^{R}\left(f(\lambda,v_{r,t})-E_t^*(f(\lambda,v_{r,t}))\right)\right|>R\varepsilon\right)\le 2\exp\left(-\frac{\varepsilon^2R^2}{CR}\right)=O\left(\exp(-R)\right).$$

For the second term first assume that for all r and t, $v_{r,t}\in\mathcal{V}$ where $\mathcal{V}$ is compact. Then, since f is a continuous function on a compact set, f is by the Heine Cantor theorem uniformly continuous. Hence, for n large enough (so large that $\sup_{\lambda\in\Lambda_i^n}||\lambda_i^n-\lambda||\le\delta$ for some small δ, but n is finite),

$$\sup_{v\in\mathcal{V}}\sup_{\lambda\in\Lambda_i^n}|f(\lambda,v)-f(\lambda_i^n,v)|\le\varepsilon/4.$$

Hence, also for all t,

$$\sup_{\lambda\in\Lambda_i^n}|E_t^*(f(\lambda_i^n,v_{r,t}))-E_t^*(f(\lambda,v_{r,t}))|\le\varepsilon/4$$

which implies that

$$\Pr_t^*\left(\frac{1}{R}\sum_{r=1}^{R}\sup_{\lambda\in\Lambda_i^n}|f(\lambda,v_{r,t})-f(\lambda_i^n,v_{r,t})+E_t^*(f(\lambda_i^n,v_{r,t}))-E_t^*(f(\lambda,v_{r,t}))|>\varepsilon/2\right)=0.$$

Alternatively assume that for all l, $v_{r,t,l}=g(\gamma_{t,l},w_{r,t,l})$ where $w_{r,t,l}\sim\tilde{P}_t\in\{P^1,\ldots,P^m\}$ where m is finite and $\gamma_{t,l}\in\Gamma$ where Γ is a compact subset of $\mathbb{R}^k$. Denote the vector of stacked elements

$v_{r,t,l}$ by $v_{r,t} = g(\gamma_t, w_{r,t})$. With abuse of notation now define $\lambda = (x, \theta, \gamma)$ and $\Lambda = \mathcal{X} \times \Theta \times \Gamma$. Furthermore, denote $f(\lambda, w) = f(x, \theta, v) = f(x, \theta, g(\gamma, w))$. Now partition Λ in $\Lambda_1^n, \ldots, \Lambda_n^n$ such that the difference between any two elements in Λ_i^n goes to 0 as $n \to \infty$ for all i. Let λ_i^n be an arbitrary element from Λ_i^n for all i. Then

$$
\begin{aligned}
&\Pr_t^* \left(\sup_{\theta \in \Theta} \max_{1 \le t \le T} \left| \frac{1}{R} \sum_{r=1}^{R} (f(x_t, \theta, v_{r,t}) - E_t^*(f(x_t, \theta, v_{r,t}))) \right| > \varepsilon \right) \\
&\quad = \Pr_t^* \left(\sup_{\theta \in \Theta} \max_{1 \le t \le T} \left| \frac{1}{R} \sum_{r=1}^{R} (f(x_t, \theta, g(\gamma_t, w_{r,t})) - E_t^*(f(x_t, \theta, g(\gamma_t, w_{r,t})))) \right| > \varepsilon \right) \\
&\quad \le \sum_{t=1}^{T} \Pr_t^* \left(\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} \sup_{\gamma \in \Gamma} \left| \frac{1}{R} \sum_{r=1}^{R} (f(x, \theta, g(\gamma, w_{r,t})) - E_t^*(f(x, \theta, g(\gamma, w_{r,t})))) \right| > \varepsilon \right) \\
&\quad = \sum_{t=1}^{T} \Pr_t^* \left(\sup_{\lambda \in \Lambda} \left| \frac{1}{R} \sum_{r=1}^{R} (f(\lambda, w_{r,t}) - E_t^*(f(\lambda, w_{r,t}))) \right| > \varepsilon \right) \\
&\quad \le \sum_{t=1}^{T} \sum_{i=1}^{n} \Pr_t^* \left(\left| \frac{1}{R} \sum_{r=1}^{R} (f(\lambda_i^n, w_{r,t}) - E_t^*(f(\lambda_i^n, w_{r,t}))) \right| > \varepsilon/2 \right) \\
&\qquad + \sum_{t=1}^{T} \sum_{i=1}^{n} \Pr_t^* \left(\frac{1}{R} \sum_{r=1}^{R} \sup_{\lambda \in \Lambda_i^n} \Big| f(\lambda, w_{r,t}) - f(\lambda_i^n, w_{r,t}) \right. \\
&\qquad\qquad \left. + E_t^*(f(\lambda_i^n, w_{r,t})) - E_t^*(f(\lambda, w_{r,t})) \Big| > \varepsilon/2 \right)
\end{aligned}
$$

Again, the first term converges to 0 if $\frac{\ln(T)}{R} \to 0$ by the Bernstein inequality for bounded random variables. For the second term define

$$h_t(\lambda, w_{r,t}) = f(\lambda, w_{r,t}) - E_t^*(f(\lambda, w_{r,t}))$$

where the expected value is with respect to $w_{r,t}$ which can only be drawn from a finite number of distributions for each t. Then for all t,

$$\lim_{n \to \infty} \sup_{\lambda \in \Lambda_i^n} |(h_t(\lambda, w_{r,t}) - h_t(\lambda_i^n, w_{r,t}))| = 0$$

for all t. Thus, by Lebesgue's dominated convergence theorem,

$$\lim_{n \to \infty} E_t^* \sup_{\lambda \in \Lambda_i^n} |(h_t(\lambda, w_{r,t}) - h_t(\lambda_i^n, w_{r,t}))| = 0$$

uniformly over i. This function depends on t since the distribution might differ. However, since only a finite number of distributions are allowed, there exists an n such that for all t

$$E_t^* \sup_{\lambda \in \Lambda_i^n} |(h(\lambda, w_{r,t}) - h(\lambda_i^n, w_{r,t}))| \le \varepsilon/4.$$

Then

$$\sum_{t=1}^{T}\sum_{i=1}^{n}\Pr^*\left(\frac{1}{R}\sum_{r=1}^{R}\sup_{\lambda\in\Lambda_i^n}|(h_t(\lambda,w_{r,t})-h_t(\lambda_i^n,w_{r,t}))|>\varepsilon/2\right)$$
$$\leq \sum_{t=1}^{T}\sum_{i=1}^{n}\Pr^*\left(\frac{1}{R}\sum_{r=1}^{R}\sup_{\lambda\in\Lambda_i^n}|(h_t(\lambda,w_{r,t})-h_t(\lambda_i^n,w_{r,t}))|\right.$$
$$\left.-E_t^*\sup_{\lambda\in\Lambda_i^n}|(h_t(\lambda,w_{r,t})-h_t(\lambda_i^n,w_{r,t}))|>\varepsilon/4\right)$$

This probability converges to 0 using the Bernstein inequality as in the first part.

