Conformant and Efficient Estimation of Discrete Choice Demand Models*

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

June 7, 2022

Abstract

We propose a likelihood-based estimator for random coefficients discrete choice demand models that is applicable in a broad range of data settings. Intuitively, it combines the likelihoods of two mixed logit estimators—one for consumer level data, and one for product level data—with product level exogeneity restrictions. Our estimator is both efficient and conformant: its rates of convergence will be the fastest possible given the variation available in the data. The researcher does not need to pre-test or adjust the estimator and the inference procedure is valid across a wide variety of scenarios. Moreover, it can be tractably applied to large datasets. We illustrate the features of our estimator by comparing it to alternatives in the literature.

1 Motivation

First introduced in Berry, Levinsohn and Pakes (1995) (henceforth BLP), random coefficients discrete choice demand models provide a tractable framework within which to flexibly estimate substitution patterns between many differentiated products in the presence of price endogeneity. Since its introduction, this model has been estimated using a wide array of datasets featuring consumer level data, product level data, or a mixture of both. We propose a likelihood-based estimator for BLP-style models that is applicable to all the above data settings. Intuitively, it combines the likelihoods of two mixed logit estimators, one for consumer level data (assuming it is available), and one for product level data with product level exogeneity restrictions. We impose no additional assumptions over those posited in Berry, Levinsohn and Pakes (1995) which are also used in other estimators extended with consumer level data (e.g., Petrin, 2002; Berry, Levinsohn and Pakes, 2004; Goolsbee and Petrin, 2004; Chintagunta and Dube, 2005).

*We thank Nikhil Agarwal, Chris Conlon, Amit Gandhi, Jessie Handbury, Sung Jae Jun, Nail Kashaev, Mathieu Marcoux, Karl Schurter, Andrew Sweeting, and seminar participants at the University of Arizona, Microsoft, MIT, Université de Montréal, the University of Pennsylvania, Rice University, the RIDGE IO Workshop, the University of Toronto, and DC IO day for helpful suggestions. Please check http://personal.psu.edu/plg15/publication/like-blp/ 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.

Researchers have applied varied approaches when confronted with different types of data (e.g., consumer choices, market shares, or a combination of both). We note that the best achievable convergence rate varies with (the relative growth rates of) data dimensions and other circumstances. We propose a single estimator that achieves the optimal rate and is efficient in a wide variety of empirical settings. We call our estimator *conformant* for its ability to achieve the optimal rate under a variety of circumstances. To our knowledge, this is a novel property in this literature.

To fix ideas, consider first the case in which a large sample of consumer purchase data is available. The basic structure of the demand model proposed in BLP is mixed (or random coefficients) multinomial logit. The standard multinomial logit MLE has nice computational properties. For example, it is globally concave in the parameters plus the gradient and Hessian have simple expressions. Therefore, with consumer level data in hand, it is natural to consider estimating a BLP model via MLE using the individual likelihood of purchase. In order to accommodate price endogeneity, the basic structure of BLP requires the estimation of product (by market) quality parameters.[1] It can be demanding of consumer level data alone to estimate such a specification due to the presence of potentially many (hundreds, or even thousands, depending on the application) product quality parameters.

To address this issue we incorporate product level data on market shares. We now view our consumer level sample as a (perhaps small) subset of the population of individual choices represented by the observed market shares. From this perspective, the loglikelihood of both individual consumer data ('micro' data) *and* market shares ('macro' data) consists of two terms: a micro term following the mixed logit and a macro term that simply integrates over the distribution of consumer characteristics in the population. This mixed-data likelihood estimator (MDLE) could be used 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. However, there are two potential drawbacks to the MDLE approach. First, identification of unobserved preference heterogeneity is dependent on sufficient exploitable demographic variation, as we describe in section 4.2. Second, this approach alone does not yield mean tastes for product characteristics, although one could incorporate a second step which accommodates endogenous characteristics (such as price).

Our estimator combines the MDLE approach with an additional term to directly incorporate information contained in the product level exogeneity restrictions of Berry, Levinsohn and Pakes (1995). The main benefit of this approach relative to MDLE alone arises when there are more exogeneity restrictions than product characteristics. In the presence of such overidentification, the extra information can be used to help identify the preference heterogeneity parameters even when they cannot be recovered using MDLE alone. Indeed, as BLP show, with sufficient exogeneity restrictions it is possible to identify all model parameters even if the consumer sample size falls to

[1] Berry, Levinsohn and Pakes (1995) and Nevo (2000) have noted that product quality parameters 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.

zero.

Our estimator can be applied to all datasets in the applied literature, in particular it is well defined with consumer samples of any size, from zero to a full census of the market. The objective function is composed of three terms that can diverge at different rates: the micro loglikelihood with the consumer sample size, the macro loglikelihood with the market size, and the GMM objective function based on the product exogeneity restrictions with the number of products. These differing rates in the objective function are what make our estimator conformant: its rates of convergence will adjust accordingly and depend on the ratio of the number of sampled consumers to the number of products, both in all markets.[2]

Indeed, our estimator incorporates two distinct sources of identification for the consumer heterogeneity parameters. As we explain in section 5, observed variation in demographics identifies both observed and unobserved taste heterogeneity as long as that variation shifts consumers' utility across products.[3] As emphasized by Gandhi and Houde (2020), overidentifying product level exclusion restrictions can also identify taste heterogeneity. If the number of sampled consumers is much larger than the number of products then exploiting the identifying information in the micro sample (if present) will produce a faster convergence rate than relying on product level exclusion restrictions. In this case, the MDLE and our estimator are asymptotically equivalent and indeed efficient. Adding the product level exclusions to the estimator is useful both when the consumer sample is small (or not present) and if its identfying demographic variation is weak (or nonexistent). Note that when this variation is nonexistent, the information used by the MDLE estimator is insufficient for identification. Our estimator on the other hand still converges at the optimal rate and is efficient because it also exploits the product level exclusions. However the rate of convergence of some parameter estimates will then be slower (though still optimal) due to the slower divergence rate of the product restrictions component compared to the micro likelihood. Our estimator also covers the intermediate cases between the above two extremes without adjustment and the case where different data is available in different markets.

In addition to being conformant to a variety of data scenarios, we show that our estimator is efficient in each of these scenarios. Efficiency depends on two features of the objective function. First, the likelihood and moments portions of the objective function are uncorrelated because the loglikelihood sums over individuals treating product qualities as parameters whereas the moments component involves sums over products where variation in product quality gives rise to the product level structural error term. So all that remains is the proper weighting of the product level moments portion of the objective function. The optimal weight matrix to use is the same as that in standard GMM estimation albeit that now the scale matters to properly weight across likelihood and GMM terms, as we describe in sections 3.2 and 4.1.

Asymptotically valid inference is done via the standard extremum estimation framework. This is an advantage over alternative methods popular in the literature, which we describe below, that

[2]The use of the plural 'rates' is due to the fact that different elements of our estimator vector converge at different rates.

[3]Berry and Haile (2020) make a similar point in a nonparametric context.

impose share constraints which complicate inference. In particular, share constrained methods require that the total number of consumers S in the micro sample across all markets is negligibly small compared to the smallest markets size $\min_m N_m$ and, if the product quality parameters are of interest, even that S is negligibly small compared to $\min_m \sqrt{N_m}$.[4] Absent these additional restrictions, the computed standard errors would be too small, as is illustrated at the end of example 1 in section 6.2. More generally, the inference procedure is robust to the source of identification, i.e. the inference procedure is valid both when the micro data provide sufficient information to recover the taste heterogeneity parameters and when such information must come from the product level exclusion restrictions: one does not have to specify or know. Given that convergence rates can vary depending on the source of identification, as mentioned above, this feature is not obvious.

While the statistical properties of our estimator make it of theoretical interest, we also argue that it is suitable for applied work. One might expect that the high dimensionality of the parameter space due to the product quality parameters would be intractable. However, we show in section 7 that the structure of the objective function simplifies the computational problem considerably. We have verified that this procedure can be used successfully for problems with over 100,000 products and millions of consumers. Another concern might be the bias due to numerical integration to compute choice probabilities. As shown by Pakes and Pollard (1989), the method of simulated moments usually has an advantage over simulated maximum likelihood in that the bias is negligible when using a fixed number of draws per observation (consumer, in our case). This critique applies to our method when Monte Carlo integration is used. However, the number of random coefficients (as opposed to coefficients on observable demographic variables) tends to be small in applied work, typically within the range that modern quadrature methods can compute with a high degree of accuracy. While this is appropriate to compute the micro loglikelihood, the macro loglikelihood requires integration over demographics and random coefficients. Here we consider using Monte Carlo integration using a large number of draws. This requires the number of draws to diverge faster than the square of the prevailing convergence rate to leave the asymptotic behavior unaffected. This is the same condition required by Berry, Levinsohn and Pakes (1995) (and other share constrained estimators) due to the use of simulation to compute the share inversion.

Our estimator is most directly comparable to GMM approaches based on micro-moments (e.g., Petrin, 2002; Berry, Levinsohn and Pakes, 2004). Beyond the conformance and efficiency benefits from (also) 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 can be thought of as setting the score of our macro loglikelihood to zero. Our objective function properly weights this term with the micro loglikelihood and product level moments to achieve efficiency, rather than giving the macro score infinite weight. 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

[4] In Berry, Levinsohn and Pakes (1995, 2004) the N_m's are assumed to be effectively infinite.

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.

Other researchers have proposed using the likelihood of consumer data in estimating BLP-style models (e.g., Goolsbee and Petrin, 2004; Chintagunta and Dube, 2005; Train and Winston, 2007; Goeree, 2008; Bachmann et al., 2019). The key difference with our approach is twofold. First, they use a two-stage procedure, and so cannot take full advantage of over-identifying product level restrictions. Second, like Petrin (2002) and Berry, Levinsohn and Pakes (2004), these papers estimate product quality parameters using the BLP inversion, whereas our approach achieves efficiency by replacing the inversion with the macro likelihood.

Our approach has broad applicability and is appropriate for many demand estimation applications where the researcher has both product level data on shares and consumer level data on purchases. Berry and Haile (2014) showed identification of objects in a nonparametric class of these models using product level data and sufficient instruments; Berry and Haile (2020) shows how observing consumer level data reduce the number of instruments required. Although Berry, Levinsohn and Pakes (2004) and Petrin (2002) are canonical examples of applications, there are many more examples of applied research where demand is estimated with product level and consumer level data. An incomplete list of examples includes Goeree (2008), Crawford and Yurukoglu (2012), Hendel and Nevo (2006), Wollmann (2018), Crawford et al. (2018), Hackmann (2019), Neilson (2019), Backus, Conlon and Sinkinson (2021), and Grieco, Murry and Yurukoglu (2021). A specific example common in economics and marketing is when researchers combine grocery store scanner data with household level data, for example as in the IRI data or the Kilts Center Nielsen data. Examples include Chintagunta and Dube (2005) (IRI) and Tuchman (2019) and Backus, Conlon and Sinkinson (2021) (Nielsen).

Finally, our problem and approach share features with several strands of the econometrics literature. For instance, Imbens and Lancaster (1994) also considers the problem of combining different sources of data albeit that there the micro data are assumed to provide identification and the different data sources are either independent with sample sizes growing at the same rate or the macro data can be considered to be of infinite size. Further, it is common in the panel data literature to have the dataset grow in different dimensions at different rates (e.g. Hahn and Newey, 2004), but we know of no examples in which there are as many growth dimensions to consider as here: the number of markets and products, the population sizes in each market, and the number of sampled consumers in the micro sample.Third, having different elements of the estimator vector converge at different rates is a common feature of the semiparametric estimation literature (e.g. Robinson, 1988). Lastly, Abadie et al. (2020) consider the case of sample size approaching population size; their problem is different from the ones studied here.

The following section reviews the random coefficients demand model and the data available in our setting. Section 3 then proposes our estimator, whose conformance and efficiency properties are described in section 4. In section 5 we explore sources of variation in the demographic data that our method exploits to identify the taste parameters. In section 6 we explore the steps needed

and applicable trade-offs going from our estimator to the GMM estimator that is currently most commonly used. We argue for the computational tractability of our estimator in section 7. In section 8 we introduce our inference procedure and section 10 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}, ξ_{jm}), where $x_{jm} = (\tilde{x}_{jm}, p_{jm})$ is a d_x-vector of observed characteristics of the product and ξ_{jm} is a scalar unobserved product attribute. The only distinction between $\tilde{x}_{jm}$ and p_{jm} (typically price) is that $\tilde{x}_{jm}$ is uncorrelated with ξ_{jm}, so we frequently refer only to x_{jm} for notational convenience. There are N_m consumers in market m. Consumers are characterized by $(z_{im}, \nu_{im}, \epsilon_{i\cdot m})$ where z_{im} is a d_z-vector of potentially observable consumer characteristics (such as income or location), and ν_{im} is an (up to) d_x-vector of unobservable consumer taste shocks to preferences for product characteristics.[5] Finally $\epsilon_{i\cdot m}$ 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. Let $y_{ijm} = 1$ if consumer i in market m chooses product j and zero otherwise. Utility of consumer i when purchasing product j in market m is,

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

where,[6]

$$\delta_{jm} = x_{jm}^{\intercal}\beta + \xi_{jm}, \tag{2}$$

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

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

represents the mean utility for product j for consumers in market m while,

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

represents deviations from mean utility due to observed demographic variables z_{im}. Typically, μ^z is a linear combination of products of elements of x_{jm} and z_{im} parameterized by θ^z. As we shall see below, some of our results depend on whether θ^z is such that $\partial_z \mu^z = 0$, i.e., when changes in observed demographics do not affect utility. For notational ease, we assume without loss of generality that this is true if and only if $\theta^z = 0$, which corresponds to the typical case just described.

Finally,

$$\mu_{jm}^{\nu_{im}} = \mu^\nu(x_{jm}, \nu_{im}; \theta^\nu) \tag{4}$$

are deviations due to taste shocks ν_{im}. Typically μ^ν is a linear combination of product characteristics and taste shocks parameterized by θ^ν.[7] Utility of the outside good is normalized to $u_{i0m} = \epsilon_{i0m}$. When convenient, we collect the consumer heterogeneity parameters into the vector $\theta = [\theta^{z\intercal}, \theta^{\nu\intercal}]^\intercal$.

The model yields choice probabilities for each consumer of selecting product j conditional on consumer characteristics z_{im} as a function of parameters,

$$\pi_{jm}^{z_{im}}(\theta, \delta) = \Pr(y_{ijm} = 1 \mid z_{im}, x_{\cdot m}; \theta, \delta) = \int \underbrace{\frac{\exp(\delta_{jm} + \mu_{ijm}^z + \mu_{ijm}^\nu)}{\sum_{\ell=0}^{J_m} \exp(\delta_{\ell m} + \mu_{i\ell m}^z + \mu_{i\ell m}^\nu)}}_{s_{jm}(z_{im}, \nu; \theta, \delta)} \mathrm{d}F_m(\nu), \tag{5}$$

where $\delta_{0m} = \mu_{0m}^{z_{im}} = \mu_{0m}^{\nu_{im}} = 0$ for all i, m.

Similarly, market shares are obtained by integrating π_{jm}^z with respect to the distribution of consumer demographics,

$$\pi_{jm}(\theta, \delta) = \Pr(y_{ijm} = 1 \mid x_{\cdot m}) = \int \pi_{jm}^z(\theta, \delta) \, \mathrm{d}G_m(z).$$

In addition to the structure imposed on choice probabilities, the model imposes product level exogeneity restrictions of the form,[8]

$$E(\xi_{jm} b_{jm}) = 0, \tag{6}$$

where b_{jm} is a vector of instruments which includes $\tilde{x}_{jm}$. Further, b_{jm} may also contain additional exogeneity restrictions. The literature has used various approaches such as cost shifters, BLP instruments, Hausman instruments, Gandhi-Houde instruments, and Waldfogel instruments (see Gandhi and Nevo, 2021). These moment restrictions will serve two purposes. First, they are needed to identify mean product utility parameters, β. Second, if $d_b > d_\beta$, where $d_\cdot$ indicates a dimension, they may provide additional information that is potentially useful in estimating other model parameters. For example Berry, Levinsohn and Pakes (1995) uses restrictions of this form to

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

[8] One could replace (6) with a conditional expectation and derive optimal instruments, which would produce a two-step procedure in which each step has a condition of the form (6), with the instruments b_{jm} in the second step generated from the first step.

recover consumer heterogeneity parameters θ in the *absence* of consumer level data.

2.2 Data

The researcher has access to two types of data on consumer choices. First, she observes market level data on the quantity of purchases, a vector of characteristics x_{jm} of each product, and the total market size, N_m.[9] Each consumer has unit demand and purchases either one of the "inside" products or the outside good. That is, the researcher can construct market shares,

$$s_{jm} = \frac{1}{N_m} \sum_{i=1}^{N_m} y_{ijm}. \tag{7}$$

Note that the observed market shares $s_{\cdot m}$ need not equal choice probabilities $\pi_{\cdot m}$ due to the finite population size and unobserved consumer heterogeneity, however $s_{\cdot m} \xrightarrow{p} \pi_{\cdot m}$ as $N_m \to \infty$.

Second, for a subset of S_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. As we will describe below, our methodology combines the micro-sample with the product shares by integrating out z_{im} in the choice probabilities when individual i is outside the micro-sample. We can accommodate several forms of selection. In appendix E we show that for random sampling and deterministic selection on choices $y_{i\cdot m}$ (e.g., administrative data when outside good purchases are not reported) no adjustments are needed. We further show how to accommodate selection on demographics z_{im}.

3 Estimator

We propose an efficient estimator which in its most general form combines the likelihood, $\hat{L}(\theta, \delta)$, of the micro and macro choice data and an efficient GMM objective function $\hat{\Pi}$ based on (6),

$$(\hat{\beta}, \hat{\theta}, \hat{\delta}) = \arg\min_{\beta,\theta,\delta} \underbrace{\Big(-\log \hat{L}(\theta, \delta) + \hat{\Pi}(\beta, \delta) \Big)}_{\hat{\Omega}(\beta,\theta,\delta)} \tag{8}$$

Notice that the likelihood is a function of (θ, δ) but not β, whereas the product level moments are functions of (β, δ) but not θ. This separability has been noted previously in the literature, but will play an important role in making our estimator computationally feasible. The following two subsections describe the two terms of the objective function in detail.

