Optimal Estimation of Discrete Choice Demand Models with Consumer and Product Data

Paul L. E. Grieco
Penn State

Charles Murry
Michigan

Joris Pinkse
Penn State

Stephan Sagl
Indiana

this version: January 17, 2025

We propose a conformant likelihood estimator with exogeneity restrictions (CLEER) 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

Demand models with endogenous prices using the discrete choice random utility framework provide a tractable framework to flexibly estimate substitution patterns between differentiated products (see e.g., Berry et al., 1995, BLP95). 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, which we term the Conformant Likelihood Estimator with Exogeneity Restrictions (CLEER). 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. It moreover recovers product quality terms as parameters of the model. We impose no additional assumptions over those posited on demand 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 Dube 2005). We show CLEER converges at the fastest possible rate, which depends on underlying data and identification strength, and always produces asymptotically valid inference using standard techniques, regardless of the data and identification strength. We term this property *conformance*, which is a novel property in this literature. We further establish that CLEER is fully efficient under the assumptions stated within.

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 (Hausman and Wise, 1978). The standard multinomial logit MLE has nice computational

We thank Nikhil Agarwal, Steve Berry, Chris Conlon, Amit Gandhi, Jeff Gortmaker, Phil Haile, Jessie Handbury, Jean-François Houde, Sung Jae Jun, Nail Kashaev, Mathieu Marcoux, Karl Schurter, Andrew Sweeting, and many seminar and conference participants for helpful discussions and comments on previous drafts. We thank Junpeng Hu and Eric San Miguel Flores for excellent research assistance. A companion `Julia` package `Grumps` is available with accompanying documentation. Previous versions of this paper were circulated with the titles "Efficient Estimation of Random Coefficients Demand Models using Product and Consumer Datasets" and "Conformant and Efficient Estimation of Discrete Choice Demand Models." Correspondence: Grieco: `paul.grieco@psu.edu`, Murry: `ctmurry@umich.edu`, Pinkse: `joris@psu.edu`, Sagl: `ssagl@iu.edu`.

properties (McFadden, 1974). 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 random coefficients demand 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 hundreds of thousands, depending on the application) product quality parameters.

To address this issue, CLEER incorporates product level data on market shares.[2] Product data can be augmented with a consumer level sample which is a—perhaps small, perhaps large—subset of the market population. 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. With the macro term, it is possible to identify product quality parameters. However, these two terms do not exploit product level exogeneity restrictions which may have power to identify preference heterogeneity. Moreover, typically researchers are interested in (potentially endogenous) explanatory factors of product quality, which are not identified by these terms alone.

The third term in the CLEER objective function directly incorporates information contained in the product level exogeneity restrictions of BLP95. These exogeneity restrictions are additional assumptions on the data-generating process. Indeed, as BLP95 show, with sufficient exogeneity restrictions it is possible to identify all model parameters even if there is *no* consumer sample. 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.

CLEER is compatible with the bulk of datasets in the applied literature.[3] In particular, it is well-behaved with consumer samples of any size. 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 estimator's rates of convergence adjust according to the relative sample sizes and strength of information from the objective functions' composite terms.[4]

As we illustrate in app. C, observed variation in demographics identifies both observed and unobserved taste heterogeneity as long as that variation shifts consumers' utility across products.[5] As emphasized by Gandhi and Houde (2020, GH20), overidentifying product level exclusion restrictions can also identify taste heterogeneity. If the number of sampled consumers grows faster than the number of markets, then exploiting the identifying information (if present) in the micro sample will produce a faster convergence rate than relying on product level exclusion restrictions. Adding the product

[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.

[2] CLEER also covers intermediate cases when different data (micro versus macro) is available in different markets.

[3] For expositional purposes, we assume that the researcher has direct access to consumer level first choice data and/or product level market share data. Although CLEER could accommodate both ranked choice data (e.g., Berry et al., 2004*a*; Grieco et al., 2023*a*) and aggregated statistics of micro data (Sweeting, 2013), we do not explore those extensions here.

[4] The use of the plural 'rates' is because different elements of our estimator vector converge at different rates.

[5] Berry and Haile (2024) make a similar point in a nonparametric context.

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 likelihood alone is insufficient for identification. Conversely, when product level instruments are few or weak, product level restrictions are insufficient for identification. In either case, CLEER converges at the optimal rate and is efficient because it exploits all available sources of identifying information. However, weak identification of either component may result in a slower (though still optimal) rate of convergence.

We formally establish consistency and asymptotic normality of CLEER in section 4. The proofs are nonstandard to accommodate CLEER's conformance features. We show that conducting inference using formulas familiar from the standard extremum estimation framework is asymptotically valid. Validity obtains regardless of the relative divergence rates 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. Finally, we describe the conditions under which CLEER obtains the semiparametric efficiency bound in section 5.2.

Section 6 provides a comparison of CLEER to other approaches. We note several features of CLEER which facilitate the optimality, conformance, and robustness to weak identification. These include: (1) utilizing the macro likelihood to avoid enforcing share constraints to achieve efficiency and simplify inference; (2) fully utilizing the score of the likelihood with respect to observed and unobserved consumer heterogeneity parameters; and (3) allowing overidentification of product level exclusion restrictions to provide an additional source of identification. While previous estimators have utilized subsets of these elements, ours is the first to deploy all to achieve full efficiency and conformance.

Our approach has broad applicability and is appropriate for many demand estimation applications when (either or) both product level data on shares and consumer level data on purchases is available.[6] Although BLP04 and Petrin (2002) are canonical examples of applications, there are many more. 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*a*), Montag (2023), and Jiménez-Hernández and Seira (2021). A common example in economics and marketing is combining grocery store scanner data with household level data, for example the datasets maintained by IRI or Nielsen. Examples include Chintagunta and Dube (2005) (IRI) and Tuchman (2019) and Backus et al. (2021) (Nielsen).

Berry and Haile (2014) showed identification of objects in a nonparametric class of discrete choice demand models using product level data and sufficient instruments; Berry and Haile (2024) shows how observing consumer level data reduces the number of instruments in these models. CLEER is applied to the most common parametric version of these models used in applied work. It 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. Other researchers have proposed using the likelihood of

[6] In app. G, we provide two algorithms to efficiently compute CLEER. These are both implemented in this paper's companion `Julia` package, `Grumps`.

