Efficient Estimation of Random Coefficients Demand Models using Product and Consumer Datasets*

Paul L. E. Grieco† Charles Murry‡ Joris Pinkse§ Stephan Sagl¶

May 26, 2021

Abstract

We propose a mixed-data likelihood estimator (MDLE) for a mixed logit demand system which makes use of product-level and consumer-level data while allowing for price endogeneity. The estimator is efficient compared to the GMM approach commonly used by applied researchers (e.g. Petrin, 2002; Berry, Levinsohn and Pakes, 2004). We show how to conduct inference on general functions of the model parameters, including elasticities. We further extend our approach to efficiently incorporate product-level exclusion restrictions commonly used to identify consumer heterogeneity when only product-level data is available (e.g., Berry, Levinsohn and Pakes, 1995; Nevo, 2001; Gandhi and Houde, 2020). These additional restrictions can improve precision when consumer-level data only weakly identifies consumer heterogeneity. We benchmark our likelihood-based estimator to the GMM approach with a Monte Carlo exercise and find superior performance in finite samples.

1 Motivation

First introduced in Berry, Levinsohn and Pakes (1995) (henceforth BLP), random coefficients demand models provide a tractable framework within which to flexibly estimate substitution patterns between many goods while controlling for endogenously determined prices. We propose a likelihood estimator for BLP-style models that incorporates both product- and consumer-level data. Intuitively, it combines the likelihoods of two mixed-logit estimators, one for consumer level-data, and one for product level-data. This estimator imposes assumptions that are identical to or weaker than GMM-based "micro-BLP" estimators (e.g., Petrin, 2002; Berry, Levinsohn and Pakes, 2004). It makes efficient use of all variation in both data sets to estimate substitution patterns and product qualities and avoids the need for the researcher to select and weight moments between datasets.

*We thank seminar participants at Université de Montréal and the University of Arizona for helpful suggestions. Please check http://joris.pinkse.org/research/grumps.pdf for the latest version.

†Department of Economics, The Pennsylvania State University, paul.grieco@psu.edu.

‡Department of Economics, Boston College, charles.murry@bc.edu.

§Department of Economics, The Pennsylvania State University, joris@psu.edu.

¶Department of Economics, The Pennsylvania State University, stephan.sagl@psu.edu.

The basic structure of the demand model proposed in BLP is mixed (or random coefficients) multinomial logit. The standard multinomial logit MLE has nice properties. For example, it is globally concave in the parameters, the gradient and Hessian have easy to compute expressions, and convergence can be fast. Therefore, with consumer-level data in hand, it is natural to consider estimating a BLP model via MLE, using the individual likelihood of purchase. However, in order to accommodate price endogeneity, the basic structure of BLP requires the estimation of product fixed effects.[1] These would be demanding to estimate using consumer-level data alone because of the presence of potentially many (hundreds, or even thousands, depending on the application) fixed effects parameters.

To address this issue, we incorporate product-level data on aggregate shares. We view our consumer-level sample as a subset of the population of individual choices represented by the observed aggregate shares. From this perspective, the loglikelihood of both individual consumer data ("micro" data) *and* aggregate product shares ("macro" data) consists of two terms: A "micro" loglikelihood following the mixed logit and a "macro" loglikelihood that simply integrates over consumer characteristics which are observed only in the "micro" data. We use this to estimate three types of parameters: (1) unobserved preference heterogeneity (often referred to as "random coefficients" in the literature); (2) observed preference heterogeneity based on individual demographics (referred to as "demographic interactions"); and (3) product-specific quality. An auxiliary regression of product-specific qualities can then be used to address product-level endogeneity. We further show how this auxiliary regression can be incorporated into a one stage estimation procedure by adding a penalization term to the standard likelihood. As we discuss below, this penalized approach is particularly useful in the presence of over-identifying restrictions.

Our estimator is most directly comparable to GMM approaches based on micro-moments (e.g., Petrin, 2002; Berry, Levinsohn and Pakes, 2004). Beyond the efficiency benefit from using the likelihood of consumer level data, our estimator has the second advantage that it does not impose that observed market shares exactly equal market-level unconditional choice probabilities of products. To be precise, the share inversion constraints of BLP appear within the first order conditions of the "macro" likelihood of our estimator, but are not forced to hold exactly because we optimize a sum of the "micro" and "macro" loglikelihoods. The additional efficiency gain can be modest when the size S of the consumer-level data set is small relative to the size $\min_m N_m$ of the data set producing the shares in the smallest market and the former data set consists of random draws without selection from the latter data set. However, the gain can be significant in situations where market shares are small and the demographic interactions are important drivers of consumer choices.

We believe our approach has other practical advantages as well. Inference is done via the standard likelihood framework and is asymptotically valid regardless of the object of

[1] Berry, Levinsohn and Pakes (2004) first noted that product fixed effects could be used to separate the estimation of "nonlinear" parameters that govern substitution patterns from the "linear" parameters of the model such as the mean price effects.

interest provided that the number of products J is negligibly small compared to S and the number of products *in each market* is negligibly small compared to the square root of the total number of consumers in that market.[2] By comparison, methods that impose share constraints (including micro BLP) require in addition that S is negligibly small compared to N and, if the product-level fixed effects are of interest, even that S is negligibly small compared to $\sqrt{N}$.[3] Absent those additional restrictions, the computed standard errors would be too small: we provide an extreme example in which they are (asymptotically) off by a factor infinity. Additionally, the ingredients needed for inference with our procedure are routinely computed as part of the optimization procedure.[4] Computationally, our approach avoids the need to employ either a nested fixed point algorithm or constrained optimization.[5]

Our baseline estimator uses product-level exclusion restrictions only to estimate mean tastes for product characteristics. While consumer data alone is able to identify consumer heterogeneity in BLP-style models, identification may be weak in practice.[6] Berry and Haile (2014) establishes formal conditions for nonparametric identification of discrete choice demand using product-level exclusion restrictions by utilizing appropriate product-level instruments.[7] However, using product-level moments to identify consumer heterogeneity will slow the rate of convergence to $\sqrt{J}$, whereas estimating heterogeneity with consumer data yields a convergence rate of $\sqrt{S}$. We extend our likelihood estimator to incorporate product-level exclusion restrictions via a penalization term. The resulting penalized likelihood estimator is adaptive: the penalty term aids estimation when consumer heterogeneity parameters are weakly (or even not) identified by the micro data, but is asymptotically negligible when they are strongly identified in order to preserve the $\sqrt{S}$ convergence rate. To our knowledge, this adaptive feature is novel in the literature.

To summarize, there are five key advantages to our likelihood based approach compared with alternatives. First, our estimator appears to retain nice properties associated with multinomial logit. In our Monte Carlo experiments, our estimator is fast and outperforms a moments-based estimator in terms of statistical properties. Second, in our framework, the researcher does not need to select a finite set of moments from the data to match in a GMM procedure. Instead all of the information contained in the likelihood is used to inform the parameters. Typically, GMM is favored over MLE because MLE requires the likelihood to be

[2] The need for the latter requirement is demonstrated in Berry, Linton and Pakes (2004).

[3] In Berry, Levinsohn and Pakes (1995, 2004) N is assumed to be effectively infinite.

[4] Assuming one uses a Newton-based approach.

[5] While our estimator is the solution to a single unconstrained optimization problem, it may still be convenient to use a profile the likelihood to take advantage of concavity in some parameters, as we detail below.

[6] Specifically, identification requires variation in utility across consumers within the same market based on observable demographics. Weak identification will result when this variation is small. Recently, Berry and Haile (2020) have confirmed the practical intuition of this approach by showing how the use of consumer-level data relaxes the requirement for instruments in a nonparametric BLP setting.

[7] Of course, finding such instruments to precisely estimate even parameterized versions of the model has proved difficult in many applications. See Gandhi and Houde (2020) for a discussion of weak instruments in the BLP context and suggestions for strong instruments.

fully specified, but in the our context, the traditional micro BLP GMM procedure actually uses all of the structure of the likelihood, so we do not impose any more structure on the problem than the commonly used GMM procedure, yet our procedure is more efficient than the dominant GMM procedure. Third, our estimator does not impose the share constraint inversion common to BLP and micro BLP estimators. This inversion prioritizes the "macro" data over the micro data when estimating product qualities resulting in a loss of efficiency. Fourth, product-level information can be introduced to buttress identification of consumer heterogeneity without the risk of slowing down the convergence rate. Finally, and most importantly, our procedure is both efficient and provides valid inference under the widest range of scenarios using standard methodology and without needing adjustments to handle special cases.

The following section reviews the random coefficients demand model and the data available in our setting. Section 3 proposes our baseline mixed data likelihood (MDLE) estimator. Section 4 describes inference for MDLE on general functions of model parameters, which include elasticities or other objects of economic interest. Section 5 compares our MDLE with alternatives and highlights the efficiency and inference implications of two distinguishing features of our estimator: avoiding imposing a constraint on aggregate shares and the use of the full micro-likelihood. Section 6 extends MDLE to incorporate product level moments in an efficient manner via a penalty term. Section 7 presents a Monte Carlo exercise to compare MDLE to alternatives in a finite samples. Finally, section 8 concludes.

2 Random Coefficients Demand Model

In this section, we briefly review the random coefficients discrete choice demand model and describe the data used by our estimator. The model matches that of Berry, Levinsohn and Pakes (1995) with slightly adjusted notation for clarity. We will assume the researcher has access to both product-level shares and a sample consumer level choices. Importantly, our estimator will assume that consumer-level choices represent a subset of consumers on which the market-level shares are based. This is in slight contrast to the previous literature, which has treated micro and macro data as different samples.

2.1 Model

The econometrician observes M markets. In each market m, J_m products are available for purchase. A product j in market m is described by the tuple $(x_{jm}, \xi_{jm}, p_{jm})$, where x_{jm} is a d_x-vector of exogenous observed characteristics of the product, ξ_{jm} is an unobserved product attribute, and p_{jm} is an endogenous characteristic (typically price) which is potentially correlated with ξ_{jm}. Consumers in market m are characterized by $(z_{im}, \nu_{im}, \varepsilon_{im})$ where $z_{i\cdot m}$ is a d_z-vector of potentially observable consumer characteristics (such as income or location), and ν_{im} is a $(d_x + 1)$-vector of unobservable consumer taste shocks to preferences for product

characteristics (including price).[8] Finally ε_{im} is a $J_m + 1$-vector of idiosyncratic product taste shocks for each product and an outside good (e.g., no purchase), which we assume is distributed according to the standard Type-I extreme value (Gumbel) distribution. In the population, both z_{im} and ν_{im} are mutually independent and distributed according to known distributions G_m and F_m, respectively. In practice, the distribution of z_{im} is typically taken from external data (such as the population census) while the distribution of ν_{im} is typically assumed to be a standard normal and independent across components of ν_{im}.

A consumer in market m maximizes (indirect) utility by choosing from the J_m available products and the outside good, indexed by zero. Utility of consumer i when purchasing product j in market m is,

$$u_{ijm} = \delta_j + \mu_{jm}^{z_{im}} + \mu_{jm}^{\nu_{im}} + \varepsilon_{ijm}, \tag{1}$$

where,[9]

$$\delta_{jm} = x'_{jm}\beta - \alpha p_{jm} + \xi_{jm}, \tag{2}$$

represents the vertical utility for product j while,

$$\mu_{jm}^{z_{im}} = \mu^z(x_{jm}, p_{jm}, z_{im}; \theta^z) \tag{3}$$

represents deviations from mean utility due to observed demographic variables $z_{i\cdot m}$ and,

$$\mu_{jm}^{\nu_{im}} = \mu^\nu(x_{jm}, p_{jm}, \nu_{im}; \theta^\nu) = \sum_{k=1}^{d_z+1} \sigma^k x_{jmk}\nu_{imk} + \sigma^\alpha p_{jm}\nu_{im\alpha}, \tag{4}$$

are deviations due to taste shocks ν_i and $\theta^\nu = (\{\sigma^k\}_{k=1}^K, \sigma^\alpha)$.[10] Utility of the outside good is normalized to $u_{i0} = \varepsilon_{i0}$. When convenient, we collect the consumer heterogeneity parameters into the vector $\theta = (\theta^z, \theta^\nu)$.

2.2 Data

The researcher has access to two types of data on consumer choices. First, market-level data is derived from all consumers in the marketplace, but with no information about consumer characteristics. That is, the researcher observes market shares,

$$s_{jm} = \frac{1}{N_m} \sum_{i=1}^{N_m} \mathbb{1}_{y_{ijm}=1}, \tag{5}$$

[8] For notational simplicity we put random coefficients on all product-level characteristics: this is neither necessary nor generally advisable.

[9] There is no real need to assume δ_j to have this linear form but this is the most common specification.

[10] Allowing more generality in μ^z, such as correlation between taste shocks, is conceptually straightforward.

where $y_{i\cdot m}$ is a $(J_m + 1)$-vector with $y_{ijm} = 1$ if consumer i purchased product j and 0 otherwise. In our setup, N_m will correspond to the total number of consumers in market m, i.e. the market size. Note that the observed market shares s_m need not equal choice probabilities π_m due to the fixed population size and unobserved consumer heterogeneity, however $s_m \xrightarrow{p} \pi_m$ as $N_m \to \infty$.

Second, for S_m out of N_m consumers, the researcher observes both the consumer's choice and their demographic characteristics. That is, the researcher observes $\{(y_{i\cdot m}, z_{im})\}$ for these consumers. We use D_{im} as a dummy variable to denote whether consumer i is in this micro-sample.[11]

3 Mixed Data Likelihood Estimator

As noted previously in the literature, we can estimate the model in two steps. First, we can estimate a more general model by imposing (1), (3), and (4), but not (2) and leaving δ as a vector of product fixed effects to estimate. Since δ_j absorbs the unobserved product quality shock, ξ_j, price endogeneity is not a concern in this step. Moreover, this model is fully parametric, enabling the use of a likelihood function to estimate (δ, θ) (cf. Berry, Levinsohn and Pakes, 2004). Once these parameters are recovered, it is straightforward to estimate (β, α) with a second-step linear IV regression of (2) with the first stage estimates of δ on the left-hand-side and x, p on the right hand side, using instruments to address endogeneity.

Notably, our approach in this section does *not* use instruments to identify the heterogeneity parameters, θ, which is different than nearly every similar study. Instead, we rely on within-market consumer variation. Specifically, we are able to vary $z_{i\cdot m}$ within a market m, thus holding product qualities δ fixed. This is analogous to the variation Berry and Haile (2020) used to show nonparametric identification of the share function in the context of a model similar to ours. We will extend our approach to incorporate product-level instruments in Section 6.

3.1 Maximum Likelihood

We now turn to the maximum likelihood estimator for $\psi = (\theta, \delta)$. The model yields choice probabilities for each consumer of selecting product j conditional on consumer characteristics

[11] We can extend our approach to accommodate consumers on whom we have micro-level data can be selected based on their purchase decisions or on their other observable characteristics For example, it may be a survey of consumers who purchase from a particular firm, or a survey of consumers who purchase a product, but excluding those consumers who purchase the outside good. Selection may require minor adjustments to the likelihood. We also note that we can extend our model to incorporate other information about individual consumers, such as so-called second-choice data.