[9] As in the previous literature, researchers will need to observe or make an assumption regarding N_m in order to compute market shares from purchase quantity data. In practice, some consumers may purchase multiple goods within the same period; one could rationalize this by allowing market size to be the number of potential purchasing events. For example, Nevo (2001) determines market size as the number of potential servings of cereal consumed in a city over a quarter.

3.1 Likelihood

The likelihood contains two parts, relating to the micro and macro data. To understand its two elements, first *suppose* that we observed $\{y_{ijm}\}$ for all N_m observations. Then the loglikelihood would be,[10]

$$\log \hat{L}(\theta, \delta) = \sum_{m=1}^{M} \sum_{j=0}^{J_m} \sum_{i=1}^{N_m} y_{ijm} \left(D_{im} \log \pi_{jm}^{z_{im}}(\theta, \delta) + (1 - D_{im}) \log \pi_{jm}(\theta, \delta) \right), \tag{9}$$

The loglikelihood sums over all N_m consumers in the market. If an observation i is in the micro data then we see z_{im} and can condition on it, whereas otherwise we integrate over the distribution of z_{im} 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,

$$\log \hat{L}(\theta, \delta) = \underbrace{\sum_{m=1}^{M} \sum_{j=0}^{J_m} \sum_{i=1}^{N_m} D_{im} y_{ijm} \log \frac{\pi_{jm}^{z_{im}}}{\pi_{jm}}}_{\text{micro}} + \underbrace{\sum_{m=1}^{M} N_m \sum_{j=0}^{J_m} s_{jm} \log \pi_{jm}}_{\text{macro}}, \tag{10}$$

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 $\log \pi_{jm}$ to control for the fact that the consumer level data represent a subset of the consumers who make up the market.

Alternatively, the estimator can be written by adjusting the macro term to avoid double counting the consumers in the micro-sample:

$$\log \hat{L}(\theta, \delta) = \underbrace{\sum_{m=1}^{M} \sum_{j=0}^{J_m} \sum_{i=1}^{N_m} D_{im} y_{ijm} \log \pi_{jm}^{z_{im}}}_{\text{micro}} + \underbrace{\sum_{m=1}^{M} \sum_{j=0}^{J_m} \left(N_m s_{jm} - \sum_{i=1}^{N_m} D_{im} y_{ijm} \right) \log \pi_{jm}}_{\text{macro}}, \tag{11}$$

These two formulations, while equivalent, emphasize different features of the estimator so we will refer to the one that is most convenient at the time.

The likelihood recalls two common estimators in the discrete choice literature. When $N_m = S_m$—so that all consumers' characteristics are observed—or when product market shares are not observed, the likelihood simplifies to the well known mixed-logit likelihood. Indeed, identification of (θ, δ) using the log-likelihood alone follows from the arguments for identification in the mixed-logit setting (Walker, Ben-Akiva and Bolduc, 2007). However, when $S_m = 0$, so only aggregate data is available, maximizing the likelihood is equivalent to imposing the share constraint from BLP and

[10] For expositional simplicity, we present notation for the cases of random selection or deterministic selection on $y_{i \cdot m}$ into the micro sample. As discussed in appendix E, selection on demographics requires an adjustment to account for sampling in π_{jm}.

related estimators, as we show in section 6.2. This leads to a second insight: without consumer level data, (θ, δ) would not be identified by the likelihood alone as there are more parameters than share constraints.

The maximum likelihood objective makes full use of the consumer choice data (micro and macro). 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. However, it does not incorporate the product level exogeneity restrictions.

3.2 Product Level Moments

The second term of our objective function penalizes violations of the product level moments,

$$\hat{\Pi}(\beta, \delta) = \frac{1}{2}\hat{m}^{\intercal}(\beta, \delta)\hat{\mathcal{W}}\hat{m}(\beta, \delta). \tag{12}$$

where for $J = \sum_{m=1}^{M} J_m$, $J\hat{\mathcal{W}}$ is the optimal GMM weight matrix for $\hat{m}$ scaled to converge to the inverse of $\mathbb{V}(b_{jm}\xi_{jm})$ and

$$\hat{m}(\delta, \beta) = \sum_{m=1}^{M}\sum_{j=1}^{J_m} b_{jm}(\delta_{jm} - \beta^{\intercal}x_{jm}). \tag{13}$$

Note that, unlike in standalone GMM estimation, the factor $1/2$ in front of the 'J statistic' in (12) matters since it affects the relative weight placed on the likelihood and moment components of the objective function: the choice $1/2$ is optimal as shall become apparent in section 4.1.

If the dimension of b_{jm} is the same as that of β, a situation we shall refer to as "exact identification of β" then θ, δ are estimated off the likelihood portion and β off the GMM portion. Our estimator is then equivalent to a two-step estimator which estimates θ, δ off the likelihood and subsequently estimates β off $\hat{\Pi}$. Additional restrictions result in overidentification of β which can be used to aid the estimation of θ. Indeed, then $\hat{\Pi}$ will generally be positive so that both $\log \hat{L}$ and $\hat{\Pi}$ contribute to the estimation of θ, δ. However, because the micro log likelihood sums over $S = \sum_{m=1}^{M} S_m$ terms whereas $\hat{\Pi}$ involves sums over J terms these additional restrictions can be asymptotically negligible for θ, δ as we discuss in section 4.1.

4 Properties

Our estimator combines two sources of information based on the model: consumer choice decisions on the individual and aggregate level, and product level exogeneity restrictions. These sources have identifying information for overlapping sets of parameters. Moreover, the empirical content of these alternative sources will vary based on the shape of the dataset and the true values of the parameters. In this section, we establish that our estimator is *conformant* in the sense that it achieves the optimal convergence rate under multiple alternative divergence rates of $\{N_m\}, S, J$ and exploitable variation in the data;[11] moreover, it is efficient in all of these settings. The conformance property

[11] We use the term 'conform' instead of 'adapt' to avoid confusion with the adaptive estimation literature.

implies that a researcher can be confident in using our estimator without knowing or testing the precise conditions she is facing.

For clarity, we first informally argue in section 4.1 that our estimator is efficient without making reference to its convergence rates.[12] Section 4.2 then establishes the convergence rates of the estimator under a wide variety of circumstances, completing the efficiency argument. Section 8 provides a valid inference procedure.

4.1 Efficiency

Our proposed estimator is efficient under a wide range of circumstances. To see this, it is convenient to first consider the gradient of our objective function,[13]

$$\begin{bmatrix} \partial_\beta \hat{m}^\intercal \hat{\mathcal{W}} \hat{m} \\ -\partial_\theta \log \hat{L} \\ -\partial_\delta \log \hat{L} + \partial_\delta \hat{m}^\intercal \hat{\mathcal{W}} \hat{m} \end{bmatrix}. \tag{14}$$

We first show asymptotic equivalence of a GMM estimator using this gradient to the GMM estimator defined as:

$$(\hat{\beta}, \hat{\theta}, \hat{\delta}) = \arg\min_{\beta,\theta,\delta} \frac{1}{2} \begin{bmatrix} \hat{m}^\intercal & \partial_{\psi^\intercal} \log \hat{L} \end{bmatrix} \begin{bmatrix} \hat{\mathcal{W}} & 0 \\ 0 & \hat{\mathcal{W}}_L \end{bmatrix} \begin{bmatrix} \hat{m} \\ \partial_\psi \log \hat{L} \end{bmatrix}, \tag{15}$$

where $\psi = [\theta^\intercal, \delta^\intercal]^\intercal$ and $\hat{\mathcal{W}}_L = \left(-\partial_{\psi\psi^\intercal} \log \hat{L}\right)^{-1}$ evaluated at the solution $\hat{\psi}$ of (8).[14] Note that in (15) there may be more moments than parameters. Specifically, (14) has $d_\beta + d_\theta + d_\delta$ moments, whereas (15) is based on $d_b + d_\theta + d_\delta$ moments. Under exact identification of (15), i.e. if $d_b = d_\beta$, both (14) and (15) are equal to zero if $\hat{m} = 0$, $\partial_\theta \log \hat{L} = 0$, and $\partial_\delta \log \hat{L} = 0$. In the case of overidentification, the gradient of the objective function in (15) is

$$\begin{bmatrix} \partial_\beta \hat{m}^\intercal \hat{\mathcal{W}} \hat{m} \\ 0_{d_\theta} \\ \partial_\delta \hat{m}^\intercal \hat{\mathcal{W}} \hat{m} \end{bmatrix} + \begin{bmatrix} 0_{d_\beta} \\ \partial_{\theta\psi^\intercal} \log \hat{L} \hat{\mathcal{W}}_L \partial_\psi \log \hat{L} \\ \partial_{\delta\psi^\intercal} \log \hat{L} \hat{\mathcal{W}}_L \partial_\psi \log \hat{L} \end{bmatrix}, \tag{16}$$

which yields (14) at the solution since $\hat{\mathcal{W}}_L = \left(-\partial_{\psi\psi^\intercal} \log \hat{L}\right)^{-1}$, establishing the equivalence of these estimators.

Next, we argue that (15) is efficient. First, by the law of iterated expectations, at the truth,

$$\mathbb{E}\Big(\partial_\psi \log \hat{L}\ \hat{m}^\intercal\Big) = \mathbb{E}\Big(\mathbb{E}(\partial_\psi \log \hat{L} \mid x, \xi)\ \hat{m}^\intercal\Big) = 0,$$

where the second equality follows from the the likelihood principle applied to the choice problem (without product level moments); see appendix G.1 for details. The intuition for this result follows from the fact the inner expectation is over the consumer level shocks ϵ, whereas ϵ does not enter

[12] As we shall see, different elements may converge at different rates.

[13] Imbens and Lancaster (1994) combine a likelihood score with moments.

[14] We define $\hat{\mathcal{W}}_L$ in terms of (8) in case its gradient (14) is zero at multiple points.

the product level moments. Moreover, $-\hat{\mathcal{W}}_L$ is the scaled inverse information matrix of the choice problem and we assumed $\hat{\mathcal{W}}$ is the appropriately scaled optimal weight matrix of the product level moments. Therefore, this choice of weight matrix is optimal.

Despite their asymptotic equivalence, there are two reasons to prefer our estimator to the GMM estimators described in (14) and (15). First, the population analog of (14) can have multiple solutions even if the population analog of our objective function (8) has a unique optimum. For example, in the typical case where the ν_{im} are independent standard normal draws and θ^ν represents scale parameters, $\partial_{\theta^\nu} \log \hat{L} = 0$ for any parameter vector where $\theta^\nu = 0$; setting $\theta^\nu = 0$, the remaining parameters can be chosen to satisfy the rest of the score, albeit that the likelihood is then not optimized. The second reason is that computing (15) would be unwieldy because of the high degree of nonlinearity and the dimension of δ. We show in section 7 that the estimator defined in (8) can be tractably computed despite the dimensionality of δ.

4.2 Conformant convergence

We now show that our estimator is conformant. The objective function in (8) is the sum of three terms that diverge at different rates. The micro loglikelihood is the sum over S consumers, the macro loglikelihood in (10) is the sum over N consumers, and $\hat{\Pi}$ is a quadratic that diverges at rate J. Moreover, as we illustrate in section 5, the identifying power of the micro data depends on the value of θ^z. As a consequence, the rates of convergence of $\hat{\theta}^z, \hat{\theta}^\nu, \hat{\delta}$ differ across cases depending on S/J and θ^z. In contrast, the convergence rate of $\hat{\beta}$ is always $\sqrt{J}$ since it is only identified off $\hat{\Pi}$.

The remainder of this section enumerates cases defined in terms of (relative) divergence rates to which our estimator conforms. Since the convergence rate of $\hat{\beta}$ is always $\sqrt{J}$ we focus on the convergence rates of $\hat{\theta}, \hat{\delta}$. We first make explicit the following assumptions, which we maintain throughout. First, the market size N_m in any given market m diverges faster than the total number of products across all markets, J, i.e. $\min_m N_m/J \to \infty$. This is to ensure that market shares can be consistently estimated. This assumption is weaker than assuming $N_m = \infty$ since N_m need not diverge faster than S and we have not specified how much faster than J. In addition, we assume that the J_m's are fixed and that $\lim_{M\to\infty} \max_m J_m < \infty$. This ensures that the choice probabilities in each market are constant as the data grows and that observed market shares vary only due to the addition of consumers (i.e., as N_m grows).[15] We will further assume that the instruments b_{jm} used in $\hat{m}$ are strong in the standard sense (Staiger and Stock, 1997) and there are enough moments to ensure identification. If b_{jm} were weak then that poses additional challenges outside the scope of our work.

We begin with the simpler cases in which the ratio S/J is allowed to vary for given values of the model parameters. It turns out that if $\theta^z = 0$ then the micro data alone is insufficient to distinguish (θ^ν, δ), which affects convergence rates.In section 4.2.2 we then cover cases in which θ^z is allowed to drift in the spirit of the weak identification literature. These cases are critical since ex ante the researcher does not know the value of θ^z: if θ^z were close to zero then it is unclear which fixed case

[15] This is in contrast to Berry, Linton and Pakes (2004) which assumes that the number of markets is fixed.

(if either) is appropriate.

For ease of exposition, we assume in the remainder of this subsection that S diverges no faster than N_m. If this assumption is not satisfied then some of the $\sqrt{S}$ rates will slow to $\sqrt{N_m}$. Section 4.3 will relax this assumption.

4.2.1 θ^z is fixed

	rate		contributing term(s)	
case	θ^z	θ^ν, δ	for θ^z	for θ^ν
$S/J \to \infty$, $\theta^z \neq 0$	$\sqrt{S}$	$\sqrt{S}$	$\log \hat{L}$	$\log \hat{L}$
$S/J \to \infty$, $\theta^z = 0$	$\sqrt{S}$	$\sqrt{J}$	$\log \hat{L}$	$\hat{\Pi}$
$S/J \to c$, $\theta^z \neq 0$	$\sqrt{J}$	$\sqrt{J}$	both	both
$S/J \to c$, $\theta^z = 0$	$\sqrt{J}$	$\sqrt{J}$	both	$\hat{\Pi}$
$S/J \to 0$	$\sqrt{J}$	$\sqrt{J}$	$\hat{\Pi}$	$\hat{\Pi}$

Table 1: Convergence rates of the proposed estimator and terms contributing to the limit distribution in addition to the macro likelihood when θ^z is fixed and there are sufficiently many moments in $\hat{\Pi}$ to ensure identification (where needed).

Table 1 lists several cases where the parameters are fixed. They are ordered by importance of the $\log \hat{L}^{\text{micro}}$ term for the asymptotic behavior of (8).

In the first two rows, the size of the micro sample S diverges faster than the number of products J, which we view as the typical case. Then the $\log \hat{L}$ term of our objective function diverges faster than $\hat{\Pi}$. If $\theta^z \neq 0$, then the likelihood provides identification and yields an efficient estimator of $(\hat{\theta}, \hat{\delta})$. So the addition of $\hat{\Pi}$ is then asymptotically irrelevant for $(\hat{\theta}, \hat{\delta})$.[16] Of course, using $\log \hat{L}$ alone, we would be unable to recover β. However, a two step estimator in which θ, δ are estimated off $\log \hat{L}$ in the first stage and β is estimated by minimizing $\hat{\Pi}(\beta, \hat{\delta})$ in the second stage, is equivalent to our estimator (and hence also efficient). This holds even in the case of overidentification in $\hat{\Pi}$ since the additional moments do not alter the fact that $\hat{\Pi}$ diverges at the slower rate J.

However, if $\theta^z = 0$ (the second row) then $\log \hat{L}$ fails to identify all the parameters. In this case utilities and hence choice probabilities do not vary with demographics z (as we illustrate in section 5). Thus, the θ^ν and δ scores of the micro likelihood are then collinear. To see this, note that if $\theta^z = 0$ then $\jmath_{jm}(z, \nu)$ is flat in z and hence the scores with respect to θ^ν, δ then depend on the micro data only through $\sum_{i=1}^{N_m} D_{im} y_{ijm}$.[17] As a result, θ^ν, δ are not identified off $\log \hat{L}$. In this case, $\hat{\Pi}$ provides identification as we have assumed the moments are sufficient to identify θ^ν. Consequently, the convergence rate of $\hat{\theta}^\nu$ and $\hat{\delta}$ slows to $\sqrt{J}$. In contrast, θ^z is still identified by the micro likelihood because the score with respect to θ^z depends on $\sum_{i=1}^{N_m} D_{im} y_{ijm} z_{im}$ when $\jmath_{jm}$ is flat

[16] We implicitly assume sufficient variation in z to identify all random coefficients; there can be intermediate cases. See the discussion at the end of section 5.

[17] The scores of $\log \hat{L}^{\text{micro}}$ with respect to θ^z and θ^ν are in (23) and (25). The score with respect to δ_{jm} is $\sum_{i=1}^{N_m} \sum_{\ell=0}^{J_m} (D_{im} y_{i\ell m} \,/\, \pi_{\ell m}^{z_{im}}) \int \jmath_{\ell m}(z_{im}, \nu) \Big(\mathbb{1}(\ell = j) - \jmath_{jm}(z_{im}, \nu) \Big) \, \mathrm{d}F(\nu)$.

case	rate θ^z	rate θ^ν, δ	contributing term(s) for θ^z	contributing term(s) for θ^ν
$\theta^z\sqrt{S/J} \to \infty$, $S/J \to \infty$	$\sqrt{S}$	$\sqrt{S}$	$\log \hat{L}$	$\log \hat{L}$
$\theta^z\sqrt{S/J} \to c$, $S/J \to \infty$	$\sqrt{S}$	$\sqrt{J}$	$\log \hat{L}$	both
$\theta^z\sqrt{S/J} \to 0$, $S/J \to \infty$	$\sqrt{S}$	$\sqrt{J}$	$\log \hat{L}$	$\hat{\Pi}$
$\theta^z\sqrt{S/J} \to c$, $S/J \to c$	$\sqrt{J}$	$\sqrt{J}$	both	both
$\theta^z\sqrt{S/J} \to 0$, $S/J \to c$	$\sqrt{J}$	$\sqrt{J}$	both	$\hat{\Pi}$
$\theta^z\sqrt{S/J} \to 0$, $S/J \to 0$	$\sqrt{J}$	$\sqrt{J}$	$\hat{\Pi}$	$\hat{\Pi}$

Table 2: Convergence rates of the proposed estimator and terms contributing to the limit distribution in addition to the macro likelihood when θ^z can drift and there are sufficiently many moments in $\hat{\Pi}$ to ensure identification (where needed).

in z, so the rate of $\hat{\theta}^z$ continues to be $\sqrt{S}$.