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).[7] Allen et al. (2019) combine the likelihood of an equilibrium search model with a penalty term of moment equalities.

Our problem and approach share features with several strands of the econometrics literature. For instance, Imbens and Lancaster (1994) 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. 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). Our paper does not assume a panel structure for either products or consumers. Moreover, we know of no examples in which there are as many growth dimensions to consider as here. Having different elements of the estimator vector converge at different rates is a common feature of the semiparametric estimation literature (e.g. Robinson, 1988). Abadie et al. (2020) consider the case of sample size approaching population size; their problem is different from that studied here. Finally, several papers cover asymptotics in 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. Moon et al. (2018) consider a BLP type model placing panel data assumptions on the product quality terms, whereas we do not impose a panel structure. 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. Section 3 introduces CLEER. We state our formal consistency and asymptotic normality results in section 4. Conformance and efficiency properties are described in section 5. Section 6 illustrates the trade-offs in going from CLEER to GMM estimators that are commonly used in applied work. Section 7 compares the finite sample performance of CLEER relative to alternative estimators in a Monte Carlo study. Section 8 concludes.

2 Random Coefficients Demand Model

This section 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 assume the researcher has access to both product level shares and a sample of consumer level choices, although we will allow this sample to be empty. Our estimator assumes that consumer level choices are drawn from a subset of consumers on which the market level shares are based. In contrast, the previous literature has treated micro and macro data as different samples (e.g., Imbens and Lancaster, 1994).

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} is a d_x-dimensional vector of observed characteristics of the product and ξ_{jm} is a scalar unobserved product attribute. We often refer to ξ_{jm} as unobserved product quality, but it is important to keep in mind that it reflects the (common component of) unobserved preference for product j in market m which may vary across markets. To

[7] 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).

allow for endogeneity in product characteristics, we specify $x_{jm} = (\tilde{x}_{jm}, \mathscr{p}_{jm})$. The only distinction between $\tilde{x}_{jm}$ and $\mathscr{p}_{jm}$ (typically price) is that $\tilde{x}_{jm}$ is uncorrelated with ξ_{jm}.

There are N_m consumers in market m drawn from a market-specific distribution described below. Consumer i is characterized by $(z_{im}, \nu_{im}, \varepsilon_{im})$ where z_{im} is a d_z-vector of potentially observable consumer characteristics (such as income), and ν_{im} is a $d_\nu \leq d_x$-vector of unobservable consumer taste shocks to preferences for product characteristics. Finally ε_{im} is a $J_m + 1$-vector of idiosyncratic product specific 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}. In section 4.3 we discuss the implications of using an estimate of G_m.

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

$$u_{ijm} = \delta^*_{jm} + \mu^{z_{im}}_{jm} + \mu^{\nu_{im}}_{jm} + \varepsilon_{ijm}, \tag{1}$$

where $\delta^*_{jm} = x^\triangledown_{jm}\beta^* + \xi_{jm}$ represents the mean utility for product j for consumers in market m, $\mu^{z_{im}}_{jm} = x^\triangledown_{jm}\boldsymbol{\Theta}^{*z} z_{im} = \sum_{k,k'} \boldsymbol{\Theta}^{*z}_{kk'} x^k_{jm} z^{k'}_{im}$ represents deviations from mean utility due to observed demographic variables z_{im}, and $\mu^{\nu_{im}}_{jm} = x^\triangledown_{jm}\boldsymbol{\Theta}^{*\nu}\nu_{im} = \sum_{k,k'} \boldsymbol{\Theta}^{*\nu}_{kk'} x^k_{jm} \nu^{k'}_{im}$ are deviations due to taste shocks ν_{im}. In most applications, several elements of $\boldsymbol{\Theta}^{*z}$ and $\boldsymbol{\Theta}^{*\nu}$ are restricted (e.g., $\boldsymbol{\Theta}^{*\nu}$ is often assumed to be diagonal), so we will refer to θ^{*z} and $\theta^{*\nu}$ as the vectors of free parameters of $\boldsymbol{\Theta}^{*z}, \boldsymbol{\Theta}^{*\nu}$ to estimate. Although there is no need to assume δ, μ^z and μ^ν have a linear form, we consider the linear form since it is the most common specification and simplifies notation.[8] As we shall see below, some of our results depend on whether $\theta^{*z} = 0$ which implies $\partial_z u_{im} = \partial_z \mu^{z_{im}}_m = 0$, i.e., when changes in observed demographics do not affect utility. Utility of the outside good is normalized to $u_{i0m} = \varepsilon_{i0m}$. When convenient, we collect the consumer heterogeneity parameters into the vector $\theta^* = [\theta^{z\triangledown}, \theta^{\nu\triangledown}]^\triangledown$.

The model yields choice probabilities for consumer i of selecting product j conditional on demographics z_{im} and product characteristics X_m, ξ_m as a function of parameters,

$$\sigma^{z_{im}}_{jm}(\theta, \delta) = \mathbb{P}(y_{ijm} = 1 \mid z_{im}, X_m, \xi_m; \theta, \delta) = \int \underbrace{\frac{\exp(\delta_{jm} + \mu^z_{ijm} + \mu^\nu_{ijm})}{\sum_{k=0}^{J_m} \exp(\delta_{km} + \mu^z_{ikm} + \mu^\nu_{ikm})}}_{\mathscr{s}_{jm}(z_{im}, \nu; \theta, \delta) =: \mathscr{s}_{ijm}(\nu; \theta, \delta)} \mathrm{d}F_m(\nu), \tag{2}$$

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

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

$$\sigma_{jm}(\theta, \delta) = \mathbb{P}(y_{ijm} = 1 \mid X_m, \xi_m) = \int \sigma^z_{jm}(\theta, \delta)\mathrm{d}G_m(z). \tag{3}$$

As noted by Berry (1994), when we fix θ, (3) defines a one-to-one mapping from δ_m to unconditional choice probabilities in a market. We denote the inversion of this mapping as $\delta_m(\theta, \pi_m)$ and let $\delta(\theta, \pi)$ represent its concatenation across markets.

