Conformant and Efficient Estimation of Discrete Choice Demand Models*

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

this version: May 25, 2023

Abstract

We propose a conformant likelihood-based estimator with exogeneity restrictions (CLER) for random coefficients discrete choice demand models that is applicable in a broad range of data settings. 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 et al. (1995, BLP95), random coefficients discrete choice demand models provide a tractable framework 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 applies 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, along with product level exogeneity restrictions. We impose no additional assumptions over those posited in BLP95, which are also used in other estimators extended with consumer level data (e.g., Petrin 2002; Berry et al. 2004*a* (BLP04); Goolsbee and Petrin 2004; Chintagunta and

*We thank Nikhil Agarwal, Steve Berry, Chris Conlon, Amit Gandhi, Jeff Gortmaker, Jessie Handbury, Phil Haile, Jean-François Houde, Sung Jae Jun, Nail Kashaev, Mathieu Marcoux, Karl Schurter, Andrew Sweeting, and many seminar and conference participants. A companion `Julia` package `Grumps` is available with accompanying documentation. A previous version of this paper was circulated with the title "Efficient Estimation of Random Coefficients Demand Models using Product and Consumer Datasets."

[†]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.

Dube 2005).

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. Conformancy 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, and 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. However, 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 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 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, the identification of unobserved preference heterogeneity depends 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

[1] BLP95 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.

endogenous characteristics (such as price).

Our full estimator extends the MDLE approach with an additional term to directly incorporate information contained in the product level exogeneity restrictions of BLP95. This estimator achieves the conformance property, so we refer to it as the Conformant Likelihood with Exogeneity Restrictions (CLER) estimator. The exogeneity restrictions are additional assumptions on the data-generating process, providing a distinct source of identifying variation beyond the likelihood. The main benefit of CLER relative to MDLE arises when there are more exogeneity restrictions than product characteristics. In the presence of such overidentification, the extra information can help identify the preference heterogeneity parameters even when they are not 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 zero. The primary contribution of this paper is to provide an estimator that fully exploits these two sources of identifying variation to achieve the fastest possible rate of convergence, efficiency, and valid inference without relying on any pre-test of the data or tuning parameters.

The CLER estimator is compatible with all datasets in the applied literature of which we are aware. In particular, it is well-behaved with consumer samples of any size, from zero to a full census of the market. The objective function comprises three terms that can diverge at different rates: the micro loglikelihood with the consumer sample size, the macro loglikelihood with the market size, and a 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: the rates of convergence will adjust accordingly and depend on the relative sample sizes and strength of information from the three terms.[2]

The conformance property results from the CLER estimator incorporating 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, GH20), 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 (if present) in the micro sample will produce a faster convergence rate than relying on product level exclusion restrictions. In this case, MDLE

[2]The use of the plural 'rates' is because different elements of our estimator vector converge at different rates.
[3]Berry and Haile (2020) make a similar point in a nonparametric context.

and CLER are asymptotically equivalent and efficient. Adding the product level exclusions to the estimator is useful when the consumer sample is small (or not present) or 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. The CLER 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.

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. In contrast, the moments component involves sums over products where variation in product quality gives rise to the product level structural error term. The optimal weight matrix to use is the same as that in standard GMM estimation, except now the scale matters to properly weight across likelihood and GMM terms, as we describe in sections 3.2 and 4.1.

We show that conducting inference using formulas familiar from the standard extremum estimation framework is asymptotically valid. We formally establish consistency and asymptotic normality in theorem 1, whose proof is nonstandard to accommodate the conformance features of the CLER estimator. Validity obtains regardless of the relative divergence rates[4] and even though the vector of product quality parameters increases in dimension. 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.

Another advantage of the CLER estimator over alternative methods popular in the literature that use an objective function with a constraint (product shares must match choice probabilities exactly) is more robust inference. In particular, methods that impose a share constraint require that the total number of consumers S in the micro sample across all markets is negligibly small

[4] E.g., the number of markets, the number of consumers in the population of each market, the number of consumers in the micro sample, and number of products.

compared to the smallest market size $\min_m N_m$ and, if the product quality parameters are of interest (e.g., as in merger simulation exercises), even that S is negligibly small compared to $\min_m \sqrt{N_m}$.[5] Absent these additional restrictions, the computed standard errors would be too small.

While the statistical properties of our estimator make it of theoretical interest, it is also 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. We discuss numerical integration in section 7.2 and provide a Monte Carlo illustrating performance in section 9.2.4.

The CLER estimator is most directly comparable to GMM approaches based on micro-moments (e.g. Petrin 2002 and BLP04). In related work, Conlon and Gortmaker (2023, CG23) provide a comprehensive discussion of best practices for incorporating moments based on a variety of types of auxiliary consumer level data into this canonical GMM-based estimation of BLP-style models.This framework does not share our properties of efficiency and conformancy. That said, the GMM approach may be better suited to certain types of data, for example, a situation when the researcher only has access to summary statistics from an individual-level survey instead of the individual responses themselves or where the model is so complex that analytic formulas for the Hessian are difficult to obtain, which would make computation of our estimator more expensive.

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; Bachmann et al., 2019).[6] The key difference with our approach is twofold. First, they use a two-stage procedure, and so cannot take full advantage of the combination of consumer choice data, product market shares, and over-identifying product level restrictions. Second, like Petrin (2002) and BLP04, these papers recover product quality parameters using the BLP inversion, whereas our approach achieves efficiency by estimating product quality parameters using the

[5] In BLP95 and BLP04 the N_m's are assumed to be effectively infinite.

[6] MLE is a popular choice for estimating discrete choice models that do not have endogenous product characteristics; see e.g. hospital choice as in Ho (2006) and urban/location models such as Bayer et al. (2007). Our framework nests these applications.

entire CLER objective function. Another related paper is Allen et al. (2019), who combine the likelihood of an equilibrium search model with a penalty term of moment equalities.

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 BLP04 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), Ciliberto and Kuminoff (2010), Crawford and Yurukoglu (2012), Starc (2014), Wollmann (2018), Crawford et al. (2018), Hackmann (2019), Neilson (2019), Backus et al. (2021), Grieco et al. (2023), Montag (2023), and Jiménez-Hernández and Seira (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 et al. (2021) (Nielsen).

Finally, our problem and approach share features with several strands of the econometrics literature. For instance, Imbens and Lancaster (1994) also consider 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. Ridder and Moffitt (2007) provide a survey of methods to combine different data sets and van den Berg and van der Klaauw (2001) combine data sets to estimate a duration model. 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.

There are several econometrics papers that cover random coefficient discrete choice models with only product-level data. The first such paper is Berry et al. (2004*b*, BLiP04). Freyberger

(2015) and Hong et al. (2021) are closer in spirit to ours in that the number of markets increases, whereas in BLiP04 the number of products increases but the number of markets is fixed. Myojo and Kanazawa (2012) show how additional moments can be constructed on the basis of consumer level data and discuss supply side restrictions.

The following section reviews the random coefficients demand model and the data available in our setting. Section 3 proposes our estimator. Conformance and efficiency properties are described in section 4. Section 5 explores the source of variation in the demographic data exploited to identify taste heterogeneity. Section 6 illustrates the trade-offs in going from the CLER estimator to the GMM estimators that are commonly used in applied work. Section 7 argues for the computational tractability of the CLER estimator; we provide a software package to demonstrate feasibility and facilitate implementation.[7] Section 8 introduces our inference procedure. Section 9 compares the finite sample properties of CLER relative to MDLE and a canonical GMM estimator. Section 10 concludes.

2 Random Coefficients Demand Model

This section briefly reviews the random coefficients discrete choice demand model and describes the data used by our estimator. The model matches that of BLP95 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-dimensional 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 $(z_{im}, \nu_{im}, \varepsilon_{i\cdot m})$ where z_{im} is a d_z-vector of potentially observable consumer characteristics (such as income or location), and ν_{im} is a $d_\nu \leq d_x$-vector of unobservable consumer taste shocks to preferences for product charac-

[7] The `Grumps` package is available at https://github.com/NittanyLion/Grumps.jl.

teristics. Finally $\varepsilon_{i\cdot m}$ is a $J_m + 1$-vector of idiosyncratic product taste shocks for each product and an outside good (e.g., no purchase) that is distributed according to the standard Type-I extreme value (Gumbel) distribution. In the population, 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

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

where $\delta_{jm} = x_{jm}^\top \beta + \xi_{jm}$ represents the mean utility for product j for consumers in market m, $\mu_{jm}^{z_{im}} = \mu^z(x_{jm}, z_{im}; \theta^z)$ represents deviations from mean utility due to observed demographic variables z_{im}, and $\mu_{jm}^{\nu_{im}} = \mu^\nu(x_{jm}, \nu_{im}; \theta^\nu)$ are deviations due to taste shocks ν_{im}. There is no real need to assume δ_j has this linear form but this is the most common specification. 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, μ^ν is typically a linear combination of product characteristics and taste shocks parameterized by θ^ν. Utility of the outside good is normalized to $u_{i0m} = \varepsilon_{i0m}$. When convenient, we collect the consumer heterogeneity parameters into the vector $\theta = [\theta^{z\top}, \theta^{\nu\top}]^\top$.

The model yields choice probabilities for consumer i of selecting product j conditional on demographics z_{im} and product characteristics $x_{\cdot m}$ as a function of parameters,

$$\pi_{jm}^{z_{im}}(\theta, \delta) = \Pr(y_{ijm} = 1 \mid z_{im}, x_{\cdot m}; \theta, \delta) = \int \underbrace{\frac{\exp(\delta_{jm} + \mu_{ijm}^z + \mu_{ijm}^\nu)}{\sum_{\ell=0}^{J_m} \exp(\delta_{\ell m} + \mu_{i\ell m}^z + \mu_{i\ell m}^\nu)}}_{s_{jm}(z_{im}, \nu; \theta, \delta)} \mathrm{d}F_m(\nu), \tag{2}$$

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

Similarly, unconditional choice probabilities, which correspond to expected market shares, are obtained by integrating π_{jm}^z with respect to the distribution of consumer demographics,

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

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

$$\mathbb{E}(\xi_{jm} b_{jm}) = 0, \tag{3}$$

where b_{jm} is a vector of instruments which includes $\tilde{x}_{jm}$. Further, b_{jm} may contain additional exogeneity restrictions. The literature has used various approaches such as cost shifters, BLP instruments, Hausman instruments, Waldfogel instruments, and differentiation 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 $d_b > d_\beta$ they may provide additional information that is potentially useful in estimating other model parameters. For example BLP95 uses restrictions of this form to 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, the 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

$$s_{jm} = \frac{1}{N_m} \sum_{i=1}^{N_m} y_{ijm}. \tag{4}$$

Note that the observed market shares $s_{\cdot m}$ need not equal choice probabilities $\pi_{\cdot m}$ due to the finite population size, 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 consumers' choices and their demographics. 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 inte-

[8] One could replace (3) 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 (3), with the instruments b_{jm} in the second step generated from the first step.

[9] As in the previous literature, researchers need to observe or make an assumption regarding N_m in order to compute market shares from purchase quantity data.

grating out z_{im} in the choice probabilities when individual i is outside the micro-sample. We can accommodate several forms of selection. In appendix A 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

This section proposes the CLER estimator which combines the likelihood, $\hat{L}(\theta, \delta)$, of the micro and macro choice data and an efficient GMM objective function $\hat{\Pi}$ based on (3),

$$(\hat{\beta}, \hat{\theta}, \hat{\delta}) = \arg\min_{\beta,\theta,\delta} \Big(-\log \hat{L}(\theta, \delta) + \hat{\Pi}(\beta, \delta) \Big) \tag{5}$$

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. The first subsection describes the mixed data likelihood, which alone is the objective function for the MDLE. The second subsection introduces the product level moments term, $\hat{\Pi}$.

3.1 Mixed Data Likelihood

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

$$\log \hat{L}(\theta, \delta) = \sum_{m=1}^{M} \sum_{j=0}^{J_m} \sum_{i=1}^{N_m} y_{ijm} \left(D_{im} \log \pi_{jm}^{z_{im}}(\theta, \delta) + (1 - D_{im}) \log \pi_{jm}(\theta, \delta) \right), \tag{6}$$

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 can be equivalently written in terms of the consumer level observations and the market level share data,

[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 A, selection on demographics requires an adjustment to account for sampling in π_{jm}.

$$\log \hat{L}(\theta, \delta) = \underbrace{\sum_{m=1}^{M} \sum_{j=0}^{J_m} \sum_{i=1}^{N_m} D_{im} y_{ijm} \log \frac{\pi_{jm}^{z_{im}}}{\pi_{jm}}}_{\text{micro}} + \underbrace{\sum_{m=1}^{M} N_m \sum_{j=0}^{J_m} s_{jm} \log \pi_{jm}}_{\text{macro}}, \tag{7}$$

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 second 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. It is convenient to refer to the two terms of the likelihood separately, so we define $\log \hat{L}^{\text{mic}}$ and $\log \hat{L}^{\text{mac}}$ as the micro and macro terms of (7), respectively. Alternatively, the estimator can be written by adjusting the macro term to avoid double counting the consumers in the micro-sample:

$$\log \hat{L}(\theta, \delta) = \underbrace{\sum_{m=1}^{M} \sum_{j=0}^{J_m} \sum_{i=1}^{N_m} D_{im} y_{ijm} \log \pi_{jm}^{z_{im}}}_{\text{micro}} + \underbrace{\sum_{m=1}^{M} \sum_{j=0}^{J_m} \left(N_m s_{jm} - \sum_{i=1}^{N_m} D_{im} y_{ijm} \right) \log \pi_{jm}}_{\text{macro}}, \tag{8}$$

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 MDLE 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 et al., 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 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 MDLE 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 CLER estimator combines the MDLE objective with an additional term that penalizes violations of the product level moments,

$$\hat{\Pi}(\beta,\delta) = \frac{1}{2}\hat{m}^{\top}(\beta,\delta)\hat{\mathcal{W}}\hat{m}(\beta,\delta) \tag{9}$$

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

$$\hat{m}(\delta,\beta) = \sum_{m=1}^{M}\sum_{j=1}^{J_m} b_{jm}(\delta_{jm} - \beta^{\top}x_{jm}). \tag{10}$$

Note that, unlike in standalone GMM estimation, the factor $1/2$ in front of the 'J statistic' in (9) matters since it affects the relative weight placed on the likelihood versus the 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 via MDLE 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 product level restrictions can be asymptotically negligible for θ, δ as we discuss in section 4.1.

4 Properties

The CLER 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 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

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

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

of the estimator under a wide variety of circumstances, completing the efficiency argument. Section 8 provides a valid inference procedure.

4.1 Efficiency

The CLER estimator is efficient under a wide range of circumstances. To see this, it is convenient to first consider the gradient of the CLER objective function (5),

$$\begin{bmatrix} \partial_\beta \hat{m}^\top \hat{\mathcal{W}} \hat{m} \\ -\partial_\theta \log \hat{L} \\ -\partial_\delta \log \hat{L} + \partial_\delta \hat{m}^\top \hat{\mathcal{W}} \hat{m} \end{bmatrix}. \tag{11}$$

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

$$\underset{\beta,\theta,\delta}{\arg\min} \frac{1}{2} \begin{bmatrix} \hat{m}^\top & \partial_{\psi^\top} \log \hat{L} \end{bmatrix} \begin{bmatrix} \hat{\mathcal{W}} & 0 \\ 0 & \hat{\mathcal{W}}_L \end{bmatrix} \begin{bmatrix} \hat{m} \\ \partial_\psi \log \hat{L} \end{bmatrix}, \tag{12}$$

where $\psi = [\theta^\top, \delta^\top]^\top$ and $\hat{\mathcal{W}}_L = (-\partial_{\psi\psi^\top} \log \hat{L})^{-1}$ evaluated at the solution $\hat{\psi}$ of (5).[13] Note that in (12) there may be more moments than parameters. Specifically, (11) has $d_\beta + d_\theta + d_\delta$ moments, whereas (12) is based on $d_b + d_\theta + d_\delta$ moments. Under exact identification of (12), i.e. if $d_b = d_\beta$, both (11) and (12) 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 (12) is

$$\begin{bmatrix} \partial_\beta \hat{m}^\top \hat{\mathcal{W}} \hat{m} \\ 0_{d_\theta} \\ \partial_\delta \hat{m}^\top \hat{\mathcal{W}} \hat{m} \end{bmatrix} + \begin{bmatrix} 0_{d_\beta} \\ \partial_{\theta\psi^\top} \log \hat{L} \hat{\mathcal{W}}_L \partial_\psi \log \hat{L} \\ \partial_{\delta\psi^\top} \log \hat{L} \hat{\mathcal{W}}_L \partial_\psi \log \hat{L} \end{bmatrix}, \tag{13}$$

which yields (11) at the solution since $\hat{\mathcal{W}}_L = (-\partial_{\psi\psi^\top} \log \hat{L})^{-1}$, establishing the equivalence of these estimators.

Next, we argue that (12) is efficient. First, by the law of iterated expectations, at the truth,

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

where the second equality follows from the the likelihood principle applied to the choice problem

[13] We define $\hat{\mathcal{W}}_L$ in terms of (5) in case its gradient (11) is zero at multiple points.

(without product level moments); see appendix B 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 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 the CLER estimator to the GMM estimators described in (11) and (12). First, the population analog of (11) can have multiple solutions even if the population analog of our objective function (5) 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 (12) would be unwieldy because of the high degree of nonlinearity and the dimension of δ. We show in section 7 that the CLER estimator can be tractably computed despite the dimensionality of δ.

4.2 Conformant convergence

We now show that the CLER estimator is conformant. The objective function in (5) is the sum of three terms that diverge at different rates. The micro loglikelihood is the sum over S consumers, the macro loglikelihood in (7) 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 $\sqrt{J}$ (assuming there are at least d_β strong product level moments) since it is only identified off $\hat{\Pi}$.

The remainder of this section enumerates cases defined in terms of (relative) divergence rates to which the CLER 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).[14] For exposition, 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 identification of θ, δ can still come from consumer level data.

To build intuition and connect our convergence results to the previous literature, note that the one-to-one mapping between shares and δ_m *as a function of* θ (Berry, 1994), can be estimated at rate $\sqrt{N_m}$ because we assume that J_m is finite. To see this, first note that the (macro) shares converge at rate $\sqrt{N_m}$. Thus, for given θ, the convergence rate of an estimator $\hat{\delta}_m(\theta)$ of $\delta_m(\theta)$ using share data alone, would converge at rate $\sqrt{N_m}$. Micro variation does not improve the convergence rate of $\hat{\delta}_m(\theta)$ for a given θ because we still only have N_m observations from market m. The convergence rate of the estimator $\hat{\delta}_m$ of the parameter δ_m can be slower than $\sqrt{N_m}$ since θ must also be estimated. Indeed, $\hat{\delta}_m - \delta_m = \hat{\delta}_m(\hat{\theta}) - \delta_m(\theta)$ so the convergence rate of $\hat{\delta}_m$ is the slower of $\sqrt{N_m}$ and the convergence rate of $\hat{\theta}$. 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.

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 (if either) is appropriate.

	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).

[14] This is in contrast to BLiP04 which assumes that the number of markets is fixed.

4.2.1 θ^z **is fixed.** Table 1 lists several cases where the parameters are fixed ordered by importance of the log $\hat{L}^{\text{mic}}$ term for the asymptotic behavior of (5).

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 $(\hat{\theta}, \hat{\delta})$.[15] 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 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 collinear. Indeed, if $\theta^z = 0$ then $\mathcal{s}_{jm}(z, \nu)$ is flat in z and the scores with respect to θ^ν and δ depend on the micro data through $\sum_{i=1}^{N_m} D_{im} y_{ijm}$ only.[16] As a result, θ^ν and 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 $\mathcal{s}_{jm}$ is flat 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 $\hat{L}^{\text{mic}}$ 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 $\hat{L}^{\text{mic}}$. 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

[15] 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.

[16] The scores of log $\hat{L}^{\text{mic}}$ with respect to θ^z and θ^ν are in (20) and (22). The score with respect to δ_{jm} is $\sum_{i=1}^{N_m} \sum_{\ell=0}^{J_m} (D_{im} y_{i\ell m} \,/\, \pi_{\ell m}^{z_{im}}) \int \mathcal{s}_{\ell m}(z_{im}, \nu) \Big(\mathbb{1}(\ell = j) - \mathcal{s}_{jm}(z_{im}, \nu) \Big) \, \mathrm{d}F(\nu)$.

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 $\log \hat{L}^{\text{mic}} = 0$. This is the environment of BLP95 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.

case	rate θ^z	rate θ^ν, δ	contributing term(s) for θ^z	contributing term(s) for θ^ν
$\sqrt{S\lambda^2/J} \to \infty, S/J \to \infty$	$\sqrt{S}$	$\sqrt{S\lambda^2}$	$\log \hat{L}$	$\log \hat{L}$
$\sqrt{S\lambda^2/J} \to c, S/J \to \infty$	$\sqrt{S}$	$\sqrt{J}$	$\log \hat{L}$	both
$\sqrt{S\lambda^2/J} \to 0, S/J \to \infty$	$\sqrt{S}$	$\sqrt{J}$	$\log \hat{L}$	$\hat{\Pi}$
$\sqrt{S\lambda^2/J} \to c, S/J \to c$	$\sqrt{J}$	$\sqrt{J}$	both	both
$\sqrt{S\lambda^2/J} \to 0, S/J \to c$	$\sqrt{J}$	$\sqrt{J}$	both	$\hat{\Pi}$
$\sqrt{S\lambda^2/J} \to 0, S/J \to 0$	$\sqrt{J}$	$\sqrt{J}$	$\hat{\Pi}$	$\hat{\Pi}$

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).

4.2.2 θ^z **can drift.** In section 4.2.1 there is a discontinuity in the asymptotic behavior of the CLER 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.[17] We denote the drifting rate by λ, so $\lambda = \|\theta^z\|$. Table 2 summarizes these cases, which are again ordered in decreasing importance of the micro likelihood for asymptotic behavior of (5).

In the first row in table 2, $\sqrt{S\lambda^2/J} \to \infty$ which is equivalent to the first row of table 1 in terms of asymptotic behavior. In the next two cases, $\log \hat{L}^{\text{mic}}$ diverges faster than $\hat{\Pi}$, but the two cases differ in the strength of identification they provide due to $\lambda \to 0$ at different rates. The knife edge case where the rate of λ is such that $\sqrt{S\lambda^2/J}$ goes to a constant has no analog in table 1. Here both $\log \hat{L}^{\text{mic}}$ and $\hat{\Pi}$ contribute to the limit distribution of $\hat{\theta}^\nu$ because the faster divergence of $\log \hat{L}^{\text{mic}}$ is just offset by the convergence of λ. The case where $\sqrt{S\lambda^2/J} \to 0$ 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.

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

Estimation method	$\hat{\delta}$	$\hat{\theta}^z$	$\hat{\theta}^\nu$	$\hat{\beta}$
CLER: $-\log \hat{L} + \hat{\Pi}$	$\min(\sqrt{S}, \sqrt{N_m})$	$\sqrt{S}$	$\sqrt{S}$	$\sqrt{J}$
MDLE: $-\log \hat{L}$, 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	$\hat{\delta}$	$\hat{\theta}^z$	$\hat{\theta}^\nu$	$\hat{\beta}$
CLER: $-\log \hat{L} + \hat{\Pi}$	$\min\{\max(\sqrt{J}, \sqrt{S\lambda^2}), \sqrt{N_m}\}$	$\sqrt{S}$	$\max(\sqrt{J}, \sqrt{S\lambda^2})$	$\sqrt{J}$
MDLE: $-\log \hat{L}$, then $\hat{\Pi}$	$\min\{\sqrt{S}\lambda, \sqrt{N_m}\}$	$\sqrt{S}$	$\sqrt{S\lambda^2}$	$\min(\sqrt{S\lambda^2}, \sqrt{J})$
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 $d_b \geq d_\beta + d_{\theta^\nu}$

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 the CLER estimator 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 MDLE 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 (12), 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 the CLER estimator, the rate λ determines which of the first three rows in table 2 applies. The MDLE two-step, 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. An estimator relying on $\hat{m}$ to estimate θ^ν is not affected by the fact that the likelihood provides less information than in the

[17] 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.

leading case, because it was not using that information anyway. The CLER estimator 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 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.