We now move to the cases where S/J converges to a nonzero constant. Here, the micro term $\log \hat{L}^{\text{micro}}$ of the loglikelihood and $\hat{\Pi}$ diverge at the same rate, and all parameter estimates converge at the same rate $\sqrt{J} \sim \sqrt{S}$. However, our estimator is still more efficient than alternatives since it combines both terms optimally. There remains a distinction when $\theta^z = 0$ since again $\log \hat{L}$ has no identifying demographic variation to pin down θ^ν and so only $\hat{\Pi}$ contributes to the limiting distribution for this parameter.

Finally, we consider the case where $S/J \to 0$. Now $\hat{\Pi}$ diverges faster than the micro loglikelihood $\log \hat{L}^{\text{micro}}$. Consequently, if $d_b \geq d_\beta + d_\theta$ then $\hat{\Pi}$ will deliver the asymptotics. However, if $d_\beta + d_{\theta^\nu} \leq d_b < d_\beta + d_{\theta^\nu} + d_{\theta^z}$ and S diverges then the micro likelihood will contribute to the limit distribution and the convergence rate will be $\sqrt{S}$ instead of the $\sqrt{J}$ rate displayed in the table. An extreme example of this case arises when $S = 0$, so $\log \hat{L}^{\text{micro}} = 0$. This is the environment of Berry, Levinsohn and Pakes (1995) and both estimators are equally efficient under the assumptions of this section, albeit that ours would be more efficient if $\min_m N_m/J \not\to \infty$ because ours does not impose the share constraint; see section 6.2.

4.2.2 θ^z can drift

Note that in section 4.2.1 there is a discontinuity in the asymptotic behavior of our estimator between the $\theta^z = 0$ and $\theta^z \neq 0$ cases. In order to address this discontinuity, we now extend our discussion by allowing θ^z to drift, i.e. to depend on S, J.[18] Table 2 summarizes these cases, which are again ordered in decreasing importance of the micro likelihood for asymptotic behavior of (8).

In the first row in table 2, $\theta^z\sqrt{S/J} \to \infty$ which is equivalent to the first row of table 1 in terms of asymptotic behavior.

In the next two cases, $\log \hat{L}^{\text{micro}}$ diverges faster than $\hat{\Pi}$, but the two cases differ in the strength of identification they provide due to $\theta^z \to 0$ at different rates. The knife edge case where the rate of

[18] We can also let σ_ξ, the standard deviation of ξ_{jm} drift, which alters the explanatory power of $\hat{\Pi}$ instead of that of $\log \hat{L}$. We believe that the θ^z close to zero case is of greater concern in applied work than σ_ξ close to zero.

θ^z is such that $\theta^z\sqrt{S/J}$ goes to a constant has no analog in table 1. Here both $\log \hat{L}^{\text{micro}}$ and $\hat{\Pi}$ contribute to the limit distribution of $\hat{\theta}^\nu$ because the faster divergence of $\log \hat{L}^{\text{micro}}$ is just offset by the convergence of θ^z. The case where $\theta^z\sqrt{S/J} \to 0$ is effectively equivalent to the second case of table 1 where $\theta^z = 0$.

The final three cases all have direct analogs in the final three rows of table 1.

4.3 Summary

What the above discussion has illustrated is that it is optimal to rely on the variation in the micro data alone to identify $\theta^z, \theta^\nu, \delta$ if the micro sample is large and demographic variation affects choice probabilities substantially. Otherwise, $\hat{\Pi}$ becomes useful. Both our estimation and inference procedures automatically conform so that one does not have to test which situation one is in.

Table 3 summarizes these ideas. We compare our method to two alternatives under the maintained assumptions that $S/J \to \infty$ and that the overidentifying moments in $\hat{\Pi}$ are sufficient to identify θ^ν (which requires $d_b \geq d_\beta + d_{\theta^\nu}$).

First consider the leading case where $\theta^z \neq 0$ is fixed. We have already described the behavior of our estimator in table 1. The first alternative in table 3 is the two-step estimator described in section 4.2.1, which in this case is asymptotically equivalent to our method. The second alternative, relying on $\hat{m}$ rather than the micro sample to provide identification for θ^ν would occur if one dropped the θ^ν gradients from (15), which had $d_b + d_{\theta^z} + d_{\theta^\nu} + d_\delta$ moments for $d_\beta + d_{\theta^z} + d_{\theta^\nu} + d_\delta$ parameters. Doing so slows down the convergence rate to $\sqrt{J}$ for $\hat{\theta}, \hat{\delta}$.

We now generalize to the case in which θ^z is drifting toward zero at rate λ, a case that was first discussed in section 4.2.2. For our estimator, the rate λ determines which of the first three rows in table 2 applies. The first alternative, on the other hand, could do poorly if λ converges to zero fast. In the extreme, i.e. if $\theta^z = 0$, this estimator is inconsistent. The second alternative estimator is not affected by the fact that the likelihood provides less information than in the leading case, because it was not using that information anyway. Our proposed method uses both sources of information and hence converges at the faster rate of the two alternative estimators, which can nevertheless be slower than in the leading case.

5 Identifying unobserved heterogeneity from micro data

Above, we have highlighted that the micro likelihood can efficiently use the information in the micro sample to estimate consumer heterogeneity parameters θ. We now turn to a specific example to illustrate the underlying variation in the micro sample that provides identification.

We begin this exercise with a simple case of a single market with two products and an outside good.[19] There is a single demographic variable, so z_i is a scalar. Utility for product j is,

$$u_{ij} = \delta_j + \theta^z x_j^{(1)} z_i + \theta^\nu x_j^{(2)} \nu_i + \varepsilon_{ij},$$

[19] Since there is a single market in this section, we drop m from the notation.

leading case ($S/J \to \infty$, $\theta^z \neq 0$)

Estimation method	$\hat{\delta}$	$\hat{\theta}^z$	$\hat{\theta}^\nu$	$\hat{\beta}$
Proposed Method: $-\log \hat{L} + \hat{\Pi}$	$\min(\sqrt{S}, \sqrt{N_m})$	$\sqrt{S}$	$\sqrt{S}$	$\sqrt{J}$
Two Step: $-\log \hat{L}$, then $\hat{\Pi}$	$\min(\sqrt{S}, \sqrt{N_m})$	$\sqrt{S}$	$\sqrt{S}$	$\sqrt{J}$
Rely on $\hat{\Pi}$ to identify θ^ν	$\sqrt{J}$	$\sqrt{S}$	$\sqrt{J}$	$\sqrt{J}$

more general case: ($S/J \to \infty$, $\theta^z \sim \lambda$)

Estimation method	$\hat{\delta}$	$\hat{\theta}^z$	$\hat{\theta}^\nu$	$\hat{\beta}$
Proposed Method: $-\log \hat{L} + \hat{\Pi}$	$\min\{\max(\sqrt{J}, \sqrt{S}\lambda), \sqrt{N_m}\}$	$\sqrt{S}$	$\max(\sqrt{J}, \sqrt{S}\lambda)$	$\sqrt{J}$
Two Step: $-\log \hat{L}$, then $\hat{\Pi}$	$\min\{\sqrt{S}\lambda, \sqrt{N_m}\}$	$\sqrt{S}$	$\sqrt{S}\lambda$	$\min(\sqrt{S}\lambda, \sqrt{J})$
Rely on $\hat{\Pi}$ to identify θ^ν	$\sqrt{J}$	$\sqrt{S}$	$\sqrt{J}$	$\sqrt{J}$

Table 3: Rates of convergence with product level moments if $d_b \geq d_\beta + d_{\theta^\nu}$

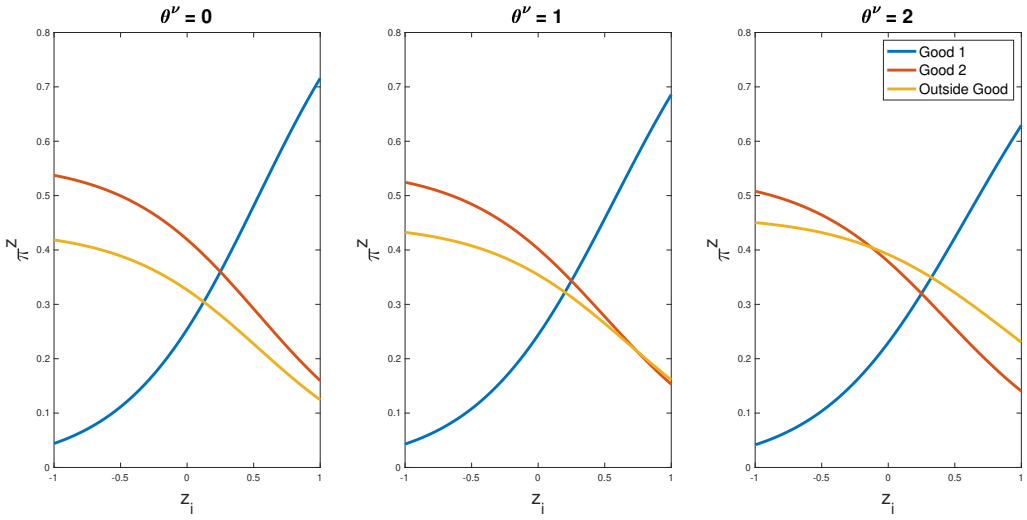

Figure 1: Conditional shares $\tilde{\pi}^z$ are identified by the micro sample.

where the product characteristics are,

$$x^{(1)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

So the demographic characteristic shifts utility of only good 1, and the single random coefficient induces correlation in the utilities of the two inside goods.[20] As is typical, in this example ν_i has a standard normal distribution.

Suppose we observe a random sample of microdata $\{y_{i\cdot}, z_i\}$. The micro data nonparametrically identifies the function $\tilde{\pi}^z = \Pr(y_{i\cdot} = 1 \mid z, x)$. We plot this function over $z \in [-1, 1]$ in figure 1 for three different parametrization of the model, namely $\theta^\nu = \{0, 1, 2\}$ with $\delta = (-.25, 25)^\intercal$ and $\theta^z = 2$ fixed. Intuitively, the share of good 1 rises with z in all three panels. However, the slope differs based on the value of θ^ν. The other notable difference is that as θ^ν increases, z has a larger impact on the share of good 2, $\tilde{\pi}^z_2$, relative to the outside good, $\tilde{\pi}^z_0$. Since the utilities of goods 1 and 2 are increasingly correlated as θ^ν grows, it becomes more likely that consumers are on the margin between the two inside goods than between good 1 and the outside good. Therefore, a slight increase in z will induce relatively more substitution away from good 2 than from the outside good.

We can also nonparametrically identify the derivatives of conditional choice probabilities with respect to z. Given our special case we have,

$$\mathrm{d}_z \tilde{\pi}^z_j = \theta^z \partial_{u_1} \pi^z_j,$$

where we employ the fact that z only affects the utility of good 1. Taking a ratio of these gives us diversions with respect to utility from good 1 to good 2 and from good 1 to the outside good for

[20] This is analogous to nesting out the outside good in a nested logit model.

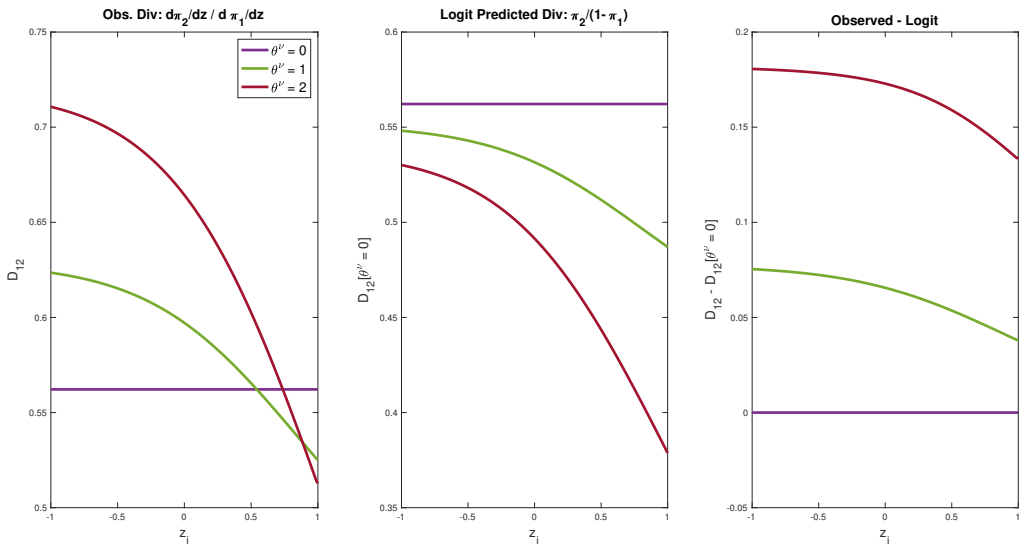

Figure 2: Diversion and Demographics

every value of z, i.e., for $j = \{0, 2\}$,

$$\frac{\mathrm{d}_z \tilde{\pi}_j^z}{\mathrm{d}_z \tilde{\pi}_1^z} = \frac{\partial_{u_1} \pi_j^z}{\partial_{u_1} \pi_1^z} = D_{1j}^z. \tag{17}$$

Equation (17) provides intuitive variation with which to identify θ^ν. To see this, recall that when $\theta^\nu = 0$ then we have multinomial logit demand. This implies that diversion is a function of conditional choice probabilities: if $\theta^\nu = 0$ then $D_{1j}^{zi} = \pi_j^z/(1 - \pi_1^z)$. Moreover, due to the independence of irrelevant alternatives property, diversion will be constant over z. Figure 2 illustrates the implications of this.

The first panel depicts diversion with respect to utility from good 1 to good 2 as a function of z, i.e. D_{12}^z. As predicted, diversion is constant in z for $\theta^\nu = 0$, yet it is decreasing for $\theta^\nu > 0$. The reason for the decline can be seen in figure 1: as z increases, the conditional share of good 2 falls more rapidly for $\theta^\nu > 0$, so a larger proportion of switchers must come from the outside good in response to an increase in z.

The second panel of figure 2 plots the logit-implied diversion ratios computed from conditional shares generated by the three parametrizations of θ^ν. If $\theta^\nu = 0$, we exactly reproduce the constant diversion rate from the first panel. For $\theta^\nu > 0$, we see decreasing functions that are below the line for $\theta^\nu = 0$. The reason these functions are decreasing is the same as for the first panel. The reason the level of the logit-predicted diversion decreases in θ^ν is that in fact diversion between goods 1 and 2 is more than proportional to shares when $\theta^\nu > 0$. An illustration of diversion between good 1 and the outside good would produce a mirror image since as θ^ν rises, this weakens diversion between these goods.

The third panel of figure 2 takes the difference of the first two panels and so provides the

difference between observed diversion and diversion implied by assuming $\theta^\nu = 0$. As θ^ν rises, the logit model under-predicts diversion between the two inside goods. Moreover, the degree of under-prediction varies in z. This suggests moments with which to identify θ^ν by comparing the estimated diversion rate to the model-predicted diversion rate. In this exercise we have fixed the values of the other parameters θ^z and δ. In practice, the described moments for θ^ν would need to be paired with commonly used moments to identify θ^z, δ; e.g., matching market shares for δ and matching correlations between demographics and product characteristics for θ^z. As always, an advantage of the likelihood approach to using moments is that it fully exploits all of the information in the micro sample.

So far, we have focused on a special case in which it is clear that the micro sample has so much valuable information to identify θ^ν that the $\hat{\Pi}$ term of our estimator would be redundant. To see a case where $\hat{\Pi}$ is necessary for identification, simply set $\theta^z = 0$ in our example. Now $\partial_z \tilde{\pi}_j^z = 0$ and we cannot use variation in demographics to recover diversion between goods. Consequently, the moments we have suggested are no longer informative. This provides a role for overidentifying restrictions in $\hat{m}$ to aid in the estimation of θ^ν, albeit that $\hat{\theta}, \hat{\delta}$ then converge at a slower rate, as discussed above.

Another complication is the richness of the demographic variation provided by z and the flexibility of the random coefficient specification. In the simple example, we specified z to shift the utility of exactly one good and restricted θ^ν to have dimension one. While diversion identified in this simple example, this may not be possible in general. For example, μ^z is typically specified as a linear combination of interactions between product characteristics and consumer demographics, e.g.,

$$\mu^z(x_j, z_i; \theta^z) = x_j^\intercal \Theta^z z_i = \sum_k \sum_d \theta^{z(k,d)} x_j^k z_i^d,$$

where Θ^z is a matrix with elements $\theta^{z(k,d)}$. With this form we have,

$$\mathrm{d}_{z^d} \tilde{\pi}_j^z = \sum_{k=1}^{K} \sum_{\ell=1}^{J} \theta^{z(k,d)} x_\ell^k \partial_{u_\ell} \pi_j^z. \tag{18}$$

In matrix notation, (18) can be written as

$$\mathrm{d}_{z^\intercal} \tilde{\pi}^z = \partial_{u^\intercal} \pi^z \partial_{z^\intercal} u = \partial_{u^\intercal} \pi^z \partial_{z^\intercal} \mu^z = \partial_{u^\intercal} \pi^z X^\intercal \Theta^z. \tag{19}$$

Thus, only if $X^\intercal \Theta^z$ has maximum column rank, does there exist a unique $\partial_{u^\intercal} \pi^z$ that solves (19). In other words, if this rank condition holds, then we can recover the substitution matrix for all z from θ^z and the data. Flexibility of the substitution matrix is the primary motivation for the introduction of random coefficients. Since the introduction of θ^ν imposes parametric structure, *nonparametric* identification of the full substitution matrix is not necessary for the identification of θ^ν.