The model also imposes product level exogeneity restrictions of the form,

[8] Additive separability of δ in ξi is essential to our approach.

$$\mathbb{E}(\xi_{jm} \mid b_{jm}) = 0, \tag{4}$$

where b_{jm} is a vector of instruments including $\tilde{x}_{jm}$. Further, b_{jm} may contain additional instruments. The literature has used various approaches such as cost shifters, BLP instruments (in various forms, including the "differentiation instruments" proposed by GH20), Hausman instruments, and Waldfogel instruments. 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 $s_{jm} = N_m^{-1} \sum_{i=1}^{N_m} y_{ijm}$. Note that the observed market shares s_m need not equal choice probabilities π_m^* due to the finite population size, however $s_m \xrightarrow{p} \pi_m^*$ as $N_m \to \infty$.

Second, for a subset of I_m consumers, the researcher observes both the consumers' choices and their demographics. That is, the researcher observes $\{(y_{im}, z_{im})\}$ for these consumers. We use D_{im} as a dummy variable to denote whether consumer i is in this micro-sample. As we will describe below, our methodology combines the micro-sample with the product shares by integrating out z_{im} in the choice probabilities when individual i is outside the micro-sample. We can accommodate several forms of selection. In app. A.2 we show that for random sampling and deterministic selection on choices y_{im} (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 CLEER, which combines the mixed data likelihood, $\hat{L}(\theta, \delta)$, of the micro and macro choice data and a GMM objective function $\hat{\chi}$ based on (4),

$$(\hat{\theta}, \hat{\delta}, \hat{\beta}) = \underset{\theta,\delta,\beta}{\arg\min}\Big(-\log \hat{L}(\theta, \delta) + \hat{\chi}(\beta, \delta)\Big). \tag{5}$$

Notice that the likelihood is a function of (θ, δ) but not β, whereas the product level moments (PLMs) 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.

3.1 Likelihood components of the objective function

The mixed data likelihood contains two parts relating to the micro and macro data. To understand its elements, first *suppose* that we observed y_{im} for all N_m individuals in market m. Recall that if $D_{im} = 1$, (y_{im}, z_{im}) are jointly observed. Then the loglikelihood would be,[10]

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

This is an extension of the standard mixed logit estimator for N_m observations where z_{im} is missing when $D_{im} = 0$. The loglikelihood sums over all N_m consumers in the market. If an observation i is

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

[10] For expositional simplicity, we consider the cases of random selection or deterministic selection on $y_{i \cdot m}$ into the micro sample. As discussed in app. A.2, selection on demographics requires an adjustment to account for sampling.

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,

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

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 $\log \sigma_{jm}$ 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 $\log \hat{L}^{\blacklozenge}$ and $\log \hat{L}^{\blacksquare}$ 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:

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

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 mixed data likelihood can be optimized in isolation to yield an estimator for (θ^*, δ^*). Since the first stage does not separate x from ξ, endogeneity concerns do not arise. This estimator can be paired with a plug-in estimator for β^*—utilizing the instruments b—to yield a two-step estimator. As it is a useful basis for comparison, we refer to this as the mixed data likelihood two-step estimator (MDLE). Under stronger assumptions than are necessary for CLEER, the two-step estimator is consistent and asymptotically normal. However, it is neither conformant nor generally efficient. The reason is straightforward: the MDLE does not incorporate information contained in the PLMs (4) when estimating θ^*.

To summarize, the mixed data likelihood makes full use of the micro and macro choice data.

3.2 Product Level Moments (PLM)

The CLEER objective function combines the mixed data likelihood with an additional term that penalizes violations of the product level exogeneity restrictions,

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

where $\hat{\mathcal{W}}$ is a positive definite weight matrix and

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

In practice, an initial choice of $\hat{\mathcal{W}}$ would be $(B^{\triangledown} B)^{-1}$, where B is the $J \times d_b$ matrix with $J = \sum_m J_m$ rows $b_{jm}^{\triangledown}$. Note that, unlike in standalone GMM estimation, the scaling factor 1/2 in front of the 'J statistic' in (9) matters since it affects the relative weight placed on $\log \hat{L}$ versus $\hat{\chi}$ in the objective function.

If the dimension of b_{jm}, d_b, is the same as that of β^*, d_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. In this special case, CLEER is equivalent to the MDLE. Additional restrictions result in overidentification of β^* which can be used to aid the estimation of θ^*. Indeed, then $\hat{\chi}$ will generally be positive (i.e.

nonzero) so that both $\log \hat{L}$ and $\hat{\chi}$ contribute to the estimation of θ^*, δ^*. However, because the micro log likelihood sums over $I = \sum_{m=1}^{M} I_m$ terms whereas $\hat{\chi}$ involves sums over J terms these additional product level restrictions can be asymptotically negligible for θ^*, δ^* as we discuss in section 5.2.

To achieve efficiency, b_{jm} should be chosen via a two-step procedure to be the optimal instruments for θ^*, β^* in the sense of Chamberlain (1987). In this case $d_b = d_\theta + d_\beta$, and generally (10) will not be zero at the optimum, so the choice $\hat{\mathcal{W}}$ matters.

4 Consistency and Asymptotic Normality

This section formally establishes the consistency and asymptotic normality of CLEER as the number of markets M grows. This is contrast to the proof of consistency Berry et al. (2004*b*) where M is fixed but the number of products within a market grows. The proof is complicated by several features of the estimator. In particular, [i] the dimension of δ, the number of observed products across all markets, is growing with M by construction; [ii] the rate of convergence of the estimator depends on the relative rates of divergence of the micro sample I, the number of markets M, and the population of each market N_m; [iii] the rate of convergence may be effected by weak identification arising from the the micro-sample, the product level exclusion restrictions, or both.

While we have presented CLEER in its most natural form for applied work in section 3, the proof of consistency is more straightforward if we write the estimator in a reparameterized and recentered, but equivalent form, as we describe in the following subsection.