Consider a simple case of a single market with two products and an outside good. There is a single demographic variable, so z_i is a scalar.[18] Utility for product j is

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

where the product characteristics are $x^{(1)} = [1 \quad 0]^\top$, $x^{(2)} = [1 \quad 1]^\top$. The demographic variable shifts utility of good 1 only, and the single random coefficient induces correlation in the utilities of the two inside goods. 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 $\tilde{\pi}^z = \Pr(y_{i\cdot} = 1 \mid z, x)$. Figure 1 plots this function over $z \in [-1, 1]$ for three different parametrization of the model, namely $\theta^\nu = \{0, 1, 2\}$ with $\delta = (-.25, 25)^\top$ and $\theta^z = 2$. 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, $\tilde{\pi}_2^z$, relative to the outside good, $\tilde{\pi}_0^z$. 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 induces relatively more substitution away from good 2 than the outside good.

We can also nonparametrically identify the derivatives of $\tilde{\pi}^z$. Given our special case we have, $\mathrm{d}_z \tilde{\pi}_j^z = \theta^z \partial_{u_1} \pi_j^z$, where we employ the fact that z only affects the utility of good 1. Taking a ratio of these gives us diversion with respect to utility from good 1 to good 2 and from good 1 to the outside good for every value of z, i.e., for $j = \{0, 2\}$,

$$\frac{\mathrm{d}_z \tilde{\pi}_j^z}{\mathrm{d}_z \tilde{\pi}_1^z} = \frac{\partial_{u_1} \pi_j^z}{\partial_{u_1} \pi_1^z} = D_{1j}^z. \tag{14}$$

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

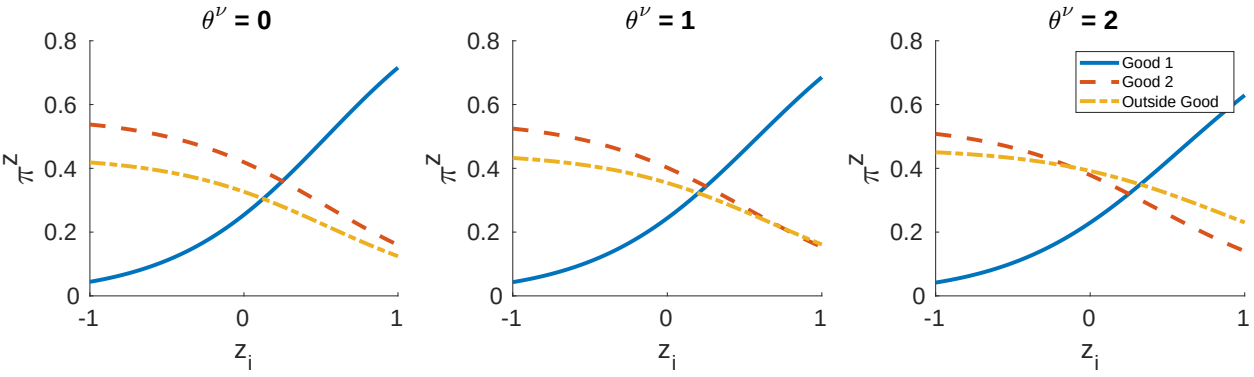

Figure 1: Conditional shares $\tilde{\pi}^z$ are identified by the micro sample.

Equation (14) 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 $D_{1j}^{z_i} = \pi_j^z/(1-\pi_1^z)$. Moreover, due to the independence of irrelevant alternatives property, diversion will be constant over z.

Figure 2 illustrates the implications of diversion for different θ^ν. 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 parameterizations 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 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 increasing θ^ν weakens diversion between these goods.

The third panel of figure 2 takes the difference of the first two panels. 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. An advantage

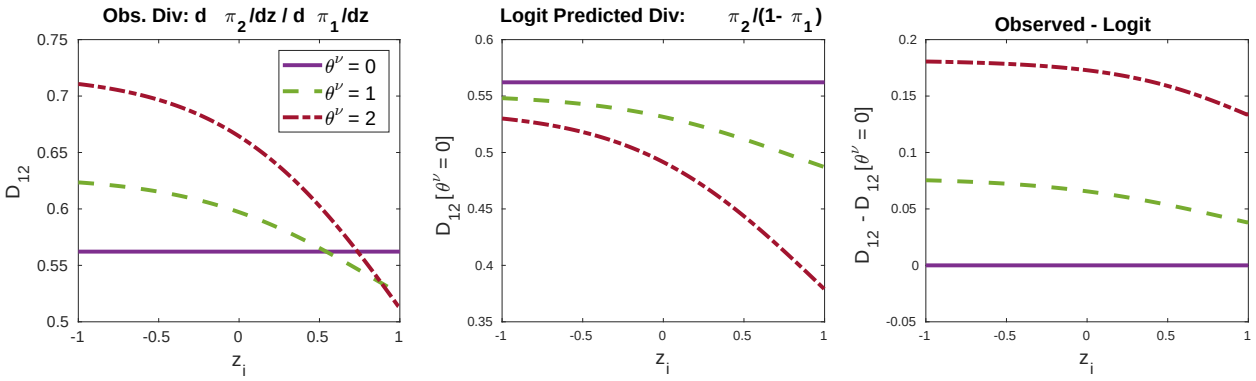

Figure 2: Diversion and Demographics

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 $\partial_z \tilde{\pi}^z_j = 0$ and the moments we have suggested are undefined and no longer informative.

In our example, we specified z to shift the utility of exactly one good and restricted θ^ν to have dimension one. There are more general conditions for identification of θ^ν from consumer demogrpahics. μ^z is typically specified as a linear combination of interactions between product characteristics and consumer demographics, e.g.,

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

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

$$\mathrm{d}_{z^d} \tilde{\pi}^z_j = \sum_{k=1}^{K} \sum_{\ell=1}^{J} \theta^{z(k,d)} x^k_\ell \partial_{u_\ell} \pi^z_j. \tag{15}$$

In matrix notation, (15) can be written as

$$\mathrm{d}_{z^\top} \tilde{\pi}^z = \partial_{u^\top} \pi^z \partial_{z^\top} u = \partial_{u^\top} \pi^z \partial_{z^\top} \mu^z = \partial_{u^\top} \pi^z X^\top \Theta^z. \tag{16}$$

Thus, only if $X^\top \Theta^z$ has maximum column rank, does there exist a unique $\partial_{u^\top} \pi^z$ that solves (16). 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 sufficient to identify 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 $J(J-1)/2$ correlations), one would have the same number of unknowns as there are restrictions in (16). 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 $X^\top \Theta^z$ fails, we still have restrictions like (16) that may 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 the CLER estimator, we now relate it to other estimators used in the discrete choice literature.

First, as noted above, with $S = N$, $\log \hat{L}$ 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 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 fig. 3 of Grieco et al. (2022).

The other major class of estimators used in applied work consists of share constrained GMM estimators (e.g., BLP04; Petrin 2002; Grieco et al. 2023).[19] The remainder of this section shows how the CLER estimator can be converted into members of this class of estimators. Since the CLER estimator is efficient, so we will point out losses of efficiency along the way. There may be a trade-off between efficiency and computational tractability that justifies using an inefficient estimator. We also discuss these trade-offs. One should keep in mind that computational resources tends to be less costly than data. We argue for the computational tractability of the CLER estimator in section 7.

Figure 3 provides a summary of the steps. The highest node in the tree represents the CLER 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

[19] 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 3.

tree, so in total the figure describes nine alternative estimators (including share constrained likelihood, see footnote 19). 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 (12) which is asymptotically equivalent to our estimator, assuming that (12) does not lose identification; as we pointed out in section 4.1. Going from minimizing the objective function (5) to setting its derivatives to zero can lose identification due to the existence of multiple (local) optima.

For equivalence to obtain, it is essential that the $\hat{\mathcal{W}}_L$ and $\hat{\mathcal{W}}$ matrices used in (12) 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 set of moments entails a sum over consumers whereas the other is a sum over products.

GMM estimators are often used to avoid parametric distributional assumptions, however 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 (5) similarly avoids distributional assumptions on ξ.

Our estimator has an important computational advantage over (12): its objective function is approximately 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 (12) one can impose share constraints $\pi = s$.[20] Following the intuition of Berry (1994) this is equivalent to treating δ as a deterministic function of θ and yields a consistent estimator as $N_m, S, J \to \infty$.

Three issues arise when imposing the share constraint. First, because it is a one to one mapping on the interior of the probability simplex, doing so rules out the presence of zero shares. While this is reasonable for conditional choice probabilities, applied cases have arisen where zero shares are observed in data. By optimizing the CLER objective rather than enforcing

[20] Share constraints can also be imposed on $\log \hat{L}$ directly, see footnote 19.

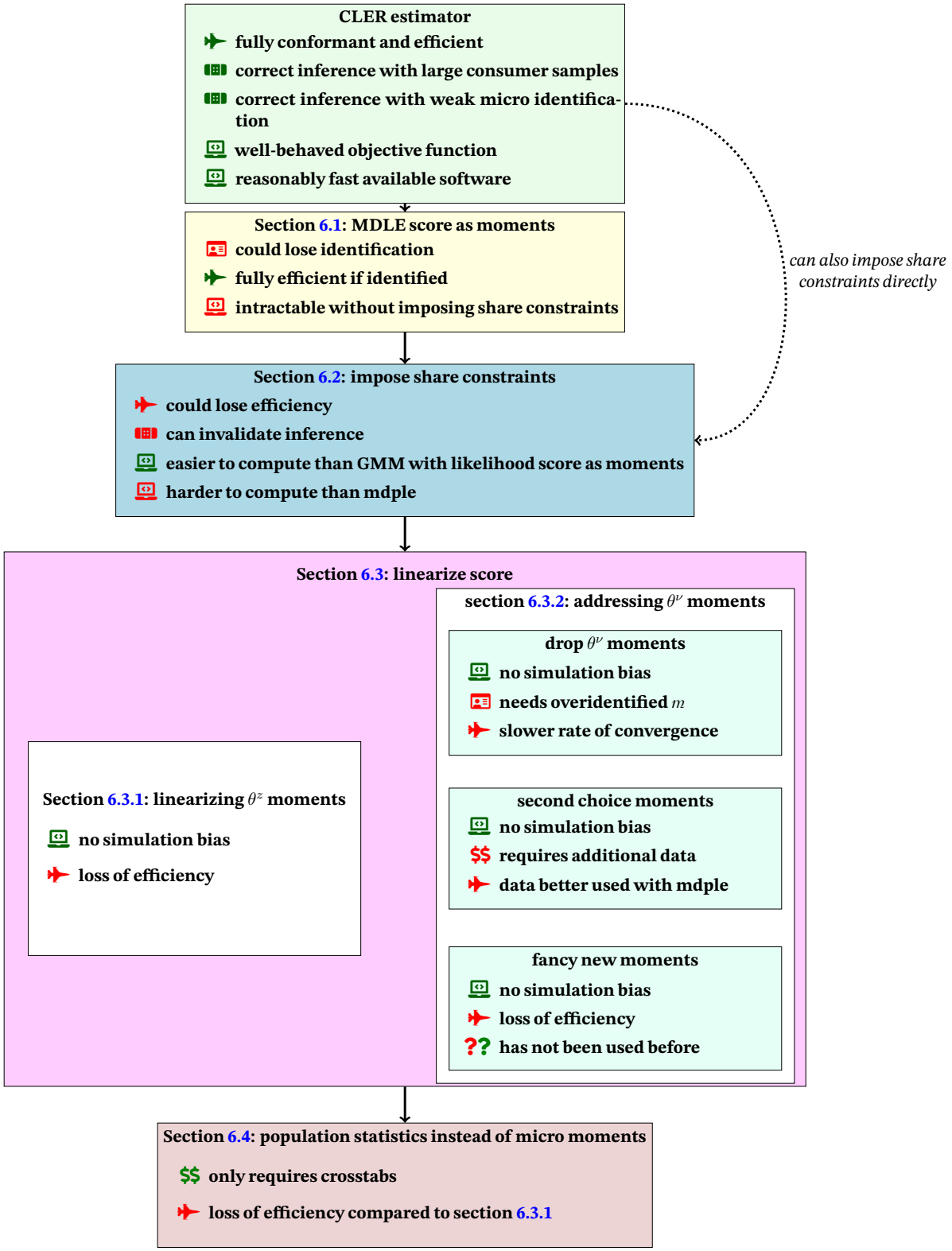

Figure 3: Schematic comparison of our estimator to alternatives. See text for details.

that unconditional choice probabilities equal market shares, the CLER estimator offers some robustness to zero or small shares in the data.

Second, and more importantly, imposing the share constraints introduce a potential loss of efficiency. Suppose that $\theta^{z} \neq 0$ such that the MDLE and the CLER 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. Lemma 1 establishes this result for the single market case.

Third, and most importantly, using the assumption $s = \pi$ in inference, also, can produce incorrect inference unless the total number of consumers in all markets is negligible compared to the square root of the population in the smallest market.

We start by demonstrating the potential efficiency loss.

Lemma 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 the MDLE (or CLER) estimator of δ, θ. □

The proof of this lemma follows immediately from the proofs of lemmas 2 and 3 in appendix C, which formally derive the asymptotic variance of the MDLE (or CLER) 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. In practice, imposing the share constraint can lead to a substantial efficiency loss as examples 1 and 2 in Grieco et al. (2022) illustrate.

For additional intution, consider 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 (7) and considers the derivative of the macro term with respect to δ, i.e. for all $m = 1, \dots, M$ and all $j = 1, \dots, J_m$,

$$\sum_{\ell=0}^{J_m} \frac{s_{\ell m}}{\pi_{\ell m}} \int \mathcal{s}_{\ell m}(z,\nu)\Big(\mathbb{1}(\ell = j) - \mathcal{s}_{jm}(z,\nu)\Big)\, \mathrm{d}F(\nu)\, \mathrm{d}G(z) = 0, \tag{17}$$

where $\mathcal{s}$ was defined in (2). If $s = \pi$, then the left hand side in (17) becomes

$$\int \mathcal{s}_{jm}(z,\nu)\, \mathrm{d}F(\nu)\, \mathrm{d}G(z) - \int \mathcal{s}_{jm}(z,\nu) \underbrace{\sum_{\ell=0}^{J_m} \mathcal{s}_{\ell m}(z,\nu)}_{=1}\, \mathrm{d}F(\nu)\, \mathrm{d}G(z). \tag{18}$$

So setting $s = \pi$ solves (17). By Berry (1994), 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 complicates inference. If one treats δ as a deterministic function of θ, one ignores the uncertainty arising from observed market shares. This will result in a downward bias in 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 $\min_m \sqrt{N_m}$.

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

$$\frac{\sqrt{S} v^\top (\hat{\delta}_{\cdot m} - \delta_{\cdot m})}{\sqrt{v^\top \partial_{\theta^\top} \hat{\delta}_{\cdot m}(\hat{\theta}) \hat{\mathcal{V}}_\theta \partial_\theta \hat{\delta}_{\cdot m}^\top(\hat{\theta}) v}} \xrightarrow{d} N(0,1), \tag{19}$$

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 $v^\top \partial_{\theta^\top} \delta_{\cdot m} = 0$,[21] where the left hand side of (19) 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.[22] We provide the correct asymptotic variance formulas for the single market case in appendix C.2. Grieco et al. (2022) provides a numerical example that shows that imposing the share constraint without adjusting the standard errors can lead to standard errors being off by an arbitrarily large factor.

To summarize, inference using the CLER estimator 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 C.2 which are based on the moments in (34), not on the more convenient formulas that obtain if N is set to ∞. This issue

[21] Indeed, then by a Taylor expansion,

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

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

[22] One would have to draw the bootstrap population from the superpopulation, which is impossible.

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 have mean zero at the truth, regardless of the number of simulation draws. Consequently, as Pakes and Pollard (1989, PP89) 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 (7) 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 (7) and (8) because the micro score has mean zero in both cases. So we will work with the micro score in (8).

6.3.1 Approximation of θ^z moments for linear simulation error. We first consider the micro score of (8) with respect to $\theta^{z(k,d)}$, i.e.

$$\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{j=0}^{J_m}\frac{D_{im}y_{ijm}}{\pi_{jm}^{z_{im}}}\int s_{jm}(z_{im},\nu)\left(x_{jm}^k z_{im}^d-\sum_{\ell=1}^{J_m}x_{\ell m}^k z_{im}^d s_{\ell m}(z_{im},\nu)\right)\mathrm{d}F(\nu), \tag{20}$$

which is a ratio of two integrals due to the presence of $\pi_{jm}^{z_i}$ in the denominator. A commonly used approximation to the score can be found by setting $\nu=0$ selectively as follows. Continuing from (20), we have

$$\begin{aligned}
&=\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{j=0}^{J_m}D_{im}y_{ijm}\frac{\int s_{jm}(z_{im},\nu)\left(x_{jm}^k z_{im}^d-\sum_{\ell=1}^{J_m}x_{\ell m}^k z_{im}^d s_{\ell m}(z_{im},\nu)\right)\mathrm{d}F(\nu)}{\int s_{jm}(z_{im},\nu)\,\mathrm{d}F(\nu)}\\
&\approx\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{j=0}^{J_m}D_{im}y_{ijm}\frac{\int s_{jm}(z_{im},0)\left(x_{jm}^k z_{im}^d-\sum_{\ell=1}^{J_m}x_{\ell m}^k z_{im}^d s_{\ell m}(z_{im},\nu)\right)\mathrm{d}F(\nu)}{\int s_{jm}(z_{im},0)\,\mathrm{d}F(\nu)}
\end{aligned}$$

$$= \sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} D_{im}(y_{ijm} - \pi_{jm}^{z_{im}}) x_{jm}^k z_{im}^d, \tag{21}$$

The final line of (21) matches the correlation of demographics and product characteristics in the micro sample to that of the model. This moment is commonly used in applied work, see CG23 for a list of examples.[23] A convenient feature of this moment is that it 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 simulation draws required in the computation of $\delta(\theta)$, which is an argument to s_{jm}, must diverge faster than S to avoid affecting efficiency and necessitating a different inference procedure,[24] 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 micro score of (8) with respect to $\theta^{\nu(k)}$ is similar to (20), replacing z_{im}^d with ν^k in the integrand, i.e.

$$\sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} D_{im} \frac{y_{ijm}}{\pi_{jm}^{z_{im}}} \int s_{jm}(z_{im}, \nu) \left(x_{jm}^k \nu^k - \sum_{\ell=1}^{J_m} x_{\ell m}^k \nu^k s_{\ell m}(z_{im}, \nu) \right) \mathrm{d}F(\nu), \tag{22}$$

However, 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 work 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 et al. (2004*a*) and Grieco et al. (2023) is introducing second choice data based on surveys of consumer purchases to construct alternative moments. The CLER estimator could accommodate second choice data efficiently by including it directly in the likelihood. There are, however, 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 more importantly, such data is often prohibitively costly to obtain.

[23] Discretizing either z_{im} or x_{jm} will lead to two other popular classes of moments discussed by CG23 namely $\mathbb{E}[z_{im}|j \in \mathbb{J}(x_{jm})]$ or $\mathbb{E}[x_{jm}|i \in \mathbb{I}(z_{im})]$ for some sets of products or consumers defined by their characteristics or demographics. The discretization may impose a further loss of information. Note that applied work often conditions these moments on making an inside purchase; alternatively, one could define $x_{0m} = 0$ and use an unconditional moment.

[24] 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 simulated and actual values of $\delta(\theta)$.

While we are unaware of its use in the literature, there is a third possibility that utilizes two independent ν draws per simulation r, as we now explain. First, note that[25] $\sum_{j=0}^{J_m} \vartheta_{jm}(x_{jm}^k \nu^k - \sum_{\ell=0}^{J_m} \vartheta_{\ell m} x_{\ell m}^k \nu^k) = 0$, such that (22) can be expressed as

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

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 sum over the product of two integrals, namely

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

Thus, approximating the integrals with sums using independent Monte Carlo draws satisfies the conditions of PP89. While utilizing this moment will result in an estimator with the same convergence rates as our estimator, and so will satisfy conformance, it will not be efficient.

6.4 Step 4: Population statistics instead of micro data

One may further alter the moment described in section 6.3.1 by integrating (21) over z,

$$\sum_{m=1}^{M} \sum_{j=0}^{J_m} \left(\frac{1}{S_m} \sum_{i=1}^{N_m} D_{im} y_{ijm} x_{jm}^k z_{im}^d - \int \pi_{jm}^z x_{jm}^k z^d \, \mathrm{d}G(z) \right). \tag{23}$$

This is the moment described in BLP04, eq. 8, and Gandhi and Nevo (2021, eq. 4.4).

There are two possible motivations using (23) over (21). 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 (23). The second is that the right hand side of (23) does not involve a sum over observed consumers. However, in view of PP89, the total number of simulation draws needed is the same in both cases. To simulate (21), 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 (23) one needs a number of

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

independent draws that is at least proportional to S.

However, using (23) over (21) has an additional efficiency cost. In particular, (23) 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 (23) weakly greater than (21). For 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 (21) and (23) (if evaluated at the truth) have the same Jacobian in expectation. The variance contribution for observation i using (23) equals

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

which is the variance contribution of observation i in (21). These two facts combined with the sandwich formula for the asymptotic variance of the GMM estimator imply that using (21) dominates (23).

7 Computation

While the CLER estimator is of theoretical interest, it must also be computationally tractable in order to be appropriate for applied use. This section discusses two critical computational aspects of our estimator. First, the CLER 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.[26] 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 the CLER estimator for which we use Newton's method with Trust Regions. Recall from (5) that our optimization problem is

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

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

Like BLP95, we start by concentrating out β which leaves

$$(\hat{\theta}, \hat{\delta}) = \arg\min_{\theta,\delta} \Big(-\log \hat{L}(\theta, \delta) + \hat{\Pi}\{\hat{\beta}(\delta), \delta\} \Big). \tag{24}$$

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 BLP95 with the important exception that the inner loop objective is (5)—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 BLP95, tractability followed from the existence of a contraction mapping to compute $\pi = s$. For our problem, first suppose that (5) 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 separable across markets in $\delta_{\cdot m}$ and is nearly 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 highly tractable.

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 (9) to be $(B^\top B)^{-1}$ where B is a $J \times d_b$ matrix with rows $b_{jm}^\top$, the instruments introduced in (10). Unfortunately, $\hat{\Pi}$ is not additively separable in $\delta_{\cdot m}$. However, there are several convenient features which make the inner loop tractable.

The first such feature is that $\hat{\beta}(\delta)$ is simply a linear IV estimator, i.e. $\hat{\beta}(\delta) = (X^\top \mathcal{P}_B X)^{-1} X^\top \mathcal{P}_B \delta$, with $\mathcal{P}_B = B(B^\top B)^{-1} B^\top$ an orthogonal projection matrix. Second, $\hat{\Pi}$ is quadratic in δ. Thus, writing $\mathcal{P}_{\mathcal{P}_B X} = \mathcal{P}_B X (X^\top \mathcal{P}_B X)^{-1} X^\top \mathcal{P}_B$, (24) becomes

$$-\log \hat{L}(\theta, \delta) + \frac{1}{2} \delta^\top (\mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X}) \delta \tag{25}$$

Third, (25) combines the computationally convenient likelihood with a convex term, so the resulting objective can be optimized over δ via Newton's method. 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 $\mathcal{K}\mathcal{K}^\top$ 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 4 in appendix D.

We now turn to the primary complication of applying Newton's method to optimize (25) 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 (25), $H + \mathcal{K}\mathcal{K}^\top$, 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 (25) can be written as

$$H^{-1} - H^{-1}\mathcal{K}(I + \mathcal{K}^\top H^{-1}\mathcal{K})^{-1}\mathcal{K}^\top H^{-1}, \tag{26}$$

where I is the identity matrix.[27]

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 $H + \mathcal{K}\mathcal{K}^\top$ to inverting a thousand 100 by 100 matrices, which is both much less demanding computationally and reduces memory demand by a factor 1,000 (i.e., $100,000^2/(100^2 \times 1,000)$). This makes the 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.

An even more computationally convenient alternative. Although computation of the CLER estimator is straightforward, there is an alternative that has the same conformance and efficiency properties as CLER and can be computed even more easily.[28] This estimator optimizes the sum of the macro and micro loglikelihoods over δ in the inner optimization (dropping the product level moments), but then optimizes the CLER objective function over θ in the outer optimization. Doing so avoids the need to compute (26) which permits the inner loop to be entirely

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

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

[28] One reason to avoid implementing CLER directly is that there is usually no convenient way to pass the information on the structure of the Hessian to packaged optimization routines. The `Grumps` package does provide this functionality.

parallelized by market. As we establish as an intermediate result of theorem 1, this estimator is asymptotically equivalent to the CLER estimator, but potentially at a lower computational and programming cost. Because of the presence of the product level moments term, the CLER estimator has some robustness against low shares which the alternative does not inherit, but that is not the focus of this paper. We have implemented both estimators as part of the `Grumps` package using the name `cheap` to denote this alternative.