The most general specification of μ^ν would let x be product dummies. Then, if ν were distributed mean zero normal such that θ^ν would be $J(J+1)/2$-dimensional (J variances and $J(J-1)/2$

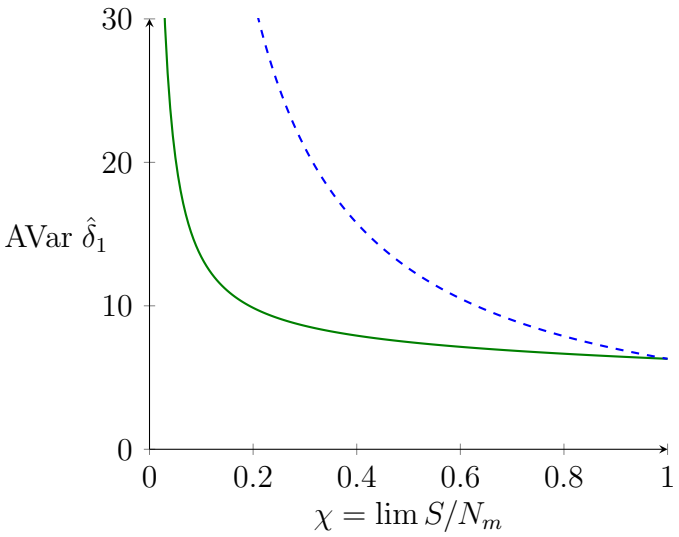

Figure 3: Asymptotic variance of $\hat{\delta}_1$ for the mixed logit estimator (dashed blue) versus our estimator (solid green) as the size of the micro sample S grows relative to the market size N_m for the specification described in example 1 with $\delta_1 = \theta^z = 1$.

correlations), one would have the same number of unknowns as there are restrictions in (19). Applied work typically imposes restrictions to reduce the dimension of θ^ν by introducing random coefficients on product characteristics instead of on products and restricting ν_i to be independent across its elements. If the rank condition on $X^\intercal \Theta^z$ fails, we still have restrictions like (19) that may or may not pin down some or all elements of θ^ν depending on the specification of μ^ν.

6 Comparison with Alternative Estimators

To clarify the contribution of our estimator, we now show how our estimator relates to two estimators used in the discrete choice literature.

First, as noted above, with $S = N$ our $\log \hat{L}$ simplifies to the mixed logit loglikelihood. If $S < N$, the only difference is that $\log \hat{L}$ exploits the market share data via the macro term. This term is particularly useful when J is large relative to S, since then there would otherwise be an incidental parameters problem in estimating δ. More generally, market share data can dramatically improve the precision of the estimator as illustrated in figure 3, which uses example 1 described below.

The other major class of estimators used in applied work consists of share constrained GMM estimators (e.g., Berry, Levinsohn and Pakes, 2004; Petrin, 2002; Grieco, Murry and Yurukoglu, 2021).[21] The remainder of this section shows how our estimator can be converted into members of this class of estimators. As we have shown above, our estimator is efficient, so we will point out

[21] An alternative class of share constrained micro likelihood estimators (e.g., Goolsbee and Petrin, 2004; Chintagunta and Dube, 2005; Train and Winston, 2007; Goeree, 2008; Bachmann et al., 2019) also derives from our estimator by only imposing share constraints on our estimator without recasting it as a GMM problem as described by the dotted line in Figure 4.

losses of efficiency along the way. There may be a tradeoff between efficiency and computational tractability that justifies using an inefficient estimator. So we also discuss these tradeoffs in this section. With that said, it is important to keep in mind that the marginal cost of computational resources tends to be less than that of data, and also decreases more quickly over time. We argue for the computational tractability of our estimator in section 7.

Figure 4 provides a summary of the steps described below. The highest node in the tree represents our estimator. Each node below represents an alteration to arrive at an alternative estimator. The large pink box representing section 6.3 proposes three alternative alterations for linearizing the score with respect to θ^ν as described in section 6.3.2. One can stop the process at any node in the tree, so in total the figure describes nine alternative estimators (including share constrained likelihood, see footnote 21). At each node, we briefly list the primary costs (red) and benefits (green) of the step relating to econometric efficiency (✈), inference (📼), computational tractability (💻), data requirements (**$$**) and experience in applied work (**??**). Each step downward in the tree leads to an estimator that is weakly less efficient than its parent. To our knowledge, all estimators that have been applied in empirical work on discrete choice demand are covered here.

6.1 Step 1: A GMM version of our estimator

In section 4.1, we presented a GMM estimator (15) which is asymptotically equivalent to our estimator, assuming that (15) does not lose identification; as we pointed out in section 4.1, going from minimizing the objective function (8) to setting its derivatives (14) (or indeed (15)) to zero can lose identification due to the existence of multiple (local) optima.

Note that for equivalence to obtain, it is essential that the $\hat{\mathcal{W}}_L$ and $\hat{\mathcal{W}}$ matrices used in (15) have the norming indicated in section 4.1: unlike in standard GMM the convergence *rate* of the GMM estimator can be affected by a poor choice of weight matrix. The reason for this is that one entails a sum over consumers whereas the other is a sum over products.

While GMM estimators are used to avoid parametric distributional assumptions, this rationale does not apply in this case. Indeed, GMM estimators discussed in this paper also use the distributional assumptions on ν, ϵ for the moments, and $\hat{\Pi}$ in (8) similarly avoids distributional assumptions on ξ.

Our estimator has an important computational advantage over (15): it is convex in δ. Since δ is high dimensional this convexity is important. In fact, the next step is driven by addressing the computational complexity introduced here.

6.2 Step 2: Imposing share constraints

To resolve the dimensionality issue in (15) one can impose share constraints $\pi = s$.[22] Following the intuition of Berry (1994) this is equivalent to treating δ as a deterministic function of θ that is easy to compute and yields a consistent estimator as $N_m, S, J \to \infty$.

However, imposing the share constraints introduces a potential loss of efficiency. Suppose that

[22] Share constraints can also be imposed on $\log \hat{L}$ directly, but doing this serves no purpose.

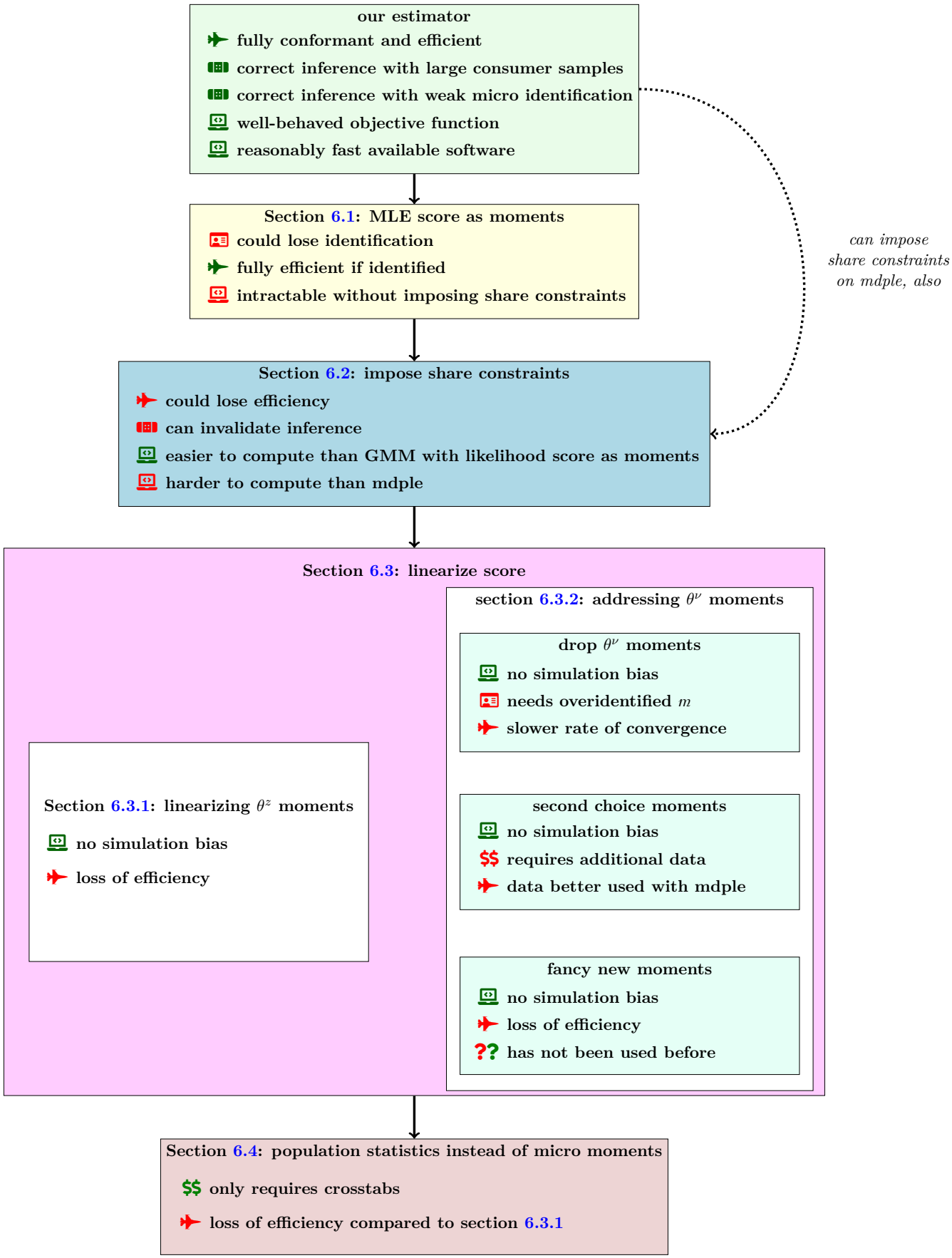

Figure 4: Schematic comparison of our estimator to alternatives. See text for details.

$\theta^z \neq 0$ such that the MDLE and our estimator are asymptotically equivalent. Then this efficiency loss occurs unless the population in the *smallest* market diverges faster than the total number of consumers in the micro sample across *all* markets S and the total number of products J. Theorem 1 establishes this result for the single market case.

Theorem 1. *Suppose that there is a single market with a finite number of products J 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 of δ, θ.* □

The 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 being 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. eIn practice, imposing the share constraint can lead to a substantial efficiency loss as examples 1 and 2 demonstrate.[23]

We can intuitively understand this result by considering the share constrained estimator as a GMM estimator with infinite weight on a subset of moments. Specifically, suppose that one separates out the micro and macro terms of $\log \hat{L}$ as specified in (10) and considers the derivative of the macro term with respect to δ, i.e. for all $m = 1, \ldots, M$ and all $j = 1, \ldots, J_m$,

$$\sum_{\ell=0}^{J_m} \frac{s_{\ell m}}{\pi_{\ell m}} \int \imath_{\ell m}(z,\nu)\Big(\mathbb{1}(\ell = j) - \imath_{jm}(z,\nu)\Big)\,\mathrm{d}F(\nu)\,\mathrm{d}G(z) = 0, \tag{20}$$

where $\imath$ was defined in (5). If $s = \pi$, then the left hand side in (20) becomes

$$\int \imath_{jm}(z,\nu)\,\mathrm{d}F(\nu)\,\mathrm{d}G(z) - \int \imath_{jm}(z,\nu) \underbrace{\sum_{\ell=0}^{J_m} \imath_{\ell m}(z,\nu)}_{=1}\,\mathrm{d}F(\nu)\,\mathrm{d}G(z). \tag{21}$$

So setting $s = \pi$ solves (20). Since the macro loglikelihood is concave in δ this solution is unique. Therefore, imposing share constraints effectively places infinite weight on this moment. It is well known from standard GMM theory that placing infinite weight on a subset of moments is generally inefficient. As noted, in our setting, there would be an efficiency loss unless S and J were negligibly small compared to N_m because then the macro score runs over more terms than the other moments.

In addition to the efficiency cost, imposing the share constraints also creates challenges for inference. If one treats δ as a deterministic function of θ, one ignores the uncertainty arising from estimating choice probabilities using observed market shares. This will result in a downward bias in

[23] Example 2 is in appendix C.

the standard errors for $\hat{\delta}$. Indeed, for some linear combinations of δ asymptotics are governed by the estimation error in market shares unless S is negligibly small compared to $\min_m \sqrt{N_m}$, as we now explain.

To illustrate, consider inference on $\delta_{\cdot m}$. With the share constraints, it would be tempting to use the delta method to conclude that for any vector $v \neq 0$

$$\frac{\sqrt{S} v^\intercal (\hat{\delta}_{\cdot m} - \delta_{\cdot m})}{\sqrt{v^\intercal \partial_{\theta^\intercal} \hat{\delta}_{\cdot m}(\hat{\theta}) \hat{\mathcal{V}}_\theta \partial_\theta \hat{\delta}_{\cdot m}^\intercal(\hat{\theta}) v}} \xrightarrow{d} N(0, I), \tag{22}$$

where $\hat{\delta}_{\cdot m}(\theta)$ is the share inversion for market m 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 $v^\intercal \partial_{\theta^\intercal} \delta_{\cdot m} = 0$.[24] In this case the left hand side of (22) diverges. The space of such vectors v is of dimension no less than $J_m - d_\theta > 0$ since $\delta_{\cdot m} : \mathbb{R}^{d_\theta} \to \mathbb{R}^{J_m}$. Using the bootstrap the way it is typically used does not solve this problem.[25] The inference problem can be avoided by using the asymptotic variance formulas in appendix B.

We now illustrate the impact of the share constraint on efficiency and inference. For simplicity, $\mu^\nu = 0$ in example 1, so consumer heterogeneity is due only to observed demographics z. While we focus on estimation of δ, with random coefficients the inefficiency issue extends to the θ^z, θ^ν coefficients as well. For additional intuition, we have included another example in appendix C.

Example 1. Consider a single market with one inside product with utility

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

where the consumer characteristic z_i is distributed standard normal. We observe the market share of good 1 and a micro sample of consumer choices of size S such that $S/N \to \chi$. Further, assume that the product level moments $\hat{m}$ just identify β and so only the loglikelihood plays a role in estimating (θ, δ).

Consider the relative asymptotic efficiency of a share constrained estimator to our efficient estimator. When $\chi = 0$, there is no efficiency loss since the micro sample is then negligibly small. As χ grows, our estimator exploits the correlation between the scores of the micro likelihood whereas

[24] Indeed, then by a Taylor expansion,

$$v^\intercal \{\hat{\delta}_{\cdot m}(\hat{\theta}) - \delta_{\cdot m}(\theta)\} \simeq \underbrace{v^\intercal \{\hat{\delta}_{\cdot m}(\hat{\theta}) - \delta_{\cdot m}(\hat{\theta})\}}_{O_p(1/\sqrt{N_m})} + \underbrace{v^\intercal \partial_{\theta^\intercal} \delta_{\cdot m}(\theta)}_{=0} (\hat{\theta} - \theta) + \frac{1}{2} \sum_j v_j \underbrace{(\hat{\theta} - \theta)^\intercal \partial_{\theta\theta^\intercal} \delta_{jm}(\theta) (\hat{\theta} - \theta)}_{O_p(1/S)},$$

such that asymptotics are governed by the first right hand side term unless $S/\sqrt{N_m}$ vanishes.

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

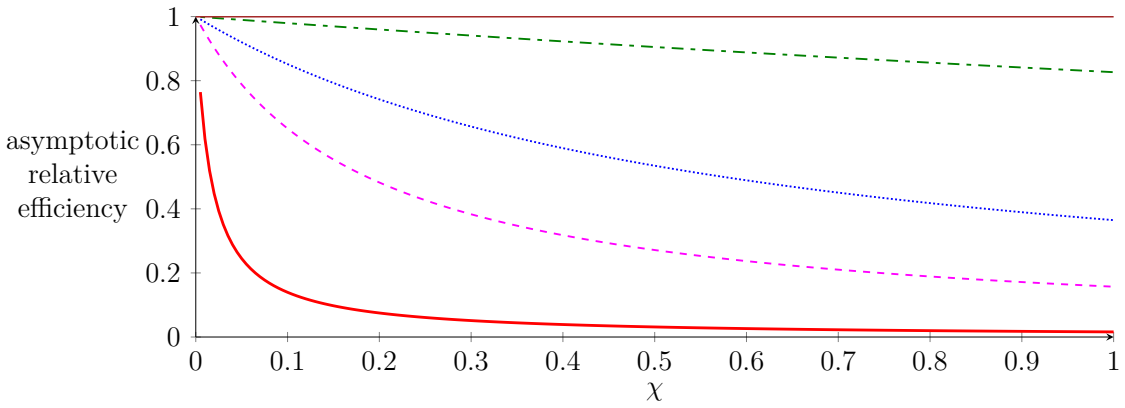

Figure 5: Asymptotic relative efficiency of our estimator of $\delta_1 = 0$ compared to imposing the share constraint as a function of χ for $\theta^z = 0, 1, 4, 10, 100$ (brown, green, blue, magenta, red).

the share constrained estimator does not.[26] Specifically, the share constrained estimator solves

$$\sum_i D_i z_i (y_{i1} - \pi_1^{z_i}) = 0, \qquad s_1 - \pi_1 = 0.$$

Note that the micro sample plays no role in the second moment. In contrast, our estimator uses (11) which incorporates information on the correlation between the scores. In particular, the derivatives of the micro term are $\sum_i D_i z_i (y_{i1} - \pi_1^{z_i})$ and $\sum_i D_i (y_{i1} - \pi_1^{z_i})$, the second of which is the analog of the share constraint in the micro sample. While integrating this term over z leads to the share constraint, doing so loses information. The information loss here is analogous to using a diagonal weight matrix in the context of GMM.

Figure 5 illustrates the impact of the share constraint on efficiency as we vary θ^z and the size of the micro sample as a fraction of the market size. We plot the asymptotic efficiency of the constrained estimator relative to our (efficient) estimator (8). If $\theta^z = 0$ then the correlation between these two moments derived from (11) is zero, so then the share constrained estimator does not lose efficiency. For $\theta^z \neq 0$, the share constrained estimator is inefficient as long as $\chi > 0$. Its inefficiency is increasing in the magnitudes of θ^z and χ. Moreover, the relative asymptotic efficiency of the constrained estimator goes to 0 as $\theta^z \to \infty$ for any $\chi > 0$.

This example also provides a stark illustration of the implications of imposing the share constraint on inference. Suppose $\theta^z = 0$, because $\partial_\theta \hat{\delta}_{.m}^{\intercal}(\hat{\theta})$ converges to $\partial_\theta \delta_{.m}^{\intercal}(\theta) = 0$ for this parameterization (since $\theta^z = 0$ and z has mean 0), the denominator in (22) converges to 0 and so the delta method is invalid.[27] Here the ratio between of the asymptotic variance implied by the delta method and the actual asymptotic variance of the share constrained estimator is 0 (see appendix B), so the standard errors using (22) will be too small. □

[26] If $\chi = 1$ then the entire population is in the micro sample and it is then clearly preferable to use the micro data than to impose share constraints from the macro data. As was illustrated in figure 3, for $\chi = 1$ our estimator and the mixed logit estimator are equally efficient.