4.1 Objective function

Section 3 defines CLEER as an estimator of $(\theta^*, \delta^*, \beta^*)$. Since there is a one-to-one mapping between mean product qualities, δ, and unconditional choice probabilities, π, for fixed consumer heterogeneity parameters θ (Berry, 1994) and moreover since β can be profiled out of the objective function, it is equivalent to write the estimator in terms of (θ, π) where $\pi = [\pi_1^\triangledown, \dots, \pi_M^\triangledown]^\triangledown$ with $\pi_m = [\pi_{1m}, \dots, \pi_{J_m m}]^\triangledown$. So the parameter vector π excludes the outside good probabilities, $\pi_{0m} = 1 - \sum_{j=1}^{J_m} \pi_{jm}$ and the dimensions of δ and π are both J.[11] This is convenient because $s_m \overset{p}{\to} \pi_m^*$ independent of θ while $\delta_m^* = \delta_m(\theta^*, \pi_m^*)$ depends on the unknown θ^*.

Following this formulation, consider the sample objective function, which consists of three terms,

$$\hat{\Omega}(\theta, \pi) = \hat{\mathcal{L}}(\theta, \pi) + \hat{\Phi}(\theta, \pi) = \hat{\mathcal{L}}^{\blacklozenge}(\theta, \pi) + \hat{\mathcal{L}}^{\blacksquare}(\pi) + \hat{\Phi}(\theta, \pi). \tag{11}$$

The first term, $\hat{\mathcal{L}}^{\blacklozenge}$, is the (negative) log likelihood of the micro data net of the contribution to market shares (to avoid double counting with the second term, as above),

$$\hat{\mathcal{L}}^{\blacklozenge}(\theta, \pi) = \log \frac{\hat{L}^{\blacklozenge}(\theta^*, \delta^*)}{\hat{L}^{\blacklozenge}[\theta, \delta(\theta, \pi)]} = \sum_m \hat{\mathcal{L}}_m^{\blacklozenge}(\theta, \pi_m) = \sum_{mij} D_{im} y_{ijm} \Big[\log \frac{\varsigma_{ijm}(\theta^*, \pi_m^*)}{\varsigma_{ijm}(\theta, \pi_m)} - \log \frac{\pi_{jm}^*}{\pi_{jm}} \Big],$$

where $\varsigma_{ijm}(\theta, \pi_m) = \sigma_{jm}^{z_{im}}[\theta, \delta_m(\theta, \pi_m)]$. The only distinction between $-\log \hat{L}^{\blacklozenge}$ and $\hat{\mathcal{L}}^{\blacklozenge}$ is that the latter is recentered by a constant such that it is zero at the truth (θ^*, π^*), as shown in the first equality. The second line reorganizes $\hat{\mathcal{L}}^{\blacklozenge}$ to introduce notation that will be useful in the proofs.

The second term is the (negative) log likelihood of the market share data recentered such that it is zero at the *observed* market shares and positive otherwise:

$$\hat{\mathcal{L}}^{\blacksquare}(\pi) = \log \frac{\hat{L}^{\blacksquare}[\theta, \delta(\theta, s)]}{\hat{L}^{\blacksquare}[\theta, \delta(\theta, \pi)]} = \sum_m \hat{\mathcal{L}}_m^{\blacksquare}(\pi_m) = \sum_m N_m \sum_j s_{jm} \log \frac{s_{jm}}{\pi_{jm}}, \tag{12}$$

which is true for all $\theta \in \Theta$. Again, the final equalities present convenient reorganization for the proofs.

[11] The same will be true for other such vectors, e.g., $\hat{\pi}, s$.

The third term $\hat{\varPhi}$ is the PLM objective function $\hat{\chi}$ evaluated at $\delta = \delta(\theta, \pi)$, profiling out β:

$$\hat{\varPhi}(\theta, \pi) = \min_{\beta} \hat{\chi}\{\beta, \delta(\theta, \pi)\} = \frac{1}{2} \min_{\beta} \|\mathcal{P}_B[\delta(\theta, \pi) - X\beta]\|^2 = \frac{1}{2}\|\mathcal{P}\delta(\theta, \pi)\|^2,$$

To simplify the presentation of the proof, we take $\hat{\mathcal{W}} = (B^{\triangledown}B)^{-1}$ in the second inequality. This is not restrictive in view of the discussion in section 4.6, which considers alternatives choices of $\hat{\mathcal{W}}$, including those needed to use an optimal weight matrix or optimal instruments as described in section 5.2. Hence $\mathcal{P}_B = B(B^{\triangledown}B)^{-1}B^{\triangledown}$ is the projection matrix onto the instruments B and $\mathcal{P} = \mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X} = \mathcal{P}_B - \mathcal{P}_B X(X^{\triangledown}\mathcal{P}_B X)^{-1}X^{\triangledown}\mathcal{P}_B$.[12]

To see that this formulation is equivalent to CLEER presented in section 3: For any (θ', π') that optimize $\hat{\varOmega}$, CLEER $(\hat{\theta}, \hat{\delta}, \hat{\beta})$ is then $[\theta', \delta(\theta', \pi'), (X^{\triangledown}\mathcal{P}_B X)^{-1}X^{\triangledown}\mathcal{P}_B\delta(\theta', \pi')]$ because $\hat{\varPhi}(\theta, \pi) = \min_{\beta} \hat{\chi}[\delta(\theta, \pi), \beta]$ and $\hat{\mathcal{L}}(\theta, \pi) = -\{\log \hat{L}[\theta, \delta(\theta, \pi)] + \log \hat{L}^{\blacklozenge}[\theta^*, \delta(\theta^*, \pi^*)] + \log \hat{L}^{\blacksquare}(s)\}$, where the final two terms are constant with respect to the model parameters. Given this, for the remainder of the paper we will refer to the maximizer of (11) as $(\hat{\theta}, \hat{\pi})$ and refer to $\hat{\pi}$ as the CLEER of market-level choice probabilities.