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 to avoid affecting the asymptotic behavior, the numerical error must 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.

The CLER 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. 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.[29]

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 examine the sensitivity of CLER's numerical performance to the number of nodes used for numerical integration in section 9.

[29] If ν is of high dimension, sparse quadrature methods can be viable alternatives. The designed quadrature approach of Bansal et al. (2021) may be particularly attractive as all nodes have positive weights.

8 Inference

This section describes inference on functions of model parameters, including elasticities and counterfactuals. As discussed above, the conformant property of the CLER estimator ensures that it can be applied under a wide variety of conditions. This also applies to our inference procedure. We first outline the intuition behind our approach to inference in section 8.1. We then move to a formal statement of our assumptions and proof in section 8.2.

8.1 Intuition

In all cases, inference will be built upon the Hessian of the CLER objective function (5),

$$\begin{array}{c} \\ \beta \\ \theta^z \\ \theta^\nu \\ \delta \end{array} \begin{bmatrix} \partial_\beta \hat{m}^\top \hat{\mathcal{W}} \partial_{\beta^\top} \hat{m} & 0 & 0 & \partial_\beta \hat{m}^\top \hat{\mathcal{W}} \partial_{\delta^\top} \hat{m} \\ 0 & -\partial_{\theta^z \theta^{z\top}} \log \hat{L} & -\partial_{\theta^z \theta^{\nu\top}} \log \hat{L} & -\partial_{\theta^z \delta^\top} \log \hat{L} \\ 0 & -\partial_{\theta^\nu \theta^{z\top}} \log \hat{L} & -\partial_{\theta^\nu \theta^{\nu\top}} \log \hat{L} & -\partial_{\theta^\nu \delta^\top} \log \hat{L} \\ \partial_\delta \hat{m}^\top \hat{\mathcal{W}} \partial_{\beta^\top} \hat{m} & -\partial_{\delta \theta^{z\top}} \log \hat{L} & -\partial_{\delta \theta^{\nu\top}} \log \hat{L} & \partial_\delta \hat{m}^\top \hat{\mathcal{W}} \partial_{\delta^\top} \hat{m} - \partial_{\delta \delta^\top} \log \hat{L} \end{bmatrix}. \quad (27)$$

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.

First consider the leading case where $S/J \to \infty$ and $\theta^z \neq 0$. In this case, the CLER estimator is asymptotically equivalent to the MDLE two-step estimator that first estimates (θ, δ) and then plugs in $\hat{\delta}$ to estimate β. With the MDLE, the information matrix for $\psi = [\theta^\top, \ \delta^\top]^\top$ is the Hessian of $-\log \hat{L}$. Notice that this is the (ψ, ψ) block of (27) with the exception of the $\partial_\delta \hat{m}^\top \hat{\mathcal{W}} \partial_{\delta^\top} \hat{m}$ term in the (δ, δ) block. However, that term diverges at rate J and is dominated by $-\partial_{\delta\delta^\top} \log \hat{L}$. Similarly, because $\hat{\psi}$ converges faster than $\hat{\beta}$, the (β, β) block in (27) 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 (27) is

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

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 lemma 5 in appendix E, the score of the objective with respect to $\psi^\nu = (\theta^\nu, \delta)$ becomes collinear, leading to a loss of rank in the Hessian of $\log \hat{L}$. However, rank is preserved in (27) 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

$$-\left(\partial_{\theta^z\theta^{z\top}} \log \hat{L} - \partial_{\theta^z\delta^\top} \log \hat{L} \left(\partial_{\delta\delta^\top} \log \hat{L}\right)^{-1} \partial_{\delta\theta^{z\top}} \log \hat{L}\right)^{-1}, \tag{28}$$

as we show in lemma 7 in appendix E. Expression (28) 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., BLP95). As we discussed in section 4.2.1, θ, δ are not identified off the likelihood alone since $\log \hat{L}^{\text{mic}} = 0$ and $\log \hat{L}^{\text{mac}}$ is maximized for any θ by choosing δ such that $\pi = s$ as we have shown in section section 6.2.[30] Consequently, $\partial_{\psi\psi^\top} \log \hat{L}$ is then singular, indeed of rank d_δ.[31] However, analogous to the $\theta^z = 0$ case, the (ψ, ψ) block in (27) 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 identification 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 BLP95 estimator which is $O_p(J^{-1})$; see lemma 8 in appendix E. Returning to the case where $S/J \to 0$ but some micro data exists, $\hat{\Pi}$ will 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 $\hat{\Pi}$ varies. However, by using (27) for inference, we include all relevant terms so that inference is valid across all these scenarios.

[30] Since $\log \hat{L}^{\text{mac}}$ integrates over z, we need not distinguish between θ^z and θ^ν in this case.

[31] 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.

A second complication is that δ grows with J, so (27) is also growing. To address this, write $\gamma = [\beta^\top, \theta^\top, \delta^\top]^\top$. Recall from section 4.2 that we assume that $\lim_{M\to\infty} \max_m J_m < \infty$. The following subsection provides an inference method for finite-dimensional linear combinations of γ.

8.2 Formal Result

We conduct inference on $\Lambda(\hat{\gamma} - \gamma_0)$, where Λ is specified in theorem 1.[32] The purpose of this theorem is to demonstrate how the CLER estimator balances different sources of identification and achieves conformance. The assumptions stated below are stronger than necessary and the key result obtains under weaker conditions with a longer proof. We will discuss some of the differences between what is covered by the assumptions and situations under which our estimator works as they arise.

Assumption A (Sampling of markets)**.** The markets in the sample, indexed by $m = 1, \dots, M$, are i.i.d.. Market m has J_m products, N_m consumers altogether, and S_m consumers in the micro sample. The N_m consumers are i.i.d. draws from a superpopulation for market m and the set of S_m micro consumers are probability $\chi_m \in [0, 1]$ i.i.d. (without replacement) draws from the N_m consumers comprising the population of market m. □

Since market selection is random, so are $\{N_m, \chi_m, J_m\}$. In addition, the distribution of N_m varies with M and the distribution of χ_m can vary with M. We define $N = \sum_{m=1}^M N_m$, $S = \sum_{m=1}^M S_m$, and $J = d_\delta = \sum_{m=1}^M J_m$. Asymptotics are in the number of markets (or equivalently, products), population sizes, and possibly micro consumer sample sizes (via N_m, χ_m), as discussed in assumption D. Note that for ease of notation thus far we have referred to limits in S and J, which under assumption D should be interpreted as limits in $M\mathbb{E}(N_m\chi_m)$ and M respectively.

Assumption B (Utility linear in parameters)**.** μ^z_{jm} and μ^ν_{jm} are for all m linear in θ^z, θ^ν respectively. □

For convenience, we assume that the heterogeneity terms of utility, in addition to mean utility, are linear in parameters. This could easily be relaxed.

Assumption C (Distribution of product characteristics)**.** (a) Observed product characteristics $x_{\cdot m}$ have bounded support; (b) for some $p_\xi > 8$, $\mathbb{E}\exp(p_\xi|\xi_{jm}|) < \infty$. □

[32] In the formal results, a zero subscript denotes the truth.

Assumption C ensures that the smallest product-level choice probability, $\min_{j,m} \pi_{jm}$, does not go to zero too fast as M grows. It is implicit in much of the literature, though it could be relaxed. Condition (b) is a restriction on the tails of the distribution of ξ. It is trivially satisfied if ξ is assumed to have bounded support or is normally distributed.

Assumption D (Rates). (a) $M \to \infty$; (b) the distribution of J_m does not depend on M and for some $\bar{J} < \infty$, $\mathbb{P}(1 \leq J_m \leq \bar{J}) = 1$; (c) $\mathbb{E}(N_m \chi_m) \succ 1$; (d) $M^2 \bar{\chi} / \mathbb{E}(N_m \chi_m) \preceq 1$ and $\mathbb{E}(1/N_m) \prec 1$, where $\bar{\chi} = \mathbb{E}\chi_m$, $a \preceq b$ means that $a/b = O_p(1)$ or $O(1)$ and where $\succeq, \prec, \succ$ are analogously defined. □

Condition (a) in assumption D deviates from BLiP04, where the number of markets is fixed and the number of products increases, but is similar to what is assumed in Hong et al. (2021); neither of these papers covers the case with consumer micro data. It is needed for consistent estimation of β_0 and, more generally, in the absence of or with poor micro consumer data. Together with (b), (a) requires that the number of markets grows but the number of products per market does not. These two conditions guarantee that J and M grow at the same rate.

Condition (c) requires the number of micro consumers S to grow with M. We make this assumption for convenience; if this were not true then the only source of identification would be the product level moments, which is covered by Hong et al. (2021). Finally, the first half of (d) can be most easily understood if one considered the case in which N_m, χ_m were independent, in which case it simplifies to $M^2 \preceq \mathbb{E}N_m$, i.e. the average market population size grows no slower than the square of the number of markets. The second half of (d) says that all markets must grow in population.

Assumption E (Parameter space). The true value θ_0 is an interior point of the compact parameter space Θ. Further, δ_{0m} is bounded away from the boundary of the parameter space $\Delta_m = \{\delta_m : \exists \theta \in \Theta : \delta_{0m}^{\text{mac}}(\theta) = \delta_m\}$, where $\delta_{0m}^{\text{mac}}(\theta)$ (formally defined in appendix F) is, for a given θ, the maximizer of the macro term of the population likelihood in market m, $\log L_m^{\text{mac}} = \mathbb{E} \log \hat{L}_m^{\text{mac}}$. □

The definition of Δ_m is explicitly specified because it depends on x and ξ, which are random at the product level. The function $\delta_{0m}^{\text{mac}}(\theta)$ is the Berry (1994) inversion, except that we invert product choice probabilities rather than observed market shares, a distinction that is assumed to be irrelevant in much of the literature. This assumption rules out parameter on the boundary and associated asymptotic size issues analyzed by Ketz (2019).

Let B_m be the submatrix of the instrument matrix B corresponding to market m, let $\mathcal{B}$ denote the sigma field generated by B, and let $\mathcal{J}$ be the sigma field generated by $B, X, \xi, \{N_m\}, \{\chi_m\}$.

Assumption F (Product Level Restrictions). (a) The elements of ξ are independent conditional on $\mathcal{B}$; (b) $\mathbb{E}(\xi \mid \mathcal{B}) = 0$; (c) $\mathbb{E}(\xi\xi^\top \mid \mathcal{B}) = I$; (d) $0 < \mathbb{E}(B_m B_m^\top / J_m) < \infty$; (e) $\mathbb{E}(B_m X_m^\top / J_m)$ has rank $d_x \leq d_b$; (f) $\min_{\|\theta-\theta_0\|\geq\epsilon} \Pi\{\delta_0^{\text{mac}}(\theta)\} \succeq M\epsilon^2$; and $\lambda_{\min}(\partial_\theta \delta_0^{\text{mac}\top}(\theta_0)\mathcal{K}\mathcal{K}^\top \partial_{\theta^\top} \delta_0^{\text{mac}\top}(\theta_0)) \succeq M$. □

Assumption F contains a number of conditions that are implicitly made in the literature. First, (a) and (b) are standard, though (a) can be relaxed at the expense of longer proofs. The extension is not theoretically interesting, so we do not pursue it here. Condition (c) may look strong, but heteroskedasticity and some dependence can be accommodated by redefining the objective function and the same goes for scaling. The only caveat there is that the optimal weight matrix can depend on unknown coefficients which would have to be estimated in a two step procedure; the same goes for the scaling parameter. That adjustment is routine and not considered here. Condition (d) is standard and says that there is no collinearity in the instruments. In the proofs, we will take $B^\top B$ to be invertible, which is not implied by (d), but is true with probability approaching one.[33] Condition (e) is a standard rank condition.

Condition (f) assumes strong identification off the product level moments. This allows us to highlight the role of the micro likelihood for identification without undue notation. Condition (f) can fail for three reasons. First, if the product level moments just identify β (e.g., $d_b = d_x$), the Π term of our estimator can be set to 0 for any (θ, δ) and so does not contribute to the estimation of (θ_0, δ_0). Consequently, the asymptotics of our estimator are then covered under standard maximum likelihood theory, albeit that the dimension of δ increases. The case of weak product level instruments is similar.[34] Finally, it is possible that the number of strong moments overidentifies β but is insufficient to also identify θ (e.g., $d_x < d_b < d_x + d_\theta$). In this case, the rates of convergence would depend on the relative divergence rates of $S\lambda^2$ and M and also on the linear combination of the parameters that is being estimated. The statement of theorem 1 is valid for all three cases, although covering each case would necessitate a longer proof.

Recall that λ is the weak micro identification parameter, i.e. the rate at which θ_0^z converges to

[33] This comment addresses the immaterial technical issue that in the presence of discrete variables $B^\top B$ is singular with positive probability, as is well-known.

[34] We are using the standard definition of weak instruments here, in which the first stage coefficients decrease to zero at rate $1/\sqrt{M}$. Then, it can be shown that our procedure reduces to maximum likelihood if $\lambda \succ \sqrt[4]{M}/\sqrt{N\chi}$ and otherwise is inconsistent.

zero; $\lambda = 1$ in the case of strong micro identification.

Assumption G (Micro identification). Let $\Delta = \prod_{m=1}^{M} \Delta_m$ and $\|\theta - \theta_0\|_\lambda^2 = \|\theta^z - \theta_0^z\|^2 + \lambda^2 \|\theta^\nu - \theta_0^\nu\|^2$. Then,

$$\inf_{\theta \in \Theta : \|\theta - \theta_0\|_\lambda > 0} \min_{\delta \in \Delta} \frac{\mathcal{L}^{\text{mic}}(\theta, \delta)}{N\chi \|\theta - \theta_0\|_\lambda^2} \succeq 1,$$

where $\mathcal{L}^{\text{mic}}(\theta, \delta) = \log L^{\text{mic}}(\theta_0, \delta_0) - \log L^{\text{mic}}(\theta, \delta)$ is minus the centered micro term of the population likelihood and is formally defined in appendix F. □

Assumption G assumes that variation in the micro data is sufficient to identify θ_0^z. It allows for no, weak, or strong identification of θ_0^ν based on the value of λ, which is fixed in the case of strong identification, drifting to zero for weak micro identification, and zero for lack of micro identification. It can be justified by a second order Taylor expansion of each $\mathcal{L}_m^{\text{mic}}$ around (θ_0, δ_{0m}).

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.

Theorem 1. *Let $\{\Lambda\}$ be a sequence of nonrandom $d_\Lambda \times (d_\beta + d_\theta + J)$ matrices for which $\Lambda\Lambda^\top$ converges to a positive definite $d_\Lambda \times d_\Lambda$ matrix. Under assumptions A to G,*

$$(\Lambda \hat{V} \Lambda^\top)^{-1/2} \Lambda (\hat{\gamma} - \gamma_0) \xrightarrow{d} N(0, I), \tag{29}$$

where $\hat{\gamma}$ is the CLER estimator and $\hat{V}$ is the inverse of (27). □

The sequence of matrices $\{\Lambda\}$ is specified such that inference is conducted on a finite dimensional vector of linear combinations of γ_0. This assumption will cover traditional counterfactual analysis (e.g., merger simulation of observed markets). Appendix F contains the proof with supporting lemmas and an informal outline of the intuition.

While the elements of $\Lambda(\hat{\gamma} - \gamma_0)$ converge at different rates and these rates themselves will depend on the identifying variation, $(\Lambda \hat{V} \Lambda^\top)^{-1/2}$ scales this vector such that the product always converges to a standard normal and can be used to conduct inference without explicit knowledge of the rates.

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 presents Monte Carlo simulations across a number of different settings to investigate the performance of our estimator relative to alternatives. We will vary (a) the amount of micro data available, (b) the degree of heterogeneity in utility due to demographics and unobserved tastes, (c) the strength of the product level instruments, (d) the accuracy of the numerical approximation of the log likelihood in our preferred estimator. Varying these settings affects the relative power of the micro observations and product level exclusion restrictions for estimation of the random coefficients θ^ν, which affects the precision of all parameters of the model.[35] Throughout, we will compare the CLER estimator, which efficiently utilizes both these sources of identification, with estimators that emphasize only one.

9.1 Design and Estimators

Our baseline empirical design includes two observable and exogenous product characteristics (x^1_{jm}, x^2_{jm}), with associated parameters (β_1, β_2); two demographic characteristics (z^1_{im}, z^2_{im}) interacted with a single corresponding product characteristic with associated parameters (θ^z_1, θ^z_2); and two random coefficients $(\theta^\nu_1, \theta^\nu_2)$.

For each specification, we draw data for 50 markets with varying and independent numbers of products, with the median market having 20 products. There are $100,000$ consumers (N_m) in each market and we vary the size of the micro dataset, with $S_m = 1,000$ the baseline. In the baseline specification, average share is roughly 2.1%, and the tenth percentile of shares is roughly 0.06%. Full details of the monte carlo design are presented in appendix G.

We compare three different estimators. First, we consider the CLER estimator, (5). Along with product characteristics, we include differentiation instruments in Π following GH20, so the Π is overidentified for β and are potentially useful to identify θ. Second, we consider the GMM estimator with the share constraint described in section section 6.3. This is a common approach used in the applied literature and we implement it using the `pyblp` package, version 0.13 (Conlon and Gortmaker, 2020, 2023). Lastly, we implement the MDLE two-step estimator

[35] We also ran an experiment where we varied the amount of variation in consumer demographics across markets. Both CLER and GMM perform better with more cross-market variation, although CLER always outperforms GMM. Results of this experiment are available from the authors.

that first estimates (θ, δ) by minimizing $\log \hat{L}$, and then estimates β by minimizing $\Pi(\beta, \hat{\delta})$. In this two-step procedure, product level moment restrictions are not used in the estimation of θ; the same set of moments as above are used to recover β.

All three estimators must integrate over both ν and z to compute π; we implement this integration using Monte Carlo simulation with $10,000$ consumer draws. The two likelihood estimators must also compute $\pi^{z_{im}}$ for each observation in the consumer sample. We use 11-point Gaussian quadrature in both dimensions of ν, but evaluate this choice in Section 9.2.4.

For all experiments, we estimate the model for each of 500 draws of the data generating process and present the plots of estimated parameter values across these draws. For CLER and MDLE we use a single, arbitrary, starting point. For GMM, which is known to have local optima, we multi-start from three values, including the truth.

To summarize, while the GMM estimator utilizes product level moments for the identification of θ^ν, it fails to incorporate all the information in the likelihood of the consumer sample. The two-step estimator does the opposite: fully utilizing micro data for the estimation of θ^ν while not leveraging the information in the product level moments. The CLER estimator fully exploits all available information from the data.

9.2 Results

9.2.1 Varying the size of the consumer sample. The first experiment varies the size of the consumer sample for the baseline data generating process. Increasing S_m should improve the precision of all estimators. However, for the GMM estimator the benefit comes only from greater precision in the estimation of the demographic micro-moment, whereas the MDLE two-step and CLER estimators fully exploit the consumer data via the micro-likelihood. Figure 4 presents results for this experiment. Each plot compares the distribution of the three estimators for a specific consumer sample size (rows) and a given parameter (columns). The CLER estimator is a solid blue line, the GMM procedure is a dashed green line, the MDLE two-step procedure is a dotted black line.

Visually, it is clear that our method dominates both the GMM procedure and the MDLE two-step when $S_m = 250$ for θ^z. At this small micro-sample size, CLER and the GMM procedure perform similarly for θ^ν and β, outperforming MDLE, which does not utilize product level moments when estimating θ instead relying exclusively on the small micro-sample. As S_m increases, there is significant improvement in the precision of both $\hat{\theta}^z$ and $\hat{\theta}^\nu$ for CLER and

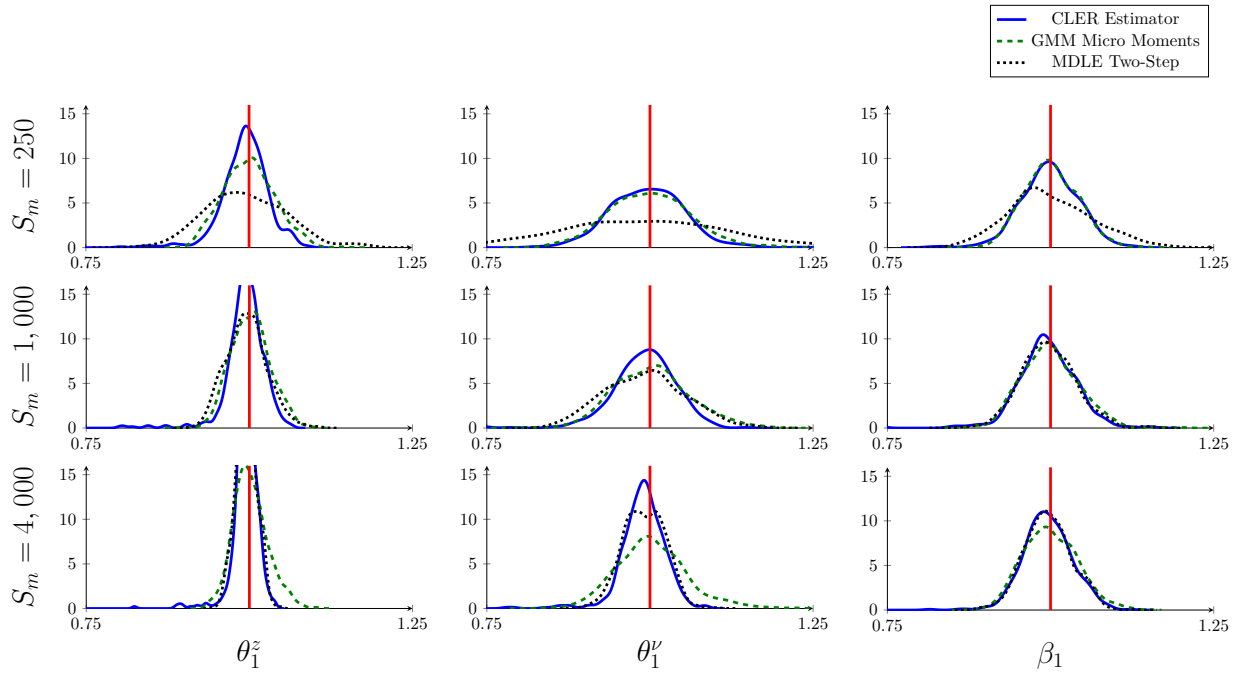

Figure 4: Distribution of parameters for different sizes of consumer sample.

the MDLE, both of which utilize the score of the likelihood with respect to θ^ν. In contrast, the GMM estimator has a smaller improvement as S_m increases, even though the micro-moment matching the covariance between demographics and purchased product characteristics is more precisely measured. At $S_m = 4,000$ the MDLE and the CLER estimator almost coincide. They outperform the GMM estimator, particularly for θ^ν. CLER and MDLE perform similarly when S_m is high because the information on θ from the likelihood dominates that of the product level moments for micro-samples of this size (in our baseline parameterization).

9.2.2 Varying consumer heterogeneity. We next consider the estimators' performance for different parameterizations of θ while fixing $S_m = 1000$. The goal of this exercise is to illustrate the estimators' performance as we change the relative power of the two sources of identifying variation for θ^ν.

As discussed in section 5, the identifying power of the consumer sample for θ^ν becomes weak as $\theta^z \to 0$. Intuitively, if changes in observable demographics do not substantially vary utility across products, then comparisons between consumers are not useful in measuring substitution. On the other hand, the product level moments will have identifying power only when the overidentifying instruments are strong in the sense of GH20.

We focus on the distribution of the estimates of the random coefficients, $\hat{\theta}^\nu$. Figure 5 plots the

distribution of $\hat{\theta}_1^\nu$ across three estimators as we vary θ^ν (rows) and θ^z (columns).[36] For reference, the central plot in this image is the baseline data generating process, which is also the DGP in the central row of Figure 4 where $S_m = 1000$. First, note the poor performance of the MDLE two-step for small values of θ^z (first column), where the identifying variation in the consumer sample is relatively weak. In contrast, the GMM and CLER estimators perform similarly when θ^z is small. Both rely on the variation from the product level moments that compose $\hat{\Pi}$ to estimate θ^ν. While CLER also incorporates information from the micro sample, this is negligible when θ^z is small.

There are also cases where GMM performs poorly but the MDLE two-step is comparable to the CLER estimator. In particular, this occurs when θ^z is large relative to θ^ν. For some intuition, note that the differentiation IVs on which the GMM estimator relies are a function of distance in characteristic space to other products and do not directly incorporate consumer demographics. Roughly, these moments target both θ^ν and θ^z, and rely on the demographic micro-moment to distinguish the two. This will be more difficult when θ^ν is large, which effectively adds noise to the micro-moment. On the other hand, the CLER estimator and the MDLE two-step efficiently use all information at the consumer level.