$z_{i\cdot m}$ as a function of parameters,[12]

$$\pi_{jm}^{z_{im}}(\psi) = \Pr(y_{ijm} = 1 \mid z_{i\cdot m}; \psi) = \int \underbrace{\frac{\exp(\delta_{jm} + \mu_{ijm}^{z} + \mu_{ijm}^{\nu})}{\sum_{g=0}^{J_m} \exp(\delta_{gm} + \mu_{igm}^{z} + \mu_{igm}^{\nu})}}_{s_{jm}(z_{im}, \nu; \psi)} \, \mathrm{d}F_m(\nu), \tag{6}$$

where $\delta_{0m} = \mu_{0m}^{z_{im}} = \mu_{0m}^{\nu_{im}} = 0$ for all i, m. The probabilities π_{jm}^{z} would form the basis of the mixed-logit likelihood if we only observed a random sample of consumer-level data

First, consider the S_m consumer-level observations. The full contribution to the log-likelihood accounts for selection into the sample, $\log \Pr(D_{im} = 1 \cap y_{ijm} = 1 | z_{im})$. If selection into the consumer sample is random, then the log-likelihood contribution is just $\log \pi_{jm}^{z_{im}}(\psi)$ (plus an irrelevant constant). If D_{im} depends only on z_{im}, then $\Pr(D_{im} = 1 \cap y_{ijm} = 1 \mid z_{im}) = \Pr(D_{im} = 1 \mid z_{i\cdot m}) \pi_{jm}^{z_{im}}(\psi)$. In this case the selection term can be ignored since it will be independent of ψ in the loglikelihood and merely contribute an additive constant. Hence, the log-likelihood contribution is again $\log \pi_{jm}^{z_{im}}(\psi)$. A more general case would arise if selection also depends on consumers' choices, $\{y_{i\cdot m}\}$. If this form of selection is deterministic then again the loglikelihood contribution is straightforward. This would occur, for example, if consumer-level data is obtained from an administrative source which records all transactions, but does not capture consumers who opt for the outside option. In this case, if there is also selection into the sample based on z_i according to D_{im}^{*}, we have $D_{im} = D_{im}^{*} \mathbb{1}(y_{i0m} = 0)$ and

$$\Pr(D_{im} = 1 \cap y_{ijm} = 1 \mid z_{im}) = \begin{cases} 0 & j = 0 \\ \Pr(D_{im}^{*} = 1 | z_{im}) \pi_{jm}^{z_{im}} & j > 0 \end{cases}.$$

Again, the loglikelihood contribution simplifies to $\log \pi_{jm}^{z_{im}}$ following the same argument as above. If selection depends on $y_{i\cdot m}$ in a stochastic way, it must be directly modeled. However, since this case does not appear commonly in applications, we do not consider it in this paper.

Now, consider the $N_m - S_m$ consumers *not* in the consumer-level sample, we effectively observe the consumer choice, but not the consumer's characteristics. Therefore, choice probabilities must integrate over the distribution of characteristics, conditioning on those consumers not appearing in the sample,

$$\pi_{jm}^{D=0}(\psi) = \int \Pr(y_{ijm} = 1 \cap D_{im} = 0 \mid z_{im} = z) \, \mathrm{d}G_m(z).$$

In the case of selection on the z_{im}'s only, the formula for $\pi_{jm}^{D=0}(\psi)$ simplifies to $\int \Pr(D_{im} = 0 | z_i = z) \pi_{jm}^{z}(\psi) \, \mathrm{d}G_m(z) = \Pr(D_{im} = 0) \int \pi_{jm}^{z}(\psi) \, \mathrm{d}G_m^{D=0}(z)$ with $G_m^{D=0}(z)$ the distribution of $z_{i\cdot m}$ in market m but *not* in the micro sample.The example with $D_{im} = D_{im}^{*} \mathbb{1}(y_{i0m} = 0)$ is

[12] For expositional purchases, we suppress conditioning on the choice set $x_{\cdot m}$ throughout this section.

more involved, but can still be calculated from the data.[13] $G_m^{D=0}(z)$ and its complement $G_m^{D=1}(z)$ are easy to compute from the consumer-level data and the known distribution of $z_{i\cdot m}$ in the population, $G_m(z)$. In general, $\pi_{jm}^{D=0}(\psi)$ will differ from the unconditional choice probability,

$$\pi_{jm}(\psi) = \Pr(y_{ijm} = 1) = \int \pi_{jm}^{z}(\psi)\, \mathrm{d}G_m(z).$$

which can be estimated by market shares defined by (5). Although the difference will be small when N_m is large relative to S_m and the consumer sample is selected at random.

We now have all the components to construct the likelihood of the observed data. First, *suppose* that we observed $\{y_{ijm}\}$ for all N_m observations. Then the loglikelihood would be,

$$\mathrm{LL}(\psi) = \sum_{m=1}^{M} \sum_{j=0}^{J_m} \sum_{i=1}^{N_m} y_{ijm} \left(D_{im} \log \pi_{jm}^{z_{im}}(\psi) + (1 - D_{im}) \log \pi_{jm}^{D=0}(\psi) \right), \tag{7}$$

The loglikelihood sums over all N_m consumers in the market. If an observation is in the micro data then we see $z_{i\cdot m}$ and can condition on it, whereas otherwise we integrate over the distribution of $z_{i\cdot m}$ conditional on this consumer not being in the consumer sample.

Of course, we do not directly observe the choices of consumers who are not in the micro sample. However, the loglikelihood function can be equivalently written in terms of the consumer-level observations and the market-level share data,

$$\mathrm{LL}(\psi) = \underbrace{\sum_{m=1}^{M} \sum_{j=0}^{J_m} \sum_{i=1}^{N_m} D_{im} y_{ijm} \log \frac{\pi_{jm}^{z_{im}}(\psi)}{\pi_{jm}^{D=1}(\psi)}}_{\text{micro}} + \underbrace{\sum_{m=1}^{M} N_m \sum_{j=0}^{J_m} s_{jm} \log \pi_{jm}(\psi)}_{\text{macro}}, \tag{8}$$

where the first term is the contribution of the consumer-level data and the second term is the contribution of the market-level data. In order to express the market-level term using observed market shares, we add and subtract $\log \pi_{jm}^{D=1}$ to control for the fact that the consumer-level data represent a subset of the consumers who make up the market.

One notable feature of the maximum likelihood estimator is that it does not explicitly impose the share inversion constraint, $s_{jm} = \pi_{jm}$ common to BLP estimators. This constraint would be implied by the macro portion of the likelihood in the limit as $N_m \to \infty$. Berry, Levinsohn and Pakes (1995) explicitly assume shares are calculated from a continuum of consumers and the assumption is implicit in much of the literature. As we argue in section 4, imposing this constraint can be inefficient when consumer-level data is available. The inefficiency of imposing the share constraint has been noted previously (Train, 2003; Berry, Levinsohn and Pakes, 2004). Perhaps more importantly, assuming that $N_m = \infty$ can have important implications for inference on δ and other statistics of interest, as we discuss below.

[13] For $j > 0$, we get $\Pr(y_{ijm} = 1 \cap D_{im} = 0 \mid z_{im} = z) = \Pr(y_{ijm} = 1 \mid z_{im} = z) - \Pr(y_{ijm} = 1 \cap D_{im} = 1 \mid z_{im} = z) = \Pr(y_{ijm} = 1 \mid z_{im} = z) \Pr(D_{im}^{*} = 0 \mid z_{im} = z)$. Further, $\Pr(y_{i0m} = 1 \cap D_{im} = 0 \mid z_{im} = z) = \Pr(y_{i0m} = 1 \mid z_{im} = z)$. Thus, the complication here is that $G_m^{D=0}$ is replaced with a different $G_m^{D=0}$ depending on j.

The likelihood estimator recalls two common estimators in the discrete choice literature. When $N_m = S_m$, so that all consumers' characteristics are observed, we have the well known mixed-logit likelihood. Therefore, identification of ψ follows from the arguments for identification in the mixed-logit setting (Walker, Ben-Akiva and Bolduc, 2007). On the other hand, when $S_m = 0$, so only aggregate data is available, the likelihood is equivalent to imposing the share constraint from BLP and related estimators. This leads to a second insight: without consumer-level data, ψ would not be identified as there are more parameters than share constraints. The same would be true if $\theta^z = 0$, as we discuss below.[14]

Identification could be restored with additional exclusion restrictions along the lines of those used in Berry, Levinsohn and Pakes (1995)—e.g., $E(\xi_{jm} \mid v_{jm}) = 0$ for some product-level instruments v_{jm} (separate from x_{jm} and the instrument for price) which necessitates the number of products, $\sum_m J_m$, goes to infinity at a slower rate than the size of the sample of consumer-level data, $\sum_m S_m$. We find below that imposing these restrictions does not improve the asymptotic efficiency of our estimates and in fact results in a slower rate of convergence. However, if $\theta^z = 0$, such that the model is not identified, these restrictions may have identifying power. Analogously to the case of weak instruments, if θ^z is close zero, then imposing these restrictions may improve precision in small samples. In extend our procedure to accommodate this possibility in section 6. However, for expositional clarity, we first analyze our estimator without these additional assumptions.

While $\psi = (\theta, \delta)$ is a high-dimensional object, optimization over $\text{LL}(\cdot)$ is computationally tractable. First, for fixed θ^ν, the log-likelihood is concave in δ. Hence, one could compute the estimator using a globally convergent inner loop with an outer loop of *smaller* dimension than a traditional BLP estimator. In contrast, the BLP estimator uses a globally convergent contraction mapping to solve for δ in an inner loop and searches over (θ^z, θ^ν) in an outer loop. Our estimator could be computed by solving a globally concave optimization problem for (δ, θ^z) in an inner loop and search over θ^ν in an outer loop. Moreover, especially in cases in which there are many markets, it can be convenient to optimize over δ in each market in an inner loop before maximizing over θ^z since δ is market–specific but θ^z is common across all markets.

Finally, because it is the maximum likelihood estimator, our approach is efficient and makes full use of the observed data. In contrast to the traditional GMM estimator, there is no need to choose which moments of the data to include in the objective function, nor to determine the weighting between moments. We compare the efficiency gain of our estimator versus alternatives in Section 5.

[14] In this case, identification of θ^ν and δ would be possible if consumers second-choice of products were observed. It is straightforward to extend our likelihood to accommodate second choices and we would expect such data to dramatically improve the precision of estimation.

3.2 Estimation of mean product-level coefficients

The above–described estimation procedure yields estimates of ψ, but not of (β, α), since these are absorbed by the product fixed effects, δ. Since the only distinction between β and α is that they multiply exogenous and endogenous regressors, respectively, we now for ease of exposition will combine both prices and exogenous characteristics into X and from here on let β be the corresponding vector of coefficients, combining β and α. Let P_v be a projection matrix based on the instruments, such that

$$\hat{\beta} = \Xi\hat{\delta}, \tag{9}$$

with $\Xi = (X^\intercal P_v X)^{-1} X^\intercal P_v$.

Because δ's in different markets are estimated with different precision, there is a case to be made to give different weights to different products.

For consistent estimation of β we need the number of products to increase, but having an increasing number of products is not necessary and can be a nuisance for studying the properties of estimators of ψ. Such issues are discussed in Section 4.

4 Inference on functions of model parameters

This section describes inference on functions of model parameters, including elasticities and counterfactuals. A formal theorem of this result that also covers the case of product–level moments can be found in section 6. We begin with the fully general inference formula, and then consider several examples. Specifically, we consider statistics of the form $\hat{\phi} = \phi(\hat{\psi}, \hat{\beta})$ where ϕ is of fixed dimension and totally differentiable and $\hat{\psi} = [\hat{\theta}^\intercal \quad \hat{\delta}^\intercal]^\intercal$ is our estimator of ψ. For ease of exposition, we initially focus on the single market case and drop market subscripts.

The following expression has a limiting standard normal distribution. It is simply the analogy to the MLE information matrix, but explicitly incorporating the delta method becuase $\hat{\beta}$ (esimated in a second stage) is a linear function of $\hat{\delta}$. We have that

$$\underbrace{\left(\partial_{\beta^\intercal}\hat{\phi}\Xi\,\mathrm{diag}(\hat{\xi})^2\Xi^\intercal\partial_\beta\hat{\phi}^\intercal + \Lambda\hat{\mathcal{V}}_\psi\Lambda^\intercal/S\right)}_{\Upsilon}{}^{-1/2}(\hat{\phi} - \phi) \tag{10}$$

has a limiting multivariate standard normal distribution, where

$$\begin{aligned}
\hat{\xi} &= \hat{\delta} - X\hat{\beta}, \\
\hat{\mathcal{V}}_\psi &= -\{\partial_{\psi\psi^\intercal}\,\mathrm{LL}(\hat{\psi})/S\}^{-1}, \\
\Lambda &= [\partial_{\theta^\intercal}\hat{\phi} \quad \partial_{\beta^\intercal}\hat{\phi}\Xi + \partial_{\delta^\intercal}\hat{\phi}], \\
\partial_{\beta^\intercal}\hat{\phi} &= \partial_{\beta^\intercal}\phi(\hat{\psi}, \hat{\beta}).
\end{aligned} \tag{11}$$

Note that $\hat{\xi}$ are just residuals in the second stage regression, $\hat{\mathcal{V}}_\psi$ is the inverse of the information matrix for ψ, and the formulas for Λ and $\partial_{\beta^\intercal}\hat{\phi}$ are derivatives associated with the delta method. The constituent terms that make up the Hessian of the loglikelihood are listed in appendix A.

These derivations would be entirely standard if J, the dimension of δ, were fixed, but to consistently estimate β, it must be that $J \to \infty$. Throughout we assume that the number of products J is small relative to the number of consumers in the population N. Indeed, we need $J_m^2/N_m \to 0$ as $N_m \to \infty$ in each market, analogous to Berry, Linton and Pakes (2004). Note that our methodology accommodates both the possibility that S diverges more slowly than J and that it diverges as fast as N, though (obviously) not simultaneously.

The convergence rate of $\hat{\phi}$ depends on whether ϕ is a function of β. If $\partial_\beta \phi = 0$ then the first term in Υ in (10) drops out and the convergence rate is then $\sqrt{S}$. Otherwise, the convergence rate is $\sqrt{\min(J,S)}$. In contrast, the mixed logit estimator of $\phi(\psi)$ would require $J/S \to 0$ for consistency and would generally converge at a slower rate. Typically, we would expect the number of products to be small relative to the number of surveyed consumers, but that is not always the case. For example, Berry, Levinsohn and Pakes (2004) have $S = 37,500$ and $J = 203$, Grieco, Murry, and Yurukoglu (2021) have S on the order of thousands for consumer demographics and J on the order of 200 per year for 40 years, Gowrisankaran and Rysman (2012) have S and J roughly 400,000 and 4,000, and Wollmann (2018) has S on the order of 130,000 and J at 70 per year. This discussion illustrates the advantages of combining consumer level and product level data.

We now turn to several examples for ϕ. Perhaps the easiest choice for $\hat{\phi}$ is $\hat{\theta}$, in which case Υ simplifies to the the top left block of $\hat{\mathcal{V}}_\psi/S$. In that case (10) reduces to

$$\left(\hat{\mathcal{V}}_\theta/S\right)^{-1/2}(\hat{\theta} - \theta), \tag{12}$$

which is a standard result in maximum likelihood estimation.

The only substantive difference between θ and δ is that the dimension of δ increases with J. If J is fixed, or if we focus on a finite–dimensional subvector of δ, then inference is analogous to that for θ. Similarly, we can conduct inference for a finite–dimensional subvector of the vector of choice probabilities π. This is notable in part because many applications treat choice probabilities as known and equal to product shares. In our approach, choice probabilities are functions of ψ and their asymptotic behavior is hence determined by that of $\hat{\psi}$. Therefore, Υ is of the form $\Lambda\hat{\mathcal{V}}_\psi\Lambda^\intercal/S$. In the case in which $\phi = \pi_j$, we have $\Lambda = \partial_{\psi^\intercal}\pi_j$.