To prove consistency of CLEER, we will show that $(\hat{\theta}, \hat{\pi})$ converges to the minimizer of the population analog, $\varOmega(\theta, \pi) = \mathcal{L}^{\blacksquare}(\pi) + \mathcal{L}^{\blacklozenge}(\theta, \pi) + \varPhi(\theta, \pi)$, where $\mathcal{L}^{\blacksquare}(\pi) = \sum_m N_m \sum_j \pi^*_{jm} \log(\pi^*_{jm} \ / \ \pi_{jm})$, $\mathcal{L}^{\blacklozenge}(\theta, \pi) = \mathbb{E}[\hat{\mathcal{L}}^{\blacklozenge}(\theta, \pi) \mid \mathbb{A}]$ with $\mathbb{A}$ the sigma algebra generated by product characteristics and the D_{im}'s, and $\varPhi(\theta, \pi) = \|\mathcal{P}[\delta(\theta, \pi) - \delta(\theta^*, \pi^*)]\|^2 \ / \ 2$. Note that $\mathcal{L}^{\blacksquare}, \mathcal{L}^{\blacklozenge}, \varPhi$ are chosen to ensure all are zero at the truth.

4.2 Identification of Demographic and Random Coefficients

Consistency of CLEER requires that the model is identified. While identification of π^* is straightforward, we define identification of θ^* in terms of $\varOmega$ evaluated at the true unconditional choice probabilities π^*, i.e.

$$\varOmega(\theta, \pi^*) = \mathcal{L}^{\blacklozenge}(\theta, \pi^*) + \varPhi(\theta, \pi^*), \tag{13}$$

since $\mathcal{L}^{\blacksquare}(\pi^*) = 0$. Clearly, $\varOmega(\theta^*, \pi^*) = 0$. While (13) uses π^*, note that $\mathcal{L}^{\blacksquare}(\pi)$ diverges for all $\pi \neq \pi^*$ and so $\varOmega(\theta, \pi)$ will as well. There are two possible sources of identification for our estimator, both may be weak. Let $\rho^{\blacklozenge}(\theta)$ be the rate of $\mathcal{L}^{\blacklozenge}(\theta, \pi^*)$ and $\rho^{\varPhi}(\theta)$ be the rate of $\varPhi(\theta, \pi^*)$. Hence, $\rho_{\mathrm{id}}(\theta) := \rho^{\blacklozenge}(\theta) + \rho^{\varPhi}(\theta)$ is the rate of $\varOmega(\theta, \pi^*)$. In the absence of weak identification, for all fixed $\theta \neq \theta^*$, $\rho^{\blacklozenge}(\theta) \sim I$ and $\rho^{\varPhi}(\theta) \sim M$, we do not assume that one of these rates is faster than the other.[13] We will allow for weak identification, which slows the rate of either or both $\rho^{\blacklozenge}(\theta)$ and $\rho^{\varPhi}(\theta)$ as described below. To obtain consistency $\rho_{\mathrm{id}}(\theta)$ must diverge (fast enough) for all θ outside a neighborhood $\varTheta_\epsilon = \{\theta \in \varTheta \colon \|\theta - \theta^*\| < \epsilon\}$ of θ^* (for any fixed $\epsilon > 0$). However, the terms of $\rho_{\mathrm{id}}(\theta)$ need not satisfy our identification condition individually, which is what gives rise to the conformance property.

Identification from PLMs occurs when $\varPhi(\theta, \pi^*) > 0$ away from θ^*. Weak product identification can slow $\rho^{\varPhi}(\theta)$ in a manner similar to traditional moment condition models (e.g., weak instruments). Identification from the micro sample occurs when $\mathcal{L}^{\blacklozenge}(\theta, \pi^*) > 0$ away from θ^*. This will fail, e.g., at $\theta^{*z} = 0$ because $\varsigma_{ijm}(0, \theta^{\nu}, \pi^*) = \pi^*$ for all z_{im} and so $\mathcal{L}^{\blacklozenge}(0, \theta^{\nu}, \pi^*) = 0$ for all θ^{ν}.[14] Weak micro identification will occur for θ^{*z} drifting towards zero.

[12] The second equality requires $X^{\triangledown}P_B X$ be invertible, which would fail if the number of instruments were less than the number of regressors or the instruments had no explanatory power. In these cases it may still be possible to identify θ^*, but not β^*. To cover this, write $\mathcal{P}$ more generally as $\mathcal{P} = \mathcal{P}_B - \mathcal{P}_B X(\mathcal{P}_B X)^+$ with $\cdot^+$ denoting a Moore Penrose inverse.

[13] Indeed, in the absence of micro data, our estimator will still be consistent even though I does not diverge.

[14] App. C provides an example and elaborates on the intuition for the identifying power of microdata within our model.

4.3 Assumptions

We are now ready to turn to our formal assumptions, which begin with those regarding micro and product level identification.

As discussed above, micro identification will always fail if $\theta^{*z} = 0$. When $\theta^{*z} \neq 0$, micro identification will fail if there exists a $\theta^{\nu} \neq \theta^{*\nu}$ such that $\mathcal{L}^{\blacklozenge}(\theta^{*z}, \theta^{\nu}, \pi^{*}) = 0$, which implies $\varsigma_{ijm}(\theta^{*z}, \theta^{\nu}, \pi^{*}) = \varsigma_{ijm}(\theta^{*z}, \theta^{*\nu}, \pi^{*})$ for all i.[15] However, this is unlikely to arise in applications as this imposes $J \times |\mathcal{Z}|$ restrictions—which is infinite if z is continuous—and θ^{ν} is low dimensional. We rule out these cases, which we interpret as model misspecification on the part of the researcher, with the following assumption.

Assumption A (Micro identification strength). Let $\lambda = \|\theta^{*z}\|$ and let $\rho^{\blacklozenge}(\theta) = I\|\theta - \theta^{*}\|_{\lambda}^{2}$ with $\|\theta - \theta^{*}\|_{\lambda}^{2} = \|\theta^{z} - \theta^{*z}\|^{2} + \lambda^{2}\|\theta^{\nu} - \theta^{*\nu}\|^{2}$. We have,

$$\inf_{\theta \in \Theta : \rho^{\blacklozenge}(\theta) > 0} \frac{\mathcal{L}^{\blacklozenge}(\theta, \pi^{*})}{\rho^{\blacklozenge}(\theta)} \succeq 1,$$

where $\succeq$ means that the left hand side is (element-wise) of greater or equal order compared to the right-hand side.[16]

Following the weak identification literature, this assumption allows θ^{*z} to drift towards zero. The parameter $I\lambda^{2}$ plays a role analogous to the concentration parameter in the weak IV setting. If $\lambda > 0$, the norm $\|\cdot\|_{\lambda}$ is zero only at θ^{*}. If $\lambda = 0$, the norm is zero whenever $\theta^{z} = \theta^{*z} = 0$ (regardless of θ^{ν}).