[27] While this example appears to be a knife edge case, this is because $J = d_\theta = 1$. In the usual case where $J > d_\theta$ the set of v's for which the denominator in (22) converges to 0 has at least dimension $J - d_\theta$.

To conclude, our estimator has no inference problems and inference can be done using standard extremum estimation techniques. By contrast, the asymptotic variance for the share constrained estimator should be based on the asymptotic variance formulas in appendix B which are based on the moments in (41), also in appendix B, 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.

6.3 Step 3: Adjustments to Likelihood-based Moments

One motivation for using a GMM estimator is to apply the method of simulated moments (MSM) rather than simulated maximum likelihood. With the MSM, the simulated moments also have mean zero at the truth, regardless of the number of simulation draws. Consequently, as Pakes and Pollard (1989) have shown, the MSM estimator has a mean zero normal limit distribution whose convergence rate is the square root of the slower of the *total* number of draws and the number of observations. For example, if the number of draws per observation were fixed then the total number of draws grows proportionally to the number of observations and the convergence rate is the square root of the number of observations, albeit that the asymptotic variance would then be greater. However, the derivatives of the simulated $\log \hat{L}$ do not have mean zero at the truth since they are nonlinear in the simulated integrals. Step 3 replaces the score of the likelihood with approximations that are able to take advantage of the linearity property. This results in a loss of efficiency in return for less computational cost for a given level of numerical (as opposed to statistical) accuracy.

We can focus on the micro score because the macro score in (10) is equal to zero if observed shares are equal to choice probabilities, which we imposed in section 6.2. We can ignore the double counting discrepancy in the micro score between (10) and (11) because the micro score has mean zero in both cases. So we will work with the micro score in (11).

6.3.1 Approximation of θ^z moments for linear simulation error

We first consider the score with respect to θ^z, i.e.

$$\partial_{\theta^{z(k,d)}} \log \hat{L} = \sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} \frac{D_{im} y_{ijm}}{\pi_{jm}^{z_{im}}} \int \imath_{jm}(z_{im}, \nu) \Big(x_{jm}^k z_{im}^d - \sum_{\ell=1}^{J_m} x_{\ell m}^k z_{im}^d \imath_{\ell m}(z_{im}, \nu) \Big) \, \mathrm{d}F(\nu), \quad (23)$$

which is a ratio of two integrals due to the presence of $\pi_{jm}^{z_i}$ in the denominator. An commonly used approximation to the score can be found by setting $\nu = 0$ selectively as follows,

$$\begin{aligned} \partial_{\theta^{z(k,d)}} \log \hat{L} &= \sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} D_{im} y_{ijm} \frac{\int \imath_{jm}(z_{im}, \nu) \Big(x_{jm}^k z_{im}^d - \sum_{\ell=1}^{J_m} x_{\ell m}^k z_{im}^d \imath_{\ell m}(z_{im}, \nu) \Big) \, \mathrm{d}F(\nu)}{\int \imath_{jm}(z_{im}, \nu) \, \mathrm{d}F(\nu)} \\ &\approx \sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} D_{im} y_{ijm} \frac{\int \imath_{jm}(z_{im}, 0) \Big(x_{jm}^k z_{im}^d - \sum_{\ell=1}^{J_m} x_{\ell m}^k z_{im}^d \imath_{\ell m}(z_{im}, \nu) \Big) \, \mathrm{d}F(\nu)}{\int \imath_{jm}(z_{im}, 0) \, \mathrm{d}F(\nu)} \end{aligned}$$

$$= \sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} D_{im}(y_{ijm} - \pi_{jm}^{z_{im}}) x_{jm}^k z_{im}^d, \tag{24}$$

The final line of (24) matches the correlation of demographics and product characteristics in the micro sample to that of the model. This moment is linear in $\pi_{jm}^{z_{im}}$, its only approximated object, so it can be approximated without simulation bias if one uses Monte Carlo integration. However, since the share inversion is a nonlinear transformation of a simulated object, the number of simulations required in the computation of $\delta(\theta)$, which is an argument to s_{jm}, requires the number of those simulation draws to diverge faster than S not to affect efficiency and to avoid having to use a different inference procedure,[28] and at at least the same rate as S in order not to affect the convergence rate.

6.3.2 Handling θ^ν moments

The score with respect to θ^ν is similar to (23), replacing z_{im}^d with ν^k in the integrand, i.e.

$$\partial_{\theta^{\nu(k)}} \log \hat{L} = \sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} D_{im} \frac{y_{ijm}}{\pi_{jm}^{z_{im}}} \int s_{jm}(z_{im}, \nu) \Big(x_{jm}^k \nu^k - \sum_{\ell=1}^{J_m} x_{\ell m}^k \nu^k s_{\ell m}(z_{im}, \nu) \Big) \, \mathrm{d}F(\nu), \tag{25}$$

Unfortunately, the above used approximation is not useful since the integral would simplify to zero.

There are at least three ways of dealing with this issue. The most common in the applied literature is to simply drop the score with respect to θ^ν and rely on product level moments for identification. As discussed above, doing so may slow the rate of convergence of $\hat{\theta}^\nu$ from $\sqrt{S}$ to $\sqrt{J}$.

A second alternative employed by e.g. Berry, Levinsohn and Pakes (2004) and Grieco, Murry and Yurukoglu (2021) is introducing second choice data based on surveys of consumer purchases. Our estimator could accommodate second choice data efficiently by including it directly in the likelihood. There are two potential issues with second choice data. First, surveys rely on consumer responses rather than revealed preference and can be sensitive to selection issues due to low response rates. Perhaps Moreover importantly, such data is often prohibitively costly to obtain.

While we are unaware of its use in the literature, there is a third possibility that requires two independent ν draws per simulation r, as we now explain. First, note that[29] $\sum_{j=0}^{J_m} s_{jm}(x_{jm}^k \nu^k - \sum_{\ell=0}^{J_m} s_{\ell m} x_{\ell m}^k \nu^k) = 0$, such that the right hand side in (25) can be expressed as

$$\sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} D_{im} \frac{y_{ijm} - \pi_{jm}^{z_{im}}}{\pi_{jm}^{z_{im}}} \int s_{jm}(z_{im}, \nu) \Big(x_{jm}^k \nu^k - \sum_{\ell=0}^{J_m} x_{\ell m}^k \nu^k s_{\ell m}(z_{im}, \nu) \Big) \, \mathrm{d}F(\nu),$$

because summing the integrand over j equals zero and $\pi_{jm}^{z_{im}} / \pi_{jm}^{z_{im}} = 1$. Noting that the conditional expectation of the last displayed equation given all z's and x's equals zero at the truth and that the denominator only depends on z's and x's, we can remove the weighting in the denominator. Removing the denominator affects efficiency but still provides a valid moment. So we are left with a

[28] Otherwise, there would be an extra term in the moment due to the error in simulating $\delta(\theta)$, i.e. there would be one term with $\delta(\theta)$ and one expansion term involving the difference between the simulated and actual values of $\delta(\theta)$.

[29] We set $x_{0m} = 0$ without loss of generality.

sum over the product of two integrals, namely

$$\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{j=0}^{J_m}\int D_{im}\{y_{ijm}-\mathcal{s}_{jm}(z_{im},\nu^*)\}\,\mathrm{d}F(\nu^*)\times \int \mathcal{s}_{jm}(z_{im},\nu)\Big(x_{jm}^k\nu^k-\sum_{\ell=0}^{J_m}\mathcal{s}_{\ell m}(z_{im},\nu)x_{\ell m}^k\nu^k\Big)\,\mathrm{d}F(\nu).$$

Thus, approximating the integrals with sums using independent Monte Carlo draws satisfies the conditions of Pakes and Pollard (1989). While utilizing this moment will result in an estimator with the same convergence rate as our estimator, it will result in a loss of efficiency.

6.4 Step 4: Population statistics instead of micro-data

One may further alter the correlation moment described in section 6.3.1 by integrating (24) over z,

$$\sum_{m=1}^{M}\sum_{j=0}^{J_m}\Big(\frac{1}{S_m}\sum_{i=1}^{N_m}D_{im}y_{ijm}x_{jm}^k z_{im}^d-\int \pi_{jm}^z x_{jm}^k z^d\,\mathrm{d}G(z)\Big). \tag{26}$$

This is the moment described in Berry, Levinsohn and Pakes (2004, equation 8) and Gandhi and Nevo (2021, equation 4.4).

There are two possible motivations using (26) over (24). The stronger is that it is less data intensive in that it may be computed using only statistics of the micro data. For example, Sweeting (2013) uses data from a survey conducted by a third party that reports averages at the market-demographic level which correspond to the first term in the summand of (26). The second is that the right hand side of (26) does not involve a sum over observed consumers. However, in view of Pakes and Pollard (1989), the total number of simulation draws needed is the same in both cases. To simulate (24), we need only a finite number of simulation draws *per consumer* in order not to affect the convergence rate, as long as all draws are independent, whereas for (26) one needs a number of independent draws that is at least proportional to S.

However, using (26) over (24) comes at an additional cost in efficiency. In particular, (26) does not exploit the consumer level data in the second term because it does not condition on z_i. It is straightforward to show that the variance of (26) is no less than that of (24). For the sake of ease of notation, consider the single market case with x, z both scalars and let $\omega_i = \sum_{j=0}^{J} D_i x_j y_{ij} z_i$. The moments in (24) and (26) (if evaluated at the truth) have the same Jacobian in expectation. The variance contribution for observation i using (26) 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 (24). These two facts combined with the sandwich formula for the asymptotic variance of the GMM estimator imply that using (24) dominates (26).

7 Computation

While the efficient estimator is of theoretical interest in its own right, it must also be computationally tractable in order to be appropriate for applied use. In this section, we discuss two critical computational aspects of our estimator.

First, our estimator involves an optimization over δ which is a vector of length J. In modern datasets, the number of products across all markets can run into the hundreds of thousands, posing a potential problem for nonlinear optimization. However, there are a number of features of our optimization problem that simplify this task considerably.

Second, any estimator must numerically approximate integrals over demographics z and taste shocks ν.[30] As discussed above, the choice of integration method will impact that accuracy of the estimator. We discuss several approaches in section 7.2.

7.1 Dimensionality

We now describe a feasible algorithm for the computation of our estimates for which we use Newton's method with Trust Regions.

Recall from (8) that our optimization problem is

$$(\hat{\beta}, \hat{\theta}, \hat{\delta}) = \arg\min_{\beta,\theta,\delta} \Big(-\log \hat{L}(\theta, \delta) + \hat{\Pi}(\beta, \delta) \Big).$$

Like Berry, Levinsohn and Pakes (1995), we start by concentrating out β which leaves

$$(\hat{\theta}, \hat{\delta}) = \arg\min_{\theta,\delta} \Big(-\log \hat{L}(\theta, \delta) + \hat{\Pi}\{\hat{\beta}(\delta), \delta\} \Big). \tag{27}$$

We then have two levels of optimization. In the inner optimization we compute $\hat{\delta}$ as a function of θ, i.e. for each candidate value θ we find a minimizer $\hat{\delta}(\theta)$. In the outer optimization we then minimize over θ. This approach is similar to that in Berry, Levinsohn and Pakes (1995) with the important exception that the inner loop objective is (8)—the same as the outer loop objective—rather than the share constraint $\pi = s$.

The high-dimensional problem is now confined to the inner loop. For Berry, Levinsohn and Pakes (1995), tractability followed from the existence of a contraction mapping to compute $\pi = s$. For our problem, first suppose that (8) is just identified. In this case, $\hat{\Pi}\{\hat{\beta}(\delta), \delta\} = 0$ for all values of δ, in which case we only need to optimize $\log \hat{L}$ in the inner loop. Conveniently, $\log \hat{L}$ is additively

[30] The exception to this is the mixed logit, which only uses micro data and hence only integrates over ν.

separable across markets in $\delta_{\cdot m}$ and is globally concave in δ for fixed θ. So we can parallelize the computation of $\hat{\delta}_{\cdot m}(\theta)$ market by market, and each computation is a globally concave problem.

The overidentified case is more complicated. To simplify exposition but without loss of generality, we will take $\hat{\mathcal{W}}$ in the definition of $\hat{\Pi}$ in (12) to be $(B^\intercal B)^{-1}$ where B is a $J \times d_b$ matrix with rows $b_{jm}^\intercal$, the instruments introduced in (13). Unfortunately, $\hat{\Pi}$ is not additively separable in $\delta_{\cdot m}$. However, there are several convenient features which make the inner loop optimization tractable.

The first such feature is that $\hat{\beta}(\delta)$ is simply a linear IV estimator, i.e. $\hat{\beta}(\delta) = (X^\intercal \mathcal{P}_B X)^{-1} X^\intercal \mathcal{P}_B \delta$, with $\mathcal{P}_B = B(B^\intercal B)^{-1} B^\intercal$ an orthogonal projection matrix. Second, $\hat{\Pi}$ is quadratic in δ. Thus, after routine matrix algebra, (27) becomes[31]

$$-\log \hat{L}(\theta, \delta) + \frac{1}{2}\delta^\intercal (\mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X})\delta \tag{28}$$

Third, (28) is convex in δ. Fourth, barring collinearities the matrix $\mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X}$ is a positive semidefinite matrix of rank $d_b - d_\beta$. Note that by the spectral decomposition, $\mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X}$ can hence be expressed in the form $\mathcal{K}\mathcal{K}^\intercal$ for a $d_\delta \times (d_b - d_\beta)$ matrix $\mathcal{K}$. This is convenient because X may include many exogenous regressors (eg., brand or product—rather than product-market—dummies) which also appear in B. Such $\mathcal{K}$ is not unique but all choices are equivalent: we derive an explicit form for $\mathcal{K}$ in lemma 1 in appendix D.

Using these features, we now focus on the primary complication of applying Newton's method to optimize (28) over δ in the inner loop: computation of the inverse of the Hessian (with respect to δ). Just storing a Hessian in 100,000 parameters would take 80Gb of memory, the computational cost of taking the inverse is cubic in d_δ, and the result could be subject to substantial numerical error.

Fortunately, we do not need to store or directly invert the full Hessian of (28), $H + \mathcal{K}\mathcal{K}^\intercal$, where H is the Hessian of $-\log \hat{L}$. Instead, we can compute the inverse Hessian exploiting the above-mentioned features. The inverse of the Hessian of (28) is then

$$H^{-1} - H^{-1}\mathcal{K}(I + \mathcal{K}^\intercal H^{-1}\mathcal{K})^{-1}\mathcal{K}^\intercal H^{-1}, \tag{29}$$

where I is the identity matrix.[32]

Since $\log \hat{L}$ is additively separable in the $\delta_{\cdot m}$'s, H is block diagonal, so H^{-1} can be efficiently computed and stored. To appreciate the importance of this feature, note that if one has 1,000 markets with 100 inside goods in each market, the problem reduces from inverting a full 100,000 by 100,000 matrix $H + \mathcal{K}\mathcal{K}^\intercal$ to inverting a thousand 100 by 100 matrices, which is both much less demanding computationally and reduces memory demand by a factor 1,000.[33] This makes the

[31] $\mathcal{P}_{\mathcal{P}_B X} = \mathcal{P}_B X (X^\intercal \mathcal{P}_B X)^{-1} X^\intercal \mathcal{P}_B$.

[32] To see this, note that for $\Delta = I + \mathcal{K}^\intercal H^{-1}\mathcal{K}$,

$$\left(H^{-1} - H^{-1}\mathcal{K}\Delta^{-1}\mathcal{K}^\intercal H^{-1}\right)(H + \mathcal{K}\mathcal{K}^\intercal) = I + H^{-1}\mathcal{K}\mathcal{K}^\intercal - H^{-1}\mathcal{K}\Delta^{-1}\mathcal{K}^\intercal - H^{-1}\mathcal{K}\Delta^{-1}\mathcal{K}^\intercal H^{-1}\mathcal{K}\mathcal{K}^\intercal =$$
$$I + H^{-1}\mathcal{K}\underbrace{\Delta^{-1}(I + \mathcal{K}^\intercal H^{-1}\mathcal{K})}_{=I}\mathcal{K}^\intercal - H^{-1}\mathcal{K}\Delta^{-1}\mathcal{K}^\intercal - H^{-1}\mathcal{K}\Delta^{-1}\mathcal{K}^\intercal H^{-1}\mathcal{K}\mathcal{K}^\intercal = I.$$

[33] $100,000^2/(100^2 \times 1,000)$

optimization step of the inner loop practical for many products.

The outer loop is over a low dimensional parameter vector, albeit computations of the derivatives involves application of the chain rule to account for inner loop optimization. We have verified that this procedure can be used successfully for problems with over 100,000 products and millions of consumers.

7.2 Numerical integration

As we have pointed out, the largest disadvantage of our estimator is that a computable version relies on numerical integration which is costly since in order not to affect the asymptotic behavior, we need the numerical error to be negligible. However, as always, we can arbitrarily reduce the numerical approximation error by incurring a higher computational cost. In contrast, the MSM can achieve the same convergence rate by averaging over noisy approximations of these integrals. But as mentioned section 6.3.1, numerical approximation of the share inversion adds an additional source of complexity for estimators in our setting that enforce share constraints.

Our estimator evaluates two types of integrals, those over ν (e.g., π^z) and those over both ν and z (e.g., π). This distinction suggests different integration methods for each type.

Quadrature methods are well suited for micro integrals over ν only. The distribution of ν is assumed known and is usually a familiar and tractable one, often normal. Moreover, ν is often of small dimension, so the curse of dimensionality associated with tensor product quadrature methods is less binding. If ν is of high dimension, sparse quadrature methods can be viable alternatives.[34]

The integrals over both z and ν are more difficult to compute. In addition to (z, ν) being higher dimensional than ν, the distribution of z is usually informed by data and so less amenable to quadrature methods (e.g., the distribution of income in the consumer population). On the other hand, they are only computed for each product (J) rather than each product-consumer pair (JS). Given this, (quasi-)Monte Carlo methods with a high number of draws are appropriate, albeit this requires the number of Monte Carlo draws to grow faster than the square of the prevailing convergence rate, which is the same number as is needed for MSM not to lose efficiency.

We will experiment with alternative numerical integration approaches in section 9 in a future version of this manuscript.