We now turn to inference on the coefficients β in the second stage regression. If $\phi = \beta$ then Υ simplifies to $\Xi\{\text{diag}(\hat{\xi})^2 + \hat{\mathcal{V}}_\delta/S\}\Xi$, where $\hat{\mathcal{V}}_\delta$ is the bottom right hand block of $\hat{\mathcal{V}}_\psi$. Consistency of $\hat{\beta}$ requires $J \to \infty$ so the dimensions of $\hat{\mathcal{V}}_\delta$ grow with J.[15] Indeed, as noted

[15]The form of Υ implicitly assumes that the ξ_j's are independent across j. If there is (weak) dependence then the diagonal matrix in the definition of Υ can be replaced with a Heteroskedasticity and Autocorrelation–Consistent (covariance matrix estimator).

earlier, the condition for (10) to be correct if $\phi = \beta$ is that $J_m^2/N_m \to 0$ in every market m. For the same total number of products, this condition is easier to satisfy if the number of markets M is large than if it is small.[16]

Finally, (10) can be used for more general function of the parameters, $\hat{\phi} = \phi(\hat{\psi}, \hat{\beta})$. One common case is the matrix of own and cross-price elasticities.[17] If μ^z in (3) is linear such that $\partial_{p_j}\mu_j^z$ can be expressed as $z^\intercal\theta_p^z$ then

$$\hat{\phi} = \hat{\eta}_{jk} = -\frac{p_k}{\pi_j(\hat{\psi})}\int\int \alpha(z,\nu;\hat{\psi},\hat{\beta})\imath_j(z,\nu;\hat{\psi})\{\mathbb{1}_{j=k} - \imath_k(z,\nu;\hat{\psi})\}\,\mathrm{d}F(\nu)\,\mathrm{d}G(z), \tag{13}$$

where the consumer–specific price sensitivity $\alpha(z,\nu;\psi,\beta) = \alpha - z^\intercal\theta_p^z - \sigma^\alpha\nu_\alpha$. Since $\hat{\beta}$ converges slowest, its contribution will dominate the asymptotics for $\hat{\eta}_{jk}$.

5 Comparison with Alternative Estimators

Our estimation method combines features of the mixed logit and micro BLP estimators. The primary difference with mixed logit is that we incorporate aggregate data. Indeed, if one had consumer–level data on the entire population then our estimator would coincide with the mixed logit estimator. Therefore we focus on the contrast between our estimator and micro BLP. There are two main differences: we use the macro likelihood contribution in lieu of the BLP–style share constraint and we use the micro maximum likelihood in lieu of micro moments.

5.1 Share constraint

Berry (1994) provided the insight that there is a one–to–one mapping between unconditional choice probabilities and δ's. In order to exploit this mapping, it has been commonly assumed that aggregate market shares are derived from a continuum of consumer choices, which is justified if $N \to \infty$ much faster than S, J such that observed market shares differ negligibly from unconditional choice probabilities. Consequently, the asymptotic variance of the shares is then zero by assumption such that δ's can then be treated as a known function of θ.

This assumption, while convenient, does have a cost in terms of both efficiency and inference. While the efficiency loss is asymptotically negligible if $S/N \to 0$, more importantly standard micro BLP inference requires that $S/\sqrt{N} \to 0$, as we show further below.[18] If there are multiple markets then the requirements are that respectively $S/\min_m N_m \to 0$ and

[16] Indeed, suppose that $\sum_{m=1}^M J_m = 1{,}000$ and $\sum_{m=1}^M N_m = 1{,}000{,}000$. With one market, $J_m^2/N_m = 1$ whereas with 1,000 equal–sized markets $J_m^2/N_m = 0.001$.

[17] Although we do not discuss supply-side assumptions in detail, own-price elasticities coupled with other assumptions (static simultaneous Nash Equilibrium in prices with single product firms) are sufficient to recover marginal costs and markups.

[18] We provide a microBLP standard errors formula that is correct as long as $S/N \to 0$ in Appendix B.

$S/\min_m \sqrt{N_m} \to 0$. With our estimator neither of these problems arises, while maintaining computational convenience.

5.1.1 Efficiency

We first discuss the efficiency issue. To see how our estimator improves efficiency, consider the simple case below in which we are only trying to estimate choice probabilities.

Example 1. Consider the case of one inside good and one outside good *without* random coefficients, but with possible selection on consumer characteristic z_i, i.e. the utility of the inside and outside goods is respectively $\delta + z_i\theta + \epsilon_{i1}$ and ϵ_{i0}, such that $\pi_1^z = \pi_1^z(\psi) = \Pr(y_i = 1 \mid z_i = z) = \exp(\delta + z\theta)/\{1 + \exp(\delta + z\theta)\}$. Selection on z_i produces a selection probability $\chi(z) = \Pr(D_i = 1 \mid z_i = z)$.

Consider the problem of estimating the logarithm of the choice probability $\pi^* = \Pr(y_i = 1) = \int \pi_1^z \, \mathrm{d}G(z)$ if π^* is close to zero. Using the share constraint equality this produces an (asymptotic) variance equal to $1/\{\pi^*(1-\pi^*)\}$, which goes to infinity as $\pi^* \to 0$.

Our estimator of π^* is $\int \pi_1^z(\hat{\psi}) \, \mathrm{d}G(z)$, where

$$\hat{\psi} = \arg\max_{\psi} \sum_{i=1}^{N} \sum_{j=0}^{1} y_{ij} \left(D_i \log \pi_j^{z_i}(\psi) + (1 - D_i) \log \int \pi_j^z(\psi) \, \mathrm{d}G(z) \right).$$

Our estimator makes use of the consumer–level data to exploit the parametric assumptions on π^z. Consequently, the variance of our estimator of π^* is less. The efficiency gain is increasing in the correlation between $\chi(z_i)$ and $\pi_1^{z_i}$, basically if the consumer–level sample is weighted towards purchasers. But even if $\chi = \chi(z) > 0$ is flat in z is our estimator more accurate.

To illustrate, suppose that z_i is binary with $0 < \Pr(z_i = 1) = p < 1$ and χ does not vary with z. Suppose further that $\delta = -\theta/2$ such that $\pi_1^1 = 1 - \pi_1^0$ and that θ is such that $\pi_1^0 = \pi^{*3}$. Then, the asymptotic variance of our estimator is

$$\frac{\pi^*(1 - \pi^{*3})}{(1 - \pi^*)\{\chi + (1 - \chi)\pi^{*2}(1 - \pi^{*3})\}} \to 0$$

as $\pi^* \to 0$, so the only case in which we do not get an improvement is $\chi = 0$, i.e. when we have no consumer–level data. Note that $\chi = E[S/N]$. Thus, in this example with significant observed consumer heterogeneity the asymptotic variance of our estimator goes to zero even though the asymptotic variance of the raw log share estimator goes to infinity. □

When estimating ψ, imposing the share constraint effectively places infinite weight on the macro portion of our likelihood, rather than N_m.

Intuitively, letting $N_m \to \infty$ is equivalent to maximizing a mixed logit likelihood function using the micro data subject to population shares that equal choice probabilities, so the estimator optimizes,[19]

$$f(\theta, \delta) = \sum_{m=1}^{M} \sum_{i=1}^{N_m} D_{im} \sum_{j=0}^{J} y_{ijm} \log \pi_{jm}^{z_{im}} \tag{14}$$

subject to the aggregate market share constraint, $s_{jm} = \pi_{jm}$ for all j, m. This is analogous to Goolsbee and Petrin (2004) and Murry and Zhou (2020), albeit that in those papers it is the aggregate market shares are themselves constructed from the micro data.

[19] For clarity in this section, we assume selection in the consumer-level dataset is random, therefore $\pi_{jm}^{D=1} = \pi_{jm}^{D=0} = \pi_{jm}$ simplifying the log likelihood formula.

The share constraint is a one–to–one mapping from s to δ for fixed θ, which has popularly been solved via a contraction map. However, to study the asymptotic variance of the share constraint estimator, it is useful to re-characterize the constrained optimization problem to account for uncertainty in the aggregate shares as,

$$\max_{\theta} f(\theta, \hat{\delta}(\theta)) \tag{15}$$

where, $\hat{\delta}(\theta)$ is given by the solution to the convex optimization problems,

$$\hat{\delta}(\theta) = \arg\max_{\{\delta_{jm}\}_{j=1}^{J_m}} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} y_{ijm} \log \pi_{jm}(\theta, \delta_{\cdot m}). \tag{16}$$

Focusing on the single market case, we can write the first order conditions of (15) and (16) as,

$$\begin{cases} \sum_{i=1}^{N} \sum_{j=0}^{J} y_{ij} D_i \left(\partial_\theta \log \pi_j^{z_{i\cdot}} + \partial_\theta \hat{\delta}^\intercal \partial_\delta \log \pi_j^{z_{i\cdot}} \right) = 0, \\ \sum_{i=1}^{N} \sum_{j=0}^{J} y_{ij} \partial_\delta \log \pi_j = 0, \end{cases} \tag{17}$$

where by the implicit function theorem,

$$\partial_\theta \hat{\delta}^\intercal = - \sum_{i=1}^{N} \sum_{j=0}^{J} y_{ij} \partial_{\theta\delta^\intercal} \log \pi_j \left(\sum_{i=1}^{N} \sum_{j=0}^{J} y_{ij} \partial_{\delta\delta^\intercal} \log \pi_j \right)^{-1}$$

which captures the impact of θ on the solution to (16).

As noted above, imposing the share constraint can be a reasonable thing to do if S/N is small. However, theorem 1 establishes that doing so is never more efficient than our procedure and is inefficient except in knife-edge cases.

Theorem 1. *Suppose that J is finite and that the micro sample consists of random draws from the population of size N, each member of the population being drawn with probability $0 < \chi_N \to \chi$ as $N \to \infty$ with $0 \leq \chi \leq 1$. Then imposing the share restriction cannot be more efficient and is generally less efficient than using our estimator.* □

Proof of this theorem follows immediately from the proofs of theorems 3 and 4 in Appendix C, which formally derive the asymptotic variance of the MDLE estimator and the share constrained likelihood estimator respectively. There are two cases in which there is no loss of efficiency. The first is if $\chi = 0$, which should in practical terms be interpreted as the size of the micro sample be negligible compared to the size of the population. The second case is if the coefficients on the observable micro regressors, θ^z, are all equal to zero. This case is not helpful since then there is no identification, so a comparison of efficiency is moot.

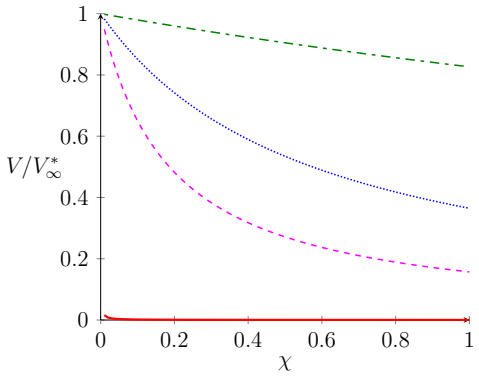
In practice, imposing the share constraint can lead to a substantial efficiency loss as

Example 2 demonstrates.

Example 2. Consider a market where there is one product ($j = 1$) and an outside good. For the inside product, utility is

$$u_{i1} = \delta_1 + \theta^z z_i + \varepsilon_{i1}$$

where the consumer characteristic z_i is distributed standard normal. As is typically the case, utility for the outside good be normalized as $u_{i0} = \varepsilon_{i0}$.



Relative asymptotic efficiency of our estimator of $\delta = 0$ compared to imposing the share constraint as a function of χ for $\theta^z = 1, 4, 10, 10000$ (green, blue, magenta, red).

The figure represents the efficiency gain of our estimator relative to imposing the share restrictions as a function of χ. The formulae for the asymptotic variances of our estimator (V) and the share constraint estimator (V^*_∞) are derived in theorems 3 and 4 respectively in appendix B. When $V/V^*_\infty = 1$, there is no efficiency loss to imposing the share restriction. Each curve corresponds to a different value of θ^z. As the figure illustrates, the efficiency gain of our estimator is negligible if both χ and θ^z are modest but can grow arbitrarily large with θ^z is large for a given $\chi > 0$.

In this simple example there is no efficiency gain for the estimation of θ. We attribute this to the simple design without random coefficients. □

The intuition for this result is straightforward. Imposing the share constraint effectively ignores the information in the micro-sample for the estimation of δ. The benefit of the consumer–level sample is that it can exploit the variation in z_i, so its value is largest when z_i induces a large variation in shares. That is, when θ^z is large.

5.1.2 Inference

There is a second drawback to using the share constraint that arises in a greater range of applications: it introduces a downward bias into the asymptotic variance matrix estimator itself. The source is again treating aggregate shares as being exactly equal to choice probabilities. The implication is that ignoring sampling error in the aggregate shares leads one to be 'overconfident' in the precision of the estimate of δ. This is necessarily true if $S/N \not\to 0$, which is the natural counterpart of the efficiency argument above. In this case,

inference for both θ and δ is incorrect; typically standard errors are too small. Our discussion is under the assumption that selection into the consumer–level sample is random. Selection can make the inference issue more severe and the requirement $S/N \to 0$ be insufficient.

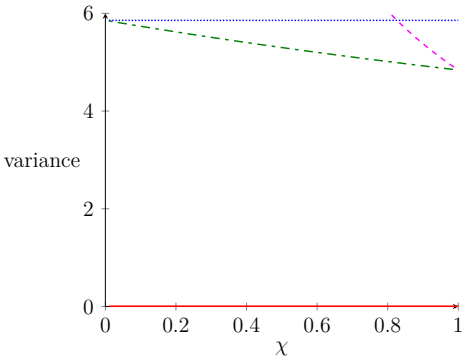
However, if one wishes to conduct inference on δ then care should be taken as soon as $S/\sqrt{N} \not\to 0$. With the share constraints, it would be tempting to use the delta method to conclude that for any vector $v \neq 0$

$$\frac{\sqrt{S}v^\intercal(\hat{\delta} - \delta)}{\sqrt{v^\intercal \partial_{\theta^\intercal}\hat{\delta}(\hat{\theta})\hat{\mathcal{V}}_\theta \partial_\theta \hat{\delta}^\intercal(\hat{\theta})v}}, \tag{18}$$

has a standard normal limiting distribution, where $\hat{\delta}(\theta)$ is the share inversion and $\mathcal{V}_\theta$ is the asymptotic variance of $\hat{\theta}$. This ignores sampling error in the aggregate data, which becomes a problem for all vectors v for which $v^\intercal \partial_{\theta^\intercal}\delta = 0$, in which case (18) diverges. The space of such vectors v is of dimension no less than $J - d_\theta > 0$ since $\delta : \mathbb{R}^{d_\theta} \to \mathbb{R}^J$. Using the bootstrap the way it is typically used does not solve this problem.[20] The inference problem can be avoided by using the asymptotic variance formulas in appendix B.

We now illustrate this problem using an example with the same design (though different parameters) as in example 2. In example 3 $J = d_\theta$ but the problem described above arises if $\theta^z = 0$ because in that case $\partial_\theta \delta(\theta) = 0$.[21]

Example 3.



asymptotic variances for our estimator (green) of $\delta = 1$ as a function of χ for $\theta^z = 0$ and those of the share–constraint estimator (blue), the mixed logit estimator (magenta), and the promised asymptotic variance of the share–constraint estimator (red)

This example contains the same design as example 2. The blue curve in the graph above corresponds to the true asymptotic variance of the share constraint estimator, derived by the GMM problem presented in (17), the green curve to the asymptotic variance of our likelihood–based estimator, and the magenta curve to that of the mixed logit estimator. This graph illustrates the efficiency advantage of our likelihood–based estimator compared to both the share constraint and mixed logit estimators. If $\chi = 1$ the mixed logit estimator coincides with ours, but if $\chi < 1$ the efficiency gain of our estimator increases steeply.