Our next assumption essentially defines the identifying power in the PLMs.

Assumption B (Product level identification strength). Define $\mathbb{D}_{\theta} = \partial_{\theta^{\triangledown}} \delta$ and $\mathbb{D}_{\pi} = \partial_{\pi^{\triangledown}} \delta$ as the derivatives of the Berry (1994) inversion $\delta(\theta, \pi)$ with respect to its arguments. Let $\rho^{\Phi}(\theta) = \|\mathcal{P}\mathbb{D}_{\theta}(\theta^{*}, \pi^{*})(\theta - \theta^{*})\|^{2}$ and assume that

$$\inf_{\theta \in \Theta : \rho^{\Phi}(\theta) > 0} \inf_{t \neq 0} \frac{\Phi\big(\theta^{*} + t(\theta - \theta^{*}), \pi^{*}\big)}{t^{2}\rho^{\Phi}(\theta)} \succeq 1. \qquad \square$$

The rate ρ^{Φ} is analogous to the concentration parameter in the weak identification literature.[17]

For identification, $\rho_{\mathrm{id}}(\theta) = \rho^{\blacklozenge}(\theta) + \rho^{\Phi}(\theta)$ must diverge for all θ outside a neighborhood $\Theta_{\epsilon} = \{\theta \in \Theta : \|\theta - \theta^{*}\| < \epsilon\}$ of θ^{*} (for any fixed $\epsilon > 0$). Observe that we allow the component responsible for $\rho_{\mathrm{id}}(\theta)$ diverging at a particular value of θ to depend on the value of θ. While identification *per se* only requires divergence, our consistency proof requires it do so at a minimum rate for all $\theta \in \Theta_{\epsilon}^{c}$. The following assumption makes this explicit.

Assumption C (Identification). Let $\kappa = \exp(-4\kappa_{\delta}^{\uparrow})$ with $\kappa_{\delta}^{\uparrow} = 2\sqrt{2c_{\xi}^{*} \log M}$ for c_{ξ}^{*} defined in G. Then, $\forall \epsilon > 0 : \breve{\rho}_{\mathrm{id}}(\epsilon) := \inf_{\theta \in \Theta_{\epsilon}^{c}} \rho_{\mathrm{id}}(\theta) \succ \kappa^{-12} \log^{2}(I + M) \succ 1.$ □

This rate should be compared to the rate that would arise under standard assumptions, i.e. $\max(M, I)$. C relaxes this significantly to accommodate a wide variety of relative rates of divergence, accounting for the size and identifying power of the micro sample, market population, and PLMs. The rate used here diverges slower than any power of M (times $\log^{2} I$) but faster than any power of $\log M$; κ is slowly

[15] In this statement we can evaluate $\mathcal{L}^{\blacklozenge}(\theta^{*z}, \theta^{\nu}, \pi^{*})$ at θ^{*z} since the variation in features is observed. If z_{im} has zero variance then identification also fails because there is then collinearity between z times x and x. This possibility is also ruled out by A.

[16] $\succ, \preceq, \prec, \simeq$ are analogously defined.

[17] There are four minor differences: (1) the model is nonlinear so now the Jacobian depends on parameters; (2) π is fixed at the truth π^{*} (there is no analog to our π in the context of a standard concentration parameter); (3) β has been concentrated out; (4) the standard concentration parameter omits the $(B^{\triangledown}B)^{-1}$ in our identity matrix and would include σ_{ξ}^{2} in the denominator.

varying.[18] So our rate condition is not restrictive.

The next three assumptions deal with sampling of markets, consumers and products. It is possible to relax these in many directions, but we focus on the standard case.

Assumption D (Markets)**.** Product characteristics, product instruments, consumer demographics and preferences $(x_m, \xi_m, b_m, z_m, \nu_m, \varepsilon_m)$ are independent across markets. □

Independence across markets could be relaxed. This would be relevant if e.g. some products are offered in multiple markets. The bulk of the demand literature assumes independent markets while incorporating similarities in products across markets by including "brand", "model" or "SKU" dummies in x_{jm}.[19]

Assumption E (Consumers). [i] Consumer demographics and preferences $(z_{im}, \nu_{im}, \varepsilon_{i\cdot m})$ in a given market are N_m i.i.d. draws from a superpopulation with known distribution $G \times F \times \text{Gumbel}^{J_m+1}$; [ii] Consumers in the micro sample are I_m independent draws without replacement from the consumer population. □

While we assume the distribution of consumer demographics is known and constant across markets, it is likely to be estimated in practice and could be allowed to vary by market. As a result, E implies that the micro sample is compatible with the population in the sense of CG23. In app. A, we discuss ways to relax this assumption to allow for estimation of G and some forms of selection. In many cases—e.g., when the sample is selected based on consumer choices y or demographics z—these can be accommodated without adjusting our inference procedure beyond utilizing sampling weights in the latter case. In other cases–e.g., when sample selection depends on unobserved tastes ν_{im}—a model of selection would be required, as it would be to incorporate selection on unobservables utilizing a GMM micro moments estimator (Petrin, 2002; Berry et al., 2004*a*).

Assumption F (Products). [i] The number of products in a market J_m satisfies: $1 \le J_m \le \bar{J} < \infty$ for $\bar{J}$ independent of M; [ii] Product characteristics and instruments in a market $(x_{jm}, \xi_{jm}, b_{jm})$ are independent across j and satisfy $\mathbb{E}(\xi_{jm} \mid b_{jm}) = 0$ and $0 < \inf_b \mathbb{V}(\xi_{jm} \mid b_{jm} = b) \le \sup_b \mathbb{V}(\xi_{jm} \mid b_{jm} = b) < \infty$; [iii] Product instruments have full rank, $\mathbb{E}(b_{jm} b_{jm}^{\triangledown}) > 0$; [iv] The rank of $\mathbb{E}(b_{jm} x_{jm}^{\triangledown})$ is at least d_x. □

The condition on the instruments b_{jm} is standard, although we do not place conditions on the dimension or strength of b_{jm} beyond what is assumed in C. If $d_b = d_x$, then $\rho^{\Phi}(\theta) = 0$ and PLMs do not contribute to the identification of θ^*, π^*. When $d_b < d_x$ (*contra* F[iv]), β^* is not identified.