8 Inference

This section describes inference on functions of model parameters, including elasticities and counterfactuals. As we discussed above the conformant property of our estimator ensures that it can be applied under a wide variety of conditions. This also applies to our inference procedure. In all cases,

[34] The designed quadrature approach of Bansal et al. (2021) may be particularly attractive as all nodes have positive weights.

inference will be built upon the Hessian of our objective function (8),

$$\begin{array}{c} \\ \beta \\ \theta^z \\ \theta^\nu \\ \delta \end{array} \begin{array}{c} \begin{array}{cccc} \beta & \theta^z & \theta^\nu & \delta \end{array} \\ \begin{bmatrix} \partial_\beta \hat{m}^\intercal \hat{\mathcal{W}} \partial_{\beta^\intercal} \hat{m} & 0 & 0 & \partial_\beta \hat{m}^\intercal \hat{\mathcal{W}} \partial_{\delta^\intercal} \hat{m} \\ 0 & -\partial_{\theta^z \theta^{z\intercal}} \log \hat{L} & -\partial_{\theta^z \theta^{\nu\intercal}} \log \hat{L} & -\partial_{\theta^z \delta^\intercal} \log \hat{L} \\ 0 & -\partial_{\theta^\nu \theta^{z\intercal}} \log \hat{L} & -\partial_{\theta^\nu \theta^{\nu\intercal}} \log \hat{L} & -\partial_{\theta^\nu \delta^\intercal} \log \hat{L} \\ \partial_\delta \hat{m}^\intercal \hat{\mathcal{W}} \partial_{\beta^\intercal} \hat{m} & -\partial_{\delta \theta^{z\intercal}} \log \hat{L} & -\partial_{\delta \theta^{\nu\intercal}} \log \hat{L} & \partial_\delta \hat{m}^\intercal \hat{\mathcal{W}} \partial_{\delta^\intercal} \hat{m} - \partial_{\delta\delta^\intercal} \log \hat{L} \end{bmatrix} \end{array}. \tag{30}$$

The Hessian alone is sufficient since our estimator is efficient so the usual sandwich formula collapses. As we will see below, the Hessian conforms to provide valid inference in each of the cases described in section 4.2. Importantly, the researcher does not need to assume or determine the rates of convergence of the estimator in her situation to conduct inference correctly.[35]

First consider the leading case where $S/J \to \infty$ and $\theta^z \neq 0$. In this case, our estimator is asymptotically equivalent to a two-step estimator that first estimates (θ, δ) and then plugs in $\hat{\delta}$ to estimate β. With the two-step estimator, the information matrix for $\psi = [\theta^\intercal,\ \delta^\intercal]^\intercal$ is the Hessian of $-\log \hat{L}$. Notice that this is the (ψ, ψ) block of (30) with the exception of the $\partial_\delta \hat{m}^\intercal \hat{\mathcal{W}} \partial_{\delta^\intercal} \hat{m}$ term in the (δ, δ) block. However, that term diverges at rate J and is dominated by $-\partial_{\delta\delta^\intercal} \log \hat{L}$. Similarly, because $\hat{\psi}$ converges faster than $\hat{\beta}$, the (β, β) block in (30) is all that matters for inference on β. To see this, note that by the partitioned inverse formula, the (β, β) block of the inverse of (30) is

$$\begin{aligned} &\Big(((\beta,\beta) \text{ block}) - ((\beta,\delta) \text{ block}) * ((\delta,\delta) \text{ block})^{-1} * ((\delta,\beta) \text{ block}) \Big)^{-1} \\ &\qquad = \Big(\partial_\beta \hat{m}^\intercal \hat{\mathcal{W}} \partial_{\beta^\intercal} \hat{m} - \partial_\beta \hat{m}^\intercal \hat{\mathcal{W}} \partial_{\delta^\intercal} \hat{m} * \left(\partial_\delta \hat{m}^\intercal \hat{\mathcal{W}} \partial_{\delta^\intercal} \hat{m} - \partial_{\delta\delta^\intercal} \log \hat{L} \right)^{-1} * \partial_\delta \hat{m}^\intercal \hat{\mathcal{W}} \partial_{\beta^\intercal} \hat{m} \Big)^{-1}. \end{aligned}$$

Again, since the loglikelihood dominates, the second term inside the outer inverse is asymptotically negligible, so the limiting distribution of $\hat{\beta}$ is determined entirely by the product level moments.

Now consider the case where $S/J \to \infty$ and $\theta^z = 0$. As we show in appendix F, the scores of the objective with respect to θ^ν and δ become collinear, leading to a loss of rank in the Hessian of $\log \hat{L}$. However, rank is preserved in (30) due to the presence of the product level moments in the (δ, δ) block. As noted above, this affects the rate of convergence as $\hat{\Pi}$ will enter the dominant term of the (ψ^ν, ψ^ν) block of the inverse Hessian. However, the rate of $\hat{\theta}^z$ is unaffected since the score with respect to θ^z is not collinear and the dominant term of the (θ^z, θ^z) block of the inverse Hessian will be

$$-\left(\partial_{\theta^z \theta^{z\intercal}} \log \hat{L} - \partial_{\theta^z \delta^\intercal} \log \hat{L} \left(\partial_{\delta\delta^\intercal} \log \hat{L} \right)^{-1} \partial_{\delta \theta^{z\intercal}} \log \hat{L} \right)^{-1}, \tag{31}$$

as we show in lemma 4 in appendix F. Expression (31) converges at rate S.

Now consider the case where $S/J \to 0$. The clearest intuition comes form the extreme case where $S = 0$ (i.e., Berry, Levinsohn and Pakes, 1995). As we discussed in section 4.2.1, θ, δ are not identified off the likelihood alone since $\log \hat{L}^{\text{micro}} = 0$ and $\log \hat{L}^{\text{macro}}$ is maximized for any θ by

[35] Recall that the use of the plural 'rates' is due to the fact that different elements of our estimator vector converge at different rates.

choosing δ such that $\pi = s$ as we have shown in section section 6.2.[36] Consequently, $\partial_{\psi\psi^\intercal} \log \hat{L}$ is then singular, indeed of rank d_δ.[37] However, analogous to the $\theta^z = 0$ case, the (ψ, ψ) block in (30) has full rank due to the product level moments entering the (δ, δ) block. Note that because here the micro data is not available to pin down θ^z, we need $d_b \geq d_\beta + d_{\theta^z} + d_{\theta^\nu}$ to preserve identication rather than $d_b \geq d_\beta + d_{\theta^\nu}$ in the $\theta^z = 0$ case above. It can be shown that the dominant term of the (θ, θ) block of the inverse Hessian has the same form as the corresponding expression for the Berry, Levinsohn and Pakes (1995) estimator which is $O_p(J^{-1})$; see lemma 5 in appendix F. Returning to the case where $S/J \to 0$ but some micro data exists. Now $\hat{\Pi}$ will dominate dominate $\log \hat{L}$ in the Hessian, and all parameters converge at rate $\sqrt{J}$. However, for the same reasons as stated above, the Hessian remains invertible.

The remaining cases are merely combinations of the above logic. If S/J converges to a non-zero constant, both $\log \hat{L}$ and $\hat{\Pi}$ contribute to the limiting distribution and both are accounted for in the Hessian with the appropriate weighting. If $\theta^z \to 0$, the contribution of $\hat{\Pi}$ to the limiting distribution of θ, δ will be non-negligible but accounted for in the Hessian. To summarize, under different scenarios the relative importance of $\log \hat{L}$ and $\hat{\Pi}$ vary. However, by using (30) for inference, we include all relevant terms so that inference is valid across all these scenarios.

A second complication is that δ grows with J, so (30) is also growing. To address this, write $\gamma = [\beta^\intercal, \theta^\intercal, \delta^\intercal]^\intercal$. Recall from section 4.2 that we assume that $\lim_{M\to\infty} \max_m J_m < \infty$. Since γ's dimension grows with the number of markets, the following theorem provides an inference method for finite-dimensional linear combinations $\Lambda\gamma$ of γ, where Λ has a fixed number of rows. Its proof is outlined in appendix F.2.

Theorem 2. *Assume that*

(I) The parameter space of $\beta, \theta, \delta_{\cdot 1}, \delta_{\cdot 2}, ...$ the Cartesian product of their individual parameter spaces and each one of $\beta, \theta, \delta_{\cdot 1}, \delta_{\cdot 2}, ...$ is bounded away from the boundary, uniformly across m.[38]

(II) (i) The population in each market consists of i.i.d. draws from a superpopulation; (ii) The consumer micro sample (if present) in each market consists of a randomly chosen subset of the population in that market; (iii) There is independence across markets;

(III) (i) The model described in section 2.1 is correctly specified; (ii) μ^z_{jm} and μ^ν_{jm} are for all m linear in θ^z, θ^ν respectively;

(IV) One of the following two scenarios applies: (i) S is fixed and the intersection Γ_M of the set satisfying the product level exclusion restrictions and the set of maximizers of the expectation of the macro loglikelihood is such that $\Gamma_M \to \Gamma$ as $M \to \infty$ where Γ consists of a singleton;

[36] Notice that since the $\log \hat{L}^{\text{macro}}$ integrates over z, we need not distinguish between θ^z and θ^ν in this case.

[37] For any given θ, there is a unique δ that maximizes $\log \hat{L}$—or equivalently satisfies the share constraint (Berry, 1994)—so the degree of underidentification is d_θ.

[38] In other words, there is a common positive minimum distance between each parameter vector and the corresponding boundary.

(ii) S grows to infinity and the intersection Γ_M of the set satisfying the product level exclusion restrictions and the set of maximizers of the expectation of the macro loglikelihood is such that $\Gamma_M \to \Gamma$ as $M \to \infty$ where Γ consists of a singleton;

(V) (i) $M \to \infty$; (ii) J_m can differ across m but does not change; (iii) $\lim_{M\to\infty} \max_{m\leq M} J_m < \infty$; (iv) $\liminf_{M\to\infty} \min_{m\leq M} N_m/M = \infty$;

(VI) $\{\Lambda\}$ is a sequence of matrices that is such that for a given fixed M^*, the first $d_\beta + d_\theta + \sum_{m=1}^{M^*} J_m$ columns of Λ are fixed and the remaining columns consist of zeros.

Then,

$$(\Lambda \hat{V} \Lambda^\intercal)^{-1/2} \Lambda (\hat{\gamma} - \gamma) \xrightarrow{d} N(0, I),$$

where $\hat{V}$ is the inverse of (30). □

Although the number of unknown coefficients increases (the number of δ's increases), it only does so as more markets are added. In other words, (subject to identification) one could estimate θ off finitely many markets with an increasing number of consumers in the micro sample. The problem is hence inherently different from that in the seminonparametric estimation literature in which there are infinitely many parameters from the outset.

To conduct inference on finite-dimensional nonlinear functions of γ one can apply the delta method. This enables the researcher to conduct inference on arbitrary differentiable functions of the model parameters, such as elasticities, pass-through rates, or counterfactual outcomes.

9 Monte Carlo Experiments

This section will appear in a future version of the manuscript.

10 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 provide a method that optimally combines the likelihood for purchase data with product level exogeneity restrictions into a unified estimator that conforms to a wide variety of data environments and achieves efficiency in each. Our estimator does not require additional parametric assumptions relative to a GMM estimator. By showing how to transform our estimator into those used previously in the literature, we illustrate several tradeoffs between statistical efficiency and other researcher concerns, such as computational tractability and data availabilty. With that said, we show that our estimator is computationally tractable, suggesting that it will be directly useful for applied work in a wide variety of settings. Indeed, our estimator has an additional advantage that inference is more straightforward and correct under more applicable assumptions than the standard approach.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica*, 88(1): 265–296.

Bachmann, Ruediger, Gabriel Ehrlich, Ying Fan, Dimitrije Ruzic, and Benjamin Leard. 2019. "Firms and collective reputation: a Study of the Volkswagen Emissions Scandal." National Bureau of Economic Research.

Backus, Matthew, Christopher Conlon, and Michael Sinkinson. 2021. "Common ownership and competition in the ready-to-eat cereal industry." National Bureau of Economic Research.

Bansal, Prateek, Vahid Keshavarzzadeh, Angelo Guevara, Shanjun Li, and Ricardo A Daziano. 2021. "Designed quadrature to approximate integrals in maximum simulated likelihood estimation." *The Econometrics Journal*, 25(2): 301–321.

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.

Chintagunta, Pradeep K, and Jean-Pierre Dube. 2005. "Estimating a stockkeeping-unit-level brand choice model that combines household panel data and store data." *Journal of Marketing Research*, 42(3): 368–379.

Crawford, Gregory S., and Ali Yurukoglu. 2012. "The Welfare Effects of Bundling in Multichannel Television Markets." *American Economic Review*, 102(2): 643–85.

Crawford, Gregory S, Robin S Lee, Michael D Whinston, and Ali Yurukoglu. 2018. "The welfare effects of vertical integration in multichannel television markets." *Econometrica*, 86(3): 891–954.

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

Gandhi, Amit, and Aviv Nevo. 2021. "Empirical models of demand and supply in differentiated products industries." In *Handbook of Industrial Organization*. Vol. 4, 63–139. Elsevier.

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

Goeree, Michelle Sovinsky. 2008. "Limited information and advertising in the US personal computer industry." *Econometrica*, 76(5): 1017–1074.

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

Grieco, Paul, Charles Murry, and Ali Yurukoglu. 2021. "The Evolution of Market Power in the U.S. Automobile Industry." *working paper*.

Hackmann, Martin B. 2019. "Incentivizing better quality of care: The role of Medicaid and competition in the nursing home industry." *American Economic Review*, 109(5): 1684–1716.

Hahn, Jinyong, and Whitney Newey. 2004. "Jackknife and analytical bias reduction for nonlinear panel models." *Econometrica*, 72(4): 1295–1319.

Hendel, Igal, and Aviv Nevo. 2006. "Measuring the implications of sales and consumer inventory behavior." *Econometrica*, 74(6): 1637–1673.

Imbens, Guido W, and Tony Lancaster. 1994. "Combining micro and macro data in microeconometric models." *The Review of Economic Studies*, 61(4): 655–680.

Neilson, Christopher. 2019. "Targeted vouchers, competition among schools, and the academic achievement of poor students." *mimeo, Princeton University*.

Nevo, Aviv. 2000. "A Practicitoner's Guide to Estimation of Random Coefficients Logit Models of Demand." *Journal of Economics & Management Strategy*, 9(4): 513–548.

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.

Robinson, Peter M. 1988. "Root-N-consistent semiparametric regression." *Econometrica: Journal of the Econometric Society*, 931–954.

Staiger, Douglas, and James H Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica*, 65(3): 557–586.

Sweeting, Andrew. 2013. "Dynamic product positioning in differentiated product markets: The effect of fees for musical performance rights on the commercial radio industry." *Econometrica*, 81(5): 1763–1803.

Train, Kenneth E, and Clifford Winston. 2007. "Vehicle choice behavior and the declining market share of US automakers." *International economic review*, 48(4): 1469–1496.

Tuchman, Anna E. 2019. "Advertising and demand for addictive goods: The effects of e-cigarette advertising." *Marketing Science*, 38(6): 994–1022.

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

$$\pi_{ijm}(\nu) = \frac{\exp(\mu^z_{ijm} + \mu^\nu_{ijm} + \delta_{jm})}{\sum_{g=0}^{G}\exp(\mu^z_{igm} + \mu^\nu_{igm} + \delta_{gm})}. \tag{32}$$

Then,

$$\partial_\theta \log \pi^{z_{i\cdot m}}_{jm} = \frac{\int \pi_{ijm}(\nu)\Delta b_{ijm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)}, \tag{33}$$

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,

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

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 (33) and (34). An analogous argument applies to the Hessians. So we only present the Hessians for the micro contributions.

They are,

$$\partial_{\theta\theta^\intercal}\log \pi^{z_{i\cdot m}}_{jm} = \frac{\int \pi_{ijm}(\nu)\Delta b_{ijm}(\nu)\Delta b^\intercal_{ijm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)} - \partial_\theta \log \pi^{z_{i\cdot m}}_{jm}\partial_{\theta^\intercal}\log \pi^{z_{i\cdot m}}_{jm} - \sum_{g=0}^{J_m}\frac{\int \pi_{ijm}(\nu)\pi_{igm}(\nu)\Delta b_{igm}(\nu)\Delta b^\intercal_{igm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)}, \tag{35}$$

$$\partial_{\delta\delta^\intercal}\log \pi^{z_{i\cdot m}}_{jm} = \frac{\int \pi_{ijm}(\nu)\Delta \mathbb{1}_{ijm}(\nu)\Delta \mathbb{1}^\intercal_{ijm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)} - \partial_\delta \log \pi^{z_{i\cdot m}}_{jm}\partial_{\delta^\intercal}\log \pi^{z_{i\cdot m}}_{jm} - \frac{\int \pi_{ijm}(\nu)\Big[\pi_{ikm}(\nu)\{\mathbb{1}(k=t) - \pi_{itm}(\nu)\}\Big]_{k,t=1,\ldots,J_m}\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)}, \tag{36}$$

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

$$\partial_{\delta\theta^\intercal}\log \pi^{z_{i\cdot m}}_{jm} = \frac{\int \pi_{ijm}(\nu)\Delta \mathbb{1}_{ijm}(\nu)\Delta b^\intercal_{ijm}(\nu)\,\mathrm{d}F(\nu)}{\int \pi_{ijm}(\nu)\,\mathrm{d}F(\nu)} - \partial_\delta \log \pi^{z_{i\cdot m}}_{jm}\partial_{\theta^\intercal}\log \pi^{z_{i\cdot m}}_{jm}. \tag{37}$$

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

$$-\left\{\chi\Omega^m_{\psi\psi} + (1-\chi)\Omega^M_{\psi\psi}\right\}^{-1}, \tag{38}$$

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 (38) 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

$$-\begin{bmatrix} \mathcal{J} \\ \partial_\theta \delta^\intercal \end{bmatrix} \Phi^{-1} \begin{bmatrix} \mathcal{J} & \partial_{\theta^\intercal}\delta \end{bmatrix} / \chi, \quad \textit{(incorrect variance)} \tag{39}$$

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 share constrained estimator is

$$-\begin{bmatrix} \chi\Phi & \chi(\Omega^m_{\theta\delta} + \partial_\theta\delta^\intercal\Omega^m_{\delta\delta}) \\ \Omega^M_{\delta\theta} & \Omega^M_{\delta\delta} \end{bmatrix}^{-1} \begin{bmatrix} \chi\Phi & 0 \\ 0 & \Omega^M_{\delta\delta} \end{bmatrix} \begin{bmatrix} \chi\Phi & \Omega^M_{\theta\delta} \\ \chi(\Omega^m_{\delta\theta} + \Omega^m_{\delta\delta}\partial_{\theta^\intercal}\delta) & \Omega^M_{\delta\delta} \end{bmatrix}^{-1}. \tag{40}$$

The formula in (40) is based on the fact that the share constrained estimator uses the following moment conditions:

$$\begin{cases} \displaystyle\sum_{i=1}^N \sum_{j=0}^J y_{ij} D_i \left(\partial_\theta \log \pi_j^{z_i} + \partial_\theta\hat{\delta}^\intercal \partial_\delta \log \pi_j^{z_i}\right) = 0, \\ \displaystyle\sum_{i=1}^N \sum_{j=0}^J y_{ij} \partial_\delta \log \pi_j = 0, \end{cases} \tag{41}$$

where

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

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

$$(-\Omega^m_{\psi\psi})^{-1}/\chi. \tag{42}$$

C Share constraints

C.1 Some theory

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

$$\Omega(\psi) = \sum_{i=1}^{\bar{I}} \{D_i L_i^m(\psi) + \omega(1 - D_i) L_i^M(\psi)\}, \tag{43}$$

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

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

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

$$\hat{\psi}_\omega^* = \arg\max_{\psi} \sum_{i=1}^{\bar{I}} \{D_i L_i^m(\psi) + \omega L_i^M(\psi)\}, \tag{44}$$

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 (43) and (44) in that in (44) 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 $U_0(\chi U_0^\intercal A U_0)^{-1} U_0^\intercal + U_1 D_1^{-1} U_1^\intercal / \omega^2$.[39] 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. □

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.