A second interesting feature of the graph above is that the asymptotic variance implied by (pretending that) $N = \infty$, as depicted by the red line, is very different from the actual asymptotic

[20] One would have to draw the population from the superpopulation for the bootstrap to be correct.

[21] We present this example for its simplicity, even though it is a knife edge case. As noted above, an identical problem arises generally if $J > d_\theta$.

variance of the share constraint estimator, i.e. the blue curve. Indeed, in this example, the promised asymptotic variance is zero. The reason for the zero variance is straightforward. If $\theta^z = 0$ then $\partial_\theta \delta = 0$ and so sampling error in the micro sample used to estimate θ does not affect the variance of δ. By assumption, there is no sampling error in the macro sample, and hence δ is recovered "exactly" by the share constraint. For $\theta^z \neq 0$ the asymptotic variance will not vanish. In this example, there would be a concern if $\sqrt{S}\theta^z$ were small.[22] □

To conclude, our estimator has no inference problems and the asymptotic variance is standard for maximum likelihood estimators. By contrast, the asymptotic variance for the share constraint estimator should be based on the asymptotic variance formulas in appendix B which are based on the moments in (17), not on the more convenient formulas that obtain if N is set to ∞. This problem extends to any estimator in which the share constraints are imposed to hold, including the microBLP estimator.

5.2 Micro-Moments versus Micro-Likelihood

The second difference between our estimator and those used in the literature is the use of the micro-likelihood in place of moments derived from the micro-data. Consider again the single market case with random selection. BLP04 proposes to use micro moments of the form

$$\frac{1}{S}\sum_{i=1}^{N}\sum_{j=0}^{J} D_i x_j \{y_{ij} z_i - \zeta_j(\hat{\delta}, \hat{\theta})\} = 0, \tag{19}$$

where $\zeta_j(\delta, \theta) = \int \zeta_j^*(z; \delta, \theta)\, \mathrm{d}G(z)$, with $\zeta_j^*(z; \delta, \theta) = z \int s_j(z, \nu; \delta, \theta)\, \mathrm{d}F(\nu)$, where other moments of z_i are of course allowed.

In practice, $\zeta_j(\delta, \theta)$ is typically evaluated using Monte Carlo integration over (z, ν). Although this entails the method of simulated moments (MSM), here one still needs the number of simulation draws to increase to infinity to achieve consistency and to increase faster than S to achieve efficiency. Intuitively, this is true because (19) can be expressed as the difference between two sample means: one over consumers in the consumer-level data and one over simulation draws.[23]

There are two sources of inefficiency in the use of (19). The first one is the familiar fact that maximum likelihood trumps GMM in terms of efficiency, which relates to the question of which moments to use. A second source of inefficiency pertains to the construction of the moments themselves.

To deal with the second issue first, the moments in (19) only use the consumer–level data for the construction of covariances to compare to covariances implied by the model. That is the second term integrates over the distribution G and is independent of the consumer–level

[22] Analogous to the weak instruments literature the asymptotic variance would be wrong unless $\sqrt{S}\theta^z \to \pm\infty$.

[23] This is analogous to the example in section 4.1 of Pakes and Pollard (1989) where (in their notation) ns simulation draws are used with n the sample size and s the number of simulation draws per observation.

observations. One could make fuller use of the consumer–level data by conditioning on z_i to construct the second term in the moment condition. I.e., it would be more efficient to replace (19) with,

$$\frac{1}{S}\sum_{i=1}^{N}\sum_{j=0}^{J} D_i x_j \{y_{ij} z_i - \zeta_j^*(z_i; \hat{\delta}, \hat{\theta})\} = 0. \tag{20}$$

The moments in (19) and (20) (if evaluated at the truth) have the same Jacobian in expectation, but (20) has a smaller variance because for $\omega_i = \sum_{j=0}^{J} D_i x_j y_{ij} z_i$, the variance contribution for observation i using (19) equals

$$\begin{aligned}\mathbb{V}\{\omega_i - \mathbb{E}(\omega_i \mid D_i, X)\} = \mathbb{E}\mathbb{V}(\omega_i \mid D_i, X) = \\ \mathbb{E}\mathbb{V}(\omega_i \mid z_i, D_i, X) + \mathbb{E}\mathbb{V}\{\mathbb{E}(\omega_i \mid z_i, D_i, X) \mid D_i, X\} \\ \geq \mathbb{E}\mathbb{V}(\omega_i \mid z_i, D_i, X) = \mathbb{V}\{\omega_i - \mathbb{E}(\omega_i \mid z_i, D_i, X)\},\end{aligned}$$

which is the variance contribution of observation i in (20).

A potential theoretical advantage of (20) is that it is linear in the integral over ν, so that if this integral is simulated by Monte Carlo integration then one only has to have the number of simulation draws to grow to infinity with N in order for the asymptotic distribution of the estimator to be unaffected by simulation error—i.e., one can use a fixed number of simulation draws per observation (Pakes and Pollard, 1989). However, this integral can often be efficiently evaluated using quadrature methods. Moreover, the numerical integration over the share constraint to determine $\hat{\delta}$ will still require the number of integration draws to grow with N.[24]

Now returning to the first issue we consider the use of the likelihood of the micro data as in our estimator. Like (20), the micro-likelihood constructs the model analogs to the data at the level of the consumer–level observations. Moreover, using the likelihood makes full use of the consumer–level data and eliminates the need to specify and weight moments.[25] This is particularly an issue for moments to identify the random coefficients, θ^ν, which is one reason many studies have relied on additional product-level moment restrictions despite these parameters being formally identified by the consumer–level data (Walker, Ben-Akiva and Bolduc, 2007). We discuss the issue of product–level instruments in section 6.

We also need to compute numerical integrals to evaluate both the micro and macro terms of (8). The micro terms contains an integral over ν, which we recommend computing using quadrature methods. However, if one chose to integrate this term using Monte Carlo integration, one would need to make the number of simulation draws *per observation* grow with the sample size, which would be a disadvantage relative to (20). The macro term is

[24] As pointed out by Berry, Levinsohn and Pakes (1995), enforcing the share constraint requires using the same Monte Carlo integration draws for each product so that shares sum to one.

[25] One could use the optimal instruments (Chamberlain, 1987) to improve on the efficiency of (19). However, with optimal instruments the linearity in the integral is lost and with it the most attractive feature of using moments.

an integral over (z, ν) where we assume the *total* number of draws grows with N at the same rate as the one used in (19) and (20). In summary, the asymptotic requirements for numerical simulation between our likelihood estimator and GMM are identical if one uses quadrature methods in the micro likelihood, as we recommend, but would be more stringent if one used monte carlo integration in the micro likelihood.

6 Incorporating Product–level Moments

The model presented in Section 2.1 requires micro-level data to identify consumer heterogeneity. As discussed above, Berry, Levinsohn and Pakes (1995) augment this model with exclusion restrictions on product qualities (ξ) to identify consumer heterogeneity with product-level data alone. While not necessary for identification in our case, these restrictions may prove useful in the case of weak or nonidentification using consumer–level data alone, as we specify below.[26] In this section, we first propose a method to augment our estimator with moments of this kind as a penalized likelihood. With this approach a model that is weakly identified (or even not identified) with only consumer-level data may have identification restored by these additional moment restrictions, albeit at a lower rate of convergence. In contrast to the standard GMM approach, our procedure adapts automatically so that convergence rates are not compromised when consumer level data alone strongly identify the model.

6.1 Description

We introduce product–level moments by inserting a penalty term $\hat{\Pi}$ into the objective function. In this section, it is convenient to rewrite the loglikelihood function, $\text{LL}(\psi) = \sum_{m=1}^{M} \hat{\Omega}_m(\theta, \delta_m)$ where $\hat{\Omega}_m$ are the market specific terms of the loglikelihood function (7). We define the penalized objective function as

$$\sum_{m=1}^{M} \hat{\Omega}_m(\theta, \delta_m) - \hat{\Pi}(\beta, \delta), \tag{21}$$

We take the penalty term to be of the form

$$\hat{\Pi}(\beta, \delta) = \frac{1}{J} \hat{m}^{\intercal}(\beta, \delta) \hat{\mathcal{W}} \hat{m}(\beta, \delta), \tag{22}$$

[26] Informally, "weak identification" is a theoretical construct in which parameters that need to satisfy some conditions (e.g., $\theta^z \neq 0$ in our case) converge to the problem point ($\theta^z = 0$) at a rate such that there there is no identification in the limit, but there is still some informative signal. This construct is useful to develop methods that perform well even when the conditions for identification are satisfied, but identification fails nearby. In our context, whereas in principle the parameter θ^z is either zero (no identification) or nonzero (identification) the situation can be murky in samples of finite size: the promised behavior using standard asymptotic analysis may not be reflected in finite samples.

where $\hat{\mathcal{W}}$ is a consistent estimator of the fixed and positive definite optimal weight matrix $\mathcal{W}$[27] and

$$\hat{m}(\beta, \delta) = \sum_{m=1}^{M} \sum_{j=1}^{J_m} b_{jm}(\delta_{jm} - \beta^{\mathsf{T}} x_{jm}), \tag{23}$$

with b_{jm} a vector of instruments for which $\mathbb{E}(b_{jm}\xi_{jm}) = 0$ for all j, m.

If the dimension of b_{jm} is the same as that of x_{jm}, a situation we shall refer to as "exact identification of β" then θ, δ are estimated off the likelihood portion and β subsequently off the GMM portion and this estimator is equivalent to the two-step estimator we described above when β is estimated from (9). Additional restrictions lead to overidentification which can be used to aid the estimation of θ. In the overidentified case, $\hat{\Pi}$ will generally be positive so that both $\hat{\Omega}$ and $\hat{\Pi}$ contribute.

We now discuss the implications of the overidentified case and the reasons for including $\hat{\Pi}$ in the objective function. To see why this is a good idea, table 1 considers rates of convergence of three different estimators of β and θ. In the first row, the penalized likelihood estimator from this section converges at the fastest possible rate under both strong and weak identification.

In the first alternative approach (Table 1, Row 2), the moments $\hat{m}$ are augmented with the gradient of the loglikelihood to form a GMM-like estimator. In this alternative approach one would have moments over different units (consumers versus products). As a result, the convergence rate of $\hat{\theta}$ is $\sqrt{J}$ rather than $\sqrt{S}$.[28] This is true regardless of the strength of identification based on the consumer–level data alone.

	$\sqrt{S/J} \times \theta^z$ **small**		$\sqrt{S/J} \times \theta^z$ **large**	
Estimation method	$\hat{\beta}$	$\hat{\theta}^{\nu}$	$\hat{\beta}$	$\hat{\theta}^{\nu}$
Penalized likelihood	$\sqrt{J}$	$\sqrt{J}$	$\sqrt{J}$	$\sqrt{S}$
GMM: $\partial\hat{\Omega}$ and $\hat{\Pi}$	$\sqrt{J}$	$\sqrt{J}$	$\sqrt{J}$	$\sqrt{J}$
LL + projection (two step)	weak identification		$\sqrt{J}$	$\sqrt{S}$

Table 1: Asymptotic rates of convergence with product-level moments

The second alternative approach entails using product–level moments only to estimate β after θ, δ have been estimated off the likelihood portion (Table 1, Row 3). This leads to the estimator presented in section 3. This approach is asymptotically equivalent to the penalized approach if θ is strongly identified using consumer–level data alone and the product level moments are correctly specified. However, adding the product-level moments offers robustness to weak and nonidentification of θ that is not available via the sequential route taken by the second alternative approach.

[27] It is not essential that an optimal weight matrix is used, but it simplifies the formulas. Further, in contrast to typical GMM, the norming of $\hat{\mathcal{W}}$ does matter here.

[28] One could recover the faster convergence rate by downweighting the product–level moments by a factor $\sqrt{J/S}$, so that (under strong identification) the product level moments are asymptotically negligible as $S \to \infty$. However, even after such a correction, this alternative approach is inferior to ours since the concavity in δ found in (21) is then lost.

In sum, our approach in (21) produces the $\sqrt{S}$ convergence rate for $\hat{\theta}$ under strong identification using consumer–level data and robustness to weak (or no) identification using consumer–level data via overidentification of the product–level moments.

6.2 Inference

Let the argmax of (21) be $\hat{\tau} = [\hat{\theta}^\intercal \ \hat{\delta}^\intercal \ \hat{\beta}]^\intercal$ and the true values of the parameters carry a zero subscript. The object of interest for inference is, as before, a finite–dimensional vector $\phi_0 = \phi(\tau_0)$, where ϕ is a continuously differentiable function. We intend to find a matrix $\hat{\mathcal{V}}_{\text{weak}}$ for which

$$\hat{\mathcal{V}}_{\text{weak}}^{-1/2}(\hat{\phi} - \phi_0) \xrightarrow{d} N(0, \mathcal{I}_{d_\phi}). \tag{24}$$

The subscript 'weak' refers to the fact that our estimation and inference procedures are robust to weak and nonidentification of θ on the basis of consumer–level data alone, which is a strength of our approach. The $\hat{\mathcal{V}}_{\text{weak}}$ component automatically adjusts to the convergence rate of $\hat{\phi}$, which can be as fast as $\sqrt{S}$ and as slow as $\sqrt{J}$.

Analogous to classical extremum estimation theory and following from the delta method, $\hat{\mathcal{V}}_{\text{weak}}$ takes the form

$$\hat{\mathcal{V}}_{\text{weak}} = \partial_{\tau^\intercal}\phi(\hat{\tau})\hat{\mathcal{V}}\partial_\tau\phi^\intercal(\hat{\tau}) \tag{25}$$

for a matrix $\hat{\mathcal{V}}$ defined below. Note, however, that $\hat{\mathcal{V}}$ is a matrix whose dimensions grow with the total number of products across all markets J.

Let $\hat{\mathcal{M}}$ denote the Jacobian of $\hat{m}$ (i.e. $\hat{\mathcal{M}} = \partial_{\tau^\intercal}\hat{m}$) and ω_{im} the (i, m) contribution to the loglikelihood. Then,

$$\hat{\mathcal{V}} = \left(\sum_{m=1}^{M}\sum_{i=1}^{N_m}\partial_\tau\omega_{im}(\hat{\tau})\partial_{\tau^\intercal}\omega_{im}(\hat{\tau}) + \frac{1}{J}\hat{\mathcal{M}}^\intercal(\hat{\tau})\hat{\mathcal{W}}\hat{\mathcal{M}}(\hat{\tau})\right)^{-1}. \tag{26}$$

Note that for ease of notation we have taken τ as the argument of both ω_{im} and $\hat{\mathcal{M}}$ even though ω_{im} is a function of θ, δ_m only and $\hat{\mathcal{M}}$ is a function of β, δ only. Further, the first term in the inverse in (26) corresponds to the outer product of the gradient formula typical of maximum likelihood estimation and could be replaced by minus the information matrix.