One can combine these results to show that

$$\sup_{\theta\in\Theta}\max_{1\leq t\leq T}\left|\frac{1}{R}\sum_{r=1}^{R}f(x_t,\theta,v_{r,t}^1,v_{r,t}^2)-\int f(x_t,\theta,v^1,v^2)dP_t^1(v_1)dP_t^2(v_2)\right|\xrightarrow{p}0 \qquad \text{as } T\to\infty$$

where v_1 satisfies condition $(iii-a)$ and v_2 satisfies condition $(iii-b)$.

□

The following lemma is Lemma A in Serfling (1980, p. 304).

Lemma A2. Let $Y_1, Y_2, \ldots$ be independent random variables with mean 0. Let v be an even integer. Then

$$E\left(\left|\sum_{r=1}^{R}Y_r\right|^v\right)\leq A_v R^{(v/2-1)}\sum_{r=1}^{R}E\left(|Y_r|^v\right)$$

where A_v is a universal constant depending only on v.

Proof. See Serfling (1980). □

Lemma A3. Suppose that $(v_1^{(t)},...,v_R^{(t)},x_t)$ with $t=1,\ldots,T$ are random vectors such that $v_r^{(t)}$ is i.i.d. across r and independent of x_t. Let

$$y_{R,t}=s(x_t)\left[\frac{1}{R}\sum_{r=1}^{R}p\left(v_r^{(t)},x_t\right)\right]^{m_1}\left[\frac{1}{R}\sum_{r=1}^{R}q\left(v_r^{(t)},x_t\right)\right]^{m_2}$$

where m_1 and m_2 are nonnegative integers and s, p and q are measurable functions such that

$$E\left(p\left(v_r^{(t)},x_t\right)|x_t\right)=E\left(q\left(v_r^{(t)},x_t\right)|x_t\right)=0$$

for all t. Also assume that if $m_1>0$ and $m_2>0$ for some a and b satisfying $\frac{1}{a}+\frac{1}{b}=1$ it holds that for some finite M

$$E\left(|s(x_t)|^{2a}p\left(v_r^{(t)},x_t\right)^{2am_1}\right)\leq M$$

and

$$E\left(|s(x_t)|^{2b}q\left(v_r^{(t)},x_t\right)^{2bm_2}\right)\leq M.$$

If $m_1>0$ and $m_2=0$ assume instead that

$$E\left(|s(x_t)|^{2}p\left(v_r^{(t)},x_t\right)^{2m_1}\right)\leq M.$$

Then

$$\frac{1}{T}\sum_{t=1}^{T}|y_{R,t}|=O_p\left(R^{-(m_1+m_2)/2}\right).$$

Proof. The proof is very similar to the proof of Lemma A.2 in Lee (1995). By Hölder's inequality

$$E\left(y_{R,t}^2\right)\leq\left(E\left(|s(x_t)|^{2a}\left[\frac{1}{R}\sum_{r=1}^{R}p\left(v_r^{(t)},x_t\right)\right]^{2am_1}\right)\right)^{1/a}\left(E\left(|s(x_t)|^{2b}\left[\frac{1}{R}\sum_{r=1}^{R}q\left(v_r^{(t)},x_t\right)\right]^{2bm_2}\right)\right)^{1/b}.$$

Lemma A2 implies that for some constant c,

$$E\left(\left[\frac{1}{R}\sum_{r=1}^{R}p\left(v_r^{(t)},x_t\right)\right]^{2am_1}\middle|x_t\right)\leq\frac{c}{R^{m_1a}}E\left(p\left(v_r^{(t)},x_t\right)^{2am_1}\middle|x_t\right)$$

and

$$E\left(\left[\frac{1}{R}\sum_{r=1}^{R}q\left(v_r^{(t)},x_t\right)\right]^{2bm_2}\middle|x_t\right)\leq\frac{c}{R^{m_2b}}E\left(q\left(v_r^{(t)},x_t\right)^{2bm_2}\middle|x_t\right).$$

It follows that

$$\begin{aligned}E\left(y_{R,t}^2\right)&\leq\frac{c}{R^{m_1+m_2}}\left(E\left(|s(x_t)|^{2a}p\left(v_r^{(t)},x_t\right)^{2am_1}\right)\right)^{1/a}\left(E\left(|s(x_t)|^{2b}p\left(v_r^{(t)},x_t\right)^{2bm_2}\right)\right)^{1/b}\\&\leq\frac{cM}{R^{m_1+m_2}}.\end{aligned}$$

By Markov's inequality it now follows that

$$\begin{aligned}P\left(R^{(m_1+m_2)/2}\frac{1}{T}\sum_{t=1}^{T}|y_{R,t}|\geq\varepsilon\right)&\leq\frac{R^{(m_1+m_2)/2}}{\varepsilon}\frac{1}{T}\sum_{t=1}^{T}E\left(|y_{R,t}|\right)\\&\leq\frac{R^{(m_1+m_2)/2}}{\varepsilon}\frac{1}{T}\sum_{t=1}^{T}\left(E\left(|y_{R,t}|^2\right)\right)^{1/2}\\&\leq\frac{R^{(m_1+m_2)/2}}{\varepsilon}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{cM}{R^{m_1+m_2}}\right)^{1/2}\\&=\frac{(cM)^{1/2}}{\varepsilon}\end{aligned}$$

which means that $R^{(m_1+m_2)/2}\frac{1}{T}\sum_{t=1}^{T}|y_{R,t}|$ is $O_p(1)$.

□

Lemma A4. Suppose $F : \mathbb{R}^J \to \mathbb{R}$ is a $k+1$ times continuously differentiable function on an open convex set $S \subseteq \mathbb{R}^J$. For $\alpha \in \mathbb{R}^J$, let $|\alpha| = \alpha_1 + \ldots + \alpha_J$ and $\alpha! = \alpha_1! * \ldots * \alpha_J!$. Furthermore, for $x \in \mathbb{R}^J$ let

$$x^{|\alpha|} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_J^{\alpha_J}$$

and

$$\frac{\partial^{|\alpha|} F(x)}{\partial x^{|\alpha|}} = \frac{\partial^{|\alpha|} F(x)}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_J^{\alpha_J}}.$$

If $a \in S$ and $a + h \in S$, then

$$F(a+h) = \sum_{\alpha \in \mathbb{R}_+ : |\alpha| \leq k} \frac{1}{\alpha!} \left(\frac{\partial^{|\alpha|} F(x)}{\partial x^{|\alpha|}} \bigg|_{x=a} h^{|\alpha|} \right) + \sum_{\alpha \in \mathbb{R}_+ : |\alpha| = k+1} \frac{1}{\alpha!} \left(\frac{\partial^{|\alpha|} F(x)}{\partial x^{|\alpha|}} \bigg|_{x=a+ch} \right) h^{|\alpha|}.$$

for some $c \in (0,1)$.