The next two assumptions are technical although (some version thereof) is implicit in much of the literature.

Assumption G (Tails)**.** [i] The x_{jm}'s are drawn from a distribution whose support $\mathcal{X}_m$ is bounded; [ii] the ξ_{jm}'s are i.i.d. subgaussian with optimal variance proxy (OVP) $c_\xi^* < \infty$;[20] [iii] the z_{im}'s are i.i.d. subgaussian with OVP $c_z^* < \infty$ (conditional on $\mathbb{A}$); [iv] the b_{jm}'s are drawn such that $\exists \mathbb{C}_b < \infty$: $\mathbb{P}[\max_m \mathbb{E}(\|b_{jm}\|^2 \mid N_m) > \mathbb{C}_b] = 0$. □

Assumption H (Parameter Space). [i] Θ is compact and θ^* is an interior point; [ii] $\mathcal{B}$, the parameter space of β^*, is compact; [iii] $\mathbb{\Pi}$, the parameter space of π^* is $\prod_{m=1}^{M} \mathbb{\Pi}_m$ with $\mathbb{\Pi}_m = \{\pi_m : \forall j =$

[18] If M were fixed, C is satisfied if $\breve{\rho}_{\text{id}}(\epsilon)$ diverges at least square-logarithmically in I.

[19] Such an approach is innocuous provided the number of brands, models, or SKUs is small relative to the number of products, which grows with M under our assumptions. A notable alternative approach is Moon et al. (2018), which assumes a factor structure on ξ.

[20] i.e. $\forall t$: $\log \mathbb{E} \exp[t(\xi_{jm} - \mathbb{E}\xi_{jm})] \le c_\xi^* t^2/2$, which implies that $\forall t \ge 0$: $\mathbb{P}(|\xi_{jm} - \mathbb{E}\xi_{jm}| > t) \le 2\exp[-t^2/(2c_\xi^*)]$.

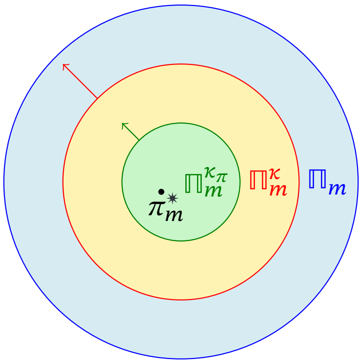


Figure 1: *Representation of the parameter space for π_m^*.*

$1, \ldots, J_m \colon 0 \leq \pi_{jm}$ and $\sum_{j=1}^{J_m} \pi_{jm} \leq 1\}$. □

Assumptions G and H imply (among many other things) that $\min_{m,j} \pi_{jm}^* \succ \kappa^{3/4}$, as we show in L9(b), which will be helpful for showing that π_m for which $\min_{m,j} \pi_{jm}$ decreases too fast cannot minimize $\hat{\Omega}$ (with probability approaching one). However, this does *not* imply that observed market shares $s_{jm} > 0$ for all j, m.

Finally, in order to estimate market shares accurately, we must assume that the population of markets grows with the data. Note that this is a significant relaxation over the standard assumption in the literature since BLP95 that unconditional choice probabilities π^* are observed and equal to s, or equivalently that $N_m = \infty$.

Assumption I (Population and sample growth). [i] As M grows, the market populations increase such that $\rho_N = \sum_m N_m^{-1}$ converges and $\rho_u = 1 \;/\; \min_m \sqrt{N_m} \prec M^{-p}$ for some (real valued) $p > 0$. Specifically, $\rho_N \prec \kappa^{12} / \log^2 \max_m (I_m + M)$. [ii] $\max_m I_m \sup_{\theta \in \Theta, \|\theta - \theta^*\|_\lambda > 0} [\|\theta - \theta^*\|_\lambda^2 \;/\; \rho_{\mathrm{id}}(\theta)] \prec 1$. □

While we do not assume that observed market shares are equal to unconditional choice probabilities, I[i] is sufficient to guarantee that $\|s - \pi^*\|$ converges to zero in the limit. It is weaker than assuming that a specific market population N_m grows faster than M.

There are several ways in which I[ii] can be satisfied. One is that the PLMs contain the dominant share of identification. Another is $\max_m I_m / I \prec 1$, i.e. an asymptotically negligible share of the micro sample is concentrated in any one market.

4.4 Consistency

We can now formally establish the asymptotic properties of the CLEER, starting with consistency.

Theorem 1. Under A to I, CLEER is consistent, i.e., (a) $\hat{\theta} - \theta^* \prec 1$; (b) $\max_m \|\hat{\pi}_m - \pi_m^*\| \prec 1$, (c) $\forall m \colon \|\hat{\delta}_m - \delta_m^*\| \prec 1$; and (d) $\hat{\beta} - \beta^* \prec 1$. □

We present the basic steps of the proof below with supporting lemmas relegated to app. B.1.

Proof. We first show consistency of $\hat{\theta}$, with the remaining parameters following straightforwardly below. For $\hat{\theta}$, it is sufficient to show $\forall \epsilon > 0 \colon \lim_{M \to \infty} \mathbb{P}\big[\inf_{\Theta_\epsilon^c \times \mathbb{\Pi}} \hat{\Omega}(\theta, \pi) - \hat{\Omega}(\theta^*, s) \leq 0\big] = 0$, since $\inf_{\mathbb{\Pi}} \hat{\Omega}(\theta^*, \pi) \leq \hat{\Omega}(\theta^*, s)$. To do so, steps *1* to *5* show that $\lim_{M \to \infty} \mathbb{P}\big[\inf_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \hat{\Omega}(\theta, \pi) - \hat{\Omega}(\theta^*, s) \leq 0\big] = 0$ for $\mathbb{\Pi}^\kappa = \prod_m \mathbb{\Pi}_m^\kappa$ where $\mathbb{\Pi}_m^\kappa = \{\pi_m \colon \min_j \pi_{jm} \geq \kappa\}$ with κ as defined in C. Step *6* extends this result from $\mathbb{\Pi}^\kappa$ to $\mathbb{\Pi}$.