The $\hat{\Pi}$ component in (26) is intuitive, also. If one assumes independence of (ξ_{jm}, b_{jm}) across products and that $\mathbb{E}(\xi_{jm} \mid b_{jm}) = 0$ then

$$\hat{\mathcal{W}} = \left(\frac{1}{J}\sum_{m=1}^{M}\sum_{j=1}^{J_m}(\hat{\delta}_{jm}^{(p)} - x_{jm}^\intercal\hat{\beta}^{(p)})^2 b_{jm}b_{jm}^\intercal\right)^{-1}, \tag{27}$$

where $\cdot^{(p)}$ denotes some pilot estimate, i.e. a consistent first step estimator in a two step procedure. If a two–step procedure is deemed cumbersome then one can use a single step procedure with a chosen positive definite matrix $\hat{\mathcal{W}}$, in which case one should adjust the matrix $\hat{\mathcal{V}}$ accordingly. For robustness to dependence across products an appropriate HAC

estimator can be used in the formulation of $\hat{\mathcal{W}}$.[29]

6.3 Theory

The inference procedure described above applies under a wide range of circumstances. Indeed, it should broadly work whenever there is identification off a combination of the consumer–level data and product–level moments,[30] which we will assume from here on.

To understand how our approach helps address possible underidentification at the consumer–level without slowing down the convergence rate of $\hat{\theta}$ if there is strong identification at the consumer–level, consider the formula for $\hat{\mathcal{V}}$ and the special case in which consumer–level data are not only independent but also identically distributed over i, m. Then with strong identification, $\mathbb{E}\{\partial_{\psi}\omega_{im}(\tau_0)\partial_{\psi^\intercal}\omega_{im}(\tau_0)\}$ will have maximum rank. Since the MLE portion in (26) sums over more terms than the GMM portion, the GMM portion in (26) is asymptotically negligible for the estimation of ψ_0. In other words, under those circumstances, $\hat{\mathcal{V}}$ can be replaced with a block diagonal matrix where the first block corresponds to ψ_0 and the second block to β_0, with the blocks diverging at different rates.

On the other hand, if θ_0^{ν} is not identified off the consumer–level data, e.g. if $\theta_0^{z} = 0$, then $\mathbb{E}\{\partial_{\psi}\omega_{im}(\tau_0)\partial_{\psi^\intercal}\omega_{im}(\tau_0)\}$ will be singular, so even though the GMM portion of (26) diverges more slowly than the MLE portion, the GMM portion will still contribute to make up for the singularity.

Although the procedure should work broadly, our theoretical results cannot cover all eventualities, so we limit our results to a well–described set of circumstances. First, we assume that there is a uniform and finite bound on the number of products per market. Thus, the total number of products across all markets, J, increases at the same rate as the number of markets M. Second, we assume that the number of consumers in the consumer–level sample, S, increases faster than M and J.We further require that the total number of consumers in the smallest market diverges faster than M and J, i.e. $M/\min_{m=1,\dots,M} N_m \to 0$.

The results of this section are summarized in the following theorem. A sketch of the proof is provided in appendix D.

Theorem 2. *Under the conditions described above,* (24) *holds.* □

7 Monte Carlo Experiments

In this section, we compare our procedure with micro-BLP and the likelihood-based estimator enforcing the share constraint in a Monte Carlo simulation study.

For our experiments, products have two observable characteristics $x_{jm} = (x_{jm}^1, x_{jm}^2)$ which are distributed Uniform$[1, 3]$ and Normal$(0, 1)$ respectively; the unobservable product

[29] $\hat{\mathcal{W}}$ should not be rescaled based on the sample size because that would amount to placing more or less weight on the GMM portion of the objective function compared to the MLE portion.

[30] Identification based on the consumer–level data alone should, absent selection, equate to identification using a combination of consumer–level and product–level data. The product–level moments buy more.

characteristic ξ_{jm} is distributed Normal$(0, 0.5)$. We do not include an endogenous product characteristic (e.g., price), since we focus on recovering "mean" product quality δ and the taste heterogeneity parameters, so the endogeneity issue is moot. Mean product quality following (2) is,

$$\delta_{jm} = \beta_0 + \beta_1 x^1_{jm} + \beta_2 x^2_{jm} + \xi_{jm},$$

where the true parameters for β are $(-6, 1, 1)$.[31] This implies that mean product quality δ will vary across replications, but we compute the root mean square error and mean absolute deviation for $\hat{\delta}$ by averaging over both products and replications.

Consumers have two observable characteristics $z_{im} = (z^1_{im}, z^2_{im})$ which are distributed Normal$(-1, 1)$ and Normal$(1, 0.5)$ respectively. Taste heterogeneity based on observable consumer characteristics is parameterized according to,

$$\mu^{z_{im}}_{jm} = \theta^z_1 z^1_{im} + \theta^z_2 z^1_{im} x^1_{jm} + \theta^z_3 z^2_{im} x^1_{jm},$$

where the true parameters for θ^z are $(1.5, .75, 1.25)$.

We will consider versions of the model with and without unobserved consumer heterogeneity. For the model with unobserved heterogeneity, consumers have unobserved characteristics $\nu_{im} = (\nu^1_{im}, \nu^2_{im})$ which are both distributed Normal$(0, 1)$, and the unobserved heterogeneity term is,

$$\mu^{\nu_{im}}_{jm} = \theta^\nu_1 \nu^1_{im} x^1_{jm} + \theta^\nu_2 \nu^2_{im} x^2_{jm},$$

where the true parameters for θ^ν are $(1.5, 0.2)$.

For each Monte Carlo replication, we simulate a total of 5 markets, each with 10 products per market, and a market size of $N_m = 100,000$ consumers. Product-level share data, s_{jm} is then calculated based on the choices of these 100,000 consumers. For the consumer-level data, we take a random sample of size $S_m \in \{250; 1,000; 4,000\}$ for the micro dataset. The micro data contains their choice, $y_{i\cdot m}$ (a vector where $y_{ijm} = 1$ if consumer i chose product j and zero otherwise) together with their observable characteristics, z_{im}. For each experiment, we draw 128 replications of the dataset.[32]

We compare three estimators: our mixed data likelihood estimator using the micro and macro likelihood (8); a likelihood estimate implementing the share constraint maximizing (15), and a GMM with share constraint estimator (micro BLP) which uses moments constructed according to (20) to identify θ^z.[33] For the exercises with random coefficients, we require

[31] These were chosen so that the share of the outside good was roughly 20 percent of the aggregate share, although this varies significantly from market to market.

[32] The number 128 is chosen because it is an integer multiple of the number of cores available in our computer.

[33] Note that our implementation of the GMM estimator uses the version of the moments that conditions on observable characteristics to construct the model counterpart of the moments, making it more efficient than (19) which is commonly used in applied work. In other words, our simulation results favor micro BLP.

additional moments to identify θ^ν. We use moments of the form,[34]

$$\frac{1}{M \cdot S_m} \sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} D_{im}(x_{jm} - \bar{x}_m)^2 \{y_{ijm} z_{im} - \zeta^*_{jm}(z_{im}; \hat{\delta}_m, \hat{\theta})\} = 0, \tag{28}$$

where $\bar{x}_m$ is the average product characteristic in market m, so that this moment takes advantage of variation in the choice set across markets. The intuition of this moment is that a random coefficient θ^ν affects the variance of product characteristics chosen conditional on z_{im}, which varies with z_{im} as long as $\theta^z \neq 0$.

We have not incorporated product level moments into the Monte Carlo study, but plan to do so in a future version of this paper. As discussed in section 6, product level moments can be most helpful when θ^ν is weakly identified, which we will see does not appear to be the case in our current design.

We optimize each estimator starting from an initial guess where all parameters are equal to 0.5.[35] We employ the analytic gradient and hessian of the log-likelihood for both our estimator and the share constraint likelihood estimator. The convergence tolerance for these methods is based on the analytic gradient and set to $1.5 * 10^{-8}$. For our micro BLP routine, we use an accelerated version nested fixed point algorithm to solve for the product mean utilities (Reynaerts, Varadhan and Nash, 2012). We use a tolerance of $2 * 10^{-10}$ for the NFP algorithm and compute numerical gradients, and optimize this function using an LBFGS algorithm, following standard BLP procedure. We found that the objective gradients rarely converge to a tolerance of 10^{-6} and therefore define convergence using the weaker criteria that the function improvement or step length are lower than 10^{-6}.

We first consider the results when the model does not include unobserved consumer heterogeneity (that is, without random coefficients). We plot the density function of $\hat{\theta}^z_1$ in fig. 1. Summary statistics for the full parameter vector are available in appendix F. The solid line plots the density of our mixed data likelihood estimator, which is converging towards the truth as S_m increases. Indeed, the observed convergence rate is consistent with the theoretical $1/\sqrt{S}$ rate. The density of the likelihood estimator imposing the share constraint is plotting with a dashed line, it is more disperse than MDLE for $S_m = 250$, but nearly identical for the large two samples. The GMM + SC estimator (aka micro BLP) is plotted using the dotted line. It also performs well for the samples of size 250 and 1,000. For the 4,000 sample size, there is some bi-modality, although his may be due to convergence to a local minimum. Overall, we find that MDLE and LL + SC estimators perform extremely similarly without random coefficients and they slightly outperform the GMM + SC estimator. These results are in line with asymptotic theory.

We now turn to the model with random coefficients. Figure 2 plots the density of $\hat{\theta}^z_1$ and

[34] All moments aggregate over markets, although we suppressed this in (20) for clarity. We make this aggregation explicit here since this moment is a function of the market-level average product characteristics.

[35] We explore robustness to the initial guess in appendix F.

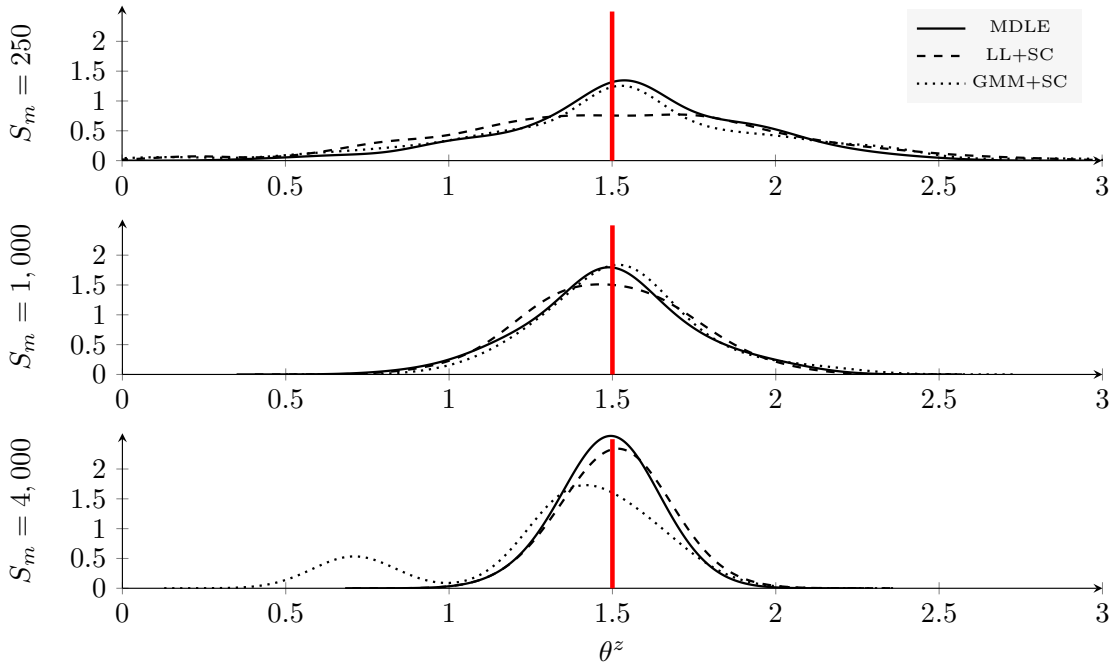


Figure 1: Densities of estimators of $\theta_1^z = 1.5$ with only observed consumer heterogeneity $\mu^\nu = 0$. Solid line is our proposed mixed-data likelihood estimator (MDLE). Dashed line is the likelihood estimator imposing the share constraint (LL + SC). The dotted line is the commonly employed GMM estimator with share constraint (GMM + SC). Solid red line delineates the true parameter value.

$\hat{\theta}_1^\nu$ for the same three estimators for the same three sample sizes. Again, summary statistics for the full parameter vector are available in appendix F. Here, the differences between the estimators are larger, especially for the random coefficient. In all cases MDLE is the most accurate, followed by LL + SC. However, at the smallest consumer sample size, $S_m = 250$, none of the estimators performs well in estimating unobserved heterogeneity. It appears that the consumer sample size is too small to precisely estimate θ^ν in this case. This may represent a situation where product level moments—if correctly specified—may serve to improve the precision of the estimator. However, for greater sample sizes, the distribution is better behaved for both MDLE and LL + SC. Both are downward biased, but appear to be converging towards the truth and exhibiting asymptotic normality, although MDLE seems superior, as suggested by our theoretical results in section 5.

The GMM + SC estimator appears to be severely biased. In fact, the mean of the GMM estimator of θ_1^ν appears to be at the start point of 0.5 rather than near the true value of 1.5. As this raises the possibility of convergence to a local minima, we re-ran the experiment for $S_m = 1,000$ with starting points of 1.5 and 2.0, comparing MDLE and GMM + SC. Results of this experiment in Figure 3 confirm that results of GMM + SC are sensitive to the starting point, while MDLE is relatively robust. Most importantly, MDLE is significantly more accurate in all three cases.

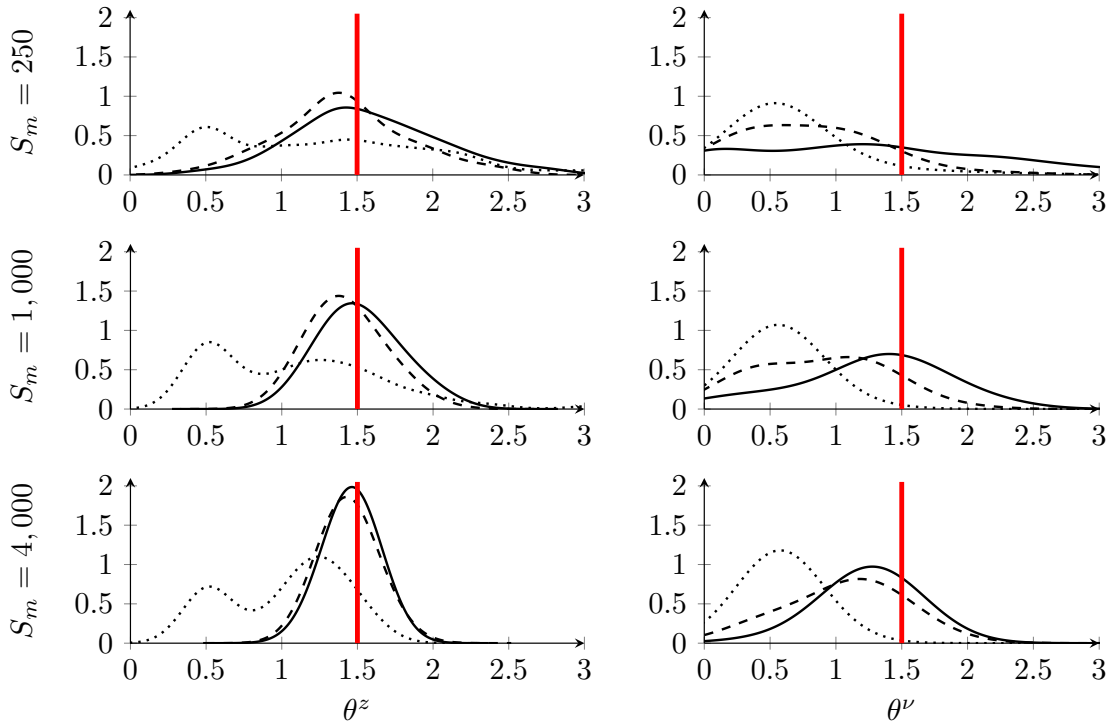


Figure 2: Densities of estimators of $\theta_1^z = 1.5$ and $\theta_1^\nu = 1.5$ with observed and unobserved consumer heterogeneity. Solid line is our proposed mixed-data likelihood estimator (MDLE). Dashed line is the likelihood estimator imposing the share constraint (LL + SC). The dotted line is the commonly employed GMM estimator with share constraint (GMM + SC). Solid red line delineates the true parameter value.