Proof. This is a standard result. □

B Proof of Lemma 1

I will first prove that the assumptions imply that $\Delta\xi(\theta, P_{0,t}, s_t, p_t, x_t)$ is an element of a bounded subset of $\mathbb{R}^J$. Let Q_t denote the distribution with compact support. Assume for simplicity that for the random coefficients with non-compact support, $v_{r,t,l} = g(a_{t,l}, w_{r,t,l})$ where $w_{r,t,l}$ has a distribution function F and $a_t \in \Upsilon$. The more general case can easily be dealt with. Now assume that $\Delta\xi(\theta, P_{0,t}, s_t, p_t, x_t)$ is not an element of a bounded subset of $\mathbb{R}^J$. Then, there exist, a sequence $\{s^n, p^n, x^n, \theta^n\}$, $n = 1, 2, \ldots$ with $\varepsilon \leq s_j^n \leq 1 - \varepsilon$, $j = 0, \ldots, J$, $\theta \in \Theta$ and $(x^n, p^n) \in \mathcal{M}$ such that for some j, $\Delta\xi_j^n \equiv \Delta\xi_j(\theta^n, P_{0,t_n}, s^n, p^n, x^n) \to \infty$ or $\Delta\xi_j^n \to -\infty$. I assume that $\Delta\xi_j^n \to \infty$. The other case is similar. Note that for all $k = 1, \ldots, J$, s_k^n satisfies

$$s_k^n = \int \frac{\exp(\gamma(x_k^n, p_k^n, \xi_k^n, \Delta\xi_k^n, v, D; a_{t_n}, \theta^n))}{1 + \sum_{m=1}^J \exp(\gamma(x_m^n, p_m^n, \xi_m^n, \Delta\xi_m^n, v, D; a_{t_n}, \theta^n))} dQ_{t_n}(D) dF(v).$$

Let Q_t be a distribution with compact support and let F be a distribution that is identical in all markets. Define

$$s_k(\Delta\xi^n) \equiv \sup_{\theta \in \Theta} \sup_{(x,p) \in \mathcal{M}} \sup_{D \in supp(Q_t)} \sup_{a \in \Upsilon} \int \frac{\exp(\gamma(x_k, p_k, \xi_k, \Delta\xi_k^n, v, D; a, \theta))}{1 + \sum_{m=1}^J \exp(\gamma(x_m, p_m, \xi_m, \Delta\xi_m^n, v, D; a, \theta))} dF(v).$$

Let $K \in \mathbb{R}$ be arbitrary and assume that $\Delta\xi_k^n \leq K$ for some $k \neq j$ and all n. Since Θ, $\mathcal{M}$, and $supp(Q_t)$ are compact sets,

$$\begin{aligned} s_k^n &\leq s_k(\Delta\xi^n) \\ &\leq \sup_{\theta \in \Theta} \sup_{(x,s) \in \mathcal{M}} \sup_{D \in supp(Q_t)} \sup_{a \in \Upsilon} \int \frac{\exp(\gamma(x_k, p_k, \xi_k, \Delta\xi_k^n, v, D; a, \theta))}{1 + \exp(\gamma(x_j, p_j, \xi_j, \Delta\xi_j^n, v, D; a, \theta))} dF(v) \\ &\to 0 \end{aligned}$$

as $n \to \infty$ because $\Delta\xi_j^n \to \infty$. Hence, for some $N \in \mathbb{N}$, $s_k^n < \varepsilon$ for all $n > N$ which contradicts Assumption 2. The previous argument also implies that if $\Delta\xi_k^{n^*} \leq K$ for some $n^* > N$, $s_k^{n^*} < \varepsilon$, which is again a contradiction. Thus $\Delta\xi_k^n > K$ for all $n > N$. Since K was arbitrary, this implies that $s_0^n \to 0$ which is a contraction.

The proof now follows from Lemma A1.

C Proof of Theorem 1

I use a simplified version of the proof of Berry, Linton, and Pakes (2004). The proof consists of two steps. I first show that an estimator defined as any sequence that satisfies

$$||G_T(\breve{\theta}, P_0)|| = \inf_{\theta \in \Theta} ||G_T(\theta, P_0)|| + o_p(1)$$

is a consistent estimator for θ. To prove this, first notice the law of large numbers implies that $||G_T(\theta_0, P_0)|| = O_p(1/\sqrt{T})$. Thus, by Theorem 1 of Pakes and Pollard (1989), It is sufficient to prove that

$$\sup_{||\theta - \theta_0|| > \delta} ||G_T(\theta, P_0)||^{-1} = O_p(1) \quad \text{for all } \delta > 0.$$

This is implied by proving that for any $(\varepsilon, \delta) > (0, 0)$ there exists $C^*(\delta) > 0$ and $T(\varepsilon, \delta)$ such that for all $T \geq T(\varepsilon, \delta)$,

$$\Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0)|| \geq C^*(\delta)\right) \geq 1 - \varepsilon.$$

Now fix $(\varepsilon, \delta) > (0, 0)$. Take $C(\delta)$ such that

$$\lim_{T \to \infty} \Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0) - G_T(\theta_0, P_0)|| \geq C(\delta)\right) = 1.$$

Now let $\varepsilon^* = \min\{\varepsilon, C(\delta)\}$, such that $0 < \varepsilon^* \leq \varepsilon$. Then by the triangle inequality

$$\begin{aligned}\Pr&\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0)|| \geq C(\delta) - ||G_T(\theta_0, P_0)||\right) \\ &\geq \Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0) - G_T(\theta_0, P_0)|| \geq C(\delta)\right).\end{aligned}$$

Now by Assumption 5, there exists $T_1(\varepsilon^*)$ such that for all $T \geq T_1(\varepsilon^*, \delta)$

$$\Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0) - G_T(\theta_0, P_0)|| \geq C(\delta)\right) \geq 1 - \varepsilon^*/2.$$

Since $||G_T(\theta_0, P_0)|| = o_p(1)$, there exists a $T_2(\varepsilon^*)$ such that for all $T \geq T_2(\varepsilon^*)$

$$\Pr\left(||G_T(\theta, P_0)|| \geq \varepsilon^*/2\right) \leq \varepsilon^*/2.$$

It follows that for all $T \geq T_2(\varepsilon^*)$,

$$\begin{aligned}
&\Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0)|| \leq C(\delta) - \varepsilon^*/2\right) \\
&\qquad \leq \Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0)|| \leq C(\delta) - ||G_T(\theta_0, P_0)||\right) + \varepsilon^*/2.
\end{aligned}$$

Thus for all $T \geq \max\{T_1(\varepsilon^*, \delta), T_2(\varepsilon^*)\}$

$$\begin{aligned}
&\Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0)|| \geq C(\delta)/2\right) \\
&\qquad \geq \Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0)|| \geq C(\delta) - \varepsilon^*/2\right) \\
&\qquad \geq \Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0)|| \geq C(\delta) - ||G_T(\theta_0, P_0)||\right) - \varepsilon^*/2. \\
&\qquad \geq \Pr\left(\inf_{\theta \notin \mathcal{N}_{\theta_0}(\delta)} ||G_T(\theta, P_0) - G_T(\theta_0, P_0)|| \geq C(\delta)\right) - \varepsilon^*/2. \\
&\qquad \geq 1 - \varepsilon^*/2 - \varepsilon^*/2 \\
&\qquad \geq 1 - \varepsilon.
\end{aligned}$$

Defining $C^*(\delta) = C(\delta)/2 > 0$ completes the proof.