Over all cases, the CLER estimator performs well. When only one source of identification is useful, it roughly matches the performance of the estimator that exploits that source. When both sources are useful, it efficiently weights the two. This exercise provides a finite sample illustration of how conformance affects estimator precision.

9.2.3 Endogenous Characteristics. So far, we have assumed x_{jt} to be exogenous, which is unlikely in empirical applications. This section makes x^1 is endogenous and varies the strength of the available instrument b^1. To facilitate comparison, we slightly alter the design to vary the strength of the instrument without altering the distribution of x^1. Specifically, let the vector of instruments b^1, random noise u, and the unobserved characteristics ξ all be drawn Normal$(0, 1)$ and then construct x^1 according to,

$$x^1_{jm} = w_a b^1_{jm} + (1 - w_a)(w_c u_{jm} + (1 - w_c)\xi_{jm})$$

[36] We maintain throughout that $\theta^z_1 = \theta^z_2$ and $\theta^\nu_1 = \theta^\nu_2$ and so drop the index subscript for legibility and plot only the distribution of $\hat{\theta}^\nu_1$. Due to the symmetry of our specification, the distribution for $\hat{\theta}^\nu_2$ is the same up to simulation error.

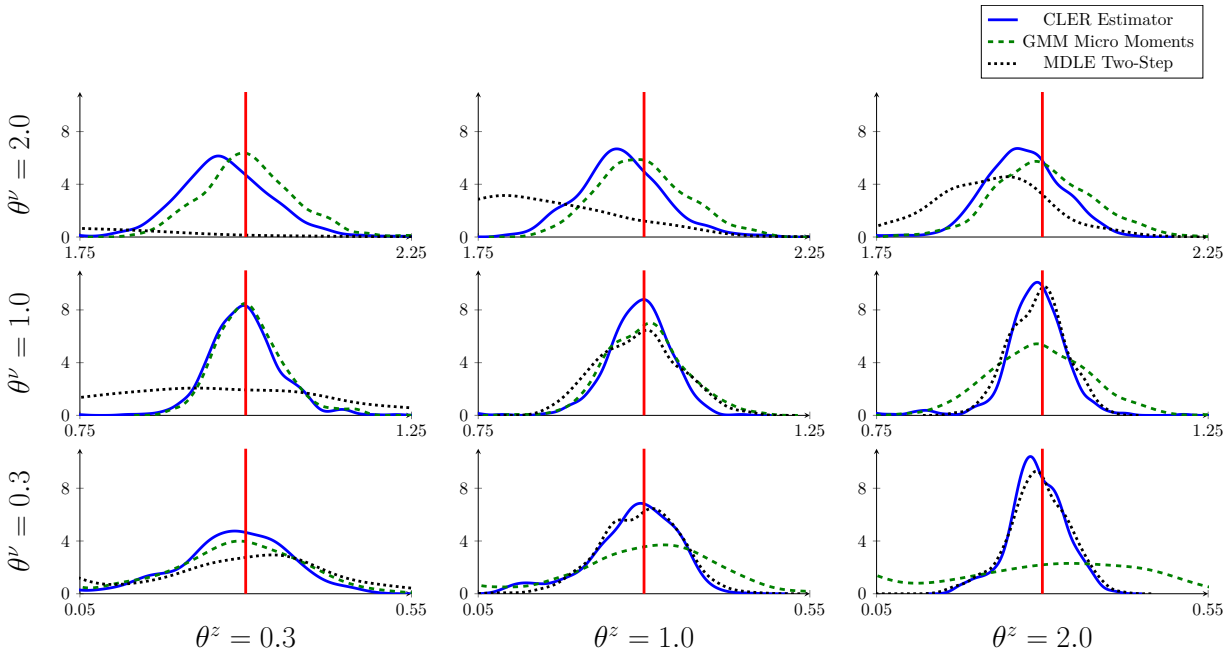

Figure 5: Distribution of $\hat{\theta}^\nu$ across three estimators as we vary θ^ν (rows) and θ^z (columns).

where $w_a = a/\sqrt{a^2 + (1-a)^2}$ for $a \in [0, 1]$ governs the strength of the instrument b^1 and $w_c = c/\sqrt{c^2 + (1-c)^2}$ for $c \in [0, 1]$ governs the degree to which the remaining variation in x is due random noise versus the product's unobserved quality. In estimation, we use b^1 as an instrument for x^1. We must also construct the differentiation instrument for x^1 using b^1 following GH20. That is, we run a first stage regression of x^1 on x^2 and b^1 and use the resulting predictions $\hat{x}^1$ to construct the differentiation IVs.

Figure 6 plots the distribution of $(\theta_1^z, \theta_1^\nu, \beta_1)$ (columns) varying a (rows), which governs the strength of the instrument.[37] When $a = 1$, x^1 is exogenous and $b^1 = x^1$. The only difference between this specification and that of our baseline (center row of figure 4) is that ξ is normally distributed here rather than Pareto. All three estimators perform well, but the CLER estimator has a slightly tighter distribution around the truth. In the $a = 0.5$ case, the instrument has moderate power. For the θ parameters, this has no effect on the MDLE two-step, which does not use the instrument to identify θ. The CLER estimator and the GMM estimator both become less precise. The biggest decline in performance comes from the GMM estimator, which ignores the micro data variation described in Section 5. As expected, all three estimators are less precise for β. Finally, when $a = 0.15$, the instrument is weak. As expected, the GMM estimator performs poorly for all three parameters. However, the distributions of CLE and the MDLE two-step estimator are

[37] For this figure, $c = 0.5$. Results varying c are available on request, but reveal little additional insight.

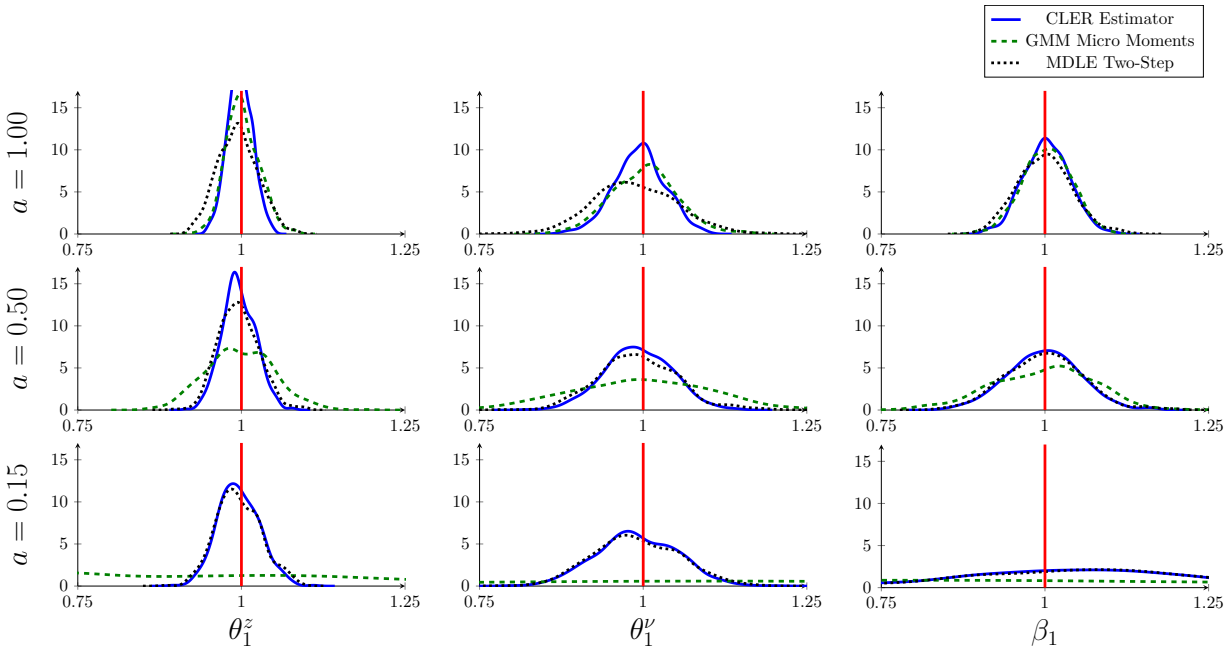

Figure 6: Distribution of parameters $(\theta_1^z, \theta_1^\nu, \beta_1)$ varying the strength of the instrument for x^1. When $a = 1$, $x_1 = b^1$, the correlation between x^1 and b^1 declines with a.

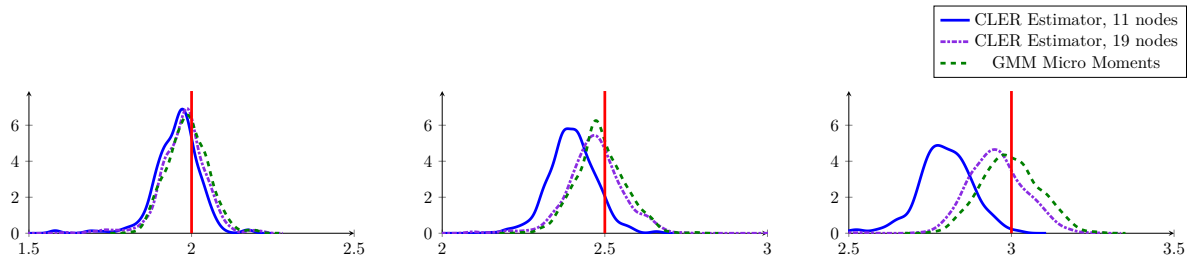

Figure 7: Distribution of $\widehat{\theta}_1^\nu$ for different values of θ^ν comparing the full estimator with 11-node and 19-node quadrature integration with the GMM estimator.

essentially identical and remain precise for θ. The MDLE two-step suffers essentially no loss of precision for the estimate of θ from the $a = 0.5$ case. The CLER estimator is no longer more precise than the MDLE owing to the fact that the product level moments are no longer adding useful information for θ, but it matches the MDLE's performance. There is also a difference for β between GMM and the two likelihood estimators. Since the MDLE two-step and CLER estimators identify θ and δ from the micro data, all the useful variation in b^1 is preserved for the estimation of β.

9.2.4 Numerical Bias. As discussed in Section 7.2, log likelihood based estimators subject to bias due to the use of numerical integration over ν. This bias will grow more severe as θ^ν rises. All the simulations above have used 11-point Gaussian quadrature (121 nodes over two dimensions of ν) to approximate the likelihood. We now compare the performance of the CLER estimator

using 19-point quadrature (361 nodes) and the GMM estimator—which does not use the log likelihood—when θ^ν is large and all other parameters of the DGP are set at our baseline.

Figure 7 displays the results. The 19-point quadrature estimator is displayed in purple. In the first panel $\theta^\nu = 2$, which corresponds to the top-center panel of figure 5. Approximation bias appears to be minimal here as all three estimates show similar results. In the center panel, $\theta^\nu = 2.5$, some bias for the 11-point quadrature estimator is apparent, but it is largely eliminated by moving to the 19-point quadrature. The GMM estimator, as expected, is unaffected. Finally, when $\theta^\nu = 3$, bias is visible for both the 11 and 19 point estimators, although it is much reduced for the more precise approximation.

Importantly, the degree of approximation bias is under the control of the researcher, and can be alleviated at the expense of more computational resources. These results suggest that the bias can be contained to acceptable levels given modern computing resources. Of course, computational demands will rise with the dimension of θ^ν. However, stipulating that the variation exists to identify a high dimensional θ^ν, one could use sparse quadrature methods to attain a high degree of accuracy with a reasonable number of integration nodes (e.g. Bansal et al., 2021).

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 the CLER estimator tthat 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. This estimator does not require additional parametric assumptions relative to a GMM estimator. By showing how to transform the CLER estimator into those used previously in the literature, we illustrate several trade-offs between statistical efficiency and other researcher concerns, such as computational tractability and data availability. With that said, we show that the CLER estimator is computationally tractable, suggesting that it will be directly useful for applied work in a wide variety of settings. Indeed, the CLER estimator has an additional advantage that inference is more straightforward and correct under more applicable assumptions than the standard approach.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica*, 88(1): 265–296.

Allen, J., R. Clark, and J.-F. Houde. 2019. "Search frictions and market power in negotiated-price markets." *Journal of Political Economy*, 127(4): 1550–1598.

Bachmann, R., G. Ehrlich, Y. Fan, D. Ruzic, and B. Leard. 2019. "Firms and collective reputation: a Study of the Volkswagen Emissions Scandal." National Bureau of Economic Research.

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

Bansal, P., V. Keshavarzzadeh, A. Guevara, S. Li, and R. A. Daziano. 2021. "Designed quadrature to approximate integrals in maximum simulated likelihood estimation." *Econometrics Journal*, 25(2): 301–321.

Bayer, P., F. Ferreira, and R. McMillan. 2007. "A unified framework for measuring preferences for schools and neighborhoods." *Journal of Political economy*, 115(4): 588–638.

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

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

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

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

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

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

Chintagunta, P. K., and J.-P. 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.

Ciliberto, F., and N. V. Kuminoff. 2010. "Public policy and market competition: how the master settlement agreement changed the cigarette industry." *BE Journal of Economic Analysis & Policy*, 10(1).

Conlon, C., and J. Gortmaker. 2020. "Best practices for differentiated products demand estimation with pyblp." *RAND Journal of Economics*, 51(4): 1108–1161.

Conlon, C., and J. Gortmaker. 2023. "Incorporating Micro Data into Differentiated Products Demand Estimation with PyBLP." *NYU working paper*.

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

Crawford, G. S., R. S. Lee, M. D. Whinston, and A. Yurukoglu. 2018. "The welfare effects of vertical integration in multichannel television markets." *Econometrica*, 86(3): 891–954.

Freyberger, J. 2015. "Asymptotic theory for differentiated products demand models with many markets." *Journal of Econometrics*, 185(1): 162–181.

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

Gandhi, A., and J.-F. Houde. 2020. "Measuring Substitution Patterns in Differentiated-Products Industries." University of Pennsylvania and UW-Madison.

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

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

Grieco, P., C. Murry, and A. Yurukoglu. 2023. "The Evolution of Market Power in the U.S. Automobile Industry." *working paper*.

Grieco, P., C. Murry, J. Pinkse, and S. Sagl. 2022. "Conformant and efficient estimation of discrete choice demand models." Penn State.

Hackmann, M. 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, J., and W. Newey. 2004. "Jackknife and analytical bias reduction for nonlinear panel models." *Econometrica*, 72(4): 1295–1319.

Ho, K. 2006. "The welfare effects of restricted hospital choice in the US medical care market." *Journal of Applied Econometrics*, 21(7): 1039–1079.

Hong, H., H. Li, and J. Li. 2021. "BLP estimation using Laplace transformation and overlapping simulation draws." *Journal of Econometrics*, 222(1): 56–72.

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

Jiménez-Hernández, D., and E. Seira. 2021. "Should the government sell you goods? Evidence from the milk market in Mexico." Stanford University Working Paper.

Ketz, P. 2019. "On asymptotic size distortions in the random coefficients logit model." *Journal of Econometrics*, 212(2): 413–432.

Montag, F. 2023. "Mergers, foreign competition, and jobs: Evidence from the US appliance industry."

Myojo, S., and Y. Kanazawa. 2012. "On Asymptotic Properties of the Parameters of Differentiated Product Demand and Supply Systems When Demographically Categorized Purchasing Pattern Data are Available." *International Economic Review*, 53(3): 887–938.

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

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

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

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

RIDDER, G., AND R. MOFFITT. 2007. "The Econometrics of Data Combination." In . Vol. 6 of *Handbook of Econometrics*, , ed. James J. Heckman and Edward E. Leamer, 5469–5547. Elsevier.

ROBINSON, P. M. 1988. "Root-N-consistent semiparametric regression." *Econometrica*, 931–954.

STAIGER, D., AND J. H. STOCK. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica*, 65(3): 557–586.

STARC, A. 2014. "Insurer pricing and consumer welfare: evidence from Medigap." *RAND Journal*, 45: 198–220.

SWEETING, A. 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, K. E., AND C. WINSTON. 2007. "Vehicle choice behavior and the declining market share of US automakers." *International economic review*, 48(4): 1469–1496.

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

VAN DEN BERG, G. J., AND B. VAN DER KLAAUW. 2001. "Combining micro and macro unemployment duration data." *Journal of Econometrics*, 102(2): 271–309.

WALKER, J. L., M. BEN-AKIVA, AND D. BOLDUC. 2007. "Identification of parameters in normal error component logit-mixture (NECLM) models." *Journal of Applied Econometrics*, 22(6): 1095–1125.

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

Appendices (for Online Publication)

A 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 loglikelihood presented in (6) (up to a constant).

Interestingly, deterministic selection based on $y_{i\cdot m}$ of the form $D_{im} = D_{im}^* \mathbb{1}(y_{i0m} \in \mathbb{J})$ where D_{im}^* is random and $\mathbb{J}$ represents a subset of products 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, $\Pr(D_{im} = 1 \cap y_{ijm} = 1 \mid z_{im}) = \Pr(D_{im}^* = 1)\pi_{jm}^{z_{im}} \mathbb{1}(j \in \mathbb{J})$, so we have

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

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,

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

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

B Weight matrix is block-diagonal

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

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

C Share constraints

C.1 Efficiency considerations: a simple example

Consider the situation in which we have 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

$$\log \hat{L}(\psi) = \sum_{i=1}^{N}\{D_i \log L_i^{\text{mic}}(\psi) + \omega(1-D_i)\log L_i^{\text{mac}}(\psi)\}, \tag{30}$$

for $\omega = 1$ and $\psi = [\theta^\top, \delta^\top]$ where $\log L_i^{\text{mic}} = \sum_j y_{ij}\log \pi_j^{z_i}$ and $\log L_i^{\text{mac}} = \sum_j y_{ij}\log \pi_j$ are contributions to the loglikelihood for observation i and D_i is the micro selection dummy which is independent of everything else and equals one with fixed 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.

Lemma 2. Under the stated assumptions we have, $\sqrt{N}(\hat{\psi} - \psi) \xrightarrow{d} N(0, V)$, where $V = (\chi A + \omega(1-\chi)B)^{-1}(\chi A + \omega^2(1-\chi)B)(\chi A + \omega(1-\chi)B)^{-1}$, with $A = -\mathbb{E}\{\partial_{\psi\psi^\top}\log L_i^{\text{mic}}(\psi)\}$ and $B = -\mathbb{E}\{\partial_{\psi\psi^\top}\log L_i^{\text{mac}}(\psi)\}$. The optimal weight ω equals one. □

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

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

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

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

$$\hat{\psi}_\omega^* = \arg\max_{\psi} \sum_{i=1}^{N}\{D_i \log L_i^{\text{mic}}(\psi) + \omega \log L_i^{\text{mac}}(\psi)\}, \tag{31}$$

as a function of ω and then letting $\omega \to \infty$. Note that imposing that the gradient of $\sum_{i=1}^{N}\log L_i^{\text{mac}}$ equal

zero is equivalent to imposing the product level share equations. Note further that there is a subtle but important difference between (30) and (31) in that in (31) we sum over all $\log L_i^{\text{mac}}$, 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.

Lemma 3. Let V_ω^* be the asymptotic variance of $\hat{\psi}_\omega^*$. Then $V_\infty^* = \lim_{\omega\to\infty} V_\omega^* = \{\chi A U_0 (U_0^\top A U_0)^{-1} U_0^\top A + B\}^{-1} \geq V$, 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 $B = U_1 D_1 U_1^\top$ 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 $U_0(\chi U_0^\top A U_0)^{-1} U_0^\top + U_1 D_1^{-1} U_1^\top/\omega^2$.[38] Pre and postmultiply by $\chi A + \omega B$ and take $\omega \to \infty$ to obtain V_∞^*. Finally, note that

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

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 Asymptotic variance comparison in a single market

This appendix provides formulas for the asymptotic variance of the MDLE of ψ and the estimator that maximizes the mixed logit objective function subject to the share constraints for a single market, i.e. $m = 1$. 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 $\mathbb{L}^{\text{mic}}$ to denote $\mathbb{E}\sum_{j=0}^J Y_{ij} \log \pi_j^{z_i}$, $\mathbb{L}_\psi^{\text{mic}}$ its gradient, $\mathbb{L}_{\psi\psi}^{\text{mic}}$ its Hessian, and $\mathbb{L}^{\text{mac}} = \mathbb{E}\sum_{j=0}^J Y_{ij} \log \pi_j$. Let similar symbols be analogously defined.

For the MDLE, if $\chi > 0$, the asymptotic variance of $\sqrt{N}(\hat{\psi} - \psi)$ and $\chi > 0$ is then

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

$$-\{\chi \mathbb{L}_{\psi\psi}^{\text{mic}} + (1-\chi)\mathbb{L}_{\psi\psi}^{\text{mac}}\}^{-1}. \tag{32}$$

For $\chi = 0$, consider the limit distribution of $\sqrt{N\chi}(\hat{\psi} - \psi)$ for $\chi > 0$, i.e. multiply (32) by χ and then let $\chi \downarrow 0$. This takes some caution since $\mathbb{L}_{\psi\psi}^{\text{mac}}$ is generally singular.

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

$$-\begin{bmatrix} I \\ \partial_\theta \delta^\top \end{bmatrix} \Phi^{-1} \begin{bmatrix} I & \partial_{\theta^\top}\delta \end{bmatrix} \Big/ \chi, \quad \textit{(incorrect variance)}$$

where $\partial_{\theta^\top}\delta = -(\mathbb{L}_{\delta\delta}^{\text{mac}})^{-1}\mathbb{L}_{\delta\theta}^{\text{mac}}$ and $\Phi = \mathbb{L}_{\theta\theta}^{\text{mic}} + \partial_\theta\delta^\top\mathbb{L}_{\delta\theta}^{\text{mic}} + \mathbb{L}_{\theta\delta}^{\text{mic}}\partial_{\theta^\top}\delta + \partial_\theta\delta^\top\mathbb{L}_{\delta\delta}^{\text{mic}}\partial_{\theta^\top}\delta$. The correct asymptotic variance formula for the share constrained estimator is

$$-\begin{bmatrix} \chi\Phi & \chi(\mathbb{L}_{\theta\delta}^{\text{mic}} + \partial_\theta\delta^\top\mathbb{L}_{\delta\delta}^{\text{mic}}) \\ \mathbb{L}_{\delta\theta}^{\text{mac}} & \mathbb{L}_{\delta\delta}^{\text{mac}} \end{bmatrix}^{-1} \begin{bmatrix} \chi\Phi & 0 \\ 0 & \mathbb{L}_{\delta\delta}^{\text{mac}} \end{bmatrix} \begin{bmatrix} \chi\Phi & \mathbb{L}_{\theta\delta}^{\text{mac}} \\ \chi(\mathbb{L}_{\delta\theta}^{\text{mic}} + \mathbb{L}_{\delta\delta}^{\text{mic}}\partial_{\theta^\top}\delta) & \mathbb{L}_{\delta\delta}^{\text{mac}} \end{bmatrix}^{-1}. \tag{33}$$

The formula in (33) is based on the fact that the share constrained estimator uses the following moment conditions:

$$\begin{cases} \displaystyle\sum_{i=1}^{N}\sum_{j=0}^{J} y_{ij} D_i \left(\partial_\theta \log \pi_j^{z_i} + \partial_\theta\hat{\delta}^\top \partial_\delta \log \pi_j^{z_i}\right) = 0, \\ \displaystyle\sum_{i=1}^{N}\sum_{j=0}^{J} y_{ij}\partial_\delta \log \pi_j = 0, \end{cases} \tag{34}$$

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

Finally, a mixed logit estimator ignoring the product share information would have asymptotic variance $(-\mathbb{L}_{\psi\psi}^{\text{mic}})^{-1}/\chi$.

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 (25) 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 4. 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,

$$\forall\delta : \{\delta - X\hat{\beta}(\delta)\}^\top\mathcal{P}_B\{\delta - X\hat{\beta}(\delta)\} = \delta^\top\mathcal{K}\mathcal{K}^\top\delta, \tag{35}$$

where $\mathcal{K} = \mathcal{U}_B\mathcal{M}_{\mathcal{U}_B^\top\mathcal{U}_X}$ 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 (35) can be expressed as $\delta^\top \mathscr{P}^* \delta$ where $\mathscr{P}^* = \mathscr{P}_B - \mathscr{P}_{\mathscr{P}_B X}$. Noting that $\mathscr{P}_B = \mathscr{P}_C + \mathscr{P}_{B^*}$ and $\mathscr{P}_{\mathscr{P}_B X} = \mathscr{P}_C + \mathscr{P}_{\mathscr{P}_{B^*} X^*}$, we have $\mathscr{P}^* = \mathscr{P}_{B^*} - \mathscr{P}_{\mathscr{P}_{B^*} X^*}$. The stated result then follows by application of the singular value decomposition to both B^* and X^*. □

E Hessian 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 5. Let $\psi_m^\nu = [\theta^{\nu\top},\ \delta_m^\top]^\top$. If $\theta^z = 0$ then $\partial_{\psi_m^\nu \psi_m^{\nu\top}} \log L$ can have rank at most J_m.