Finally, it is well known that simulated likelihood estimators will exhibit bias. Our implementation of MDLE uses Gaussian quadrature to approximate the integral over ν in the micro portion of the likelihood and Monte Carlo integration to approximate the integral over (z, ν) in the macro likelihood.[36] This approach corresponds to what would most likely be used in application, since the distribution of ν is likely to be assumed, while the distribution of z_{im} will be estimated (or sampled) from data. For the Monte Carlo integration in the macro likelihood, we use 10,000 draws from the distribution of (z_{im}, ν_{im}) in our baseline results. To explore the sensitivity of our estimator to the number of draws, we repeat the experiment with $S_m = 1,000$ varying the number of draws. Figure 4 displays the results of this experiment, where we slightly under-smooth the resulting density functions for illustrative purposes. When the number of Monte Carlo draws is only 1,000, the estimator does not perform well. In fact, there is an atom in this distribution at 0. However, for 10,000 draws performance improves substantially, although some bias is still apparent. As expected, performance improves further for 50,000 draws, although it is clear that the gains

[36] Since ν is two dimensional in our example, we use a tensor product quadrature with 11 quadrature nodes for either dimension for a total of 121 nodes. For higher dimensions, we would suggest using sparse quadrature to approximate this integral.

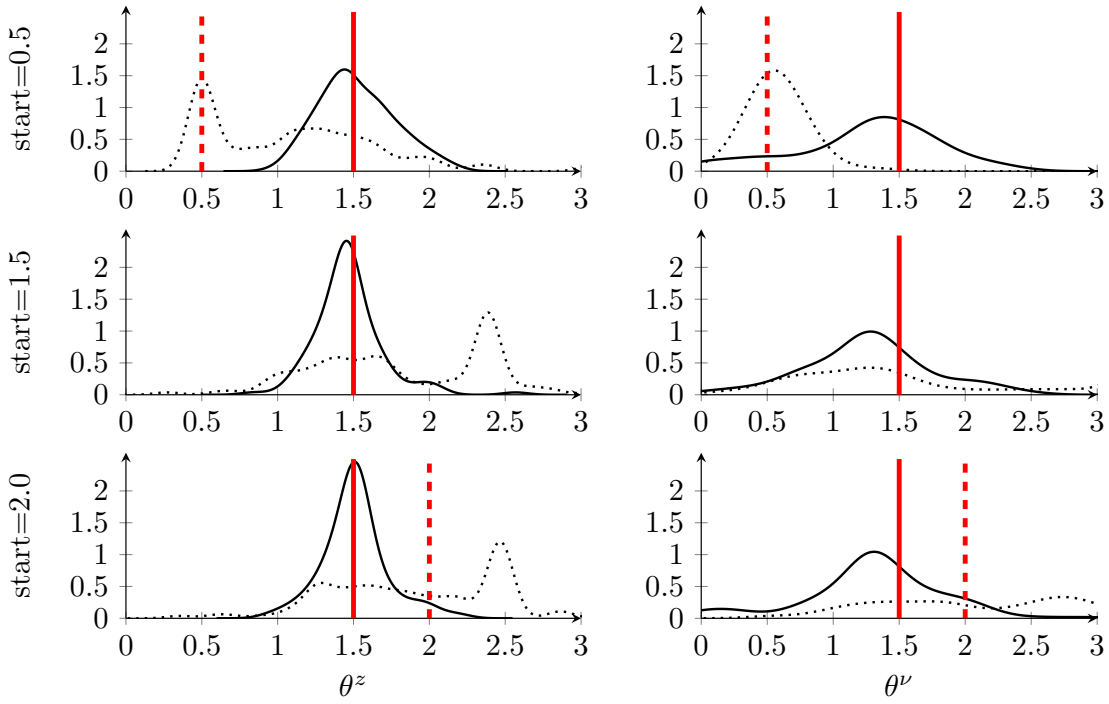


Figure 3: Densities of estimators of $\theta_1^z = 1.5$ and $\theta_1^\nu = 1.5$ varying the initial guess when $S_m = 1,000$. Solid line is our proposed mixed-data likelihood estimator (MDLE). The dotted line is the commonly employed GMM estimator with share constraint (GMM + SC). Solid red line delineates the true parameter value, dashed red line delineates initial guess.

are diminishing. We conclude from this experiment that although one should be mindful of simulation bias, modern computing allows for expanding the number of draws to adequately address this concern.

8 Conclusion

Random coefficients discrete choice demand models are a workhorse of applied industrial organization. GMM-based estimators have combined data at the consumer and product level to enhance the precision of estimates of substitution patterns. In this paper, we investigate how the use of a likelihood estimator simplifies the incorporation of consumer-level and product-level data. Our estimator does not require additional parametric assumptions relative to a GMM estimator. In addition to efficiency, it has several attractive features including computational tractability and straightforward inference. Moreover, we are able to incorporate product-level exclusion restrictions into the estimator with out reducing the rate of convergence. The estimator can be adapted to accommodate additional sources of information, such as "second-choice data" which has been shown to be useful in identifying substitution patterns between products. Our preliminary Monte Carlo results indicate substantial benefits too employing the likelihood approach even before we incorporate

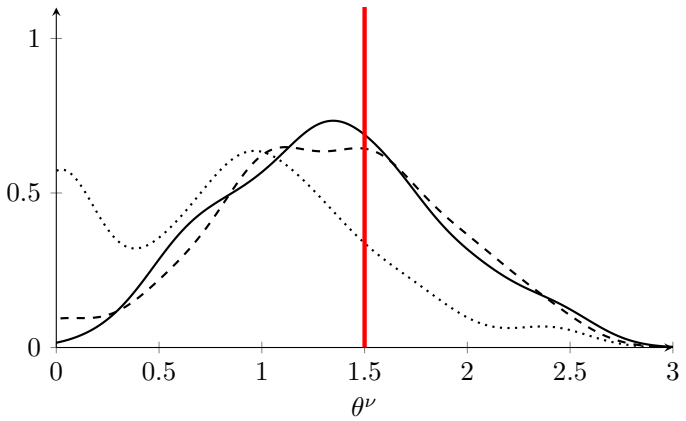


Figure 4: Densities of estimators of $\theta_1^\nu = 1.5$ varying the number of draws used in Monte Carlo integration for the macro term of MDLE. Dotted line uses 1,000 draws. Dashed line is 10,000 draws. Solid line is 50,000 draws. $S_m = 1,000$. Solid red line delineates the true parameter value.

product-level moments.

References

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile prices in market equilibrium." *Econometrica*, 841–890.

Berry, Steven, James Levinsohn, and Ariel Pakes. 2004. "Differentiated products demand systems from a combination of micro and macro data: The new car market." *Journal of Political Economy*, 112(1): 68–105.

Berry, Steven T. 1994. "Estimating discrete-choice models of product differentiation." *The RAND Journal of Economics*, 242–262.

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

Berry, Steven T., and Philip A. Haile. 2020. "Nonparametric identification of differentated products demand using micro data." Yale University.

Berry, Steve, Oliver B Linton, and Ariel Pakes. 2004. "Limit theorems for estimating the parameters of differentiated product demand systems." *The Review of Economic Studies*, 71(3): 613–654.

Chamberlain, Gary. 1987. "Asymptotic efficiency in estimation with conditional moment restrictions." *Journal of Econometrics*, 34(3): 305–334.

Davidson, James. 1994. *Stochastic limit theory: An introduction for econometricians.* OUP Oxford.

Gandhi, Amit, and Jean-Francois Houde. 2020. "Measuring Substitution Patterns in Differentiated-Products Industries." University of Pennsylvania and UW-Madison.

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

Gowrisankaran, Gautam, and Marc Rysman. 2012. "Dynamics of consumer demand for new durable goods." *Journal of political Economy*, 120(6): 1173–1219.

Grieco, Paul, Charles Murry, and Ali Yurukoglu. 2020. "The Evolution of Market Power in the US Automobile Industry." *mimeo*.

Murry, Charles, and Yiyi Zhou. 2020. "Consumer search and automobile dealer colocation." *Management Science*, 66(5): 1909–1934.

Nevo, Aviv. 2001. "Measuring Market Power in the Ready-to-Eat Cereal Industry." *Econometrica*, 69(2): 307–342.

Pakes, Ariel, and David Pollard. 1989. "Simulation and the Aymptotics of Optimization Estimators." *Econometrica*, 57(5): 1027–1057.

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

Reynaerts, J., R. Varadhan, and J. C. Nash. 2012. "Enhancing the Convergence Properties of the BLP (1995) Contraction Mapping." KU Leuven.

Train, Kenneth. 2003. *Discrete Choice Methods with Simulation.* Cambridge, UK:Cambridge University Press.

Walker, Joan L., Moshe Ben-Akiva, and Denis Bolduc. 2007. "Identification of parameters in normal error component logit-mixture (NECLM) models." *Journal of Applied Econometrics*, 22(6): 1095–1125.

Wollmann, Thomas G. 2018. "Trucks without bailouts: Equilibrium product characteristics for commercial vehicles." *American Economic Review*, 108(6): 1364–1406.

Appendix

A Gradients and Hessians

The derivations below assume that $\mu^z_{ijm}, \mu^\nu_{ijm}$ are linear in θ. Define $b_{ijm} = b_{ijm}(\nu) = \partial_\theta(\mu^z_{ijm} + \mu^\nu_{ijm})$, which does not depend on θ by construction. Thus, $\pi^{z_{i\cdot m}}_{jm} = \int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)$ where

$$\pi_{ijm}(\nu) = \frac{\exp(\mu^z_{ijm} + \mu^\nu_{ijm} + \delta_{jm})}{\sum_{g=0}^{G} \exp(\mu^z_{igm} + \mu^\nu_{igm} + \delta_{gm})}. \tag{29}$$

Then,

$$\partial_\theta \log \pi^{z_{i\cdot m}}_{jm} = \frac{\int \pi_{ijm}(\nu)\Delta b_{ijm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)}, \tag{30}$$

where $\Delta b_{ijm}(\nu) = b_{ijm}(\nu) - \bar{b}_{i\cdot m}(\nu)$ with $\bar{b}_{i\cdot m}(\nu) = \sum_{j=0}^{J_m} \pi_{ijm}(\nu) b_{ijm}(\nu)$. Further,

$$\partial_\delta \log \pi^{z_{i\cdot m}}_{jm} = \frac{\int \pi_{ijm}(\nu)\Delta \mathbb{1}_{ijm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)}, \tag{31}$$

where the k-th element of $\Delta\mathbb{1}_{ijm}$ equals $\mathbb{1}(j=k) - \pi_{ikm}(\nu)$ for $k = 1, \ldots, J_m$.

To obtain the gradient of LL we moreover need the gradient of $\log \pi_{jm}$. But since π_{jm} is simply an integral of π^z_{jm} over z, the gradient of $\log \pi_{jm}$ is identical to that of $\log \pi^{z_{i\cdot m}}_{jm}$ except that z is integrated out in both numerator and denominator in (30) and (31). An analogous argument applies to the Hessians. So we only present the Hessians for the micro contributions.

They are,

$$\begin{aligned}\partial_{\theta\theta^\intercal} \log \pi^{z_{i\cdot m}}_{jm} = {} & \frac{\int \pi_{ijm}(\nu)\Delta b_{ijm}(\nu)\Delta b^\intercal_{ijm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)} - \partial_\theta \log \pi^{z_{i\cdot m}}_{jm} \partial_{\theta^\intercal} \log \pi^{z_{i\cdot m}}_{jm} \\ & - \sum_{g=0}^{J_m} \frac{\int \pi_{ijm}(\nu)\pi_{igm}(\nu)\Delta b_{igm}(\nu)\Delta b^\intercal_{igm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)},\end{aligned} \tag{32}$$

$$\begin{aligned}\partial_{\delta\delta^\intercal} \log \pi^{z_{i\cdot m}}_{jm} = {} & \frac{\int \pi_{ijm}(\nu)\Delta \mathbb{1}_{ijm}(\nu)\Delta \mathbb{1}^\intercal_{ijm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)} - \partial_\delta \log \pi^{z_{i\cdot m}}_{jm} \partial_{\delta^\intercal} \log \pi^{z_{i\cdot m}}_{jm} \\ & - \frac{\int \pi_{ijm}(\nu)\Big[\pi_{ikm}(\nu)\{\mathbb{1}(k=t) - \pi_{itm}(\nu)\}\Big]_{k,t=1,\ldots,J_m}\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)},\end{aligned} \tag{33}$$

where the notation $[\cdot]_{k,t=\ldots}$ means a matrix whose (k,t) element is given by the argument in square brackets and, finally,

$$\partial_{\delta\theta^\intercal} \log \pi^{z_{i\cdot m}}_{jm} = \frac{\int \pi_{ijm}(\nu)\Delta \mathbb{1}_{ijm}(\nu)\Delta b^\intercal_{ijm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)} - \partial_\delta \log \pi^{z_{i\cdot m}}_{jm} \partial_{\theta^\intercal} \log \pi^{z_{i\cdot m}}_{jm}. \tag{34}$$

The $\delta\theta^\intercal$ Hessian term has one fewer term because it is zero.

B Asymptotic variances

This appendix provides formulas for the asymptotic variance of our estimator and the estimator that maximizes the mixed logit objective function subject to the share constraints for a single market, i.e. $m = 1$; the multimarket case is an obvious extension. The formulas below are valid for the case in which selection is random; otherwise an adjustment should be made, e.g. $\pi_j^{D=0}$ should replace π_j and some cancellations do then not obtain.

We use $\psi = [\theta^\intercal, \delta^\intercal]^\intercal$ and use Ω^m to denote $\mathbb{E}\sum_{j=0}^J Y_{ij} \log \pi_j^{z_i \cdot}$, Ω^m_ψ its gradient, $\Omega^m_{\psi\psi}$ its Hessian, and $\Omega^M = \mathbb{E}\sum_{j=0}^J Y_{ij} \log \pi_j$. Let similar symbols be analogous defined. Formulas for these gradients and Hessians can be found in appendix A.

The asymptotic variance for our estimator is then

$$-\left\{\chi\Omega^m_{\psi\psi} + (1-\chi)\Omega^M_{\psi\psi}\right\}^{-1}, \tag{35}$$

where $\chi = \lim_{N\to\infty}(S/N)$. This is for $\sqrt{N}(\hat{\psi} - \psi)$ and $\chi > 0$. Fr $\chi = 0$, consider the limit distribution of $\sqrt{S}(\hat{\psi} - \psi)$ for $\chi > 0$, i.e. multiply (35) by χ and then let $\chi \downarrow 0$. This takes some caution since $\Omega^M_{\psi\psi}$ is generally singular.