In the second part I show that $\sup_{\theta \in \Theta} ||G_T(\theta, P_R) - G_T(\theta, P_0)||$ converges to 0 in probability. This implies by the triangle inequality that for any sequence $\theta_T \in \Theta$ we have

$$\left|||G_T(\theta_T, P_R)|| - ||G_T(\theta_T, P_0)||\right| \leq ||G_T(\theta_T, P_R) - G_T(\theta_T, P_0)|| = o_p(1)$$

Denote

$$\tilde{\theta} = \arg\inf_{\theta \in \Theta} ||G_T(\theta, P_0)||.$$

Then we get that the actual estimator $\hat{\theta}$ satisfies

$$||G_T(\hat{\theta}, P_0)|| = \inf_{\theta \in \Theta} ||G_T(\theta, P_0)|| + o_p(1)$$

because

$$\begin{aligned}
0 \leq ||G_T(\hat{\theta}, P_0)|| - \inf_{\theta \in \Theta} ||G_T(\theta, P_0)|| &= ||G_T(\hat{\theta}, P_0)|| - ||G_T(\tilde{\theta}, P_0)|| \\
&= ||G_T(\hat{\theta}, P_R)|| - ||G_T(\tilde{\theta}, P_0)|| + o_p(1) \\
&\leq ||G_T(\tilde{\theta}, P_R)|| - ||G_T(\tilde{\theta}, P_0)|| + o_p(1) \\
&= o_p(1).
\end{aligned}$$

Hence, by the first step, proving that

$$\sup_{\theta\in\Theta}||G_T(\theta,P_R)-G_T(\theta,P_0)||=o_p(1)$$

is sufficient for consistency. Now by the Cauchy Schwarz inequality,

$$\begin{aligned}||G_T(\theta,P_R)-G_T(\theta,P_0)||^2 &= \frac{1}{T^2}||Z'(\Delta\xi(\theta,P_R)-\Delta\xi(\theta,P_0))||^2\\ &\leq \frac{1}{T}||Z'Z||\times\frac{1}{T}||\Delta\xi(\theta,P_R)-\Delta\xi(\theta,P_0)||^2.\end{aligned}$$

Since $\frac{1}{T}||Z'Z||=O_p(1)$ by Assumption 3, it suffices to prove that

$$\sup_{\theta\in\Theta}\frac{1}{T}||\Delta\xi(\theta,P_R)-\Delta\xi(\theta,P_0)||^2=\sup_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}||\Delta\xi_t(\theta,P_{R,t})-\Delta\xi_t(\theta,P_{0,t})||^2=o_p(1).$$

By Assumption 1 we have

$$\begin{aligned}&\sup_{\theta\in\Theta}\max_{1\leq t\leq T}||\sigma_t(\Delta\xi_t(\theta,P_{R,t}),\theta,P_{R,t})-\sigma_t(\Delta\xi_t(\theta,P_{0,t}),\theta,P_{R,t})||\\ &\quad=\sup_{\theta\in\Theta}\max_{1\leq t\leq T}||\sigma_t(\Delta\xi_t(\theta,P_{0,t}),\theta,P_{0,t})-\sigma_t(\Delta\xi_t(\theta,P_{0,t}),\theta,P_{R,t})||\\ &\quad=o_p(1).\end{aligned}$$

This then implies that

$$\sup_{\theta\in\Theta}\max_{1\leq t\leq T}||\Delta\xi_t(\theta,P_{R,t})-\Delta\xi_t(\theta,P_{0,t})||^2=o_p(1)$$

because by Assumption 4, if instead

$$\sup_{\theta\in\Theta}\max_{1\leq t\leq T}||\Delta\xi_t(\theta,P_{R,t})-\Delta\xi_t(\theta,P_{0,t})||^2>\delta,$$

then

$$\sup_{\theta\in\Theta}\max_{1\leq t\leq T}||\sigma_t(\Delta\xi_t(\theta,P_{R,t}),\theta,P_{R,t})-\sigma_t(\Delta\xi_t(\theta,P_{0,t}),\theta,P_{R,t})||>\varepsilon$$

with probability approaching 1 which is a contradiction.

D Proof of Theorem 2

The objective is to minimize $G_T(\theta,P_R)'W_TG_T(\theta,P_R)$ where

$$G_T(\theta,P_R) \equiv \frac{1}{T}\sum_{t=1}^{T}z_t'\Delta\xi_t(\theta,P_{R,t}).$$

The first order condition is

$$\left(\frac{\partial}{\partial\theta}G_T(\hat{\theta},P_R)'\right)W_TG_T(\hat{\theta},P_R)=0.$$

Define

$$D_T(\hat{\theta}, P_R) = \frac{\partial}{\partial \theta} G_T(\hat{\theta}, P_R).$$

Using a first order expansion of $G_T(\hat{\theta}, P_R)$ around $\theta = \theta_0$ yields

$$D_T(\hat{\theta}, P_R)' W_T \left(G_T(\theta_0, P_R) + D_T(\tilde{\theta}, P_R) \left(\hat{\theta} - \theta_0 \right) \right) = 0$$

where $\tilde{\theta}$ is between θ_0 and $\hat{\theta}$. Thus

$$\sqrt{T} \left(\hat{\theta} - \theta_0 \right) = \left(D_T(\hat{\theta}, P_R)' W_T D_T(\tilde{\theta}, P_R) \right)^{-1} D_T(\hat{\theta}, P_R)' W_T \sqrt{T} G_T(\theta_0, P_R).$$

The proof now consists of two parts. First I show that

$$\sqrt{T} G_T(\theta_0, P_R) \quad = \quad \sqrt{T} G_T(\theta_0, P_0) + O_p\left(\frac{1}{\sqrt{R}} \right) + O_p\left(\frac{\sqrt{T}}{R} \right) + o_p\left(\frac{\sqrt{T}}{R} \right)$$

and I derive an expression for the third term. Next I prove that for any consistent estimator $\breve{\theta}$ of θ it holds that $\hat{D}_T(\hat{\theta}, P_R)$ converges to Γ in probability. Combining these results yields the conclusion of the theorem.

Let $F : \mathbb{R}^J \to \mathbb{R}^J$ be a three times continuously invertible function. The inverse function is defined by $F^{-1} : \mathbb{R}^J \to \mathbb{R}^J$. Then by Lemma A4 for any $s_1, s_0 \in \mathbb{R}^J$ there exists a $c \in (0, 1)$ such that

$$\begin{aligned} F_j^{-1}(s_1) \quad = \quad & F_j^{-1}(s_0) + \left. \frac{\partial F_j^{-1}(s)}{\partial s} \right|_{s=s_0} (s_1 - s_0) + \frac{1}{2} (s_1 - s_0)' \left. \frac{\partial F_j^{-1}(s)}{\partial s' \partial s} \right|_{s=s_0} (s_1 - s_0) \\ & + \sum_{\alpha \in \mathbb{R}_+ : |\alpha| = 3} \frac{1}{\alpha!} \left(\left. \frac{\partial^{|\alpha|} F_j^{-1}(s)}{\partial s^{|\alpha|}} \right|_{s = s_0 + c(s_1 - s_0)} \right) (s_1 - s_0)^{|\alpha|} \end{aligned}$$

where F_j^{-1} denotes the jth element of vector F^{-1}.