Proof. Consider the case in which $\chi_m = 1$, 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, recalling from section 8 that $\mathscr{J}$ is the sigma field generated by $B, X, \xi, \{N_m\}, \{\chi_m\}$,

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

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

Lemma 6.

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

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 7. If $N\chi/M \to \infty$ and $\theta^z = 0$ then the dominant term of the (θ^z, θ^z) block of the inverse

Hessian evaluated at the truth is

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

Proof. First, note that by partitioned inverses,

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

where $\partial_{\delta\delta^\top}\hat{\Pi}^*=\partial_{\delta\delta^\top}\hat{\Pi}-\partial_{\delta\beta^\top}\hat{\Pi}(\partial_{\beta\beta^\top}\hat{\Pi})^{-1}\partial_{\beta\delta^\top}\hat{\Pi}$. Since $-\mathbb{E}(\partial_{\psi^\nu\psi^{\nu\top}}\log\hat{L}\mid\mathcal{J})$ is positive semidefinite with rank J by lemma 5, we can replace $\partial_{\psi^\nu\psi^{\nu\top}}\log\hat{L}$ with $\mathcal{A}\partial_{\delta\delta^\top}\log\hat{L}\mathcal{A}^\top$. Thus, by partitioned inverses we get

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

as asserted. □

Lemma 8. Absent consumer data and evaluated at the truth, $\hat{\Omega}^{\theta^\nu\theta^\nu}\simeq\left(\partial_\theta\delta^\top\partial_{\delta\delta^\top}\hat{\Pi}^*\partial_{\theta^\top}\delta\right)^{-1}$, where $\partial_{\delta\delta^\top}\hat{\Pi}^*$ was defined in lemma 7 and where $\delta(\theta)$ solves the (expectation) share constraint.

Proof. By lemma 6 we get,

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

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 Proof of Theorem 1

The proof requires the introduction of some notation. First, we define a recentered version of our log likelihood,

$$\begin{aligned}\hat{\mathcal{L}}_m^{\text{mac}}(\theta,\delta_m) &= \log\frac{\hat{L}^{\text{mac}}(\theta_0,\delta_0)}{\hat{L}^{\text{mac}}(\theta,\delta)} = N_m\sum_{j=0}^{J_m} s_{jm}\log\frac{\pi_{jm}(\psi_{0m})}{\pi_{jm}(\theta,\delta_m)},\\ \hat{\mathcal{L}}_m^{\text{mic}}(\theta,\delta_m) &= \log\frac{\hat{L}^{\text{mic}}(\theta_0,\delta_0)}{\hat{L}^{\text{mic}}(\theta,\delta)} = \sum_{i=1}^{N_m} D_{im}\sum_{j=0}^{J_m} Y_{ijm}\log\frac{\pi_{jm}^{z_{im}}(\psi_{0m})\pi_{jm}(\theta,\delta_m)}{\pi_{jm}(\psi_{0m})\pi_{jm}^{z_{im}}(\theta,\delta_m)},\end{aligned} \tag{36}$$

and use $\hat{\mathcal{L}}_m = \hat{\mathcal{L}}_m^{\text{mac}} + \hat{\mathcal{L}}_m^{\text{mic}}$, $\hat{\mathcal{L}} = \sum_{m=1}^M \hat{\mathcal{L}}_m$. The definitions $\hat{\mathcal{L}}_m^{\text{mac}}$, $\hat{\mathcal{L}}_m^{\text{mic}}$ are recenterings of the macro and micro terms of $-\log\hat{L}_m$ to make them equal to zero if evaluated at the truth. Their population analogs have no hat, e.g. $\mathcal{L}_m^{\text{mic}}(\theta,\delta_m) = \mathbb{E}\{\hat{\mathcal{L}}_m^{\text{mic}}(\theta,\delta_m)\mid\mathcal{J}\}$. Subscripts corresponding to parameters denote partial derivatives, e.g. $\mathcal{L}_{m\psi\psi}^{\text{mic}}$ is the Hessian of the micro likelihood for market m, where we use the subscripts z,ν to denote partial derivatives with respect to θ^z,θ^ν.

Let $\hat{\Omega}(\psi) = \hat{\mathcal{L}}(\psi) + \Pi(\delta)$ with $\hat{\mathcal{L}}(\psi) = -\log\hat{L}(\psi)$ and $\Pi(\delta) = \hat{\Pi}\{\hat{\beta}(\delta),\delta\} = \delta^\top\mathcal{K}\mathcal{K}^\top\delta/2$. This is our objective function (5) after concentrating out β and the above-mentioned recentering. We further define $\Omega(\psi) = \mathcal{L}(\psi) + \Pi(\delta)$.

Next, we define the objects, $\delta_0(\theta) = \delta_0^\Omega(\theta) \in \arg\min_\delta \Omega(\theta,\delta)$, $\delta_0^{\mathcal{L}}(\theta) \in \arg\min_\delta \mathcal{L}(\theta,\delta)$, and $\delta_0^{\text{mac}}(\theta) = \arg\min_\delta \mathcal{L}^{\text{mac}}(\theta,\delta)$ and their sample analogs which receive hats. $\delta_0^{\text{mac}}(\theta)$ is unique since there is a one to one mapping from choice probabilities to δ as shown by Berry (1994). We will show that all three are equal at θ_0.[39] We further show that the difference between $\delta_0^{\mathcal{L}}(\theta)$ and $\delta_0(\theta)$ is negligible at any $\theta\in\Theta$.

Finally, let $\hat{\Omega}^*(\theta,\delta) = \hat{\mathcal{L}}^{\text{mic}}(\theta,\delta) + \hat{\mathcal{L}}^{\text{mac*}}(\theta,\delta) + \Pi(\delta)$, where

$$\hat{\mathcal{L}}_m^{\text{mac*}}(\theta,\delta_m) = N_m\sum_{j=0}^{J_m} s_{jm}\log\frac{s_{jm}}{\pi_{jm}(\theta,\delta_m)},$$

so we replaced the true choice probabilities in the numerator of $\hat{\mathcal{L}}_m^{\text{mac}}$ with observed market shares, s_{jm}. Define further $\hat{\mathcal{R}}(\theta,\delta) = \mathcal{L}^{\text{mic}}(\theta,\delta) + \hat{\mathcal{L}}^{\text{mac*}}(\theta,\delta) + \Pi\{\delta_0^{\text{mac}}(\theta)\}$. where the sample micro likelihood is replaced with its population analog and the argument of Π is now $\delta_0^{\text{mac}}(\theta)$ instead of δ.

Proof of theorem 1. Because $\hat{\beta}$ is a linear combination of $\hat{\delta}$, we only establish asymptotic normality of $\Lambda(\hat{\psi}-\psi_0)$ to reduce notation. Now, the (ψ,ψ) block of $\hat{V}$ is $\hat{\Omega}_{\psi\psi}^{-1}(\hat{\psi})$ so we will show asymptotic normality of $\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda(\hat{\psi}-\psi_0)$ where $\hat{\mathcal{D}}_\Lambda = \Lambda\hat{\Omega}_{\psi\psi}^{-1}(\hat{\psi})\Lambda^\top$.

Lemma 9 in appendix F.2 establishes consistency of $\hat{\theta}$ for θ_0, which guarantees that $\hat{\Omega}(\hat{\theta},\delta)$ is convex in δ with probability approaching one. Asymptotic normality is then established by lemmas 26 and 34. □

Before presenting the supporting lemmas, we outline the intuition of the proof.

[39] In fact, the steeper gradient of δ_0 at θ_0 relative to $\delta_0^{\text{mac}}(\theta)$ is the source of our estimator's efficiency gains relative to share constrained methods discussed in section 6.2.

F.1 Intuitive outline of the proof

F.1.1 Consistency of $\hat{\theta}$. We establish consistency of $\hat{\theta}$ for θ_0 in lemma 9 in appendix F.2. We do so in two steps, first obtaining an upper bound to the rate at which (a centered version of) the profiled objective function at the truth (θ_0), $\min_\delta \hat{\Omega}^*(\theta_0, \delta)$ diverges, then obtaining a lower bound to the rate at which the profiled objective function outside an ϵ-neigborhood of θ_0 diverges. The lower bound outside the neighborhood diverges faster than the upper bound inside. Since $\hat{\theta}$ minimizes $\min_\delta \hat{\Omega}^*(\theta, \delta)$, it must be true that $\hat{\theta} \xrightarrow{p} \theta_0$. The following two paragraphs describe these steps in more detail.

We achieve the first step by obtaining an upper bound on the divergence rate of $\hat{\Omega}^*\{\theta_0, \hat{\delta}^{\text{mac}}(\theta_0)\}$, which is trivially above the minimized profiled likelihood. This is (up to a constant shift) equal to the profiled objective function of a share constrained estimator. Lemma 10 establishes this bound. Noting that $\hat{\Omega}^*$ consists of three terms, $\hat{\Omega}^*(\theta, \delta) = \hat{\mathcal{L}}^{\text{mic}}(\theta, \delta) + \hat{\mathcal{L}}^{\text{mac*}}(\theta, \delta) + \Pi(\delta)$, the lemma proceeds by bounding the individual terms: The first term is bounded by lemma 11, which establishes a bound of $\lambda M\sqrt{\overline{\chi}}$. The second term, $\hat{\mathcal{L}}^{\text{mac*}}\{\theta_0, \hat{\delta}^{\text{mac}}(\theta_0)\}$, equals 0 by construction. The final term is shown to be bounded in probability by lemma 12.

Lemma 18 establishes the second step by providing a lower bound on the rate of the profiled objective outside a neighborhood of θ_0. Here we make use of a surrogate $\hat{\mathcal{R}}$ of the unprofiled objective $\hat{\Omega}^*$ which (a) replaces $\hat{\mathcal{L}}^{\text{mic}}$ with $\mathcal{L}^{\text{mic}}$, its expectation condtional on $\mathcal{J}$, and (b) replaces $\Pi(\delta)$ with $\Pi\{\delta_0^{\text{mac}}(\theta)\}$. Lemma 20 shows that $\hat{\Omega}^*$ is well approximated by $\hat{\mathcal{R}}$ in the sense that $|\hat{\Omega}^* - \hat{\mathcal{R}}|/\hat{\mathcal{R}}$ is uniformly small. Lemma 19 shows that $\hat{\mathcal{R}}$ diverges at at least the rate $N\chi\lambda^2 + M$ uniformly in (θ, δ) for θ away from θ_0.

With consistency established, lemma 25 provides a lower bound on the rate of convergence using the same machinery.

F.1.2 Asymptotic normality. With consistency established, we show asymptotic normality in two (large) steps in appendix F.3. First, we show in lemma 26 that for any vector v, the estimation error $v^\top(\hat{\psi} - \psi_0)$ is equal to $-v^\top \Omega_{\psi\psi}^{-1}(\psi_0)\hat{\Omega}_\psi(\psi_0)$ plus asymptotically negligible terms. In the second step, we apply a central limit theorem for martingale difference sequences to show asymptotic normality of $\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda\Omega_{\psi\psi}^{-1}(\psi_0)\hat{\Omega}_\psi(\psi_0)$ in lemma 34. Recalling that $\hat{\mathcal{D}}_\Lambda = \Lambda\hat{\Omega}_{\psi\psi}^{-1}(\hat{\psi})\Lambda^\top$, this fact, together with the first step, establishes asymptotic normality of $\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda(\hat{\psi} - \psi_0)$ and completes the proof.

The first step has three parts which correspond to δ, θ^z and θ^ν respectively. In the δ part, we show that $\hat{\delta}_m - \delta_m$ as a process of θ can be approximated by a linear combination of the gradient of the likelihood function, $\hat{\mathcal{L}}_\delta$, uniformly in m and θ in a neighborhood of θ_0. This is established in lemma 27.[40] The θ^z part, established in lemma 32, shows that uniformly in a neighborhood of θ_0^ν, $\hat{\theta}^z(\theta^\nu) - \theta_0^z(\theta^\nu)$ can approximated by a linear combination of the gradient of the objective function $\hat{\Omega}$ which uses the approximation of δ. The θ^ν part, shown in lemma 33, establishes that $\hat{\theta}^\nu - \theta_0^\nu$ can be approximated by a linear combination of the gradient of the objective function and the previous approximations. Finally, these results are collected in lemma 26.

In the second step, there are three challenges to applying a central limit theorem to

[40] This approximation involves the likelihood rather than the objective function $\hat{\Omega}$ because $\hat{\delta}_m(\theta)$ is a function of θ; for fixed θ the influence of the product level moments is negligible analagous to the Berry (1994) share inversion.

$$\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda\Omega_{\psi\psi}^{-1}(\psi_0)\hat{\Omega}_\psi(\psi_0) \simeq -\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda(\hat{\psi}-\psi_0).$$

First, the dimension of δ is growing with M, which is why we focus on a finite dimensional linear combination of $\hat{\psi}$. The second challenge is that δ acts as a product quality parameter in the likelihood and as a random variable in Π (e.g., due to its dependence on ξ). Third, the number of consumers in the micro sample, the number of consumers in the population, and the number of markets all diverge at different rates. Moreover, the information contained in the micro sample is allowed to decrease in the case of weak micro identification.

We overcome these challenges by using a martingale difference central limit theorem. This allows for the varying rates of the third challenge. However, $\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda\Omega_{\psi\psi}^{-1}(\psi_0)\hat{\Omega}_\psi(\psi_0)$ itself is not the sum of martingale differences due to endogeneity arising from the second challenge. We address this issue by showing that the difference between $\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda\Omega_{\psi\psi}^{-1}(\psi_0)\hat{\Omega}_\psi(\psi_0)$ and its analog replacing some objects with their expectation conditional on either $\mathcal{J}$ or $\mathcal{B}$—which is a sum of martingale differences—is asymptotically negligible.

Lemma 35 uses lemma 49 to establish normality of the martingale difference sum analog. Lemma 36, supported by lemmas 37 to 39, shows that the difference between $\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda\Omega_{\psi\psi}^{-1}(\psi_0)\hat{\Omega}_\psi(\psi_0)$ and the analog is asymptotically irrelevant.

F.2 Consistency

Let $\Theta_\epsilon = \{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon\}$.

Lemma 9. $\hat{\theta} \xrightarrow{p} \theta_0$

Proof. For consistency, we only need to take a fixed $\epsilon > 0$. Then,

$$\min_{\theta\in\Theta_\epsilon}\min_{\delta\in\Delta}\hat{\Omega}^*(\theta,\delta) \succeq \min_{\theta\in\Theta_\epsilon}\min_{\delta\in\Delta}\hat{\mathcal{R}}(\theta,\delta) \succeq$$
$$N\chi\lambda^2 + M \succ M\sqrt{\overline{\chi}}\lambda + 1 \succeq \hat{\Omega}^*\{\theta_0, \hat{\delta}^{\text{mac}}(\theta_0)\} \succeq \min_{\delta\in\Delta}\hat{\Omega}^*(\theta_0,\delta), \quad (37)$$

where the rate inequalities follow from lemmas 10 and 18. Hence, with probability approaching one, the minimizer of $\hat{\Omega}^*$, and hence the minimizer of $\hat{\Omega}$, will not be an element of $\Theta_\epsilon \times \Delta$. □

F.2.1 Showing $\min_{\delta\in\Delta}\hat{\Omega}^*(\theta_0,\delta) \preceq \hat{\Omega}^*\{\theta_0, \hat{\delta}^{\text{mac}}(\theta_0)\} \preceq M\sqrt{\overline{\chi}}\lambda + 1$**.** Let $a_{ijm}(\psi_m) = \log \pi_{jm}^{z_{im}}(\psi_m) - \log \pi_{jm}(\psi_m)$ and let additional subscripts denote partial derivatives.

Lemma 10. $\hat{\Omega}^*\{\theta_0, \hat{\delta}^{\text{mac}}(\theta_0)\} \preceq \lambda M\sqrt{\overline{\chi}} + 1.$

Proof. Follows from lemmas 11 and 12 and the fact that $\hat{\mathcal{L}}^{\text{mac}*}\{\theta_0, \hat{\delta}^{\text{mac}}(\theta_0)\} = 0$ by construction. □

Lemma 11. $\hat{\mathcal{L}}^{\text{mic}}\{\theta_0, \hat{\delta}^{\text{mac}}(\theta_0)\} \preceq \lambda M\sqrt{\overline{\chi}}.$

Proof. The left hand side is $\sum_{m=1}^{M}\sum_{i=1}^{N_m} D_{im}\sum_{j=0}^{J_m} Y_{ijm}\{a_{ijm}(\psi_{0m}) - a_{ijm}(\psi_m)\}$. Use the mean value

theorem to obtain

$$-\sum_{mij} D_{im}Y_{ijm}a_{\delta ijm}^{\top}(\psi_{0m})(\hat{\delta}_m^{\text{mac}}(\theta_0)-\delta_{0m}) \\ -\frac{1}{2}\sum_{mij}(\hat{\delta}_m^{\text{mac}}(\theta_0)-\delta_{0m})^{\top}D_{im}Y_{ijm}a_{\delta\delta ijm}(\theta_0,\delta_m^*)(\hat{\delta}_m^{\text{mac}}(\theta_0)-\delta_{0m}),$$

The square of the first order term is by the Schwarz inequality bounded above by

$$-\sum_m N_m^{-1}\|\sum_{ij} D_{im}Y_{ijm}a_{\delta ijm}(\psi_{0m})\|^2\sum_m N_m\|\hat{\delta}_m^{\text{mac}}(\theta_0)-\delta_{0m}\|^2 \preceq M\bar{\chi}\lambda^2 M = M^2\bar{\chi}\lambda^2,$$

by lemmas 14 and 16.

Now the second order term. It is bounded above by a half times the square root of

$$\sum_m N_m^2\|\hat{\delta}_m^{\text{mac}}(\theta_0)-\delta_{0m}\|^4\sum_m\Big\|N_m^{-1}\sum_{ij} D_{im}Y_{ijm}a_{\delta\delta ijm}(\theta_0,\delta_m^*)\Big\|^2. \tag{38}$$

By lemmas 14 and 17, the right hand side in (38) is $\preceq M\times\lambda^2 M\bar{\chi} = \lambda^2 M^2\bar{\chi}$. □

Lemma 12. $\Pi\{\hat{\delta}^{\text{mac}}(\theta_0)\}\preceq 1$.

Proof. We have

$$2\Pi\{\hat{\delta}^{\text{mac}}(\theta_0)\} = \{\hat{\delta}^{\text{mac}}(\theta_0)-\delta_{0m}\}^{\top}\mathcal{K}\mathcal{K}^{\top}\{\hat{\delta}^{\text{mac}}(\theta_0)-\delta_{0m}\}+2\{\hat{\delta}^{\text{mac}}(\theta_0)-\delta_{0m}\}^{\top}\mathcal{K}\mathcal{K}^{\top}\xi+\xi^{\top}\mathcal{K}\mathcal{K}^{\top}\xi\preceq 1,$$

by lemmas 13 and 15. □

Lemma 13. $\Pi(\delta_{0m})\preceq 1$.

Proof. Follows from the orthogonality of B,ξ and the definition of $\mathcal{K}$. □

Lemma 14. For $t=1,2$, $\sum_{m=1}^M N_m^t\|\hat{\delta}_m^{\text{mac}}(\theta_0)-\delta_{0m}\|^{2t}\preceq M$.

Proof. We show the result for $t=1$, where the result for $t=2$ is similar. From lemma 15, it follows that

$$\sum_{m=1}^M N_m\|\hat{\delta}_m^{\text{mac}}(\theta_0)-\delta_{0m}\|^2\simeq\sum_{m=1}^M N_m\Big\|\Big(\partial_{\delta_m^{\top}}\pi_m(\psi_{0m})\Big)^{-1}\{s_m-\pi_m(\psi_{0m})\}\Big\|^2.$$

Now take the expectation of the right hand side conditional on $\mathcal{J}$ to obtain

$$\sum_{m=1}^M \text{tr}\Big\{\Big(\partial_{\delta_m^{\top}}\pi_m(\psi_{0m})\Big)^{-1}\{\Pi_m^*(\psi_{0m})-\pi_m(\psi_{0m})\pi_m^{\top}(\psi_{0m})\}\Big(\partial_{\delta_m^{\top}}\pi_m(\psi_{0m})\Big)^{-1}\Big\},$$

where $\Pi_m^* = \text{diag}(\pi_m)$. Now take its expectation and use lemma 48 and assumption C to conclude that a sufficient condition is that

$$\max_m \mathbb{E} \max_{j=1,\dots,J_m} \exp(5|\xi_{jm}|) \leq \max_m J_m \mathbb{E} \exp(5|\xi_{jm}|) < \infty,$$

which is implied by assumptions C and D and lemma 52. □

Lemma 15.

$$\max_{m=1,\dots,M} \sqrt{N_m} \sup_{\theta \in \Theta} \left\| \hat{\delta}_m^{\text{mac}}(\theta) - \delta_{0m}^{\text{mac}}(\theta) - \left(\partial_{\delta_m^\top} \pi_m \{\theta, \delta_{0m}^{\text{mac}}(\theta)\} \right)^{-1} (s_m - \pi_m(\psi_{0m})) \right\| \prec 1.$$

Proof. This is an application of lemma 53. □

Lemma 16. $\sum_{m=1}^M N_m^{-1} \left\| \sum_{ij} D_{im} Y_{ijm} a_{\delta ijm}(\psi_{0m}) \right\|^2 \preceq M \bar{\chi} \lambda^2$.

Proof. Note that the inner summand has mean zero conditional on $\mathcal{J}$. Take the expectation of the left hand side conditional on $\mathcal{J}$ to obtain $\sum_{mj} \chi_m \mathbb{E}(\pi_{jm}^{z_{im}}(\psi_{0m}) \|a_{\delta ijm}(\psi_{0m})\|^2 \mid \mathcal{J}_m)$. Expand the norm around $\theta_0^z = 0$ and apply lemma 47. □

Lemma 17. $\sum_{m=1}^M \sup_{\delta_m} \left\| N_m^{-1} \sum_{ij} D_{im} Y_{ijm} a_{\delta\delta ijm}(\theta_0, \delta_m) \right\|^2 \preceq M \bar{\chi} \lambda^2$.

Proof. The left hand side is bounded above by

$$\sum_{m=1}^M \left(\frac{1}{N_m} \sum_{ij} D_{im} Y_{ijm} \sup_{\delta_m} \|a_{\delta\delta ijm}(\theta_0, \delta_m)\| \right)^2 \preceq \sum_{m=1}^M \chi_m \lambda^2 = M \bar{\chi} \lambda^2,$$

by applying lemma 47, taking an expectation, and expanding $a_{\delta\delta ijm}(\theta_0, \delta_m)$ around $\theta_0^z = 0$. □

F.2.2 Showing $\min_{\theta \in \Theta_\epsilon} \min_{\delta \in \Delta} \hat{\Omega}^*(\theta, \delta) \succeq \min_{\theta \in \Theta_\epsilon} \min_{\delta \in \Delta} \hat{\mathcal{R}}(\theta, \delta) \succeq N\chi\lambda^2 + M$.

Lemma 18. $\min_{\theta \in \Theta_\epsilon, \delta \in \Delta} \hat{\Omega}^*(\theta, \delta) \succeq (N\chi\lambda^2 + M)\epsilon^2$

Proof. Follow from lemmas 19 and 20. □

Lemma 19. For all (possibly decreasing) $\epsilon > 0$, $\min_{\theta \in \Theta_\epsilon, \delta \in \Delta} \hat{\mathcal{R}}(\theta, \delta) \succeq (N\chi\lambda^2 + M)\epsilon^2$.