C.2 Another share constraint example

Example 2. 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

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

[39] 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.

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 $\chi = \lim(S/N)$. 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. □

D Computation

The following lemma shows that the $d_\delta \times d_\delta$ matrix $\mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X}$ used in (28) can be expressed as the product of a $d_\delta \times (d_b - d_\beta)$ matrix with its transpose. Note than when computing $\mathcal{K}$ it is useful to first project out all exogenous regressors that appear in both X and B because it is less expensive to compute the singular value decomposition of a matrix of lower rank.

Lemma 1. Let $X = [C\ \tilde{X}]$ and $B = [C\ \tilde{B}]$, i.e. C are the columns shared by X and B. Let further $X^* = \mathcal{M}_C \tilde{X}$ and $B^* = \mathcal{M}_C \tilde{B}$ with $\mathcal{M}_C$ an annihilator matrix (for C). Then,

$$\forall \delta : \{\delta - X\hat{\beta}(\delta)\}^\intercal \mathcal{P}_B \{\delta - X\hat{\beta}(\delta)\} = \delta^\intercal \mathcal{K}\mathcal{K}^\intercal \delta, \tag{45}$$

where $\mathcal{K} = \mathcal{U}_B \mathcal{M}_{\mathcal{U}_B^\intercal \mathcal{U}_X}$ with $\mathcal{U}_B, \mathcal{U}_X$ matrices with orthonormal columns spanning exactly the column spaces of B^* and X^*, respectively.

Proof. Recall from the text in section 7 that (45) can be expressed as $\delta^\intercal \mathcal{P}^* \delta$ where $\mathcal{P}^* = \mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X}$. Noting that $\mathcal{P}_B = \mathcal{P}_C + \mathcal{P}_{B^*}$ and $\mathcal{P}_{\mathcal{P}_B X} = \mathcal{P}_C + \mathcal{P}_{\mathcal{P}_{B^*} X^*}$, we have $\mathcal{P}^* = \mathcal{P}_{B^*} - \mathcal{P}_{\mathcal{P}_{B^*} X^*}$. The stated result then follows by application of the singular value decomposition to both B^* and X^*. □

E Selection

Our methodology combines the micro-sample with the product shares by integrating out z_{im} in the choice probabilities when individual i is outside the micro-sample, yielding

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

This allows for a variety of forms of selection. Clearly, random selection poses no difficulty as in this case $\pi_{jm}^{D=0} = Pr(D_{im} = 0)\pi_{jm}$, leading to the logliklihood prested in (9) (up to a constant).

Interestingly, deterministic selection based on $y_{i\cdot m}$ of the form $D_{im} = D_{im}^* \mathbb{1}(y_{i0m} \in \mathcal{J})$ where D_{im}^* is random is also straightforward. This case is common, for example with vehicle registration data, administrative data of regulated industries, or data on sales of a particular subset of firms. In this case,

$$\Pr(D_{im} = 1 \cap y_{ijm} = 1 \mid z_{im}) = \begin{cases} 0 & j \notin \mathcal{J} \\ \Pr(D_{im}^* = 1)\pi_{jm}^{z_{im}} & j \in \mathcal{J} \end{cases},$$

so we have,

$$\pi_{jm}^{D=0} = \begin{cases} \pi_{jm} & j \notin \mathcal{J} \\ \Pr(D_{im}^* = 0)\pi_{jm} & j \in \mathcal{J} \end{cases}.$$

Moreover, in both of the above cases, because only logarithms of the choice probabilities appear in the loglikelihood, the $\Pr(D_{im}^* = 0)$ factor only adds a constant to the loglikelihood and is hence irrelevant.

Selection dependent on z_{im} can be accommodated by accounting for selection when integrating over the distribution of demographics. $G_m^{D=0}(z)$, the distribution of z_{im} in market m but *not* in the

micro sample, and its complement $G_m^{D=1}(z)$ are easy to compute from the consumer-level data and the known distribution of z_{im} in the population, $G_m(z)$. If selection does not depend on $y_{i\cdot m}$ except through z_{im} then,

$$\pi_{jm}^{D=0} = \int \Pr(D_{im} = 0 \mid z_i = z)\pi_{jm}^z \, \mathrm{d}G_m(z) = \Pr(D_{im} = 0) \int \pi_{jm}^z(\delta, \theta) \, \mathrm{d}G_m^{D=0}(z).$$

More general forms would have to be explicitly modeled and are outside the scope of this paper.

F Inference

F.1 Technical lemmas

This appendix shows several statements asserted in section 8. First we show that the scores of the objective function with respect to θ^ν and δ are collinear if $\theta^z = 0$, for which the following lemma suffices.

Lemma 2. Let $\psi_m^\nu = [\theta^{\nu\intercal}, \, \delta_m^\intercal]^\intercal$. If $\theta^z = 0$ then $\partial_{\psi_m^\nu \psi_m^{\nu\intercal}} \log L$ can have rank at most J_m.

Proof. Consider the case in which $S = N$, which is no less favorable than any other case. Then, since $\theta^z = 0$, π_{jm}^z is flat in z and hence at the truth,

$$\partial_{\psi_m^\nu} \log \hat{L} = \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} y_{ijm} v_{jm},$$

for some $(J_m + d_{\theta^\nu})$-dimensional vectors $\{v_{jm}\}_{j=0}^{J_m}$. Now, because the expectation of the score is zero at the truth, $\sum_{j=0}^{J_m} \pi_{jm} v_{jm} = 0$, so $v_{0m} = -\sum_{j=1}^{J_m} \pi_{jm} v_{jm} / \pi_{0m}$ is a linear combination of the remaining v_{jm}'s, so the $\{v_{jm}\}$ span a space of dimension no greater than J_m. Further,

$$\mathbb{E}\Big(\partial_{\psi_m^\nu} \log \hat{L} \, \partial_{\psi_m^{\nu\intercal}} \log \hat{L}\Big) = \mathbb{E}\left(\sum_{j=0}^{J_m} \sum_{j^*=0}^{J_m} Y_{ijm} v_{jm} Y_{ij^*m} v_{j^*m}^\intercal\right) = \sum_{j=0}^{J_m} \pi_{jm} v_{jm} v_{jm}^\intercal,$$

which hence has rank no greater than J_m. Apply the information matrix equality. □

Lemma 3.

$$\begin{aligned} -\partial_{\theta\theta^\intercal} \log \hat{L} - \partial_{\theta\delta^\intercal} \log \hat{L} \Big(-\partial_{\delta\delta^\intercal} \log \hat{L} + \partial_{\delta\delta^\intercal} \hat{\Pi}\Big) \partial_{\delta\theta^\intercal} \log \hat{L} \simeq \\ - \Big(\partial_{\theta\theta^\intercal} \log \hat{L} - \partial_{\theta\delta^\intercal} \log \hat{L} (\partial_{\delta\delta^\intercal} \log \hat{L})^{-1} \partial_{\delta\theta^\intercal} \log \hat{L}\Big) + \\ \partial_{\theta\delta^\intercal} \log \hat{L} (\partial_{\delta\delta^\intercal} \log \hat{L})^{-1} \partial_{\delta\delta^\intercal} \hat{\Pi} (\partial_{\delta\delta^\intercal} \log \hat{L})^{-1} \partial_{\delta\theta^\intercal} \log \hat{L}. \end{aligned}$$

Proof. Simply uses $(A + B)^{-1} \approx A^{-1} - A^{-1} B A^{-1}$ for A dominating B. □

From here on, we use the convention that superscripts to a matrix indicate the corresponding block of the inverse of the matrix.

Lemma 4. If $S/J \to \infty$ and $\theta^z = 0$ then the dominant term of the (θ^z, θ^z) block of the inverse Hessian evaluated at the truth is

$$-\Big(\partial_{\theta^z \theta^{z\intercal}} \log \hat{L} - \partial_{\theta^z \delta^\intercal} \log \hat{L} \, (\partial_{\delta\delta^\intercal} \log \hat{L})^{-1} \, \partial_{\delta\theta^{z\intercal}} \log \hat{L}\Big)^{-1},$$

Proof. First, note that by partitioned inverses,

$$\hat{\Omega}^{\psi\psi} = \begin{bmatrix} -\partial_{\theta^z\theta^{z\intercal}} \log \hat{L} & -\partial_{\theta^z\theta^{\nu\intercal}} \log \hat{L} & -\partial_{\theta^z\delta^\intercal} \log \hat{L} \\ -\partial_{\theta^\nu\theta^{z\intercal}} \log \hat{L} & -\partial_{\theta^\nu\theta^{\nu\intercal}} \log \hat{L} & -\partial_{\theta^\nu\delta^\intercal} \log \hat{L} \\ -\partial_{\delta\theta^{z\intercal}} \log \hat{L} & -\partial_{\delta\theta^{\nu\intercal}} \log \hat{L} & -\partial_{\delta\delta^\intercal} \log \hat{L} + \partial_{\delta\delta^\intercal}\hat{\Pi}^* \end{bmatrix}^{-1},$$

where $\partial_{\delta\delta^\intercal}\hat{\Pi}^* = \partial_{\delta\delta^\intercal}\hat{\Pi} - \partial_{\delta\beta^\intercal}\hat{\Pi}(\partial_{\beta\beta^\intercal}\hat{\Pi})^{-1}\partial_{\beta\delta^\intercal}\hat{\Pi}$. Since $-\mathbb{E}\partial_{\psi^\nu\psi^{\nu\intercal}} \log \hat{L}$ is positive semidefinite with rank J by lemma 2, we can replace $\partial_{\psi^\nu\psi^{\nu\intercal}} \log \hat{L}$ with $\mathcal{A}\partial_{\delta\delta^\intercal} \log \hat{L}\mathcal{A}^\intercal$. Thus, by partitioned inverses we get

$$\begin{aligned}\hat{\Omega}^{\theta^z\theta^z} &\simeq \left(\partial_{\theta^z\delta^\intercal} \log \hat{L} - \partial_{\theta^z\delta^\intercal} \log \hat{L}\mathcal{A}^\intercal\left(-\mathcal{A}\partial_{\delta\delta^\intercal} \log \hat{L}\mathcal{A}^\intercal + \partial_{\delta\delta^\intercal}\hat{\Pi}^*\right)^{-1}\mathcal{A}\partial_{\delta\theta^{z\intercal}} \log \hat{L}\right)^{-1} \\ &\simeq \left(-\partial_{\theta^z\theta^{z\intercal}} \log \hat{L} + \partial_{\theta^z\delta^\intercal} \log \hat{L}\left(\partial_{\delta\delta^\intercal} \log \hat{L}\right)^{-1}\partial_{\delta\theta^{z\intercal}} \log \hat{L}\right),\end{aligned}$$

as asserted. □

Lemma 5. Absent consumer data and evaluated at the truth,

$$\hat{\Omega}^{\theta^\nu\theta^\nu} \simeq \left(\partial_\theta\delta^\intercal\partial_{\delta\delta^\intercal}\hat{\Pi}^*\partial_{\theta^\intercal}\delta\right)^{-1},$$

where $\partial_{\delta\delta^\intercal}\hat{\Pi}^*$ was defined in lemma 4 and where $\delta(\theta)$ solves the (expectation) share constraint.

Proof. By lemma 3 we get,

$$\begin{aligned}\hat{\Omega}^{\theta^\nu\theta^\nu} = \left(-\partial_{\theta^\nu\theta^{\nu\intercal}} \log \hat{L} - \partial_{\theta^\nu\delta^\intercal} \log \hat{L}\left(-\partial_{\delta\delta^\intercal} \log \hat{L} + \partial_{\delta\delta^\intercal}\hat{\Pi}^*\right)^{-1}\partial_{\delta\theta^{\nu\intercal}} \log \hat{L}\right)^{-1} \simeq \\ \left(-\partial_{\theta^\nu\theta^{\nu\intercal}} \log \hat{L} + \partial_{\theta^\nu\delta^\intercal} \log \hat{L}\left(\partial_{\delta\delta^\intercal} \log \hat{L}\right)^{-1}\partial_{\delta\theta^{\nu\intercal}} \log \hat{L}+\right. \\ \left.\partial_{\theta^\nu\delta^\intercal} \log \hat{L}\left(-\partial_{\delta\delta^\intercal} \log \hat{L}\right)^{-1}\partial_{\delta\delta^\intercal}\hat{\Pi}^*\left(\partial_{\delta\delta^\intercal} \log \hat{L}\right)^{-1}\partial_{\delta\theta^{\nu\intercal}} \log \hat{L}\right)^{-1} \simeq \left(\partial_\theta\delta^\intercal\partial_{\delta\delta^\intercal}\hat{\Pi}^*\partial_{\theta^\intercal}\delta\right)^{-1},\end{aligned}$$

where the last step follows by Khinchine's weak law of large numbers and the implicit function theorem. Note that the right hand side in the lemma statement is exactly the θ^ν component of the asymptotic variance matrix of a BLP GMM estimator. □

F.2 Outline of the proof of theorem 2

The sketch of the proof uses fairly standard arguments. However, there are a few unusual features to the problem that need to be addressed. We will throughout consider the case where S grows; the fixed S case is simpler.

First, there is the issue that even if $\theta^z \neq 0$ and S, N_m grow fast relative to M, there are different convergence rates, namely $\sqrt{J}$ for $\hat{\beta}$ and $\sqrt{S}$ for $\hat{\theta}, \hat{\delta}$ (unless S increases faster than N_m). In the general case, there are more scenarios. The differing convergence rates are not themselves a major issue, but the fact that the Hessian of the loglikelihood can have reduced rank means that we also have to consider the contribution of the $\hat{\Pi}$ component of the objective function.

A second issue is that the dimension of δ grows. However, the dimension of δ grows only because markets are added. And if a market m is added then there are N_m additional consumers in the

population. Since we have assumed J_m to be finite, having N_m increase is sufficient to recover $\delta_{\cdot m}$ for a given value of θ. We will focus on inference for θ, and show that the increasing dimensions become sums over markets, with the $\delta_{\cdot m}$ parameter vectors only entering the terms for market m in those sums.

The sketch of the proof follows a standard pattern. First, we take a standard extreme value theory expansion and then strip out all dominated terms. What is left has a limiting normal distribution.

F.2.1 Expansion

Let $\hat{\Omega} = \hat{\Psi} + \delta^\intercal \mathcal{K}\mathcal{K}^\intercal \delta/2$, with $\hat{\Psi} = -\log \hat{L}$, denote the objective function after partialing out β and let subscripts to $\hat{\Omega}, \hat{\Psi}$ denote partial derivatives (omitting the transpose on the second) and m-subscripts to all objects indicate markets. In what follows it will be easiest to think of $\mathcal{K}\mathcal{K}^\intercal$ as a projection matrix, because that is what is up to finite scale. Indeed, in what is below we will follow the example of section 7 and assume without loss of generality that the optimal weight matrix is $(B^\intercal B)^{-1}$. Omitting a hat means the population equivalent with the same norming, i.e. Ψ is also a sum, with the understanding that everything is conditional on exogenous product-level objects, including instruments.[40] In the discussion below, we will think of S_m as being deterministic and let $\chi_m = S_m/N_m$, which can vary with the sample size.[41]

We focus on asymptotics for $\hat{\theta}$, which has the most interesting features, particularly $\hat{\theta}^\nu$. A standard extremum estimation expansion suggests[42]

$$\hat{\theta} - \theta \simeq -\begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \Omega_{\theta\theta} & \Omega_{\theta\delta} \\ \Omega_{\delta\theta} & \Omega_{\delta\delta} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\Omega}_\theta \\ \hat{\Omega}_\delta \end{bmatrix}, \tag{46}$$

such that we need to show that[43]

$$-\left(\Omega_{\theta\theta} - \Omega_{\theta\delta}\Omega_{\delta\delta}^{-1}\Omega_{\delta\theta}\right)^{-1/2}\left(\hat{\Omega}_\theta - \Omega_{\theta\delta}\Omega_{\delta\delta}^{-1}\hat{\Omega}_\delta\right) \xrightarrow{d} N(0, I). \tag{47}$$

The matrix $[I\ 0]$ serves the same role as Λ in the statement of theorem 2, with the only distinction being that here we have already concentrated out β.

F.2.2 Eliminating endogeneity

Recall from lemma 1 that $\mathcal{K}\mathcal{K}^\intercal = \mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X}$ where X can have endogenous elements. Now, $\mathcal{P}_B$ is exogenous, but

$$\mathcal{P}_{\mathcal{P}_B X} = B(B^\intercal B)^{-1}B^\intercal X\{X^\intercal B(B^\intercal B)^{-1}B^\intercal X\}^{-1}X^\intercal B(B^\intercal B)^{-1}B^\intercal.$$

To address the endogeneity issue, one simply replaces $X^\intercal B$ with $\bar{X}^\intercal B$ where $\bar{X}$ has its rows replaced with $\mathbb{E}(x_{jm} \mid B)$. This can be done without affecting asymptotics since the difference between the terms involving $X^\intercal B$ and $\bar{X}^\intercal B$ is dominated by the term involving $\bar{X}^\intercal B$. So from here on in this proof sketch, we will take X to be exogenous without loss of generality.