Next we derive an expression for the first and second derivative of the inverse function. First note that

$$F^{-1}(F(x)) = x.$$

It follows that

$$\frac{\partial F^{-1}(F(x))}{\partial x} = I \Leftrightarrow \frac{\partial F^{-1}(F(x))}{\partial F(x)} \frac{\partial F(x))}{\partial x} = I$$

and hence with $s = F(x)$

$$\frac{\partial F^{-1}(s)}{\partial s} = \left(\frac{\partial F(x)}{\partial x} \right)^{-1}.$$

The jth row of this matrix is $\frac{\partial F_j^{-1}(s)}{\partial s}$ which implies that

$$\frac{\partial F_j^{-1}(s)}{\partial s} = e_j' \left(\frac{\partial F(x)}{\partial x} \right)^{-1}$$

and

$$\frac{\partial F_j^{-1}(s)}{\partial s'} = \left(\frac{\partial F(x)}{\partial x'}\right)^{-1} e_j$$

where e_j is a $J \times 1$ vector of zero with a 1 at the jth element. Next notice that

$$\frac{\partial F_j^{-1}(F(x))}{\partial F(x)' \partial x_i} = \frac{\partial F_j^{-1}(F(x))}{\partial F(x)' \partial F(x)} \frac{\partial F(x)}{\partial x_i}$$

which implies that

$$\frac{\partial F_j^{-1}(F(x))}{\partial F(x)' \partial x} = \frac{\partial F_j^{-1}(F(x))}{\partial F(x)' \partial F(x)} \frac{\partial F(x)}{\partial x}$$

or

$$\frac{\partial F_j^{-1}(F(x))}{\partial F(x)' \partial F(x)} = \frac{\partial F_j^{-1}(F(x))}{\partial F(x)' \partial x} \left(\frac{\partial F(x)}{\partial x}\right)^{-1}.$$

Finally

$$\frac{\partial F_j^{-1}(F(x))}{\partial F(x)' \partial x_i} = \frac{\partial \left(\frac{\partial F(x)}{\partial x'}\right)^{-1} e_j}{\partial x_i} = -\left(\frac{\partial F(x)}{\partial x'}\right)^{-1} \left(\frac{\partial^2 F(x)}{\partial x' \partial x_i}\right) \left(\frac{\partial F(x)}{\partial x'}\right)^{-1} e_j$$

It now follows that

$$\begin{aligned}
\frac{\partial F_j^{-1}(F(x))}{\partial F(x)' \partial F(x)} &= -\left(\left(\tfrac{\partial F(x)}{\partial x'}\right)^{-1} \tfrac{\partial^2 F(x)}{\partial x' \partial x_1} \left(\tfrac{\partial F(x)}{\partial x'}\right)^{-1} e_j \quad \cdots \quad \left(\tfrac{\partial F(x)}{\partial x'}\right)^{-1} \tfrac{\partial^2 F(x)}{\partial x' \partial x_J} \left(\tfrac{\partial F(x)}{\partial x'}\right)^{-1} e_j\right) \left(\frac{\partial F(x)}{\partial x'}\right)^{-1} \\
&= -\sum_{k=1}^{J} \left(\frac{\partial F(x)}{\partial x'}\right)^{-1} \frac{\partial^2 F(x)}{\partial x' \partial x_k} \left(\frac{\partial F(x)}{\partial x'}\right)^{-1} e_j e_k' \left(\frac{\partial F(x)}{\partial x'}\right)^{-1}.
\end{aligned}$$

The previous expansion implies that there exists $c_t \in (0,1)$ such that

$$\begin{aligned}
\Delta\xi_{j,t}(\theta_0, P_{R,t}) &= \Delta\xi_{j,t}(\theta_0, P_{0,t}) \\
&\quad -e_j' H_{R,t}^{-1} e_{R,t} \\
&\quad +\frac{1}{2} e_{R,t}' \left(-\sum_{k=1}^{J} H_{R,t}^{-1} \left(\frac{\partial^2 \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial \Delta\xi \partial \Delta\xi_k}\right) H_{R,t}^{-1} e_j e_k' H_{R,t}^{-1}\right) e_{R,t} \\
&\quad +\sum_{\alpha \in \mathbb{R}_+ : |\alpha| = 3} \frac{1}{\alpha!} \left(\left.\frac{\partial^{|\alpha|} \sigma_{j,t}^{-1}(s, \theta_0, P_{R,t})}{\partial s^{|\alpha|}}\right|_{s = c_t s_t + (1-c_t)\sigma_t\left(\Delta\xi(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}\right) e_{R,t}^{|\alpha|} \\
&= \Delta\xi_{j,t}(\theta_0, P_{0,t}) \\
&\quad -e_j' H_{0,t}^{-1} e_{R,t} \\
&\quad +e_j' H_{0,t}^{-1} \frac{\partial e_{R,t}}{\partial \Delta\xi} H_{0,t}^{-1} e_{R,t} \\
&\quad -\frac{1}{2} e_{R,t}' \left(\sum_{k=1}^{J} H_{0,t}^{-1} \left(\frac{\partial^2 \sigma_t\left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial \Delta\xi \partial \Delta\xi_k}\right) H_{0,t}^{-1} e_j e_k' H_{0,t}^{-1}\right) e_{R,t} \\
&\quad +error_{j,t}
\end{aligned}$$

where

$$\begin{aligned} error_{j,t} &= e_j' \left(H_{R,t}^{-1} - H_{0,t}^{-1}\right) \left(H_{R,t} - H_{0,t}\right) H_{0,t}^{-1} e_{R,t} \\ &\quad -\frac{1}{2} e_{R,t}' \left(\sum_{k=1}^{J} \left(H_{R,t}^{-1} \left(\frac{\partial^2 \sigma_t \left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)}{\partial \Delta\xi \partial \Delta\xi_k} \right) H_{R,t}^{-1} e_j e_k' H_{R,t}^{-1} \right.\right. \\ &\qquad \left.\left. -H_{0,t}^{-1} \left(\frac{\partial^2 \sigma_t \left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial \Delta\xi \partial \Delta\xi_k} \right) H_{0,t}^{-1} e_j e_k' H_{0,t}^{-1} \right) \right) e_{R,t} \\ &\quad + \sum_{\alpha \in \mathbb{R}_+ : |\alpha| = 3} \frac{1}{\alpha!} \left(\left. \frac{\partial^{|\alpha|} \sigma_{j,t}^{-1}\left(s, \theta_0, P_{R,t}\right)}{\partial s^{|\alpha|}} \right|_{s = c_t s_t + (1-c_t)\sigma_t\left(\Delta\xi(\theta_0, P_{0,t}), \theta_0, P_{R,t}\right)} \right) e_{R,t}^{|\alpha|}. \end{aligned}$$

Now define

$$K_{0,t,k} \equiv \left(\frac{\partial^2 \sigma_t \left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t}\right)}{\partial \Delta\xi \partial \Delta\xi_k} \right)$$

and

$$I_{0,t,j} \equiv \sum_{k=1}^{J} H_{0,t}^{-1} K_{0,t,k} H_{0,t}^{-1} e_j e_k' H_{0,t}^{-1}.$$

Then

$$\begin{aligned} \Delta\xi_{j,t}(\theta_0, P_{R,t}) &= \Delta\xi_{j,t}(\theta_0, P_{0,t}) \\ &\quad -e_j' H_{0,t}^{-1} e_{R,t} \\ &\quad +e_j' H_{0,t}^{-1} \frac{\partial e_{R,t}}{\partial \Delta\xi} H_{0,t}^{-1} e_{R,t} - \frac{1}{2} e_{R,t}' I_{0,t,j} e_{R,t} \\ &\quad +error_{j,t}, \end{aligned}$$