Proof. By condition (f) of assumption F, $\min_{\theta \in \Theta_\epsilon} \text{II}\{\delta_0^{\text{mac}}(\theta)\} \succeq M\epsilon^2$. Further, by assumption G, for some $C^{\text{mic}} > 0$, expanding $\mathcal{L}^{\text{mic}}$ around θ_0 and applying lemma 47 yields

$$\min_{\theta \in \Theta_\epsilon} \min_{\delta \in \Delta} \mathcal{L}^{\text{mic}}(\theta, \delta) \succeq N\chi \min_{\theta \in \Theta_\epsilon} \|\theta - \theta_0\|_\lambda^2 = N\chi \min_{\|\theta^z - \theta_0^z\| \leq \epsilon} \{\|\theta^z - \theta_0^z\|^2 (1 - \lambda^2) + \lambda^2 \epsilon^2\} \succeq N\chi\lambda^2\epsilon^2,$$

as asserted. □

Lemma 20. $\max_{\theta \in \Theta_\epsilon, \delta \in \Delta} \left| \{\hat{\Omega}^*(\theta, \delta) - \hat{\mathcal{R}}(\theta, \delta)\} \,/\, \hat{\mathcal{R}}(\theta, \delta) \right| \prec 1$.

Proof. Follows from lemmas 21 and 22. □

Lemma 21. $\sup_{\theta\in\Theta_\epsilon}\sup_{\delta\in\Delta}\left|\{\hat{\mathcal{L}}^{\text{mic}}(\theta,\delta)-\mathcal{L}^{\text{mic}}(\theta,\delta)\}\ /\ \hat{\mathcal{R}}(\theta,\delta)\right|\prec 1.$

Proof. Combine lemmas 23 and 24. □

Lemma 22. $\max_{\theta\in\Theta_\epsilon,\delta\in\Delta}\left|[\Pi(\delta)-\Pi\{\delta_0^{\text{mac}}(\theta)\}]\ /\ \hat{\mathcal{R}}(\theta,\delta)\right|\prec 1.$

Proof. We have

$$2|\Pi(\delta)-\Pi\{\delta_0^{\text{mac}}(\theta)\}|\le 2\|\mathcal{K}^\top\{\delta-\delta_0^{\text{mac}}(\theta)\}\|\|\mathcal{K}^\top\delta_0^{\text{mac}}(\theta)\|+\|\mathcal{K}^\top\{\delta-\delta_0^{\text{mac}}(\theta)\}\|^2,$$

uniformly in θ,δ. Now, $\sup_{\theta\in\Theta}\|\mathcal{K}^\top\delta_0^{\text{mac}}(\theta)\|\le\sup_{\theta\in\Theta}\|\delta_0^{\text{mac}}(\theta)\|\preceq\sqrt{M}$. Further,

$$\max_{\theta\in\Theta_\epsilon,\delta\in\Delta}\frac{\|\mathcal{K}^\top\{\delta-\delta_0^{\text{mac}}(\theta)\}\|^2}{\hat{\mathcal{R}}(\theta,\delta)}\le \max_{\theta\in\Theta_\epsilon,\delta\in\Delta}\frac{\|\mathcal{K}^\top\{\delta-\hat{\delta}^{\text{mac}}(\theta)\}\|^2}{\hat{\mathcal{R}}(\theta,\delta)}+\max_{\theta\in\Theta_\epsilon,\delta\in\Delta}\frac{\|\mathcal{K}^\top\{\hat{\delta}^{\text{mac}}(\theta)-\delta_0^{\text{mac}}(\theta)\}\|^2}{\hat{\mathcal{R}}(\theta,\delta)}\prec 1, \quad (39)$$

by lemma 15 and the definition of $\hat{\mathcal{R}}$. □

Lemma 23.

$$\sup_{\theta^z\neq 0}\sup_{\delta}\frac{\left[\hat{\mathcal{L}}^{\text{mic}}(\theta,\delta)-\hat{\mathcal{L}}^{\text{mic}}\{\theta,\delta_0^{\text{mac}}(\theta)\}-\mathcal{L}^{\text{mic}}(\theta,\delta)+\mathcal{L}^{\text{mic}}\{\theta,\delta_0^{\text{mac}}(\theta)\}\right]^2}{\hat{\mathcal{R}}^2(\theta,\delta)}\prec 1.$$

Proof. Let $\tilde{a}_{im}(\theta,\delta_m)=\sum_{j=0}^{J_m}[D_{im}Y_{ijm}a_{ijm}(\theta,\delta_m)-\mathbb{E}\{D_{im}Y_{ijm}a_{ijm}(\theta,\delta_m)\mid\mathcal{J}\}]$. We have by the mean value theorem[41] that

$$\begin{aligned}&\hat{\mathcal{L}}^{\text{mic}}(\theta,\delta)-\hat{\mathcal{L}}^{\text{mic}}\{\theta,\delta_0^{\text{mac}}(\theta)\}-\mathcal{L}^{\text{mic}}(\theta,\delta)+\mathcal{L}^{\text{mic}}\{\theta,\delta_0^{\text{mac}}(\theta)\}\\&=\sum_{mi}\tilde{a}_{\delta im}^\top\{\theta,\delta_{0m}^{\text{mac}}(\theta)\}\{\delta_m-\delta_{0m}^{\text{mac}}(\theta)\}+\frac{1}{2}\sum_{mi}\{\delta_m-\delta_{0m}^{\text{mac}}(\theta)\}^\top\tilde{a}_{\delta\delta im}(\theta,\delta_m^*)\{\delta_m-\delta_{0m}^{\text{mac}}(\theta)\}. \quad (40)\end{aligned}$$

Square each right hand side term in (40) and apply the Schwarz inequality. For the first order term, we get an upper bound equal to

$$\sum_m\frac{1}{N_mC_m^{\text{mac}}}\Big\|\sum_i\tilde{a}_{\delta im}\{\theta,\delta_{0m}^{\text{mac}}(\theta)\}\Big\|^2\sum_m N_mC_m^{\text{mac}}\|\delta_m-\delta_{0m}^{\text{mac}}(\theta)\|^2. \quad (41)$$

Now,

[41] If one applies the mean value theorem to a vector-valued function then the 'mean value' can be different for each element of the vector. That distinction is immaterial here, so we ignore it in our notation.

$$\sup_{\theta\in\Theta_\epsilon,\delta\in\Delta}\sum_m N_m C_m^{\text{mac}}\frac{\|\delta_m-\delta_{0m}^{\text{mac}}(\theta)\|^2}{\hat{\mathcal{R}}(\theta,\delta)}\leq$$
$$\sup_{\theta\in\Theta_\epsilon,\delta\in\Delta}\sum_m N_m C_m^{\text{mac}}\left(\frac{\|\delta_m-\hat{\delta}_m^{\text{mac}}(\theta)\|^2}{\hat{\mathcal{R}}(\theta,\delta)}+\frac{\|\hat{\delta}_m^{\text{mac}}(\theta)-\delta_{0m}^{\text{mac}}(\theta)\|^2}{\hat{\mathcal{R}}(\theta,\delta)}\right)\preceq 1. \quad (42)$$

Further, by lemma 47,

$$\sum_m\frac{1}{N_m C_m^{\text{mac}}}\sup_{\theta\in\Theta:\theta^z\neq 0}\mathbb{E}\left(\left\|\sum_i\frac{\tilde{a}_{\delta im}\{\theta,\delta_{0m}^{\text{mac}}(\theta)\}}{\|\theta^z\|}\right\|^2\,\middle|\,\mathcal{J}\right)\leq C^*\sum_m\frac{\chi_m}{C_m^{\text{mac}}},$$

for some $C^*<\infty$. Take the expectation conditional on the χ_m's and divide by $\hat{\mathcal{R}}(\theta,\delta)$ to obtain for some constant $C^*<\infty$ an upper bound equal to

$$\sup_{\theta\in\Theta_\epsilon,\delta\in\Delta}\frac{C^*\sum_m\chi_m\|\theta^z\|^2}{\hat{\mathcal{R}}(\theta,\delta)}=\sup_{\theta\in\Theta_\epsilon,\delta\in\Delta}\frac{C^*M\bar{\chi}\|\theta^z\|^2}{\hat{\mathcal{R}}(\theta,\delta)}\prec\sup_{\theta\in\Theta_\epsilon,\delta\in\Delta}\frac{N\chi\|\theta-\theta_0\|_\lambda^2}{\hat{\mathcal{R}}(\theta,\delta)}\preceq 1. \quad (43)$$

Now,

$$\sup_{\theta\in\Theta}\left|\sum_m\frac{1}{N_m C_m^{\text{mac}}}\left(\left\|\sum_i\tilde{a}_{\delta im}\{\theta,\delta_{0m}^{\text{mac}}(\theta)\}\right\|^2-\mathbb{E}\left(\left\|\sum_i\tilde{a}_{\delta im}\{\theta,\delta_{0m}^{\text{mac}}(\theta)\}\right\|^2\,\middle|\,\mathcal{J}\right)\right)\right|. \quad (44)$$

which can be expressed as

$$\max_{\theta\in\Theta}\left|\sum_m A_m(\theta)\right|\leq\max_{t=1,\ldots,T}\max_{\theta\in\Theta_t}\left|\sum_m\{A_m(\theta)-A_m(\theta_t)\}\right|+\max_{t=1,\ldots,T}\left|\sum_m A_m(\theta_t)\right|,$$

for a partition $\{\Theta_t\}$ of Θ consisting of a given T sets for which distances in each Θ_t are minimized and where θ_t is an arbitrary point of Θ_t. Now for any $c>0$,

$$\mathbb{P}\left(\max_{t=1,\ldots,T}\left|\sum_m A_m(\theta_t)\right|\geq cM\right)\leq\sum_{t=1}^T\frac{\sum_m\mathbb{E}(A_m^2(\theta_t)\mid\mathcal{J})}{c^2M^2}\prec 1. \quad (45)$$

That leaves us with

$$\max_{t=1,\ldots,T}\max_{\theta\in\Theta_t}\left|\sum_m\{A_m(\theta)-A_m(\theta_t)\}\right|\leq C_d T^{-1/d}\max_{\theta\in\Theta}\sum_m\|A_{\theta m}(\theta)\|\prec M, \quad (46)$$

where C_d is a constant only depending on d. From (45) and (46) it follows that (44) is $\prec M\leq\hat{\mathcal{R}}(\theta,\delta)$ for all θ,δ. Combining the rate for (44) with (42) and the rate in (43) establishes that (41) divided by $\hat{\mathcal{R}}^2(\theta,\delta)$ vanishes, uniformly in $\theta\in\Theta_\epsilon,\delta\in\Delta$.

The second order term in (40) is similar but easier. □

Lemma 24.

$$\sup_{\theta\in\Theta_\epsilon,\delta\in\Delta}\left|\frac{\hat{\mathcal{L}}^{\text{mic}}\{\theta,\delta_0^{\text{mac}}(\theta)\}-\mathcal{L}^{\text{mic}}\{\theta,\delta_0^{\text{mac}}(\theta)\}}{\hat{\mathcal{R}}(\theta,\delta)}\right|\preceq 1.$$

Proof. This is a simple albeit messy application of lemma 51, noting that $a_{\psi^\nu ijm}(0, \psi^\nu_m) = 0$ for all ψ^ν_m and that the moment restrictions are satisfied by lemma 47. □

F.2.3 Lower bound on rate.

Lemma 25. $\hat{\theta} - \theta_0 \preceq 1/\sqrt{M}$

Proof. The proof is identical to the proof of lemma 9, but needs one additional step. Indeed, if one allows ϵ to vary then the rate inequality in (37) becomes $(N\chi\lambda^2 + M)\epsilon^2 \succ M\sqrt{\bar{\chi}}\lambda + 1$, for 'signal' to dominate 'noise.' Hence, the convergence rate of $\hat{\theta}$ is $\sqrt[4]{M^2\bar{\chi}\lambda^2 + 1} \;/\; \sqrt{N\chi\lambda^2 + M}$. First consider the possibility that $N\chi\lambda^2 \preceq M$. This implies that

$$M\sqrt{\bar{\chi}}\lambda \preceq M\sqrt{\frac{\bar{\chi}}{N\chi/M}} = M\sqrt{\frac{\sum_m \chi_m}{\sum_m N_m\chi_m}} \simeq M\sqrt{\frac{\mathbb{E}\chi_m}{\mathbb{E}(N_m\chi_m)}} \preceq 1,$$

by assumption D. Finally, if $N\chi\lambda^2 \succ M$ then

$$\frac{M\sqrt{\bar{\chi}}\lambda}{N\chi\lambda^2} \prec \frac{M\sqrt{\bar{\chi}}}{N\chi\sqrt{M/(N\chi)}} = \sqrt{\frac{M\bar{\chi}}{N\chi}} \simeq \sqrt{\frac{\mathbb{E}\chi_m}{\mathbb{E}(N_m\chi_m)}} \preceq \frac{1}{M}.$$ □

F.3 Step one of asymptotic normality

Lemma 26. For any fixed vector v with $\|v\| = 1$, $v^\top(\hat{\psi} - \psi_0) \simeq -v^\top\Omega^{-1}_{\psi\psi}(\psi_0)\hat{\Omega}_\psi(\psi_0)$.

Proof. Follows from lemmas 27, 32 and 33. □

F.3.1 $\hat{\delta}$ as a process indexed by θ.

Let $\hat{\delta}^{\mathcal{L}}_m(\theta)$ denote the minimizer of $\hat{\mathcal{L}}_m(\theta, \delta)$ with respect to δ and let $\hat{\delta}^{\Omega} = \hat{\delta}$.

Lemma 27.

$$\max_{m=1,\ldots,M} \sup_{\theta\in\Theta^*_{\epsilon_M}} N_m\left\|\hat{\delta}_m(\theta) - \delta_{0m}(\theta) + \mathcal{L}^{-1}_{m\delta\delta}\{\theta, \delta_{0m}(\theta)\}\hat{\mathcal{L}}_{m\delta}\{\theta, \delta_{0m}(\theta)\}\right\| \preceq M.$$

Proof. By lemma 30 and the triangle inequality, it suffices to show that

$$\max_{m=1,\ldots,M} \sup_{\theta\in\Theta^*_{\epsilon_M}} N_m\|\hat{\delta}_m(\theta) - \hat{\delta}^{\mathcal{L}}_m(\theta)\| \preceq M.$$

First, since $\hat{\Omega}_\delta\{\theta, \hat{\delta}(\theta)\} = 0$ for all θ by definition, we have $\hat{\mathcal{L}}_{m\delta}\{\theta, \hat{\delta}_m(\theta)\} = -\mathcal{K}_m\mathcal{K}^\top\hat{\delta}(\theta) \preceq \sqrt{M}$, uniformly in m, θ by lemmas 29 and 31. Thus,

$$\sqrt{M} \succeq \max_{m=1,\ldots,M} \sup_{\theta\in\Theta^*_{\epsilon_M}} \left\|\hat{\mathcal{L}}_{m\delta}\{\theta, \hat{\delta}_m(\theta)\}\right\| = \max_{m=1,\ldots,M} \sup_{\theta\in\Theta^*_{\epsilon_M}} \left\|\sum_j N_m^{-1}\hat{\mathcal{L}}_{m\delta\delta_j}\{\theta, \tilde{\delta}_{m;j}(\theta)\}N_m\{\hat{\delta}_{jm}(\theta) - \hat{\delta}^{\mathcal{L}}_{jm}(\theta)\}\right\|,$$

where $\tilde{\delta}_{m;j}(\theta)$ lies between $\hat{\delta}_m(\theta)$ and $\hat{\delta}^{\mathcal{L}}_m(\theta)$ with only the j–th element different. The stated result then follows from the fact that (conditional on $\mathcal{J}$) $\hat{\mathcal{L}}_{m\delta\delta}/N_m$ is an i.i.d. average that is boundedly differentiable in δ_m, that $\mathcal{L}_{m\delta\delta}/N_m$ has eigenvalues bounded away from zero by lemma 28, and that the function $\mathcal{L}_{m\delta\delta}/N_m$ only depends on $J_m \leq \bar{J}$ a.s. by assumption D. □

Lemma 28. For some sequence $\{\epsilon_M\}$ for which $1/\sqrt{M} \prec \epsilon_M \prec 1$ and $\Theta^*_\epsilon = \{\theta \in \Theta : \|\theta - \theta_0\| \leq \epsilon\}$,

$$\mathbb{P}\Big(\exists m = 1, ..., M : \min_{\theta\in\Theta^*_{\epsilon_M}} \lambda_{\min}[\mathcal{L}_{\delta\delta m}\{\theta, \delta_{0m}(\theta)\}] \leq \lambda_{\min}\{\mathcal{L}_{\delta\delta m}(\psi_{0m})\}/2\Big) \prec 1.$$

Proof. Let $a_m(\theta) = \lambda_{\min}[\mathcal{L}_{\delta\delta m}\{\theta, \delta_{0m}(\theta)\}]/N_m$ and note that $\sum_{j=1}^{J_m}\|\mathcal{L}_{\delta\delta\theta_j m}\|$ is bounded by lemma 47. Thus, noting that $a_m(\theta_0) \geq \lambda_{\min}\{\mathcal{L}^{\text{mac}}_m(\psi_{0m})\}$, we have for some fixed $C^* \leq \infty$,

$$\mathbb{P}\Big(\exists m = 1, ..., M : \min_{\theta\in\Theta^*_{\epsilon_M}} a_m(\theta) \leq a_m(\theta_0)/2\Big) \leq \sum_{m=1}^M \mathbb{P}\Big(\min_{\theta\in\Theta^*_{\epsilon_M}} 2a_m(\theta) \leq a_m(\theta_0)\Big)$$
$$\leq \sum_{m=1}^M \mathbb{P}\Big(\min_{\theta\in\Theta^*_{\epsilon_M}} 2\{a_m(\theta) - a_m(\theta_0)\} \leq -a_m(\theta_0)\Big) \leq \sum_{m=1}^M \mathbb{P}\Big(C^*\epsilon_M \geq a_m(\theta_0)\Big) \leq \sum_{m=1}^M C^{*p}\epsilon_M^p \mathbb{E}a_m^{-p}(\theta_0),$$

for any $p > 0$ for which the expectation exists, where the last inequality follows from the Markov inequality. Choose $2 < p < p_\xi/4$. It then suffices to show that $\mathbb{E}a_m^{-p}(\theta_0) < \infty$. Now, by (52) and assumption C, for some fixed $C_2 < \infty$,

$$\mathbb{E}a_m^{-p}(\theta_0) \leq \mathbb{E}\left(\max_{j=1,\ldots,J_m} \int \vartheta_{jm}(z, \nu; \psi_{0m})\vartheta_{0m}(z, \nu; \psi_{0m})\, \mathrm{d}F(\nu)\, \mathrm{d}G(z)\right)^{-2p} \leq$$
$$C_2\mathbb{E}\left(\max_{j=1,\ldots,J_m} \frac{\{\sum_{t=0}^{J_m}\exp(\delta_{tm})\}^2}{\exp(\delta_{jm})}\right)^{-2p}$$
$$\leq C_2\mathbb{E}\sum_{j,t=0}^{J_m}\{\exp(4p\delta_{tm} - 2p\delta_{jm})\} \leq C_2^2\mathbb{E}\exp(4p|\delta_{tm}|)\mathbb{E}\exp(2p|\delta_{jm}|) < \infty,$$

by assumption C. □

Lemma 29. Let ϵ_M be as in lemma 28. For any $\epsilon^*_M \preceq \epsilon_M$, $\sup_{\theta\in\Theta^*_{\epsilon^*_M}}\|\mathcal{K}^\top\hat{\delta}^{\mathcal{L}}(\theta)\| \preceq 1 + \sqrt{M}\varepsilon^*_M$.

Proof. By the triangle inequality,

$$\|\mathcal{K}^\top\hat{\delta}^{\mathcal{L}}(\theta)\| \leq \|\mathcal{K}^\top\{\hat{\delta}^{\mathcal{L}}(\theta) - \delta_0(\theta)\}\| + \|\mathcal{K}^\top\{\delta_0(\theta) - \delta_0(\theta_0)\}\| + \|\mathcal{K}^\top\xi\|. \tag{47}$$

The first right hand side term in (47) is $\prec 1$ by lemma 30 and the last term is $\preceq 1$ since $\mathcal{K}\mathcal{K}^\top$ is an orthogonal projection matrix of the form $B \times$ something $\times B^\top$. Finally, note that the middle term squared is bounded above by $\sum_{m=1}^M \|\partial_\theta \delta_{0m}^\top \partial_{\theta^\top} \delta_{0m}\| \, \|\theta - \theta_0\|^2$. □

Lemma 30. Let ϵ_M be as in lemma 28. Then

$$\max_{m=1,\ldots,M} \sup_{\theta \in \Theta^*_{\epsilon_M}} N_m \left\| \hat{\delta}_m^{\mathcal{L}}(\theta) - \delta_{0m}(\theta) + \mathcal{L}_{m\delta\delta}^{-1}\{\theta, \delta_{0m}(\theta)\} \hat{\mathcal{L}}_{m\delta}\{\theta, \delta_{0m}(\theta)\} \right\| \preceq M.$$

Proof. We first obtain results for fixed m and use lemma 53 with $\hat{f} = \hat{\mathcal{L}}_m / N_m$ and $f = \mathcal{L}_m / N_m$. Since $\hat{\mathcal{L}}_m$ is convex in δ_m on $\Theta^*_{\epsilon_M}$ with probability approaching one, *(i)* is satisfied. Because $\hat{\mathcal{L}}_m / N_m$ is an i.i.d. mean of convex differentiable functions, the remaining requirement of lemma 53 are straightforward to verify for $\rho_n = \rho_{2n} = N_m^{-1/2}$ and $\rho_{3n} = 1$. Finally, note that for any random sequence $\{A_m\}$ and any $\varepsilon > 0$, $\mathbb{P}(\max_{m=1}^M A_m > \varepsilon) \leq \sum_{m=1}^M \mathbb{P}(A_m > \varepsilon)$. Since there are only finitely different types of markets ($J_m \leq \bar{J}$ by assumption D), uniformity of $\mathbb{P}(A_m > \varepsilon)$ over m can be obtained by a finite sum over all possible values of J_m. □

Lemma 31. $\forall \theta \in \Theta : \|\mathcal{K}^\top \hat{\delta}(\theta)\| \leq \|\mathcal{K}^\top \hat{\delta}^{\mathcal{L}}(\theta)\|$.

Proof. We have

$$\frac{1}{2}(\|\mathcal{K}^\top \hat{\delta}(\theta)\|^2 - \|\mathcal{K}^\top \hat{\delta}^{\mathcal{L}}(\theta)\|^2) = [\hat{\Omega}\{\theta, \hat{\delta}(\theta)\} - \hat{\Omega}\{\theta, \hat{\delta}^{\mathcal{L}}(\theta)\}] + [\hat{\mathcal{L}}\{\theta, \hat{\delta}^{\mathcal{L}}(\theta)\} - \hat{\mathcal{L}}\{\theta, \hat{\delta}(\theta)\}] \leq 0,$$

because $\hat{\delta}$ minimizes $\hat{\Omega}$ and $\hat{\delta}^{\mathcal{L}}$ minimizes $\hat{\mathcal{L}}$. □

F.3.2 $\hat{\theta}^z$ as a process indexed by θ^ν. Define $\hat{\theta}^z(\theta^\nu) = \arg\min_{\theta^z} \hat{\Omega}\{\theta^z, \theta^\nu, \hat{\delta}^\Omega(\theta^z, \theta^\nu)\}$, let $\rho_z = \sqrt{N\chi + M}$, $\rho_\nu = \sqrt{N\chi\lambda^2 + M}$, and for $a, b \in \{\theta^z, \theta^\nu\}$ define $\mathcal{Q}_{ab} = \Omega_{ab} - \Omega_{a\delta}\Omega_{\delta\delta}^{-1}\Omega_{\delta b}$ and $\hat{q}_a = \hat{\Omega}_a - \Omega_{a\delta}\Omega_{\delta\delta}^{-1}\hat{\Omega}_\delta$.

Lemma 32.

$$\sup_{\theta^\nu \in \{\theta^\nu : \exists \theta^z : (\theta^z, \theta^\nu) \in \Theta^*_{\epsilon_M}\}} \rho_z^2 \left\| \hat{\theta}^z(\theta^\nu) - \theta_0^z(\theta^\nu) + \mathcal{Q}_{zz}^{-1}(\theta^\nu) \hat{q}_z(\theta^\nu) \right\| \preceq 1,$$

where $\mathcal{Q}_{zz}(\theta^\nu) = \mathcal{Q}_{zz}[\theta_0^z(\theta^\nu), \theta^\nu, \delta_0^\Omega\{\theta_0^z(\theta^\nu), \theta^\nu\}]$ and likewise for $\hat{q}_z(\theta^\nu)$.