The *promised* but incorrect asymptotic variance for the share constraint estimator is

$$-\begin{bmatrix} \mathcal{J} \\ \partial_\theta \delta^\intercal \end{bmatrix} \Phi^{-1} \begin{bmatrix} \mathcal{J} & \partial_{\theta^\intercal}\delta \end{bmatrix} / \chi, \quad \textit{(incorrect variance)} \tag{36}$$

where $\partial_{\theta^\intercal}\delta = -(\Omega^M_{\delta\delta})^{-1}\Omega^M_{\delta\theta}$ and $\Phi = \Omega^m_{\theta\theta} + \partial_\theta\delta^\intercal\Omega^m_{\delta\theta} + \Omega^m_{\theta\delta}\partial_{\theta^\intercal}\delta + \partial_\theta\delta^\intercal\Omega^m_{\delta\delta}\partial_{\theta^\intercal}\delta$. The correct asymptotic variance formula for the shape constraint estimator is

$$-\begin{bmatrix} \chi\Phi & \chi(\Omega^m_{\theta\delta} + \partial_\theta\delta^\intercal\Omega^m_{\delta\delta}) \\ \Omega^M_{\delta\theta} & \Omega^M_{\delta\delta} \end{bmatrix}^{-1} \begin{bmatrix} \chi\Phi & 0 \\ 0 & \Omega^M_{\delta\delta} \end{bmatrix} \begin{bmatrix} \chi\Phi & \Omega^M_{\theta\delta} \\ \chi(\Omega^m_{\delta\theta} + \Omega^m_{\delta\delta}\partial_{\theta^\intercal}\delta) & \Omega^M_{\delta\delta} \end{bmatrix}^{-1}. \tag{37}$$

Finally, a mixed logit estimator ignoring the product share information would have asymptotic variance

$$(-\Omega^m_{\psi\psi})^{-1}/\chi. \tag{38}$$

C Efficiency

Consider the situation in which we a randomly selected consumer–level sample from a single market in addition to product–level data including shares. Then the objective function can be written as

$$\Omega(\psi) = \sum_{i=1}^{\bar{I}} \left\{D_i L_i^m(\psi) + \omega(1 - D_i)L_i^M(\psi)\right\}, \tag{39}$$

for $\omega = 1$ where L_i^m, L_i^M are the likelihood for the data on which we have detailed and less detailed information respectively and D_i is the micro selection dummy which is independent of everything else and equals one with probability χ. We allow for $0 \leq \omega < \infty$ to incorporate the possibility of unequal weighting. Both intuition and mathematics indicate that choosing $\omega = 1$ is optimal.

Theorem 3. *Under the stated assumptions we have,* $\sqrt{\bar{I}}(\hat{\psi}-\psi) \xrightarrow{d} N(0,V)$*, where* $V = (\chi A+\omega(1-\chi)B)^{-1}(\chi A+\omega^2(1-\chi)B)(\chi A+\omega(1-\chi)B)^{-1}$*, with* $A = -\mathbb{E}\{\partial_{\psi\psi^\intercal} L_1^m(\psi)\}$ *and* $B = -\mathbb{E}\{\partial_{\psi\psi^\intercal} L_1^M(\psi)\}$*. The optimal weight* ω *equals one.* □

Proof. The asymptotic distribution is an immediate consequence of standard extremum estimation theory. Since both $A, B \geq 0$, the first derivative of V with respect to ω equals zero at $\omega = 1$ and the second derivative of V with respect to ω equals

$$\chi C^{-1}BC^{-1} + \chi^2 C^{-1}BC^{-1}BC^{-1} + 3\omega\chi^3 C^{-1}BC^{-1}BC^{-1}BC^{-1} \geq 0,$$

where $C = \chi A + \omega(1-\chi)B$, which follows from tedious but simple calculus. □

We now turn to the possibility that one maximizes the consumer–level likelihood subject to the product–level shares matching the choice probabilities. We do so by considering the asymptotic variance of

$$\hat{\psi}_\omega^* = \arg\max_\psi \sum_{i=1}^{\bar{I}} \{D_i L_i^m(\psi) + \omega L_i^M(\psi)\}, \tag{40}$$

as a function of ω and then letting $\omega \to \infty$. Note that imposing that the gradient of $\sum_{i=1}^{\bar{I}} L_i^M$ equal zero is equivalent to imposing the product–level share equations. Note further that there is a subtle but important difference between (39) and (40) in that in (40) we sum over all L_i^M, not only over those we lack consumer–level data on. Finally, using only the product–level likelihood is insufficient for identification since all first order conditions are satisfied by setting shares equal to choice probabilities.

Theorem 4. *Let* V_ω^* *be the asymptotic variance of* $\hat{\psi}_\omega^*$*. Then*

$$V_\infty^* = \lim_{\omega\to\infty} V_\omega^* = \{\chi A U_0 (U_0^\intercal A U_0)^{-1} U_0^\intercal A + B\}^{-1} \geq V,$$

where U_0 *contains a full set of orthogonal unit length eigenvectors of the null space of* B*.* □

Proof. Standard extremum estimation theory yields

$$V_\omega^* = (\chi A+\omega B)^{-1}\{\chi A + (2\chi\omega+\omega^2)B\}(\chi A+\omega B)^{-1}.$$

Taking $\omega \to \infty$ means that the $2\chi\omega B$ term is negligible compared to $\omega^2 B$. The same is not true for χA since B does not have full rank. Use the spectral decomposition $B = U_1 D_1 U_1^\intercal$ where U_1 contains orthogonal eigenvectors corresponding to nonzero eigenvalues. It is straightforward to verify that the inverse of $\chi A+\omega^2 B$ is (up to terms that vanish as $\omega \to \infty$) equal to $U_0(\chi U_0^\intercal A U_0)^{-1} U_0^\intercal + U_1 D_1^{-1} U_1^\intercal/\omega^2$.[37] Pre and postmultiply by $\chi A + \omega B$ and take $\omega \to \infty$ to obtain V_∞^*. Finally, note that

$$\begin{aligned} V_\infty^{*\,-1} - V^{-1} = \chi A U_0 (U_0^\intercal A U_0)^{-1} U_0^\intercal A + B - \{\chi A + (1-\chi)B\} = \\ \chi\{A U_0 (U_0^\intercal A U_0)^{-1} U_0^\intercal A - A + B\} = \\ \chi[(A-B)U_0\{U_0^\intercal (A-B) U_0\}^{-1} U_0^\intercal (A-B) - (A-B)] \leq 0, \end{aligned}$$

since the right hand side is minus an annihilator matrix. □

[37] Just premultiply by $U_0^\intercal, U_1^\intercal$ and postmultiply by U_0, U_1 (four combinations) noting that $U_0^\intercal U_0$ and $U_1^\intercal U_1$ are the identity matrix and the other products are zero matrices.

The proof shows that equality of the asymptotic variance only obtains if $A - B$ is in the null space of B, which would happen if the coefficients on all consumer–level regressors equaled zero. Conversely, one would expect the difference to be large if the consumer–level regressors are informative.

A second consequence is that the efficiency improvement is greatest for the estimation of the δ coefficients. The intuition for this finding is that imposing the aggregate share equations does not limit the exploitation of variation in the micro level regressors, but it does suggest that information contained only in the consumer–level sample is not used to recover coefficients on product–level coefficients.

D Sketch of proof

D.1 Bird's eye view

A wrinkle on standard extremum estimation theory with generous use of partitioned inverses suggests that

$$\hat{\theta} - \theta_0 \simeq -\big(\Gamma_{\theta\theta} - \Gamma_{\theta\delta}\Gamma_{\delta\delta;\beta}^{-1}\Gamma_{\delta\theta}\big)^{-1}\{\hat{\Gamma}_\theta - \Gamma_{\theta\delta}\Gamma_{\delta\delta;\beta}^{-1}(\hat{\Gamma}_\delta - \Gamma_{\delta\beta}\Gamma_{\beta\beta}^{-1}\hat{\Gamma}_\beta)\}, \tag{41}$$

where $\Gamma(\theta, \delta, \beta) = \Omega(\theta, \delta) - \Pi(\beta, \delta)$ is the expectation of $\hat{\Gamma}$ (both are sums), subscripts denote partial derivatives and $\Gamma_{\delta\delta;\beta} = \Gamma_{\delta\delta} - \Gamma_{\delta\beta}\Gamma_{\beta\beta}^{-1}\Gamma_{\beta\delta}$ is the Hessian of Γ with respect to δ after concentrating out β. Note that $\Gamma_\theta = \Omega_\theta$ since Π is a function of (δ, β) only and that $\Omega = \sum_{m=1}^{M} \Omega_m$ is the sum over $N = \sum_{m=1}^{M} N_m$ terms whereas Γ_β like Π is on net a sum over J terms.[38]

Now, note that $\Gamma_{\delta\delta;\beta} = \Omega_{\delta\delta} - \Pi_{\delta\delta;\beta}$ since Ω does not depend on β. Thus, $\Gamma_{\delta\delta;\beta}^{-1} = (\Omega_{\delta\delta} - \Pi_{\delta\delta;\beta})^{-1} \simeq \Omega_{\delta\delta}^{-1} + \Omega_{\delta\delta}^{-1}\Pi_{\delta\delta;\beta}\Omega_{\delta\delta}^{-1}$, which implies that

$$\Gamma_{\theta\theta} - \Gamma_{\theta\delta}\Gamma_{\delta\delta;\beta}^{-1}\Gamma_{\delta\theta} = (\Omega_{\theta\theta} - \Omega_{\theta\delta}\Gamma_{\delta\delta;\beta}^{-1}\Omega_{\delta\theta}) \simeq \Omega_{\theta\theta;\delta} - \mathcal{X}_{\theta;\delta}^{\intercal}\Pi_{\delta\delta;\beta}\mathcal{X}_{\theta;\delta}, \tag{42}$$

with $\Omega_{\theta\theta;\delta} = \Omega_{\theta\theta} - \Omega_{\theta\delta}\Omega_{\delta\delta}^{-1}\Omega_{\delta\theta}$ and $\mathcal{X}_{\theta;\delta} = \Omega_{\delta\delta}^{-1}\Omega_{\delta\theta}$. Further,

$$\hat{\Gamma}_\theta - \Gamma_{\theta\delta}\Gamma_{\delta\delta;\beta}^{-1}(\hat{\Gamma}_\delta - \Gamma_{\delta\beta}\Gamma_{\beta\beta}^{-1}\hat{\Gamma}_\beta) \simeq \underbrace{\hat{\Omega}_\theta - \mathcal{X}_{\theta;\delta}^{\intercal}\hat{\Omega}_\delta}_{\sim\text{sum over } N \text{ terms}} + \underbrace{\mathcal{X}_{\theta;\delta}^{\intercal}(\hat{\Pi}_\delta - \Pi_{\delta\beta}\Pi_{\beta\beta}^{-1}\hat{\Pi}_\beta)}_{\sim\text{sum over } J \text{ terms}}, \tag{43}$$

where the sum counts are *net,* i.e. the ratio of two sums over N terms is counted as a sum over one term.

Note that $-\Omega_{\theta\theta;\delta}$ is the variance of $\hat{\Omega}_\theta - \mathcal{X}_{\theta;\delta}^{\intercal}\hat{\Omega}_\delta$ and that $\Omega_{\theta\theta;\delta}$ has full rank under strong identification using the consumer-level data (SIC) only. So for the SIC case (θ_0^z is fixed and nonzero) the stated result then follows from the equivalence of Hessian and outer product of the gradient formulas for the inverse of the variance, ignoring the minor complication that the dimension of δ increases in J.

More generally, the variance of $\hat{\Omega}_\theta - \mathcal{X}_{\theta;\delta}^{\intercal}\hat{\Omega}_\delta + \mathcal{X}_{\theta;\delta}^{\intercal}(\hat{\Pi}_\delta - \Pi_{\delta\beta}\Pi_{\beta\beta}^{-1}\hat{\Pi}_\beta)$ is $-(\Omega_{\theta\theta;\delta} - \mathcal{X}_{\theta;\delta}^{\intercal}\Pi_{\delta\delta;\beta}\mathcal{X}_{\theta;\delta})$, whose second half only contributes in the null space of $\Omega_{\theta\theta;\delta}$ since it is a sum over J terms, not N. The same comments pertaining to Hessian and outer product of the gradient apply here.

[38] Two moments with J terms and a weight matrix as an inverse of a sum with J terms.

D.2 Greater detail — no incidental parameters problem

Since the vector δ increases in dimension, it may appear as though there is an incidental parameters problem here, but there is not. Suppose that $\delta_m(\theta)$, which is of finite dimension, is for any θ the solution to $\Omega_{\delta m}\{\theta, \delta_m(\theta)\} = 0$. Then all N_m observations in market m can be used to estimate δ_m which suggests a convergence rate equal to $1/\sqrt{N_m}$. To estimate $\delta_m(\theta_0)$, however, adds a $1/\sqrt{S}$ term due to the estimation of θ_0 (in SIC case). Note that S is the total number of consumers in the consumer-level sample.

D.3 Greater detail — strong identification

The leading term of the expression that must be shown to be normal can in the SIC case be expressed as $\sum_{m=1}^{M}\sum_{i=1}^{N_m}\Omega_{\theta\theta;\delta m}^{-1/2}(\omega_{\theta im} - \Omega_{\theta\delta m}\Omega_{\delta\delta m}^{-1}\omega_{\delta im}) = \sum_{m=1}^{M}\sum_{i=1}^{N_m}\zeta_{im}$. By theorem 24.3 in Davidson (1994) two conditions need to be satisfied: *(i)* $\max_{m=1,\ldots,M}\max_{i=1,\ldots,N_m}\|\zeta_{im}\| = o_p(1)$; *(ii)* $\sum_{m=1}^{M}\sum_{i=1}^{N_m}\mathbb{E}(\zeta_{im}\zeta_{im}^\intercal) - \mathcal{I} = o_p(1)$. The second condition is satisfied by construction. For the first condition, Markov's theorem implies that a sufficient condition is that $\sum_{m=1}^{M}\sum_{i=1}^{N_m}\mathbb{E}\|\zeta_{im}\|^p = o(1)$ for some $p > 2$.[39] Taking $p = 4$ we get $\sum_{m=1}^{M}\sum_{i=1}^{N_m}\mathbb{E}\|\zeta_{im}\|^4 = \sum_{m=1}^{M}\sum_{i=1}^{N_m}(C_m^4/N_m^2) = \sum_{m=1}^{M}(C_m^4/N_m)$ for some $C_m < \infty$ independent of N_m, S.[40] It should be apparent that the right hand side sum is all but guaranteed to be $o(1)$.

D.4 Greater detail — adaptation to identification situation

We now explain how our procedure adapts to different identification situations. In particular, we discuss how the convergence rate of our estimator is affected by identification strength in the consumer-level sample.

Recall from (42) that $\Gamma_{\theta\theta} - \Gamma_{\theta\delta}\Gamma_{\delta\delta;\beta}^{-1}\Gamma_{\delta\theta} \simeq \Omega_{\theta\theta;\delta} - \mathcal{X}_{\theta;\delta}^\intercal\Pi_{\delta\delta;\beta}\mathcal{X}_{\theta;\delta}$, where $\Omega_{\theta\theta;\delta}$ would be singular if $\theta_0^z = 0$. To allow for weak or nonidentification on the basis of consumer-level data alone, we write $\mathcal{A} = \Omega_{\theta\theta;\delta}/N$, $\mathcal{B} = -\mathcal{X}_{\theta;\delta}^\intercal\Pi_{\delta\delta;\beta}\mathcal{X}_{\theta;\delta}/N$.