It now follows that

$$\begin{aligned} \sqrt{T} G_T(\theta_0, P_R) &= \sqrt{T} G_T(\theta_0, P_0) + \sqrt{T} \left(G_T(\theta_0, P_R) - G_T(\theta_0, P_0) \right) \\ &= \sqrt{T} G_T(\theta_0, P_0) + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_t' \left(\Delta\xi_t(\theta_0, P_{R,t}) - \Delta\xi_t(\theta_0, P_{0,t}) \right) \\ &= \sqrt{T} G_T(\theta_0, P_0) - \frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_t' H_{0,t}^{-1} e_{R,t} \\ &\quad + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sum_{j=1}^{J} z_{j,t}' \left(e_j' H_{0,t}^{-1} \frac{\partial e_{R,t}}{\partial \Delta\xi} H_{0,t}^{-1} e_{R,t} - \frac{1}{2} e_{R,t}' I_{0,t,j} e_{R,t} \right) \\ &\quad + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_t' error_t. \end{aligned}$$

Now we have four terms. The first term is $O_p(1)$ and belongs to the GMM objective function without simulation error. Therefore, by Assumption B9 it converges to a normally distributed random variable. The second term is $O_p\left(\frac{1}{\sqrt{R}}\right)$ and converges to a normally distributed random variable as well when multiplied by $\sqrt{R}$ by Assumption B9. These two normal terms are asymptotically

independent.

The third term does not have means 0 because

$$\begin{aligned}
&E\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sum_{j=1}^{J}z'_{j,t}\left(e'_jH_{0,t}^{-1}\left(\frac{\partial e_{R,t}}{\partial\Delta\xi}\right)H_{0,t}^{-1}e_{R,t}-\frac{1}{2}e'_{R,t}I_{0,t,j}e_{R,t}\right)\right)\\
&\quad=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sum_{j=1}^{J}E\left(z'_{j,t}\left(\frac{1}{R}e'_jH_{0,t}^{-1}E_t^*\left(d\varepsilon_{r,0,t}H_{0,t}^{-1}\varepsilon_{r,0,t}\right)-\frac{1}{2}\frac{1}{R}E_t^*\left(\varepsilon'_{r,0,t}I_{0,t,j}\varepsilon_{r,0,t}\right)\right)\right)\\
&\quad=\frac{\sqrt{T}}{R}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J}E\left(z'_{j,t}\left(e'_jH_{0,t}^{-1}E_t^*\left(d\varepsilon_{r,0,t}H_{0,t}^{-1}\varepsilon_{r,0,t}\right)-\frac{1}{2}E_t^*\left(\varepsilon'_{r,0,t}I_{0,t,j}\varepsilon_{r,0,t}\right)\right)\right).
\end{aligned}$$

By Assumption B10

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J}E\left(z'_{j,t}\left(e'_jH_{0,t}^{-1}E_t^*\left(d\varepsilon_{r,0,t}H_{0,t}^{-1}\varepsilon_{r,0,t}\right)-\frac{1}{2}E_t^*\left(\varepsilon'_{r,0,t}I_{0,t,j}\varepsilon_{r,0,t}\right)\right)\right)=\bar{\mu}.$$

By a weak law of large numbers

$$\frac{R}{\sqrt{T}}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sum_{j=1}^{J}z'_{j,t}\left(e'_jH_{0,t}^{-1}\left(\frac{\partial e_{R,t}}{\partial\Delta\xi}\right)H_{0,t}^{-1}e_{R,t}-\frac{1}{2}e'_{R,t}I_{0,t,j}e_{R,t}\right)\xrightarrow{p}\bar{\mu}$$

which implies that the third term is $O_p\left(\frac{\sqrt{T}}{R}\right)$ and converges in probability to a constant term when multiplied by $\frac{R}{\sqrt{T}}$.

Finally I need to prove that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}z'_t error_{R,t} \quad = \quad o_p\left(\frac{\sqrt{T}}{R}\right).$$

Recall that

$$\begin{aligned}
&\frac{1}{\sqrt{T}}\sum_{t=1}^{T}z'_t error_{j,t}\\
&\quad=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}z'_te'_j\left(H_{R,t}^{-1}-H_{0,t}^{-1}\right)de_{R,t}H_{0,t}^{-1}e_{R,t}\\
&\qquad-\frac{1}{2}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}z'_te'_{R,t}\left(\sum_{k=1}^{J}\left(H_{R,t}^{-1}\left(\frac{\partial^2\sigma_t\left(\Delta\xi_t(\theta_0,P_{0,t}),\theta_0,P_{R,t}\right)}{\partial\Delta\xi\partial\Delta\xi_k}\right)H_{R,t}^{-1}e_je'_kH_{R,t}^{-1}\right.\right.\\
&\qquad\qquad\left.\left.-H_{0,t}^{-1}\left(\frac{\partial^2\sigma_t\left(\Delta\xi_t(\theta_0,P_{0,t}),\theta_0,P_{0,t}\right)}{\partial\Delta\xi\partial\Delta\xi_k}\right)H_{0,t}^{-1}e_je'_kH_{0,t}^{-1}\right)\right)e_{R,t}\\
&\qquad+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}z'_t\sum_{\alpha\in\mathbb{R}_+:|\alpha|=3}\frac{1}{\alpha!}\left(\left.\frac{\partial^{|\alpha|}\sigma_{j,t}^{-1}\left(s,\theta_0,P_{R,t}\right)}{\partial s^{|\alpha|}}\right|_{s=c_ts_t+(1-c_t)\sigma_t\left(\Delta\xi(\theta_0,P_{0,t}),\theta_0,P_{R,t}\right)}\right)e_{R,t}^{|\alpha|}.
\end{aligned}$$

Furthermore, by Lemma A1 and Assumption B6

$$H_{R,t}^{-1} = H_{0,t}^{-1} + o_p(1)$$

where the $J \times J$ $o_p(1)$ term does not depend on t. Similarly

$$\begin{aligned}\sum_{k=1}^{J} H_{R,t}^{-1} \left(\frac{\partial^2 \sigma_t \left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t} \right)}{\partial \Delta\xi \partial \Delta\xi_k} \right) H_{R,t}^{-1} e_j e_k' H_{R,t}^{-1} \\ = \sum_{k=1}^{J} H_{0,t}^{-1} \left(\frac{\partial^2 \sigma_t \left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t} \right)}{\partial \Delta\xi \partial \Delta\xi_k} \right) H_{0,t}^{-1} e_j e_k' H_{0,t}^{-1} + o_p(1)\end{aligned}$$

where the $J \times J$ $o_p(1)$ term does not depend on t. It now follows from Assumptions B5 and B6 as well as Lemma A3 that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_t' e_j' \left(H_{R,t}^{-1} - H_{0,t}^{-1} \right) d e_{R,t} H_{0,t}^{-1} e_{R,t} = o_p \left(\frac{\sqrt{T}}{R} \right)$$

and

$$\begin{aligned}\frac{1}{2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_t' e_{R,t}' \Bigg(\sum_{k=1}^{J} \Bigg(H_{R,t}^{-1} \left(\frac{\partial^2 \sigma_t \left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{R,t} \right)}{\partial \Delta\xi \partial \Delta\xi_k} \right) H_{R,t}^{-1} e_j e_k' H_{R,t}^{-1} \\ - H_{0,t}^{-1} \left(\frac{\partial^2 \sigma_t \left(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0, P_{0,t} \right)}{\partial \Delta\xi \partial \Delta\xi_k} \right) H_{0,t}^{-1} e_j e_k' H_{0,t}^{-1} \Bigg) \Bigg) e_{R,t} = o_p \left(\frac{\sqrt{T}}{R} \right).\end{aligned}$$