Proof. The proof is similar to that of lemma 30 except that we need not establish uniformity in m. So we omit a proof and only note that $\mathcal{Q}_{zz}$ is the Hessian of $\Omega\{\theta^z, \theta^\nu, \delta_0^\Omega(\theta^z, \theta^\nu)\}$ with respect to θ^z because $\partial_{\theta^\top} \delta_0^\Omega = -\Omega_{\delta\delta}^{-1}\Omega_{\delta\theta}$ by the implicit function theorem and that $\hat{q}_z$ is the gradient of $\hat{\Omega}\{\theta^z, \theta^\nu, \delta_0^\Omega(\theta^z, \theta^\nu)\}$ with respect to θ^z. □

F.3.3 Approximation for $\hat{\theta}^\nu - \theta_0^\nu$. Let

$$\mathcal{Z} = -\Omega_{\psi\psi}^{-1} = \begin{bmatrix} \mathcal{Z}_{zz} & \mathcal{Z}_{z\nu} & \mathcal{Z}_{z\delta} \\ \mathcal{Z}_{\nu z} & \mathcal{Z}_{\nu\nu} & \mathcal{Z}_{\nu\delta} \\ \mathcal{Z}_{\delta z} & \mathcal{Z}_{\delta\nu} & \mathcal{Z}_{\delta\delta} \end{bmatrix}, \tag{48}$$

with

$$\begin{aligned}
\mathcal{Z}_{z\psi} &= -(\mathcal{Q}_{zz} - \mathcal{Q}_{z\nu}\mathcal{Q}_{\nu\nu}^{-1}\mathcal{Q}_{\nu z})^{-1}\left[I \quad -\mathcal{Q}_{z\nu}\mathcal{Q}_{\nu\nu}^{-1} \quad -(\Omega_{z\delta} - \mathcal{Q}_{z\nu}\mathcal{Q}_{\nu\nu}^{-1}\Omega_{\nu\delta})\Omega_{\delta\delta}^{-1}\right], \\
\mathcal{Z}_{\nu\psi} &= -(\mathcal{Q}_{\nu\nu} - \mathcal{Q}_{\nu z}\mathcal{Q}_{zz}^{-1}\mathcal{Q}_{z\nu})^{-1}\left[-\mathcal{Q}_{\nu z}\mathcal{Q}_{zz}^{-1} \quad I \quad -(\Omega_{\nu\delta} - \mathcal{Q}_{\nu z}\mathcal{Q}_{zz}^{-1}\Omega_{z\delta})\Omega_{\delta\delta}^{-1}\right], \\
\mathcal{Z}_{\delta\psi} &= -\Omega_{\delta\delta}^{-1}\left(\Omega_{\delta\theta}\mathcal{Z}_{\theta\psi} + \begin{bmatrix} 0 & 0 & I \end{bmatrix}\right),
\end{aligned}$$

where everything is evaluated at ψ_0. Let further $\rho_z = \sqrt{N\chi + M}$, $\rho_\nu = \sqrt{N\chi\lambda^2 + M}$, $\rho_m = \min(\rho_\nu, \sqrt{N_m})$, and let $\tilde{\mathcal{Z}} = \mathfrak{q}_\psi \mathcal{Z}$ with

$$\mathfrak{q}_\psi = \begin{bmatrix} I_{d_z}/\rho_z & 0 & 0 & \cdots & 0 \\ 0 & I_{d_\nu}/\rho_\nu & \ddots & \ddots & 0 \\ \vdots & \ddots & I_{J_1}/\rho_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & I_{J_M}/\rho_M \end{bmatrix}, \qquad \mathfrak{q}_\theta = \begin{bmatrix} I_{d_z}/\rho_z & 0 \\ 0 & I_{d_\nu}/\rho_\nu \end{bmatrix}.$$

Write $\tilde{\mathcal{Q}}_{\theta\theta} = \mathfrak{q}_\theta \mathcal{Q}_{\theta\theta}$ and let similar symbols be analogously defined.

Lemma 33. $\rho_\nu \succeq \sqrt{M}$ by lemma 25 and $\hat{\theta}^\nu - \theta_0^\nu - \mathcal{Z}_{\nu\psi}\hat{\Omega}_\psi(\psi_0) \simeq \rho_\nu^{-2}$.

Proof. We first show that

$$\hat{\theta}^\nu - \theta_0^\nu + (\mathcal{Q}_{\nu\nu} - \mathcal{Q}_{\nu z}\mathcal{Q}_{zz}^{-1}\mathcal{Q}_{z\nu})^{-1}\left(\hat{\Omega}_\nu + \mathcal{Q}_{\nu z}\{\hat{\theta}^z(\theta_0^\nu) - \theta_0^z(\theta_0^\nu)\} + \Omega_{\nu\delta}\{\hat{\delta}^\Omega(\theta_0) - \delta_0^\Omega(\theta_0)\}\right) \simeq \rho_\nu^{-2}, \tag{49}$$

where all Ω's and $\mathcal{Q}$'s are evaluated at the truth. The proof of (49) is analogous to that of lemmas 30 and 32 except that there is no uniformity issue here. Since this proof is simpler we omit it, except to note that it is based on the expansion

$$\begin{aligned}
0 = \hat{\Omega}_\nu[\hat{\theta}^\nu, \hat{\theta}^z(\hat{\theta}^\nu), \hat{\delta}^\Omega\{\hat{\theta}^z(\hat{\theta}^\nu), \hat{\theta}^\nu\}] \simeq \\
\hat{\Omega}_\nu + \underbrace{\left(\Omega_{\nu\nu} + \Omega_{\nu z}\partial_{\theta^{\nu\top}}\theta_0^z(\theta_0^\nu) + \Omega_{\nu\delta}\{\partial_{\theta^{\nu\top}}\delta_0^\Omega(\theta_0) + \partial_{\theta^{z\top}}\delta_0(\theta_0)\partial_{\theta^{\nu\top}}\theta_0^z(\theta_0^\nu)\}\right)}_{\mathcal{Q}_{\nu\nu} - \mathcal{Q}_{\nu z}\mathcal{Q}_{zz}^{-1}\mathcal{Q}_{z\nu}}(\hat{\theta}^\nu - \theta_0^\nu) \\
+ \underbrace{(\Omega_{\nu z} + \Omega_{\nu\delta}\partial_{\theta^{z\top}}\delta_0^\Omega(\theta_0))}_{\mathcal{Q}_{\nu z}}\{\hat{\theta}^z(\theta_0^\nu) - \theta_0^z(\theta_0^\nu)\} + \Omega_{\nu\delta}\{\hat{\delta}^\Omega(\theta_0) - \delta_0^\Omega(\theta_0)\},
\end{aligned}$$

where all right hand side $\Omega, \hat{\Omega}$s are evaluated at the truth. Note that $\partial_{\theta^\top}\delta_0^\Omega = -\Omega_{\delta\delta}^{-1}\Omega_{\delta\theta}$ and $\partial_{\theta^{\nu\top}}\theta_0^z = -\mathcal{Q}_{zz}^{-1}\mathcal{Q}_{z\nu}$ by applying the implicit function theorem to the first order conditions that define $\delta_0^\Omega(\theta)$ and $\theta_0^z(\theta^\nu)$. Rearrange to obtain (49). The lemma statement then follows from applying lemmas 27 and 32 and the delta method. □

F.4 Step two of asymptotic normality

Define $\mathcal{P} = \mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X} = \mathcal{K}\mathcal{K}^\top$ and $\bar{\mathcal{P}} = \mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B \bar{X}} = \bar{\mathcal{K}}\bar{\mathcal{K}}^\top$, where $\bar{X} = \mathbb{E}(X \mid \mathcal{B})$,

$$\bar{\mathcal{Q}}_{\theta\theta} = \mathbb{E}(\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1}\mathcal{L}_{\delta\theta} \mid \mathcal{B}) + \mathbb{E}(\mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1} \mid \mathcal{B})P_B\bar{\mathcal{K}}\bar{\mathcal{K}}^\top P_B\mathbb{E}(\mathcal{L}_{\delta\delta}^{-1}\mathcal{L}_{\delta\theta} \mid \mathcal{B}),$$

let $\bar{\mathcal{Z}}$ be defined as $\mathcal{Z}$ but with the $\mathcal{Q}$'s replaced with $\bar{\mathcal{Q}}$'s and let $\bar{\bar{\mathcal{Z}}}$ be as $\bar{\mathcal{Z}}$ with $\Omega_{\theta\delta}\Omega_{\delta\delta}^{-1}$ replaced with $\mathbb{E}(\mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1} \mid \mathcal{B})$. Then define

$$A = \bar{\mathcal{Z}}^\top\Lambda^\top\bar{\mathcal{D}}_\Lambda^{-1/2}, \quad C = B^+\bar{\mathcal{K}}\bar{\mathcal{K}}^\top\bar{\bar{\mathcal{Z}}}_{\psi\delta}^\top\Lambda^\top\bar{\mathcal{D}}_\Lambda^{-1/2}, \quad \bar{\mathcal{D}}_\Lambda = \Lambda\{\mathbb{E}(\bar{\mathcal{Z}}\mathcal{L}_{\psi\psi}\bar{\mathcal{Z}}^\top \mid \mathcal{B}) + \bar{\bar{\mathcal{Z}}}_{\psi\delta}\bar{K}\bar{K}^\top\bar{\bar{\mathcal{Z}}}_{\psi\delta}^\top\}\Lambda^\top.$$

Lemma 34. $\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda\Omega_{\psi\psi}^{-1}(\psi_0)\hat{\Omega}_\psi(\psi_0) \xrightarrow{d} N(0, I)$.

Proof. Follows from lemmas 35 and 36. □

F.4.1 Normality.

Lemma 35. $\bar{\mathcal{D}}_\Lambda^{-1/2}\Lambda(\bar{\mathcal{Z}}\hat{\mathcal{L}}_\psi + \bar{\bar{\mathcal{Z}}}_{\psi\delta}\bar{\mathcal{K}}\bar{\mathcal{K}}^\top\xi) \xrightarrow{d} N(0, 1)$.

Proof. Use lemma 49 with $A = \bar{\mathcal{Z}}^\top\Lambda^\top\bar{\mathcal{D}}_\Lambda^{-1/2}$ and $C = B^+\bar{\mathcal{K}}\bar{\mathcal{K}}^\top\bar{\bar{\mathcal{Z}}}_{\psi\delta}^\top\Lambda^\top\bar{\mathcal{D}}_\Lambda^{-1/2}$. □

Lemma 36. $\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda\Omega_{\psi\psi}^{-1}(\psi_0)\hat{\Omega}_\psi(\psi_0) - \bar{\mathcal{D}}_\Lambda^{-1/2}\Lambda(\bar{\mathcal{Z}}\hat{\mathcal{L}}_\psi + \bar{\bar{\mathcal{Z}}}_{\psi\delta}\bar{\mathcal{K}}\bar{\mathcal{K}}^\top\xi) \prec 1$.

Proof. Follows from lemmas 38, 39 and 43. □

Lemma 37. $\bar{\mathcal{D}}_\Lambda^{-1/2}\Lambda(\mathcal{Z} - \bar{\mathcal{Z}})\hat{\mathcal{L}}_\psi \prec 1$.

Proof. By lemma 40, part (e),

$$\mathbb{E}\{\|\bar{\mathcal{D}}_\Lambda^{-1/2}\Lambda(\mathcal{Z} - \bar{\mathcal{Z}})\hat{\mathcal{L}}_\psi\|^2 \mid \mathcal{J}\} = \operatorname{tr}(\bar{\mathcal{D}}_\Lambda^{-1}\Lambda(\mathcal{Z} - \bar{\mathcal{Z}})\mathcal{L}_{\psi\psi}(\mathcal{Z} - \bar{\mathcal{Z}})^\top\Lambda^\top) \leq$$
$$\operatorname{tr}\left\{\bar{\mathcal{D}}_\Lambda^{-1}\Lambda\begin{bmatrix} I & 0 \\ -\Omega_{\delta\delta}^{-1}\Omega_{\delta\theta} & 0 \end{bmatrix}\begin{bmatrix} \bar{\mathcal{Z}}_{\theta\psi}\Omega_{\psi\theta} + I & 0 \\ 0 & 0 \end{bmatrix}\Omega_{\psi\psi}^{-1}\begin{bmatrix} \bar{\mathcal{Z}}_{\theta\psi}\Omega_{\psi\theta} + I & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} I & 0 \\ -\Omega_{\delta\delta}^{-1}\Omega_{\delta\theta} & 0 \end{bmatrix}^\top\Lambda^\top\right\} \prec 1. \quad \square$$

Lemma 38. $\bar{\mathcal{D}}_\Lambda^{-1/2}\Lambda(\mathcal{Z}_{\psi\delta}\mathcal{K}\mathcal{K}^\top - \bar{\bar{\mathcal{Z}}}_{\psi\delta}\bar{\mathcal{K}}\bar{\mathcal{K}}^\top)\xi \prec 1$.

Proof. There are three components: (a) $\bar{\mathcal{D}}_\Lambda^{-1/2}\Lambda(\mathcal{Z}_{\psi\delta} - \bar{\mathcal{Z}}_{\psi\delta})\mathcal{K}\mathcal{K}^\top\xi \prec 1$; (b) $\bar{\mathcal{D}}_\Lambda^{-1/2}\Lambda(\bar{\mathcal{Z}}_{\psi\delta} - \bar{\bar{\mathcal{Z}}}_{\psi\delta})\mathcal{K}\mathcal{K}^\top\xi \prec 1$; (c) $\bar{\mathcal{D}}_\Lambda^{-1/2}\Lambda\bar{\bar{\mathcal{Z}}}_{\psi\delta}(\mathcal{K}\mathcal{K}^\top - \bar{\mathcal{K}}\bar{\mathcal{K}}^\top)\xi \prec 1$. We omit showing (a) since the proof is similar to that of lemma 37. For (b), note that $\mathcal{K}\mathcal{K}^\top = P_B\mathcal{K}\mathcal{K}^\top P_B$, $\bar{\mathcal{D}}_\Lambda^{-1/2} \preceq 1/\sqrt{M}$, $\Lambda(\bar{\mathcal{Z}}_{\psi\delta} - \bar{\bar{\mathcal{Z}}}_{\psi\delta})B \preceq \sqrt{M}$ (by lemma 43), $B^+\mathcal{K}\mathcal{K}^\top B^{+\top} \leq (B^\top B)^{-1} \simeq 1/M$, and $B^\top\xi \simeq \sqrt{M}$, such that the left hand side in result (b) is $\preceq M^{-1/2}M^{1/2}M^{-1}M^{1/2} \simeq M^{-1/2} \prec 1$. The arguments for (c) are similar except that now we use

$$B^+(\mathcal{K}\mathcal{K}^\top - \bar{\mathcal{K}}\bar{\mathcal{K}}^\top)(B^+)^\top =$$
$$(B^\top B)^{-1}\Big(B^\top\bar{X}\{\bar{X}^\top B(B^\top B)^{-1}B^\top\bar{X}\}^{-1}\bar{X}^\top B - B^\top X\{X^\top B(B^\top B)^{-1}B^\top X\}^{-1}X^\top B\Big)(B^\top B)^{-1} \preceq M^{-3/2}. \quad \square$$

Lemma 39. $\hat{\mathcal{D}}_\Lambda^{-1/2}\Lambda\Omega_{\psi\psi}^{-1}(\psi_0)\hat{\Omega}_\psi(\psi_0) - \bar{\mathcal{D}}_\Lambda^{-1/2}\Lambda\Lambda\Omega_{\psi\psi}^{-1}(\psi_0)\hat{\Omega}_\psi(\psi_0) \prec 1.$

Proof. Follows from lemma 44. $\square$

F.4.2 Removing Endogeneity from $\mathcal{Q}$.

Lemma 40. (a) $\bar{\mathcal{Z}}_{\theta\psi}\Omega_{\psi\theta}+I \prec 1$; (b) $\bar{\mathcal{Z}}_{\theta\psi}\Omega_{\psi\delta} = 0$; (c) $\bar{\mathcal{Z}}_{\delta\psi}\Omega_{\psi\theta} = -\Omega_{\delta\delta}^{-1}\Omega_{\delta\theta}(\bar{\mathcal{Z}}_{\theta\psi}\Omega_{\psi\theta}+I)$; (d) $\bar{\mathcal{Z}}_{\delta\psi}\Omega_{\psi\delta}+I = 0$; (e)

$$I + \bar{\mathcal{Z}}\Omega_{\psi\psi} = \begin{bmatrix} I & 0 \\ -\Omega_{\delta\delta}^{-1}\Omega_{\delta\theta} & 0 \end{bmatrix}\begin{bmatrix} \bar{\mathcal{Z}}_{\theta\psi}\Omega_{\psi\theta}+I & 0 \\ 0 & 0 \end{bmatrix}.$$

Proof. Tedious linear algebra shows that the left hand side in (a) is

$$I - \begin{bmatrix} (\bar{\mathcal{Q}}_{zz} - \bar{\mathcal{Q}}_{z\nu}\bar{\mathcal{Q}}_{\nu\nu}^{-1}\bar{\mathcal{Q}}_{\nu z})^{-1} & 0 \\ 0 & (\bar{\mathcal{Q}}_{\nu\nu} - \bar{\mathcal{Q}}_{\nu z}\bar{\mathcal{Q}}_{zz}^{-1}\bar{\mathcal{Q}}_{z\nu})^{-1} \end{bmatrix} \times \begin{bmatrix} \mathcal{Q}_{zz} - \bar{\mathcal{Q}}_{z\nu}\bar{\mathcal{Q}}_{\nu\nu}^{-1}\mathcal{Q}_{\nu z} & \mathcal{Q}_{z\nu} - \bar{\mathcal{Q}}_{z\nu}\bar{\mathcal{Q}}_{\nu\nu}^{-1}\mathcal{Q}_{\nu\nu} \\ \mathcal{Q}_{\nu z} - \bar{\mathcal{Q}}_{\nu z}\bar{\mathcal{Q}}_{zz}^{-1}\mathcal{Q}_{zz} & \mathcal{Q}_{\nu\nu} - \bar{\mathcal{Q}}_{\nu z}\bar{\mathcal{Q}}_{zz}^{-1}\mathcal{Q}_{z\nu} \end{bmatrix}.$$

Apply lemma 41. Showing results (b) to (d) then just entails multiplying out the matrices and result (e) reformulating. $\square$

Lemma 41. (a) $\tilde{\mathcal{Q}}_{\theta\theta} \simeq 1$; (b) $\tilde{\mathcal{Q}}_{\theta\theta} - \tilde{\bar{\mathcal{Q}}}_{\theta\theta} \prec 1$.

Proof. First (a):

$$\mathcal{Q}_{\theta\theta} = \mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1}\mathcal{L}_{\delta\theta} + \mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1}\mathcal{K}(I + \mathcal{K}^\top\mathcal{L}_{\delta\delta}^{-1}\mathcal{K})^{-1}\mathcal{K}^\top\mathcal{L}_{\delta\delta}^{-1}\mathcal{L}_{\delta\theta} \simeq$$
$$(\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1}\mathcal{L}_{\delta\theta}) + \mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1}\Pi_{\delta\delta}\mathcal{L}_{\delta\delta}^{-1}\mathcal{L}_{\delta\theta}, \quad (50)$$

by lemma 42. The second right hand side term in (50) is $\simeq M$ by condition (f) of assumption F. Now, the first right hand side term in (50). Note that $\mathcal{L}_{\psi\psi} \geq \mathcal{L}_{\psi\psi}^{\text{mic}}$ and hence $\mathcal{L}_{\psi\psi}^{-1} \leq \mathcal{L}_{\psi\psi}^{\text{mic}-1}$, which in turn implies (using partitioned matrices) that

$$\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1}\mathcal{L}_{\delta\theta} = \left(\begin{bmatrix} I & 0 \end{bmatrix}\mathcal{L}_{\psi\psi}^{-1}\begin{bmatrix} I \\ 0 \end{bmatrix}\right)^{-1} \geq \left(\begin{bmatrix} I & 0 \end{bmatrix}\mathcal{L}_{\psi\psi}^{\text{mic}-1}\begin{bmatrix} I \\ 0 \end{bmatrix}\right)^{-1} = \mathcal{L}_{\theta\theta}^{\text{mic}} - \mathcal{L}_{\theta\delta}^{\text{mic}}\mathcal{L}_{\delta\delta}^{\text{mic}-1}\mathcal{L}_{\delta\theta}^{\text{mic}}.$$

Thus, $\tilde{\mathcal{Q}}_{\theta\theta} \geq \mathfrak{q}_\theta(\mathcal{L}_{\theta\theta}^{\text{mic}} - \mathcal{L}_{\theta\delta}^{\text{mic}}\mathcal{L}_{\delta\delta}^{\text{mic}-1}\mathcal{L}_{\delta\theta}^{\text{mic}} + MI)\mathfrak{q}_\theta \succeq 1$, by assumption G. So we have shown that $\tilde{\mathcal{Q}}_{\theta\theta} \succeq 1$. We now show that it is $\preceq 1$, also, for which it remains to be shown that $\mathfrak{q}_\theta(\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1}\mathcal{L}_{\delta\theta})\mathfrak{q}_\theta \preceq 1$. Since $\mathcal{L}_{\theta\theta}^{\text{mac}} - \mathcal{L}_{\theta\delta}^{\text{mac}}\mathcal{L}_{\delta\delta}^{\text{mac}-1}\mathcal{L}_{\delta\theta}^{\text{mac}} = 0$, we have

$$\mathfrak{q}_\theta(\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1}\mathcal{L}_{\delta\theta})\mathfrak{q}_\theta = \mathfrak{q}_\theta(\mathcal{L}_{\theta\theta}^{\text{mic}} - \mathcal{L}_{\theta\delta}^{\text{mic}}\mathcal{L}_{\delta\delta}^{-1}\mathcal{L}_{\delta\theta})\mathfrak{q}_\theta$$

$$+ \mathfrak{q}_\theta \mathcal{L}_{\theta\delta}^{\text{mac}} \mathcal{L}_{\delta\delta}^{\text{mac}-1} \mathcal{L}_{\delta\delta}^{\text{mic}} \mathcal{L}_{\delta\delta}^{-1} \mathcal{L}_{\delta\theta} \mathfrak{q}_\theta - \mathfrak{q}_\theta \mathcal{L}_{\theta\delta}^{\text{mac}} \mathcal{L}_{\delta\delta}^{\text{mac}} \mathcal{L}_{\delta\theta}^{\text{mic}} \mathfrak{q}_\theta$$
$$\leq \frac{1}{2} \mathfrak{q}_\theta \Bigg((\mathcal{L}_{\theta\theta}^{\text{mic}} + \mathcal{L}_{\theta\delta}^{\text{mic}} \partial_{\theta^\top} \delta + \partial_\theta \delta^\top \mathcal{L}_{\delta\theta} + \partial_\theta \delta^\top \mathcal{L}_{\delta\delta} \partial_{\theta^\top} \delta)$$
$$+ (\mathcal{L}_{\theta\theta}^{\text{mic}} + \mathcal{L}_{\theta\delta}^{\text{mic}} \partial_{\theta^\top} \delta^{\text{mac}} + \partial_\theta \delta^{\text{mac}^\top} \mathcal{L}_{\delta\theta} + \partial_\theta \delta^{\text{mac}^\top} \mathcal{L}_{\delta\delta} \partial_{\theta^\top} \delta^{\text{mac}}) \Bigg) \mathfrak{q}_\theta \leq \mathfrak{q}_\theta \mathcal{L}_{\theta\theta}^{\text{mic}} \mathfrak{q}_\theta, \preceq 1,$$

by assumption G, where the penultimate inequality follows from the theory of partitioned matrices.

Now result (b). First,

$$\mathfrak{q}_\theta \{ (\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\delta} \mathcal{L}_{\delta\delta}^{-1} \mathcal{L}_{\delta\theta}) - \overline{\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\delta} \mathcal{L}_{\delta\delta}^{-1} \mathcal{L}_{\delta\theta}} \} \mathfrak{q}_\theta$$
$$= \sum_{m=1}^{M} \mathfrak{q}_\theta \{ (\mathcal{L}_{m\theta\theta} - \mathcal{L}_{m\theta\delta} \mathcal{L}_{m\delta\delta}^{-1} \mathcal{L}_{m\delta\theta}) - \overline{\mathcal{L}_{m\theta\theta} - \mathcal{L}_{m\theta\delta} \mathcal{L}_{m\delta\delta}^{-1} \mathcal{L}_{m\delta\theta}} \} \mathfrak{q}_\theta \prec 1,$$

by a weak law of large numbers for triangular arrays, e.g. Davidson (1994, theorem 19.7). Now,

$$\mathfrak{q}_\theta (\mathcal{L}_{\theta\delta} \mathcal{L}_{\delta\delta}^{-1} \mathcal{K} \mathcal{K}^\top \mathcal{L}_{\delta\delta}^{-1} \mathcal{L}_{\delta\theta} - \overline{\mathcal{L}_{\theta\delta} \mathcal{L}_{\delta\delta}^{-1}} \bar{\mathcal{K}} \bar{\mathcal{K}}^\top \overline{\mathcal{L}_{\delta\delta}^{-1} \mathcal{L}_{\delta\theta}}) \mathfrak{q}_\theta,$$

because $\mathcal{K}\mathcal{K}^\top = P_B \mathcal{K}\mathcal{K}^\top P_B$, $P_B = BB^+$, $(\mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1} B - \overline{\mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1}} B) \prec M$ by a weak law of large numbers, $B^+ \mathcal{K}\mathcal{K}^\top B^{+\top} \preceq M^{-1}$, and

$$B^+ (\mathcal{K}\mathcal{K}^\top - \bar{\mathcal{K}}\bar{\mathcal{K}}^\top) B^{+\top} = (B^\top B)^{-1} \Big(B^\top \bar{X} (\bar{X}^\top P_B \bar{X})^{-1} \bar{X}^\top B - B^\top X (X^\top P_B X)^{-1} X^\top B \Big) (B^\top B)^{-1} \prec M^{-1},$$

since $B^\top (X - \bar{X})/M \prec 1$. □

Lemma 42. $\mathcal{K}^\top \mathcal{L}_{\delta\delta}^{-1} \mathcal{K} \prec 1$.