Start with the spectral decomposition $\mathcal{A} = \mathcal{U}\mathcal{D}\mathcal{U}^\intercal + \mathcal{U}_0\mathcal{D}_0\mathcal{U}_0^\intercal$, where $\mathcal{D}$ is full rank and $\mathcal{D}_0$ may or may not be full rank. It is incorrect to think of $\mathcal{D}$ as the eigenvalues for the estimation of θ_0^z and $\mathcal{D}_0$ as the eigenvalues for the estimation of θ_0^ν unless $\theta_0^z = 0$, but it works for intuition. Now, let $\mathcal{B} = \mathcal{P}\mathcal{B}\mathcal{P} + \mathcal{P}\mathcal{B}\mathcal{P}_0 + \mathcal{P}_0\mathcal{B}\mathcal{P} + \mathcal{P}_0\mathcal{B}\mathcal{P}_0$, where $\mathcal{P} = \mathcal{U}\mathcal{U}^\intercal$ and $\mathcal{P}_0 = \mathcal{U}_0\mathcal{U}_0^\intercal$. The idea is that $\mathcal{B}$ is vanishing but that the matrix $\mathcal{D}_0$ may be full rank, may be zero, or may be decreasing with the sample size; $\mathcal{B}$ is full rank.

We now solve for the inverse of $\mathcal{A} + \mathcal{B}$, which we express as $\mathcal{U}\mathcal{Y}_{11}\mathcal{U}^\intercal + \mathcal{U}\mathcal{Y}_{10}\mathcal{U}_0^\intercal + \mathcal{U}_0\mathcal{Y}_{01}^\intercal\mathcal{U}^\intercal + \mathcal{U}_0\mathcal{Y}_{00}\mathcal{U}_0^\intercal$. Then,

$$(\mathcal{A} + \mathcal{B})(\mathcal{U}\mathcal{Y}_{11}\mathcal{U}^\intercal + \mathcal{U}\mathcal{Y}_{10}\mathcal{U}_0^\intercal + \mathcal{U}_0\mathcal{Y}_{01}^\intercal\mathcal{U}^\intercal + \mathcal{U}_0\mathcal{Y}_{00}\mathcal{U}_0^\intercal) = \mathcal{I}. \tag{44}$$

We are only going to enforce (44) up to nonnegligible terms, which produces

$$(\mathcal{A} + \mathcal{B})^{-1} \simeq \begin{bmatrix}\mathcal{U} & \mathcal{U}_0\end{bmatrix}\begin{bmatrix}\mathcal{D}^{-1} & -\mathcal{D}^{-1}\mathcal{U}^\intercal\mathcal{B}\mathcal{U}_0(\mathcal{D}_0 + \mathcal{U}_0^\intercal\mathcal{B}\mathcal{U}_0)^{-1} \\ . & (\mathcal{D}_0 + \mathcal{U}_0^\intercal\mathcal{B}\mathcal{U}_0)^{-1}\end{bmatrix}\begin{bmatrix}\mathcal{U}^\intercal \\ \mathcal{U}_0^\intercal\end{bmatrix}. \tag{45}$$

Consider three cases: *(1)* There is strong identification, i.e. $\mathcal{D}_0$ does not vanish. In this case $\mathcal{U}_0^\intercal\mathcal{B}\mathcal{U}_0$ and $\mathcal{U}^\intercal\mathcal{B}\mathcal{U}_0$ are negligible compared to $\mathcal{D}_0$ and $(\mathcal{A} + \mathcal{B})^{-1}$ becomes $\mathcal{A}^{-1}$ plus negligible terms: the product-level moments do not affect $\hat{\theta}$ in large samples. *(2)* There is

[39] For $p = 2$ the sum would be equal to one, by construction.

[40] Powers less than four can be used via the Marcinkiewicz inequality.

no identification, i.e. $\mathcal{D}_0 = 0$, such that the blocks in the central matrix in (45) become $\mathcal{D}^{-1}$, $-\mathcal{D}^{-1}\mathcal{U}^\intercal\mathcal{B}\mathcal{U}_0(\mathcal{U}_0^\intercal\mathcal{B}\mathcal{U}_0)^{-1}$, its transpose, and $(\mathcal{U}_0^\intercal\mathcal{B}\mathcal{U}_0)^{-1}$. In this case, it can be shown that[41]

$$\Omega_{\theta\theta;\delta} = \begin{bmatrix} \Omega_{\theta^z\theta^z} & 0 \\ 0 & 0 \end{bmatrix},$$

whence

$$\hat{\theta} - \theta_0 \simeq -\begin{bmatrix} \mathcal{D}^{-1}\hat{\Omega}_{\theta^z} \\ (\mathcal{X}_{\theta^\nu;\delta}^\intercal \Pi_{\delta\delta;\beta}\mathcal{X}_{\theta^\nu;\delta})^{-1}\mathcal{X}_{\theta^\nu\delta}^\intercal(\hat{\Pi}_\delta - \Pi_{\delta\beta}\Pi_{\beta\beta}^{-1}\hat{\Pi}_\beta) \end{bmatrix}. \tag{46}$$

In other words, the consumer-level data are used to estimate θ_0^z and the product-level moments are used to estimate θ_0^ν but only the information beyond what is needed to recover β_0 can be used. *(3)* The knife–edge case in which $\mathcal{D}_0$ vanishes at exactly the same rate as $\mathcal{B}$ in which case all terms contribute.

E Second Choice Data

Data on consumers second choices has been found to be particularly useful in precisely estimating unobserved heterogeneity in tastes θ^ν, which is a significant driver substitution patterns (Berry, Levinsohn and Pakes, 2004). Grieco, Murry and Yurukoglu (2020) incorporate this data by constructing moments based on the correlation of observable characteristics between first and second choices. Accommodating consumer-level data on second (or ranked) choices is quite straightforward following the extension of the mixed logit to ranked data (Train, 2003). That is, if we observe a consumers first choice and second choices (j, k), we can replace $s_{jm}(z_{i\cdot m}, \nu; \psi)$ in (6) with its ranked choice counterpart,

$$s_{(j,k)m}(z_{i\cdot m}, \nu; \psi) = \frac{\exp(\delta_{jm} + \mu_{ijm}^z + \mu_{ijm}^\nu)}{\sum_{g\in\mathcal{J}_m}\exp(\delta_{gm} + \mu_{igm}^z + \mu_{igm}^\nu)} \frac{\exp(\delta_{km} + \mu_{ikm}^z + \mu_{ikm}^\nu)}{\sum_{g\in\mathcal{J}_m\setminus j}\exp(\delta_{gm} + \mu_{igm}^z + \mu_{igm}^\nu)}$$

where $\mathcal{J}_m$ represents the product set (including the outside good).

F Additional Monte Carlo Results

This appendix contains the results of our Monte Carlo simulation for (θ, δ) we discussed in section 7. While we focused on a select subset of parameters in section 7, for completeness, we report the bias, the root-mean-square error (RMSE) as well as the mean absolute error (MAE) for all of our estimated parameters here.

Tables A1 to A4 correspond to the results presented in figure 1 in which we only allow for observable consumer heterogeneity. The reported statistics are qualitatively similar across parameters highlighting the advantage of using a likelihood based approach over using GMM.

Tables A5 to A10 correspond to the results presented in figure 2 in which we introduce unobserved consumer heterogeneity via random coefficients. Again, the computed statistics are qualitatively similar across parameters. The results underscore the advantage of the MDLE estimator compared to the likelihood plus share constraint estimator and the traditional GMM with share constraint estimator. We do not report tables for figure 3 since the results are qualitatively very similar to tables A5 to A10.

[41] This is most apparent if one normalizes z_i to have mean zero.

Table A1: MC Simulation Results, no random coefficients - θ_1^z

Micro Sample	Estimator	bias	rmse	mae	percent converged
$S_m = 250$	MDLE	0.049	0.411	0.289	100.0
$N_m = 100$k	LL + Share Constraint	0.006	0.499	0.329	100.0
	GMM + Share Constraint (micro BLP)	0.016	0.563	0.351	100.0
$S_m = 1$k	MDLE	-0.010	0.254	0.176	100.0
$N_m = 100$k	LL + Share Constraint	-0.012	0.220	0.151	100.0
	GMM + Share Constraint (micro BLP)	0.025	0.248	0.159	100.0
$S_m = 4$k	MDLE	-0.010	0.116	0.072	100.0
$N_m = 100$k	LL Likelihood + Share Constraint	0.008	0.120	0.082	100.0
	GMM + Share Constraint (micro BLP)	-0.167	0.389	0.054	100.0

Table A2: MC Simulation Results, no random coefficients - θ_2^z

Sample	Estimator	bias	rmse	mae	percent converged
$S_m = 250$	MDLE	-0.011	0.168	0.110	100.0
$N_m = 100$k	LL + Share Constraint	0.009	0.211	0.134	100.0
	GMM + Share Constraint (micro BLP)	0.009	0.214	0.132	100.0
$S_m = 1$k	MDLE	0.005	0.098	0.057	100.0
$N_m = 100$k	LL + Share Constraint	0.004	0.090	0.069	100.0
	GMM + Share Constraint (micro BLP)	-0.003	0.102	0.061	100.0
$S_m = 4$k	MDLE	0.001	0.047	0.027	100.0
$N_m = 100$k	LL + Share Constraint	-0.003	0.049	0.030	100.0
	GMM + Share Constraint (micro BLP)	0.061	0.143	0.049	100.0

Table A3: MC Simulation Results, no random coefficients - θ_3^z

Sample	Estimator	bias	rmse	mae	percent converged
$S_m = 250$	MDLE	-0.004	0.108	0.067	100.0
$N_m = 100$k	LL + Share Constraint	0.010	0.103	0.073	100.0
	GMM + Share Constraint (micro BLP)	-0.003	0.241	0.164	100.0
$S_m = 1$k	MDLE	0.004	0.043	0.028	100.0
$N_m = 100$k	LL + Share Constraint	-0.000	0.045	0.027	100.0
	GMM + Share Constraint (micro BLP)	0.009	0.122	0.090	100.0
$S_m = 4$k	MDLE	-0.007	0.028	0.019	100.0
$N_m = 100$k	LL + Share Constraint	0.001	0.025	0.014	100.0
	GMM + Share Constraint (micro BLP)	-0.087	0.246	0.054	100.0

Table A4: MC Simulation Results, no random coefficients - δ

Sample	Estimator	rmse	mae	percent converged
$S_m = 250$	MDLE	10.528	6.489	100.0
$N_m = 100$k	LL + Share Constraint	10.788	7.767	100.0
	GMM + Share Constraint (micro BLP)	25.009	15.641	100.0
$S_m = 1$k	MDLE	5.439	3.498	100.0
$N_m = 100$k	LL + Share Constraint	5.457	3.439	100.0
	GMM + Share Constraint (micro BLP)	13.271	7.647	100.0
$S_m = 4$k	MDLE	4.041	2.397	100.0
$N_m = 100$k	LL + Share Constraint	4.010	2.351	100.0
	GMM + Share Constraint (micro BLP)	25.552	6.402	100.0

Table A5: MC Simulation Results - θ_1^z

Micro Sample	Estimator	bias	rmse	mae	percent converged
$S_m = 250$k	MDLE	0.076	0.480	0.314	100.0
$N_m = 100$k	LL + Share Constraint	-0.089	0.433	0.242	100.0
	GMM + Share Constraint (micro BLP)	0.137	2.943	0.632	99.2
$S_m = 1$k	MDLE	0.021	0.248	0.166	100.0
$N_m = 100$k	LL + Share Constraint	-0.074	0.239	0.188	100.0
	GMM + Share Constraint (micro BLP)	-0.349	0.884	0.561	98.4
$S_m = 4$k	MDLE	-0.041	0.124	0.078	100.0
$N_m = 100$k	LL + Share Constraint	-0.061	0.154	0.110	100.0
	GMM + Share Constraint (micro BLP)	-0.473	0.602	0.342	98.4

Table A6: MC Simulation Results - θ_2^z

Micro Sample	Estimator	bias	rmse	mae	percent converged
$S_m = 250$k	MDLE	-0.144	0.560	0.463	100.0
$N_m = 100$k	LL + Share Constraint	-0.407	0.507	0.382	100.0
	GMM + Share Constraint (micro BLP)	0.197	5.337	0.431	99.2
$S_m = 1$k	MDLE	-0.146	0.3530	0.197	100.0
$N_m = 100$k	LL + Share Constraint	-0.366	0.441	0.340	100.0
	GMM + Share Constraint (micro BLP)	-0.286	0.801	0.371	98.4
$S_m = 4$k	MDLE	-0.156	0.233	0.148	100.0
$N_m = 100$k	LL + Share Constraint	-0.265	0.344	0.244	100.0
	GMM + Share Constraint (micro BLP)	-0.409	0.425	0.427	98.4

Table A7: MC Simulation Results - θ_3^z

Micro Sample	Estimator	bias	rmse	mae	percent converged
$S_m = 250$	MDLE	-0.116	0.585	0.492	100.0
$N_m = 100$k	LL + Share Constraint	-0.425	0.515	0.508	100.0
	GMM + Share Constraint (micro BLP)	0.050	5.337	0.634	99.2
$S_m = 1$k	MDLE	-0.130	0.361	0.220	100.0
$N_m = 100$k	LL + Share Constraint	-0.373	0.456	0.354	100.0
	GMM + Share Constraint (micro BLP)	-0.457	0.927	0.621	98.4
$S_m = 4$k	MDLE	-0.163	0.242	0.160	100.0
$N_m = 100$k	LL + Share Constraint	-0.270	0.354	0.237	100.0
	GMM + Share Constraint (micro BLP)	-0.593	0.602	0.591	98.4

Table A8: MC Simulation Results - θ_1^ν

Micro Sample	Estimator	bias	rmse	mae	percent converged
$S_m = 250$k	MDLE	-0.266	0.965	0.726	100.0
$N_m = 100$k	LL + Share Constraint	-0.710	0.897	0.707	100.0
	GMM + Share Constraint (micro BLP)	0.176	8.453	0.970	99.2
$S_m = 1$k	MDLE	-0.247	0.602	0.253	100.0
$N_m = 100$k	LL + Share Constraint	-0.626	0.766	0.499	100.0
	GMM + Share Constraint (micro BLP)	-0.710	1.429	0.985	98.4
$S_m = 4$k	MDLE	-0.254	0.381	0.229	100.0
$N_m = 100$k	LL + Share Constraint	-0.436	0.579	0.352	100.0
	GMM + Share Constraint (micro BLP)	-0.916	0.925	0.939	98.4

Table A9: MC Simulation Results - θ_2^{ν}

Micro Sample	Estimator	bias	rmse	mae	percent converged
$S_m = 250$k	MDLE	0.220	0.566	0.199	100.0
$N_m = 100$k	LL + Share Constraint	0.202	0.378	0.172	100.0
	GMM + Share Constraint (micro BLP)	0.060	0.551	0.297	99.2
$S_m = 1$k	MDLE	0.096	0.360	0.199	100.0
$N_m = 100$k	LL + Share Constraint	0.132	0.276	0.140	100.0
	GMM + Share Constraint (micro BLP)	0.104	0.460	0.297	98.4
$S_m = 4$k	MDLE	-0.011	0.229	0.199	100.0
$N_m = 100$k	LL + Share Constraint	0.073	0.216	0.148	100.0
	GMM + Share Constraint (micro BLP)	0.006	0.317	0.295	98.4

Table A10: MC Simulation Results - δ

Sample	Estimator	rmse	mae	percent converged
$S_m = 250$	MDLE	42.095	17.212	100.0
$N_m = 100$k	LL + Share Constraint	29.514	15.272	100.0
	GMM + Share Constraint (micro BLP)	285.691	23.661	99.2
$S_m = 1$k	MDLE	22.396	9.605	100.0
$N_m = 100$k	LL + Share Constraint	23.226	11.226	100.0
	GMM + Share Constraint (micro BLP)	42.974	19.378	98.4
$S_m = 4$k	MDLE	13.363	6.044	100.0
$N_m = 100$k	LL + Share Constraint	17.336	7.672	100.0
	GMM + Share Constraint (micro BLP)	29.295	17.718	98.4