Finally consider

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_t' \sum_{\alpha \in \mathbb{R}_+ : |\alpha| = 3} \frac{1}{\alpha!} \left(\left. \frac{\partial^{|\alpha|} \sigma_{j,t}^{-1} \left(s, \theta_0, P_{R,t} \right)}{\partial s^{|\alpha|}} \right|_{s = c_t s_t + (1 - c_t) \sigma_t \left(\Delta\xi(\theta_0, P_{0,t}), \theta_0, P_{R,t} \right)} \right) e_{R,t}^{|\alpha|}.$$

Similar as for the second derivative, we can write the third partial derivatives of the inverse function as a function of partial derivatives of $\sigma_t(\Delta\xi, \theta_0, P_{0,t})$ evaluated at $\widetilde{\Delta\xi_t}$ that satisfies

$$\sigma_t \left(\widetilde{\Delta\xi_t}, \theta_0, P_{R,t} \right) = \sigma_t \left(\Delta\xi(\theta_0, P_{0,t}, \theta_0, P_{R,t}) \right) + c_t \left(s_t - \sigma_t \left(\Delta\xi(\theta_0, P_{0,t}), \theta_0, P_{R,t} \right) \right)$$

But since

$$\max_{1 \leq t \leq T} \left| s_t - \sigma_t \left(\Delta\xi(\theta_0, P_{0,t}), \theta_0, P_{R,t} \right) \right| = o_p(1)$$

it follows that

$$\max_{1 \leq t \leq T} \left| s_t - \sigma_t \left(\widetilde{\Delta\xi_t}, \theta_0, P_{R,t} \right) \right| = o_p(1)$$

or

$$\max_{1 \leq t \leq T} \left| \sigma_t \left(\Delta\xi(\theta_0, P_{R,t}), \theta_0, P_{R,t} \right) - \sigma_t \left(\widetilde{\Delta\xi_t}, \theta_0, P_{R,t} \right) \right| = o_p(1).$$

This implies by Assumption A4 that

$$\max_{1\leq t\leq T}\left|\Delta\xi(\theta_0,P_{R,t})-\widetilde{\Delta\xi_t}\right|=o_p(1)$$

which in turn means that

$$\max_{1\leq t\leq T}\left|\Delta\xi(\theta_0,P_{0,t})-\widetilde{\Delta\xi_t}\right|=o_p(1).$$

Next, it easy to verify that

$$\left.\frac{\partial^{|\alpha|}\sigma_{j,t}^{-1}(s,\theta_0,P_{R,t})}{\partial s^{|\alpha|}}\right|_{s=c_ts_t+(1-c_t)\sigma_t\left(\Delta\xi(\theta_0,P_{0,t}),\theta_0,P_{R,t}\right)}$$

is a function of

$$\left(\left.\frac{\partial\sigma_t(\Delta\xi,\theta_0,P_{R,t})}{\partial\Delta\xi}\right|_{\Delta\xi_t=\widetilde{\Delta\xi_t}}\right)^{-1}$$

as well as partial derivatives of $\sigma_t(\Delta\xi,\theta_0,P_{R,t})$ up to order 3 evaluated at $\widetilde{\Delta\xi_t}$. But since

$$\max_{1\leq t\leq T}\left|\Delta\xi(\theta_0,P_{0,t})-\widetilde{\Delta\xi_t}\right|=o_p(1)$$

and since $\Delta\xi(\theta_0,P_{0,t})$ is in a compact set it follows from Lemma A1 as well Assumption B7 that

$$\left.\frac{\partial^{|\alpha|}\sigma_{t}^{-1}(s,\theta_0,P_{R,t})}{\partial s^{|\alpha|}}\right|_{s=c_ts_t+(1-c_t)\sigma_t\left(\Delta\xi(\theta_0,P_{0,t}),\theta_0,P_{R,t}\right)}=Q_{t,0,|\alpha|}+o_p(1)$$

where $Q_{t,0,|\alpha|}$ is bounded for each t and $|\alpha|$ and the $o_p(1)$ term does not depend on t. As a consequence, it follows from Lemma A3 that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}z_t'\sum_{\alpha\in\mathbb{R}_+:|\alpha|=3}\frac{1}{\alpha!}\left(\left.\frac{\partial^{|\alpha|}\sigma_{j,t}^{-1}(s,\theta_0,P_{R,t})}{\partial s^{|\alpha|}}\right|_{s=c_ts_t+(1-c_t)\sigma_t\left(\Delta\xi(\theta_0,P_{0,t}),\theta_0,P_{R,t}\right)}\right)e_{R,t}^{|\alpha|}=o_p\left(\frac{\sqrt{T}}{R}\right).$$

Next consider

$$D_T(\hat{\theta},P_R)=\frac{1}{T}\sum_{t=1}^{T}z_t'\frac{\partial\Delta\xi_t(\theta,P_{R,t})}{\partial\theta}$$

We have to prove that

$$D_T(\theta,P_R)\xrightarrow{p}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\frac{\partial G(\theta,P_{0,t})}{\partial\theta}$$

uniformly over θ. If this holds then for any consistent estimator, $\check{\theta}$, of θ, it holds that

$$D_T(\check{\theta},P_R)\xrightarrow{p}\Gamma.$$

It is sufficient to prove that

$$D_T(\theta,P_R)-\frac{1}{T}\sum_{t=1}^{T}\frac{\partial G_T(\theta,P_{0,t})}{\partial\theta}\xrightarrow{p}0$$

uniformly over θ because

$$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial G_T(\theta, P_{0,t})}{\partial\theta} - \frac{\partial G(\theta, P_{0,t})}{\partial\theta}\right) \xrightarrow{p} 0$$

uniformly over θ by Assumption B8 and the uniform law of large numbers (see for example Amemiya (1985)). Now

$$\begin{aligned}
&D_T(\theta, P_R) - \frac{1}{T}\sum_{t=1}^{T}\frac{\partial G_T(\theta, P_{0,t})}{\partial\theta} \\
&\quad= \frac{1}{T}\sum_{t=1}^{T} z_t'\left(\frac{\partial \Delta\xi_t(\theta, P_{R,t})}{\partial\theta} - \frac{\partial \Delta\xi_t(\theta, P_{0,t})}{\partial\theta}\right) \\
&\quad= -\frac{1}{T}\sum_{t=1}^{T} z_t'\left(\left(\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial\Delta\xi_t}\right)^{-1}\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial\theta}\right. \\
&\qquad\qquad\left. - \left(\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial\Delta\xi_t}\right)^{-1}\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial\theta}\right)
\end{aligned}$$