[40] So by population 'objects,' we mean that the various expectations are taken over everything except B, assuming that B includes the exogenous components of X.

[41] Random S_m are not usually a serious complication.

[42] Such an expansion does not automatically obtain here because δ grows in dimension, but that is a minor nuisance given the features of the model, as we will argue in appendix F.2.8.

[43] This requires the application of partitioned inverses to (46).

F.2.3 Everything is a sum over markets

The second thing to notice is that everything entails sums over markets, not some complicated inverses. This is obvious for $\hat{\Omega}_\theta = \hat{\Psi}_\theta = \sum_{m=1}^M \hat{\Psi}_{m\theta}$, but it also applies to the monstrosity

$$\Omega_{\theta\delta}\Omega_{\delta\delta}^{-1}\hat{\Omega}_\delta = \Psi_{\theta\delta}(\Psi_{\delta\delta}+\mathcal{K}\mathcal{K}^\intercal)^{-1}(\hat{\Psi}_\delta+\mathcal{K}\mathcal{K}^\intercal\delta) = \sum_{m=1}^M \Psi_{m\theta\delta\cdot m}\Psi_{m\delta\cdot m\delta\cdot m}^{-1}\left(\hat{\Psi}_{m\delta\cdot m}+\mathcal{K}_m\sum_{m^*=1}^M \mathcal{K}_{m^*}^\intercal\delta_{\cdot m^*}\right)$$
$$-\sum_{m=1}^M \Psi_{m\theta\delta\cdot m}\Psi_{m\delta\cdot m\delta\cdot m}^{-1}\mathcal{K}_m\Delta^{-1}\sum_{m=1}^M \mathcal{K}_m^\intercal\Psi_{m\delta\cdot m\xi\cdot m}^{-1}\left(\hat{\Psi}_{m\delta\cdot m}+\mathcal{K}_m\sum_{m^*=1}^M \mathcal{K}_{m^*}^\intercal\xi_{\cdot m^*}\right), \quad (48)$$

which follows from (29) plus $\mathcal{K}^\intercal X = 0$, and where $\Delta = \left(I + \sum_{m=1}^M \mathcal{K}_m^\intercal\Psi_{m\delta\cdot m\delta\cdot m}^{-1}\mathcal{K}_m\right)^{-1}$. The fact that everything is a sum over markets means that the fact that the dimension of δ increases only adds terms to each of these sums, so that the concern about the increasing matrix dimensions raised in section 8 is addressed.

F.2.4 Ugliness vanishes

The right hand side in (48) can be rewritten as $\mathcal{A}^\intercal(I - \mathcal{Z}(I+\mathcal{Z}^\intercal\mathcal{Z})^{-1}\mathcal{Z}^\intercal)a$, where the $\mathcal{A}^\intercal a$ portion corresponds to the first right hand side term and the remainder to the second. The relevant part of this reinterpretation is the fact that $\mathcal{Z}$ is a matrix consisting of vertically stacked blocks $\mathcal{Z}_m = \Psi_{m\delta\cdot m\delta\cdot m}^{-1/2}\mathcal{K}_m$ and that $\mathcal{Z}^\intercal\mathcal{Z}$ is negligible compared to I. This is so, because

$$\mathrm{tr}(\mathcal{Z}^\intercal\mathcal{Z}) = \sum_{m=1}^M \mathrm{tr}(\mathcal{K}_m^\intercal\Psi_{m\delta\cdot m\delta\cdot m}^{-1}\mathcal{K}_m) \leq \max_{m=1,\ldots,M}\frac{\mathrm{tr}(\mathcal{K}^\intercal\mathcal{K})}{\Psi_{\min}(\Psi_{m\delta\cdot m\delta\cdot m})} = \max_{m=1,\ldots,M}\frac{d_b - d_\beta}{\Psi_{\min}(\Psi_{m\delta\cdot m\delta\cdot m})} = o(1),$$

under mild regularity conditions. Note that if all N_m's diverge at the same rate (which they need not) then the order would typically be M/N.[44] Thus, we can ignore the second right hand side term in (48) from here on.

Thus, up to asymptotically negligible terms, the 'numerator' in (47) equals

$$\sum_{m=1}^M\left(\hat{\Psi}_{m\theta} - \Psi_{m\theta\delta\cdot m}\Psi_{m\delta\cdot m\delta\cdot m}^{-1}\hat{\Psi}_{m\delta\cdot m}\right) - \sum_{m=1}^M \Psi_{m\theta\delta\cdot m}\Psi_{m\delta\cdot m\delta\cdot m}^{-1}\mathcal{K}_m\sum_{m=1}^M \mathcal{K}_m^\intercal\xi_{\cdot m}. \quad (49)$$

Analogously, the 'denominator' portion of (47) is (up to asymptotically negligble) terms equal to

$$\left(\sum_{m=1}^M(\Psi_{m\theta\theta} - \Psi_{m\theta\delta\cdot m}\Psi_{m\delta\cdot m\delta\cdot m}^{-1}\Psi_{m\delta\cdot m\theta}) + \sum_{m=1}^M \Psi_{m\theta\delta\cdot m}\Psi_{m\delta\cdot m\delta\cdot m}^{-1}\mathcal{K}_m\sum_{m=1}^M \mathcal{K}_m^\intercal\Psi_{m\delta\cdot m\delta\cdot m}^{-1}\Psi_{m\delta\cdot m\theta}\right)^{-1/2}. \quad (50)$$

F.2.5 Separation of micro and macro likelihoods

Because of the separation into sums over markets described above and in view of the independence across markets, we now first look at the likelihoods in individual markets, before summing. Let $\mathcal{R}_{jm\theta}^{z_i} = \partial_\theta \log \pi_{jm}^{z_{im}} = \partial_\theta \pi_{jm}^{z_{im}}/\pi_{jm}^{z_{im}}$, $\mathcal{R}_{jm\theta} = \partial_\theta \log \pi_{jm}$, and let $\mathcal{R}$ with other subscripts be

[44] There are N_m identical terms in $\Psi_{m\delta\cdot m\delta\cdot m}$ and if all N_m's were equal then $1/N_m = M/N$.

analogously defined. Then by (10), $\Psi_{m\psi_m\psi_m}$ is the sum of

$$\begin{cases} \Psi^{\text{micro}}_{m\psi_m\psi_m} = S_m \sum_{j=0}^{J_m} \mathbb{E}\Big(\pi^{z_{im}}_{jm}(\mathcal{R}^{z_{im}}_{jm\psi_m} - \mathcal{R}_{jm\psi_m})(\mathcal{R}^{z_{im}}_{jm\psi_m} - \mathcal{R}_{jm\psi_m})^\intercal\Big), \\ \Psi^{\text{macro}}_{m\psi_m\psi_m} = N_m \sum_{j=0}^{J_m} \pi_{jm}\mathcal{R}_{jm\psi_m}\mathcal{R}^\intercal_{jm\psi_m}. \end{cases}$$

As noted in lemma 2, $\Psi^{\text{macro}}_{m\psi_m\psi_m}$ has rank J_m, which is consistent with the intuition of recovering $\delta_{\cdot m}$ once θ is known. But the upshot is that $\Psi_{m\psi_m\psi_m}$ has J_m eigenvalues diverging at rate N_m and d_θ eigenvalues diverging at a possibly slower rate: S_m if $\theta^z \neq 0$ is fixed. However, it is easy to show that $\Psi_{m\delta_{\cdot m}\delta_{\cdot m}}$ is invertible and diverges at rate N_m.

F.2.6 The expectations meet expectations

It is straightforward to show that

$$\begin{cases} \mathbb{E}\Big(\hat{\Psi}^{\text{micro}}_{m\psi_m}(\hat{\Psi}^{\text{micro}}_{m\psi_m})^\intercal \,\Big|\, x\Big) = \Psi^{\text{micro}}_{m\psi_m\psi_m}, \\ \mathbb{E}\Big(\hat{\Psi}^{\text{micro}}_{m\psi_m}(\hat{\Psi}^{\text{macro}}_{m\psi_m})^\intercal \,\Big|\, x\Big) = 0, \\ \mathbb{E}\Big(\hat{\Psi}^{\text{macro}}_{m\psi_m}(\hat{\Psi}^{\text{macro}}_{m\psi_m})^\intercal \,\Big|\, x\Big) = \Psi^{\text{macro}}_{m\psi_m\psi_m}. \end{cases}$$

Indeed, this is essentially the information matrix equality. For instance, by the law of iterated expectations (conditioning on x, ξ),

$$\mathbb{E}\Big(\hat{\Psi}^{\text{micro}}_{m\psi_m}(\hat{\Psi}^{\text{macro}}_{m\psi_m})^\intercal \,\Big|\, x\Big) = S_m \sum_{j=0}^{J_m} \mathbb{E}\Big(\mathbb{E}\Big(\pi^{z_{im}}_{jm}(\mathcal{R}^{z_{im}}_{jm\psi_m\psi_m} - \mathcal{R}_{jm\psi_m\psi_m}) \,\Big|\, x, \xi\Big)\mathcal{R}^\intercal_{jm\psi_m\psi_m} \,\Big|\, x\Big) =$$
$$S_m \sum_{j=0}^{J_m} \mathbb{E}\Big(\mathbb{E}(\partial_{\psi_m}\pi^{z_{im}}_{jm} - \partial_{\psi_m}\pi_{jm} \mid x, \xi)\mathcal{R}^\intercal_{jm\psi_m\psi_m} \,\Big|\, x\Big) = 0.$$

Consequently, $\mathbb{E}(\hat{\Psi}_{m\theta}\hat{\Psi}^\intercal_{m\theta} \mid x) = \Psi_{m\theta\theta}$, $\mathbb{E}(\hat{\Psi}_{m\theta}\hat{\Psi}^\intercal_{m\delta_{\cdot m}} \mid x) = \Psi_{m\theta\delta_{\cdot m}}$, and $\mathbb{E}(\hat{\Psi}_{m\delta_{\cdot m}}\hat{\Psi}^\intercal_{m\delta_{\cdot m}} \mid x) = \Psi_{m\delta_{\cdot m}\delta_{\cdot m}}$. Now, note that $\mathcal{K}^\intercal\delta = \mathcal{K}^\intercal\xi$ because β has been concentrated out. Further, by the assumptions at the beginning of this section, $\mathcal{K}\mathcal{K}^\intercal\mathbb{E}(\xi\xi^\intercal \mid b)\mathcal{K}\mathcal{K}^\intercal = \mathcal{K}\mathcal{K}^\intercal$. So the variation in the $\mathcal{K}\mathcal{K}^\intercal\xi$ term in the second expression of (48) is accounted for by the $\mathcal{K}\mathcal{K}^\intercal$ term inside the inverse in the same expression.

In sum, conditional on x, b, the expectation of the outer product of the product of (49) and (50) is the identity matrix. This is tedious, but uneventful, to show.

F.2.7 Applying limit results

Note that (49) is the sum of two components. The first term is i.i.d. conditional on the second and hence has a limiting normal distribution conditional on the second by e.g. Eicker's central limit theorem. The second term consists of martingale difference sequences and since both conditions in Davidson (1994, theorem 24.3) are satisfied, the sum of the two components converges to a standard normal.

F.2.8 About that extremum estimation expansion

In (48) we ignored higher order terms. This would be uneventful in a standard setting. We now explore a bit more why this is reasonable here. First, consider the estimation of $\delta(\tilde{\theta})$ for arbitrary fixed $\tilde{\theta}$. The first order condition is

$$0 = \hat{\Psi}_\delta\{\tilde{\theta}, \hat{\delta}(\tilde{\theta})\} + \mathcal{K}\mathcal{K}^\intercal \hat{\delta}(\tilde{\theta}). \tag{51}$$

Note that (51) holds for any value of $\tilde{\theta}$, but we only need to consider values of $\tilde{\theta}$ near the truth. We first focus on the likelihood component of (51). Note that

$$\hat{\Psi}_\delta\{\tilde{\theta}, \hat{\delta}(\tilde{\theta})\} = \sum_{m=1}^{M} \hat{\Psi}_{m\delta_{\cdot m}}\{\tilde{\theta}, \hat{\delta}_{\cdot m}(\tilde{\theta})\}.$$

Given that $\hat{\delta}_{\cdot m}$ is finite-dimensional and $\Psi_{m\delta_{\cdot m}\delta_{\cdot m}}$ is positive definite, standard extreme value theory would suggest that

$$\hat{\Psi}_{m\delta_{\cdot m}}\{\tilde{\theta}, \hat{\delta}_{\cdot m}(\tilde{\theta})\} \simeq \hat{\Psi}_{m\delta_{\cdot m}}\{\tilde{\theta}, \delta_{\cdot m}(\tilde{\theta})\} + \Psi_{m\delta_{\cdot m}\delta_{\cdot m}}\{\tilde{\theta}, \delta_{\cdot m}(\tilde{\theta})\}\{\hat{\delta}_{\cdot m}(\tilde{\theta}) - \delta_{\cdot m}(\tilde{\theta})\}, \tag{52}$$

where $\simeq$ again means that omitted terms are negligible. Note that $\Psi_{m\delta_{\cdot m}\delta_{\cdot m}}$ is positive definite and generally has minimum eigenvalue diverging at rate N_m, such that plugging (52) back into (51), solving for $\hat{\delta}$, and ignoring dominated terms yields

$$\hat{\delta}_{\cdot m}(\tilde{\theta}) - \delta_{\cdot m}(\tilde{\theta}) \simeq -\left[\Psi_{m\delta_{\cdot m}\delta_{\cdot m}}\{\tilde{\theta}, \delta_{\cdot m}(\tilde{\theta})\}\right]^{-1} \hat{\Psi}_{m\delta_{\cdot m}}\{\tilde{\theta}, \delta_{\cdot m}(\tilde{\theta})\} = O_p(N_m^{-1/2}). \tag{53}$$

Note that the convergence rate in (53) is the rate at a given $\tilde{\theta}$, so it is different from, indeed no slower than, the rate at which $\hat{\delta}_{\cdot m}(\hat{\theta}) - \delta_{\cdot m}(\theta)$ converges.

There are two issues that we have glossed over getting from (51) to (53). The first issue is that we are doing everything at a fixed value of $\tilde{\theta}$ instead of as a function. However, given the degree of smoothness, this will not be a problem. The second issue is that the number of markets M, and hence the number of $\delta_{\cdot m}$'s, increases with M. However, in view of how these differences are to be used below, the rate of the omitted term in (53), and our conditions on the various rates, this too will be of secondary concern.

Now, note that

$$\begin{aligned}
0 = \hat{\Omega}_\theta\{\hat{\theta}, \hat{\delta}(\hat{\theta})\} &\simeq \hat{\Omega}_\theta(\theta, \delta) + \Omega_{\theta\theta}(\theta, \delta)(\hat{\theta} - \theta) + \sum_{m=1}^{M} \Omega_{\theta\delta_{\cdot m}}(\theta, \delta)\{\hat{\delta}_{\cdot m}(\hat{\theta}) - \delta_{\cdot m}(\theta)\} \simeq \\
&\hat{\Omega}_\theta(\theta, \delta) + \Omega_{\theta\theta}(\theta, \delta)(\hat{\theta} - \theta) + \sum_{m=1}^{M} \Omega_{\theta\delta_{\cdot m}}(\theta, \delta)\{\hat{\delta}_{\cdot m}(\theta) - \delta_{\cdot m}(\theta)\} + \sum_{m=1}^{M} \Omega_{\theta\delta_{\cdot m}}(\theta, \delta)\{\delta_{\cdot m}(\hat{\theta}) - \delta_{\cdot m}(\theta)\} \\
&\simeq \hat{\Omega}_\theta(\theta, \delta) - \sum_{m=1}^{M} \Omega_{\theta\delta_{\cdot m}}(\theta, \delta)\{\Omega_{\delta_{\cdot m}\delta_{\cdot m}}(\theta, \delta)\}^{-1} \hat{\Omega}_{\delta_{\cdot m}} + \left(\Omega_{\theta\theta}(\theta, \delta) + \sum_{m=1}^{M} \Omega_{\theta\delta_{\cdot m}}(\theta, \delta)\partial_{\theta^\intercal}\delta_{\cdot m}(\theta)\right)(\hat{\theta} - \theta),
\end{aligned}$$

which by the implicit function theorem takes us to (46) up to terms that are shown to be negligble in appendix F.2.4.[45]

[45] The implicit function theorem is needed to obtain $\partial_{\theta^\intercal}\delta = -\Omega_{\delta\delta}^{-1}\Omega_{\delta\theta}$.

F.2.9 Final comments

The results above are different from those in Berry, Linton and Pakes (2004) because here we assume that $M \to \infty$ and $\max_m J_m$ is finite, whereas there $S = 0$, $N = \infty$, $M = 1$, $J_1 \to \infty$. So our results neither imply nor are implied by those there.

G Miscellanea

G.1 Weight matrix is block-diagonal

Note that the expectation of the score of $\log \hat{L}$ given x, ξ is for $\gamma = [\beta^\intercal, \theta^\intercal, \delta^\intercal]^\intercal$ under random sampling equal to

$$\mathbb{E}\left(\sum_{m=1}^{M} \sum_{i=1}^{N_m} D_{im} \sum_{j=0}^{J_m} \frac{Y_{ijm}}{\pi_{jm}^{z_{im}}} \partial_\gamma \pi_{jm}^{z_{im}} + \sum_{m=1}^{M} \sum_{i=1}^{N_m} (1 - D_{im}) \sum_{j=0}^{J_m} (1 - D_{im}) \frac{Y_{ijm}}{\pi_{jm}} \partial_\gamma \pi_{jm} \,\middle|\, x, \xi \right) =$$

$$\mathbb{E}\left(\sum_{m=1}^{M} \sum_{i=1}^{N_m} D_{im} \partial_\gamma \underbrace{\sum_{j=0}^{J_m} \pi_{jm}^{z_{im}}}_{=1} + \sum_{m=1}^{M} \sum_{i=1}^{N_m} (1 - D_{im}) \partial_\gamma \underbrace{\sum_{j=0}^{J_m} \pi_{jm}}_{=1} \,\middle|\, x, \xi \right) = 0.$$