Proof. The trace of the left hand side is

$$\text{tr}(\mathcal{K}^\top \mathcal{P}_B \mathcal{L}_{\delta\delta}^{-1} \mathcal{P}_B \mathcal{K}) = \text{tr} \Big\{ (B^\top B)^{-1} \sum_{m=1}^{M} B_m^\top \mathcal{L}_{m\delta\delta}^{-1} B_m (B^\top B)^{-1} B^\top \mathcal{K}\mathcal{K}^\top B \Big\} \prec M^{-1} \times M \times M^{-1} \times M = 1.$$

□

Lemma 43. $\{\mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1} - \mathbb{E}(\mathcal{L}_{\theta\delta}\mathcal{L}_{\delta\delta}^{-1} \mid \mathcal{B})\} B \preceq \sqrt{M}$.

Proof. Take any linear combination, square, and take expectations to obtain the square of the promised rate. □

F.4.3 Estimated covariance matrix.

Let $\hat{f}_m$ be an element of $\hat{\mathcal{L}}_{m\theta\theta} - \hat{\mathcal{L}}_{m\theta\delta} \hat{\mathcal{L}}_{m\delta\delta}^{-1} \hat{\mathcal{L}}_{m\delta\theta}$.

Lemma 44. $\mathcal{D}_\Lambda^{-1} \hat{\mathcal{D}}_\Lambda - I \prec 1$.

Proof. If Λ selects only elements of θ then this follows from lemma 45. For δ's the proof is analogous. □

Lemma 45. $\hat{\tilde{Q}}_{\theta\theta} - \tilde{Q}_{\theta\theta} \prec 1$.

Proof. The proof follows the same steps as that of lemma 41 with the main difference that now $\hat{\mathcal{L}}^{\text{mac}}_{\theta\theta}\{\theta, \hat{\delta}^{\text{mac}}(\theta)\} - \hat{\mathcal{L}}^{\text{mac}}_{\theta\delta}\{\theta, \hat{\delta}^{\text{mac}}(\theta)\}\hat{\mathcal{L}}^{\text{mac}-1}_{\delta\delta}\{\theta, \hat{\delta}^{\text{mac}}(\theta)\}\hat{\mathcal{L}}^{\text{mac}}_{\delta\theta}\{\theta, \hat{\delta}^{\text{mac}}(\theta)\} = 0$ for all θ and noting that by lemma 46, $\mathfrak{q}_\theta\Big(\hat{\mathcal{L}}^{\text{mac}}_{\theta\theta}\{\hat{\theta}, \hat{\delta}(\hat{\theta})\} - \hat{\mathcal{L}}^{\text{mac}}_{\theta\delta}\{\hat{\theta}, \hat{\delta}(\hat{\theta})\}\hat{\mathcal{L}}^{\text{mac}-1}_{\delta\delta}\{\hat{\theta}, \hat{\delta}(\hat{\theta})\}\hat{\mathcal{L}}^{\text{mac}}_{\delta\theta}\{\hat{\theta}, \hat{\delta}(\hat{\theta})\}\Big)\mathfrak{q}_\theta \prec 1$. □

Lemma 46. $\sum_{m=1}^{M}[\hat{f}_m\{\hat{\theta}, \hat{\delta}_m(\hat{\theta})\} - \hat{f}_m\{\hat{\theta}, \hat{\delta}^{\text{mac}}_m(\hat{\theta})\}] \prec \rho_\nu^2$.

Proof. Using arguments similar to those in lemma 27, it can be shown that

$$\max_{m=1,\ldots,M} N_m \Bigg\| \hat{\delta}_m(\hat{\theta}) - \hat{\delta}^{\text{mac}}_m(\hat{\theta}) \\ - \mathcal{L}^{-1}_{m\delta\delta}\Big[\mathcal{L}^{\text{mic}}_{m\delta\delta}\mathcal{L}^{\text{mac}-1}_{m\delta\delta}\{\hat{\mathcal{L}}^{\text{mac}}_{m\delta} + \mathcal{L}^{\text{mac}}_{m\delta\theta}(\hat{\theta} - \theta_0)\} - \{\hat{\mathcal{L}}^{\text{mic}}_{m\delta} + \mathcal{L}^{\text{mic}}_{m\delta\theta}(\hat{\theta} - \theta_0)\}\Big]\Bigg\| \preceq 1,$$

Then,

$$\sum_{m=1}^{M}[\hat{f}_m\{\hat{\theta}, \hat{\delta}_m(\hat{\theta})\} - \hat{f}_m\{\hat{\theta}, \hat{\delta}^{\text{mac}}_m(\hat{\theta})\}] \\ \simeq \sum_{m=1}^{M} \hat{f}_{m\delta}\mathcal{L}^{-1}_{m\delta\delta}\Big[\mathcal{L}^{\text{mic}}_{m\delta\delta}\mathcal{L}^{\text{mac}-1}_{m\delta\delta}\{\hat{\mathcal{L}}^{\text{mac}}_{m\delta} + \mathcal{L}^{\text{mac}}_{m\delta\theta}(\hat{\theta} - \theta_0)\} - \{\hat{\mathcal{L}}^{\text{mic}}_{m\delta} + \mathcal{L}^{\text{mic}}_{m\delta\theta}(\hat{\theta} - \theta_0)\}\Big] \\ \preceq \sqrt{\sum_m N_m\chi_m^2\lambda^4} + \frac{N\chi\lambda^2}{\rho_\nu} + \sqrt{\sum_m N_m\chi_m\lambda^2} + \sqrt{\sum_m N_m\chi_m\lambda^2} \prec \rho_\nu^2,$$

as asserted. □

F.5 Auxiliary results

Let X_m be the matrix with rows $x_{jm}^\top$ and define $\|X_m\|_X = \sum_{j=0}^{J_m}\|x_{jm}\|$.

Lemma 47. All elements of the ℓ-th partial derivative of $\log \pi_{jm}^{z_{im}}(\psi_m)$ with respect to ψ_m are for all $\ell \geq 1$ bounded in norm by $C_\ell\|X_m\|_X^{2\ell}\|z_{im}\|^\ell$, where C_ℓ is a constant independent of $z_{im}, x_{\cdot m}, \psi_m$.

Proof. We show the result for $\ell = 1$, where the result for higher order derivatives is a trivial extension. First, for any k,

$$\partial_{\delta_{km}} \log \pi_{jm}^{z_{im}}(\theta, \delta_m) = \mathbb{1}(j = k) - \frac{\int \mathfrak{s}_{jm}(z_{im}, \nu)\mathfrak{s}_{km}(z_{im}, \nu)\varphi(\nu)\,\mathrm{d}\nu}{\pi_{jm}^{z_{im}}},$$

which is bounded above in absolute value by 1, because $\mathfrak{s}_{km} \leq 1$.

Let $\maltese_{ijmk}$ denote the variable that multiplies θ_k^z in the numerator of $\mathfrak{s}_{jm}(z_{im}, \nu; \theta, \delta_m)$, typically the product of an element in z_{im} and an element in x_{jm}. Then

$$\partial_{\theta_k^z} \log \pi_{jm}^{z_{im}}(\theta, \delta_m) = \cancel{x}_{ijmk} - \sum_{t=0}^{J_m} \cancel{x}_{itmk} \frac{\int \imath_{jm}(z_{im}, \nu)\imath_{tm}(z_{im}, \nu)\varphi(\nu)\,\mathrm{d}\nu}{\pi_{jm}^{z_{im}}},$$

where the ratio is bounded above by one. Finally, assume without loss of generality that the regressor multiplying θ_k^ν is $x_{jmk}\nu_k$.Then,

$$\partial_{\theta_k^\nu} \log \pi_{jm}^{z_{im}}(\theta, \delta_m) = \frac{x_{jmk}\int \imath_{jm}(z_{im}, \nu)\nu_k\varphi(\nu)\,\mathrm{d}\nu}{\pi_{jm}^{z_{im}}} - \sum_{t=0}^{J_m} \frac{x_{tmk}\int \imath_{jm}(z_{im}, \nu)\imath_{tm}(z_{im}, \nu)\nu\varphi(\nu)\,\mathrm{d}\nu}{\pi_{jm}^{z_{im}}}. \tag{51}$$

Now, by integration by parts,

$$\frac{x_{jmk}\int \imath_{jm}(z_{im}, \nu)\nu_k\varphi(\nu)\,\mathrm{d}\nu}{\pi_{jm}^{z_{im}}} = \theta_k^\nu x_{jmk} - \theta_k^\nu \sum_{t=0}^{J_m} x_{tmk} \frac{\int \imath_{jm}(z_{im}, \nu)\imath_{tm}(z_{im}, \nu)\varphi(\nu)\,\mathrm{d}\nu}{\pi_{jm}^{z_{im}}},$$

where we can again use the fact that $\imath_{tm}$ is bounded above by 1 to achieve the desired bound. Repeat for the second right hand side term in (51). □

Lemma 48. For some constant $c > 0$ and all m,

$$\forall m : \forall \theta, \delta_m : \lambda_{\min}\left(\partial_{\delta_m^\top} \pi_m(\theta, \delta_m)\right) \geq c \min_{j=1,\ldots,J_m} \frac{\exp(\delta_{jm})}{\left\{1 + \sum_{t=1}^{J_m} \exp(\delta_{tm})\right\}^2}.$$

Proof. Let $\mathcal{S}_m = \mathrm{diag}(\imath_m)$. Then,

$$\partial_{\delta_m^\top} \pi_m(\theta, \delta_m) = \int (\mathcal{S}_m - \imath_m \imath_m^\top)\phi = \int \mathcal{S}_m^{1/2}(I - \mathcal{S}_m^{-1/2}\imath_m\imath_m^\top\mathcal{S}_m^{-1/2})\mathcal{S}_m^{1/2}\phi. \tag{52}$$

The smallest eigenvalue of $I - \mathcal{S}_m^{-1/2}\imath_m\imath_m^\top\mathcal{S}_m^{-1/2}$ is $1 - \imath_m^\top\mathcal{S}_m^{-1}\imath_m = \imath_{0m}$, such that the right hand side is bounded below by $\int \mathcal{S}_m\imath_{0m}\phi = \mathrm{diag}(\int \imath_m\imath_{0m}\phi)$, whose smallest eigenvalue is

$$\min_{j=1,\ldots,J_m} \int \imath_{jm}\imath_{0m}\phi \geq \min_{j=1,\ldots,J_m} \frac{\exp(\delta_{jm})}{\left\{1 + \sum_{t=1}^{J_m} \exp(\delta_{tm})\right\}^2} \int \imath_{jm}(z, \nu; \theta, 0)\imath_{0m}(z, \nu; \theta, 0)\phi(\nu)\,\mathrm{d}G(z).$$

The stated result then follows from assumption C. □

Lemma 49. Let $\{A_m\}$ be a sequence of matrices with a fixed number c of columns and for which $A_m \in \mathbb{R}^{d_{\psi m} \times c}$ be measurable with respect to $\mathcal{J}_m$. Define $A_m = [A_{\theta m}, A_{\delta m}]$ and

$$A = \begin{bmatrix} A_{\theta 1} & A_{\delta 1} & 0 & \ldots & 0 \\ A_{\theta 2} & 0 & A_{\delta 2} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ A_{\theta M} & 0 & \ldots & 0 & A_{\delta M} \end{bmatrix}.$$

Let further $C \in \mathbb{R}^{d_b \times c}$ be measurable with respect to $\mathcal{B}$. If for some $\varepsilon > 1$, (a) $\mathbb{E}(A^\top \mathcal{L}_{\psi\psi} A + C^\top B^\top BC) = I$; (b) $\sum_{m=1}^M \sum_{i=1}^M \mathbb{E}\|A_m^\top \hat{\mathcal{L}}_{\psi im}\|^{2\varepsilon} = o(1)$; (c) $\sum_{m=1}^M \sum_{j=1}^{J_m} \mathbb{E}\|C^\top b_{jm}\xi_{jm}\|^{2\varepsilon} = o(1)$, then $A^\top \hat{\mathcal{L}}_\psi + C^\top B^\top \xi \xrightarrow{d} N(0, I)$.

Proof. Let $v \in \mathbb{R}^c$ with $\|v\| = 1$. Then $\zeta_{im} = v^\top (A_m^\top \hat{\mathcal{L}}_{\psi im} + \mathbb{1}(i = 1) C^\top \sum_{j=1}^{J_m} b_{jm}\xi_{jm})$ is a martingale difference sequence if the observations are ordered by market and then by consumer, i.e. $(i, m) = (N_1, 1)$ precedes $(2, 2)$. By Davidson (1994, theorem 24.3), we need to show that (1) $\sum_{m=1}^M \sum_{i=1}^{N_m} (\zeta_{im}^2 - \mathbb{E}\zeta_{im}^2) = o_p(1)$ and (2) $\max_{m=1,\dots,M} \max_{i=1,\dots,N_m} |\zeta_{im}| = o_p(1)$. First (1). For a generic constant C^*, we have by the Burkholder and c_r inequalities that

$$\mathbb{E}\left|\sum_{m=1}^M \sum_{i=1}^{N_m} (\zeta_{im}^2 - \mathbb{E}\zeta_{im}^2)\right|^\varepsilon = \mathbb{E}\mathbb{E}\left\{\left|\sum_{m=1}^M \sum_{i=1}^{N_m} (\zeta_{im}^2 - \mathbb{E}\zeta_{im}^2)\right|^\varepsilon \middle| \mathcal{J}\right\} \leq C^* \sum_{m=1}^M \sum_{i=1}^{N_m} \mathbb{E}|\zeta_{im}^2 - \mathbb{E}\zeta_{im}^2|^\varepsilon,$$

which is $o(1)$ by conditions (b) and (c). Finally, (2) follows from the Markov inequality. □

Lemma 50. Lemma 49 still holds if condition (a) is instead $\mathbb{E}(A^\top \mathcal{L}_{\psi\psi} A + C^\top B^\top BC) - I \prec 1$.

Proof. If $\zeta \xrightarrow{d} N(0, I)$ and $U \to I$ then $U\zeta = (U - I)\zeta + \zeta \xrightarrow{d} N(0, I)$. □

F.6 Generic lemmas

Lemma 51. For a generic i.i.d. sample $\{x_i(\theta)\}$ of size n, let

$$S_n(\theta) = \begin{cases} \dfrac{1}{\sqrt{n}} \sum_{i=1}^n \dfrac{x_i(\theta) - x_i(\theta_0)}{\|\theta - \theta_0\|}, & \theta \neq \theta_0, \\ 0, & \theta = \theta_0. \end{cases}$$

Suppose that $\forall \theta \in \Theta : \mathbb{E}x_{\theta i}(\theta) = 0$ for compact Θ and $\mathbb{E}\sup_{\theta\in\Theta}\|x_{\theta\theta i}(\theta)\| < \infty$. Then $\sup_{\theta\in\Theta}|S_n(\theta)| \preceq 1$.

Proof. Apply the mean value theorem to obtain that for some $\theta^* \in \Theta$,

$$S_n(\theta) = \begin{cases} \left(\dfrac{1}{\sqrt{n}} \sum_{i=1}^n x_{\theta i}(\theta^*)\right)^\top \dfrac{\theta - \theta_0}{\|\theta - \theta_0\|}, & \theta \neq \theta_0, \\ 0, & \theta = \theta_0. \end{cases} \tag{53}$$

By example 19.7 of van der Vaart (2000), $\sum_{i=1}^n x_{\theta i}/\sqrt{n}$ converges weakly to a Gaussian process and hence $\sup_{\theta\in\Theta}\|x_{\theta i}(\theta)\| \preceq 1$. □

Lemma 52. If $x \sim E(\lambda)$ then $\mathbb{E}\exp(cx) = \lambda/(\lambda - c) < \infty$ for all $c < \lambda$.

Proof. A change of variables shows that the density of $y = \exp(cx)$ is $(\lambda/c)y^{-(1+\lambda/c)}\mathbb{1}(y \geq 1)$. □

Lemma 53. Suppose that

(i) for some functions $\hat{f}$, f and all α, $\hat{\beta}(\alpha)$ minimizes $\hat{f}(\alpha,\beta)$ and $\beta_0(\alpha)$ is the unique minimizer of $f(\alpha,\beta)$ where f is continuous, the parameter space of α,β is the Euclidean product of their respective compact parameter spaces, $\hat{f}$ converges to f in probability uniformly in α,β;

(ii) $\hat{\beta}(\alpha)$ is a solution to

$$0 = \hat{f}_\beta\{\alpha,\hat{\beta}(\alpha)\}, \tag{54}$$

and $\beta_0(\alpha)$ is a unique solution to $0 = f_\beta\{\alpha,\beta_0(\alpha)\}$;

(iii) for some $\rho_n \prec 1$, $\sup_\alpha \|f_{\beta\beta}^{-1}\{\alpha,\beta_0(\alpha)\}\hat{f}_\beta\{\alpha,\beta_0(\alpha)\}\| \preceq \rho_n$;

(iv) for some open neighborhood $\aleph(\alpha)$ of $\beta_0(\alpha)$, $\sup_{\alpha,\beta\in\aleph(\alpha)}(\lambda_{\max}\{f_{\beta\beta}(\alpha,\beta)\}/\lambda_{\min}\{f_{\beta\beta}(\alpha,\beta)\}) \preceq 1$;

(v) for some $\rho_{2n} \prec 1$, $\sup_\alpha \|f_{\beta\beta}^{-1}\{\alpha,\beta_0(\alpha)\}\hat{f}_{\beta\beta}\{\alpha,\beta_0(\alpha)\} - I\| \preceq \rho_{2n}$;

(vi) $\sup_{\alpha,\beta\in\aleph(\alpha)} \|f_{\beta\beta}^{-1}(\alpha,\beta)\hat{f}_{\beta\beta}(\alpha,\beta) - I\| \prec 1$;

(vii) for some $\rho_{3n} \prec \rho_n^{-1}$, the third partial derivatives of $\hat{f}$ with respect to β are $\preceq \rho_{3n}$ uniformly in α,β.

Then, $\sup_\alpha \|\hat{\beta}(\alpha) - \beta_0(\alpha) + f_{\beta\beta}^{-1}\{\alpha,\beta_0(\alpha)\}\hat{f}_\beta\{\alpha,\beta_0(\alpha)\}\| \preceq \rho_n(\rho_{2n} + \rho_n\rho_{3n})$.

Proof. We first show that $\hat{\beta}(\alpha) - \beta_0(\alpha)$ is $O_p(\rho_n)$, uniformly in α. Consistency, uniformly in α, follows from *(i)*. Applying the mean value theorem to (54) in *(ii)*, we have for some $\beta^*(\alpha)$ that

$$\sup_\alpha \|\hat{\beta}(\alpha) - \beta_0(\alpha)\| = \sup_\alpha \|\hat{f}_{\beta\beta}^{-1}\{\alpha,\beta^*(\alpha)\}\hat{f}_\beta\{\alpha,\beta_0(\alpha)\}\|.$$

The ρ_n rate then follows from *(iii)*, *(iv)* and *(vi)*.

Now, premultiply (54) by $f_{\beta\beta}^{-1}\{\alpha,\beta_0(\alpha)\}$ to obtain by the mean value theorem and triangle inequality that

$$\sup_\alpha \|\hat{\beta}(\alpha) - \beta_0(\alpha) + f_{\beta\beta}^{-1}\{\alpha,\beta_0(\alpha)\}\hat{f}_\beta\{\alpha,\beta_0(\alpha)\}\| \leq \sup_\alpha \|(f_{\beta\beta}^{-1}\{\alpha,\beta_0(\alpha)\}\hat{f}_{\beta\beta}\{\alpha,\beta_0(\alpha)\} - I)(\hat{\beta}(\alpha) - \beta_0(\alpha))\| + O_p(\rho_{3n}) \sup_\alpha \|\hat{\beta}(\alpha) - \beta_0(\alpha)\|^2,$$

by *(vii)*. Apply *(v)* and the ρ_n rate obtained above to obtain the stated result. □

G Monte Carlo Details

In this appendix, we provide additional details about the Monte Carlo specifications.

Mean product quality is specfied as $\delta_{jm} = \beta_c + \beta_1 x_{jm}^1 + \beta_2 x_{jm}^2 + \xi_{jm}$, where the true parameters for β are $(-6,1,1)$. These were chosen so that the share of the outside good was roughly 20 percent of the aggregate share, although this varies significantly from market to market. When exogenous, x_{jm} are distributed i.i.d. according to the standard normal distribution. The unobservable product characteristic ξ_{jm} is distributed Pareto(2.3). We choose the Pareto distribution because the resulting

shares mimic real world data where there are a few large-share products and many very small share products.[42]

Consumers have observable characteristics, $z_{im} = (z^1_{im}, z^2_{im})$ that are drawn independently from the standard normal distribution. Preference heterogeneity based on observable consumer characteristics is parameterized according to $\mu^{z_{im}}_{jm} = \theta^z_1 z^1_{im} x^1_{jm} + \theta^z_2 z^2_{im} x^2_{jm}$, where the true values of θ^z in the baseline specification are $(1, 1)$. Altering θ^z affects the strength of identification of θ^ν via the micro data by increasing the variation in utility across consumers.

Consumers have unobserved characteristics $\nu_{im} = (\nu^1_{im}, \nu^2_{im})$ which are both distributed $N(0, 1)$, and the unobserved heterogeneity term is $\mu^{\nu_{im}}_{jm} = \theta^\nu_1 \nu^1_{im} x^1_{jm} + \theta^\nu_2 \nu^2_{im} x^2_{jm}$, where the true parameters for θ^ν are $(1, 1)$ in the baseline.

For each specification, we draw data for 50 markets. Products in each market are independent of other markets. We vary the number of products in each market with five markets each of {10, 12, 14, 16, 18, 20, 22, 24, 26, 28}. There are $100,000$ consumers (N_m) in each market. For the consumer level data, we take a random sample of size S_m for the micro dataset. In the baseline case, $S_m = 1,000$. The micro data contains a consumer choice, $y_{i \cdot m}$ (a vector where $y_{ijm} = 1$ if consumer i chose product j and zero otherwise) together with their observable characteristics, z_{im}. In the baseline specification, average share is roughly 2.1%, and the tenth percentile of shares is roughly 0.06%.

In specifying Π for the CLER estimator, we include the following elements in the instrument vector b: a constant, product characteristics x_{jm}, differentiation IVs following Gandhi and Houde (2020), where the (j, m) element is $d_{jm} = \sum_{j' \in J_m j} (x^k_{jm} - x^k_{j'm})^2$, and the number of products in the market J_m. Since $d_b = 6 > d_\beta = 3$, Π is overidentified for β and the extra exclusion restrictions are potentially useful to identify θ. We include the same IVs in the alternative GMM comparison.

References cited in the appendices

BERRY, S. T. 1994. "Estimating discrete-choice models of product differentiation." *RAND Journal*, 242–262.

DAVIDSON, J. 1994. *Stochastic limit theory: An introduction for econometricians.* OUP Oxford.

GANDHI, A., AND J.-F. HOUDE. 2020. "Measuring Substitution Patterns in Differentiated-Products Industries." University of Pennsylvania and UW-Madison.

VAN DER VAART, A. 2000. *Asymptotic statistics.* Vol. 3, Cambridge university press.

[42] The Pareto distribution has thicker tails than allowed by assumption C, but our assumptions are not minimal. This choice also results in a bias which is visible for some simulations. Results for the case where ξ is distributed Normal, which do not exhibit this bias and satisfies assumption C, are available upon request.