where

$$\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial\theta}$$

denotes the derivative only with respect to the second element of the function. Similar to before

$$\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial\Delta\xi_t} = \frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial\Delta\xi_t} + o_p(1)$$

in a neighborhood of θ_0 where the $o_p(1)$ term does not depend θ or t. Now by Assumption B6

$$\left(\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial\Delta\xi_t}\right)^{-1} = \left(\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial\Delta\xi_t}\right)^{-1} + o_p(1)$$

where the $o_p(1)$ term does not depend θ or t. As a consequence

$$\begin{aligned}
&D_T(\theta, P_R) - \frac{1}{T}\sum_{t=1}^{T}\frac{\partial G_T(\theta, P_{0,t})}{\partial\theta} \\
&= -\frac{1}{T}\sum_{t=1}^{T} z_t'\left(\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial\Delta\xi_t}\right)^{-1}\left(\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial\theta} - \frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial\theta}\right) \\
&\quad + o_p(1)\frac{1}{T}\sum_{t=1}^{T} z_t'\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial\theta}.
\end{aligned}$$

But Assumptions B7 and B8 imply that

$$\frac{1}{T}\sum_{t=1}^{T} z_t'\frac{\partial\sigma_t\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial\theta} = O_p(1)$$

uniformly over θ. Hence,

$$
\begin{aligned}
&D_T(\theta, P_R) - \frac{1}{T}\sum_{t=1}^{T} \frac{\partial G_T(\theta, P_{0,t})}{\partial \theta} \\
&= -\frac{1}{T}\sum_{t=1}^{T} z_t' \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial \Delta\xi_t} \right)^{-1} \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial \theta} - \frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial \theta} \right) \\
&\quad + o_p(1) \\
&= -\frac{1}{T}\sum_{t=1}^{T} z_t' \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial \Delta\xi_t} \right)^{-1} \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial \theta} - \frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{R,t}\right)}{\partial \theta} \right) \\
&\quad - \frac{1}{T}\sum_{t=1}^{T} z_t' \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial \Delta\xi_t} \right)^{-1} \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{R,t}\right)}{\partial \theta} - \frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial \theta} \right) \\
&\quad + o_p(1).
\end{aligned}
$$

But for all j it follows from Assumption B8 that

$$
\begin{aligned}
&\left| \frac{\partial \sigma_{j,t}\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial \theta'} - \frac{\partial \sigma_{j,t}\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{R,t}\right)}{\partial \theta'} \right| \\
&\qquad \leq \frac{1}{R}\sum_{r=1}^{R} \left|H(v_{r,t})\right| \left|\Delta\xi_t(\theta, P_{R,t}) - \Delta\xi_t(\theta, P_{0,t})\right| = o_p(1)
\end{aligned}
$$

uniformly over θ and t. Thus,

$$
\frac{1}{T}\sum_{t=1}^{T} z_t' \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial \Delta\xi_t} \right)^{-1} \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{R,t}), \theta, P_{R,t}\right)}{\partial \theta} - \frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{R,t}\right)}{\partial \theta} \right) = o_p(1)
$$

independent of θ and t. Finally write

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} z_t' \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial \Delta\xi_t} \right)^{-1} \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{R,t}\right)}{\partial \theta} - \frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial \theta} \right) \\
&= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{R}\sum_{r=1}^{R} z_t' \left(\frac{\partial \sigma_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial \Delta\xi_t} \right)^{-1} \\
&\qquad\qquad \times \left(\frac{\partial \nu_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, v_{r,t}\right)}{\partial \theta} - E_t^* \left(\frac{\partial \nu_t\left(\Delta\xi_t(\theta, P_{0,t}), \theta, P_{0,t}\right)}{\partial \theta} \right) \right) \\
&= o_p(1)
\end{aligned}
$$

uniformly over x_t, p_t and θ by the uniform law of large numbers. Hence

$$
\sup_{\theta \in \Theta} \left| D_T(\theta, P_R) - \frac{1}{T}\sum_{t=1}^{T} \frac{\partial G_T(\theta, P_{0,t})}{\partial \theta} \right| = o_p(1).
$$

E Proof of Theorem 3

By the proof of Theorem 2 it suffices to prove that

$$
\hat{C}_{R,t} = E_t^* \left(H_{0,t}^{-1} d\varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0) H_{0,t}^{-1} \varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0) \right) + o_p(1)
$$

and

$$\hat{S}_{R,j,t} = E_t^* \left(\varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0)' I_{0,t,j} \varepsilon_{r,t}(\Delta\xi_t(\theta_0, P_{0,t}), \theta_0) \right) + o_p(1)$$

where the $o_p(1)$ terms do not depend on t. These two results follow from identical arguments as the last part of the proof of Theorem 2.

References

Abito, J.-M. (2011). A note on weak identification in the Berry, Levinsohn and Pakes (1995) model. Working paper.

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, Massachusetts: Harvard University Press.

Andrews, D. and X. Cheng (2010). Estimation and inference with weak, semi-strong, and strong identification. Working paper.

Arellano, M. and J. Hahn (2007). Understanding bias in nonlinear panel models: Some recent developments. In R. Blundell, W. Newey, and T. Persson (Eds.), *Advances in Economics and Econometrics*, Volume 3. Cambridge University Press.

Armstrong, T. (2012). Large market asymptotics for differentiated product demand estimators with economic models of supply. Working paper.

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

Berry, S. and P. A. Haile (2010). Identification in differentiated products markets using market level data. Working paper.

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

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

Dubé, J. P., J. T. Fox, and C. L. Su (2009). Improving the numerical performance of BLP static and dynamic discrete choice random coefficients demand estimation. Working paper.

Gandhi, A. and A. Kim, K. abd Petrin (2011). Identification and estimation in discrete choice demand models when endogenous variables interact with the error. Working paper.

Hahn, J. and W. Newey (2004). Jackknife and analytical bias reduction for nonlinear panelmodels. *Econometrica 72*(4), 12951319.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimation. *Annals of Mathematical Statistics 40*, 633–643.

Judd, K. and B. Skrainka (2011). High performance quadrature rules: how numerical integration affects a popular model of product differentiation. Working paper.

Kim, D. (2004). Estimation of the effects of new brands on incumbents profits and consumer welfare: The U.S. processed cheese market case. *Review of Industrial Organization 25*, 275293.

Kristensen, D. and B. Salanie (2010). Higher order improvements for approximate estimators. Working paper.

Lee, L. (1995). Asymptotic bias in simulated maximum likelihood estimation of discrete choice models. *Econometric Theory 11*(3), 437–483.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica 57*(5), 995–1026.

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

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*(5), 1027–1057.

Romeo, C. J. (2010). Filling out the instrument set in mixed logit demand systems for aggregate data. Working paper.

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

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Villas-Boas, S. (2007). Vertical relationships between manufacturers and retailers: Inference with limited data. *Review of Economic Studies 74*, 625652.

Xu, H. (2010). Uniform exponential convergence of sample average random functions under general sampling with applications in stochastic programming. *Journal of Mathematical Analysis and Applications 368*, 692–710.