Step 1. Decompose $\hat{\Omega}(\theta, \pi) - \hat{\Omega}(\theta^*, s)$ by adding and subtracting to yield four terms,

$$\hat{\Omega}(\theta, \pi) - \hat{\Omega}(\theta^*, s) = \overbrace{\Omega^\blacklozenge(\theta, \pi^*)}^{Ⓐ} + \overbrace{\Omega^\blacklozenge(\theta, \pi) - \Omega^\blacklozenge(\theta, \pi^*)}^{Ⓑ} + \overbrace{\hat{\mathcal{L}}^\blacksquare(\pi) - \hat{\mathcal{L}}^\blacksquare(s)}^{Ⓒ} + \overbrace{\Delta\hat{\Omega}^\blacklozenge(\theta, \pi) - \hat{\Omega}^\blacklozenge(\theta^*, s)}^{Ⓓ}, \tag{14}$$

where $\Delta\hat{\Omega}^{\blacklozenge} = \Delta\hat{\mathcal{L}}^{\blacklozenge} + \Delta\hat{\Phi}$ with $\Delta\hat{\mathcal{L}} = \hat{\mathcal{L}}^{\blacklozenge} - \mathcal{L}^{\blacklozenge}$ and $\Delta\hat{\Phi} = \hat{\Phi} - \Phi$.

Step 2. The first two terms on the right-hand side in (14) consist exclusively of population objects. We note the following properties of these terms at all $\theta \in \Theta_\epsilon^c$ and all $\pi \in \mathbb{\Pi}^\kappa$: [i] because $\Omega^{\blacklozenge} = \mathcal{L}^{\blacklozenge} + \Phi$ is nonnegative, Ⓐ ≥ 0; [ii] moreover, by definition, Ⓐ $\sim \rho_{\text{id}}(\theta)$; [iii] Ⓐ + Ⓑ $= \Omega^{\blacklozenge}(\theta, \pi) \geq 0$.

Step 3. Term Ⓒ is the difference between the macro likelihood term at the candidate parameter π and observed market shares s. We show that this diverges rapidly for π far from s. Let $\rho^{\blacksquare}(\pi) = \sum_m \rho^{\blacksquare}_m(\pi_m) = \sum_m N_m \|s_m - \pi_m\|^2$. By the mean value theorem (MVT), Ⓒ $= \hat{\mathcal{L}}^{\blacksquare}(\pi) \geq \rho^{\blacksquare}(\pi) \,/\, 2 \geq 0$. This follows from expanding $\hat{\mathcal{L}}^{\blacksquare}_m(\pi_m) \,/\, N_m = \sum_j s_{jm} \log(s_{jm} \,/\, \pi_{jm})$ as a function of s_m around π_m:

$$\sum_j s_{jm} \log \frac{s_{jm}}{\pi_{jm}} = \sum_j 1 \cdot (s_{jm} - \pi_{jm}) + \frac{1}{2} \sum_j \frac{(s_{jm} - \pi_{jm})^2}{\mathring{\pi}_{jm}} \geq \frac{\|\pi_m - s_m\|^2}{2}. \tag{15}$$

Step 4. We will need to show that the nuisance term Ⓓ diverges sufficiently slowly relative to Ⓐ + Ⓑ + Ⓒ. This is accomplished by showing sufficiently slow divergence relative to Ⓐ + Ⓑ or Ⓒ. Divergence rate of Ⓓ relative to Ⓐ + Ⓑ is controlled by C, while divergence relative to Ⓒ is controlled by I. It will be sufficient that $\sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} |\text{Ⓓ}/\rho_D| \prec 1$, where $\rho_D(\theta, \pi) = \eta \max\{\eta \rho_{\text{id}}(\theta), \rho^{\blacksquare}(\pi)\}$ with $\eta = \kappa^3$, as we show here. Split the nuisance term Ⓓ by likelihood and product level moment terms and deal with them separately:

$$\text{Ⓓ} = \Delta\hat{\Omega}^{\blacklozenge}(\theta, \pi) - \hat{\Omega}^{\blacklozenge}(\theta^*, s) = \Delta\hat{\mathcal{L}}^{\blacklozenge}(\theta, \pi) - \hat{\mathcal{L}}^{\blacklozenge}(\theta^*, s) + \Delta\hat{\Phi}(\theta, \pi) - \hat{\Phi}(\theta^*, s).$$

We show convergence results for each term of Ⓓ in supporting lemmas: • L2(b) shows $\sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} |\Delta\hat{\Phi}(\theta, \pi) \,/\, \rho_D(\theta, \pi)| \prec 1$; • L2(c) shows $\hat{\Phi}(\theta^*, s) \preceq 1$. • L3(b) shows $\sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} |\Delta\hat{\mathcal{L}}^{\blacklozenge}(\theta, \pi) \,/\, \rho_D(\theta, \pi)| \prec 1$. • L3(c) shows $\sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} |\hat{\mathcal{L}}^{\blacklozenge}(\theta^*, s) \,/\, \rho_D(\theta, \pi)| \prec 1$. Taken together,

$$\sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \frac{\text{Ⓓ}}{\eta \max\{\eta \rho_{\text{id}}(\theta), \rho^{\blacksquare}(\pi)\}} = \sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \frac{\text{Ⓓ}}{\rho_D(\theta, \pi)} \prec 1.$$

Step 5. We now consider two cases, which correspond to whether π is close to s or not. In the first case we consider, π far from s, and the macro likelihood—$\hat{\mathcal{L}}^{\blacksquare}(\pi)$ in term Ⓒ—will diverge fast and dominate the nuisance term Ⓓ for all values of θ, thus guaranteeing that $\hat{\Omega}(\theta, \pi)$ diverges for this case. We then consider the case where π is near s. Here, Ⓐ will dominate Ⓓ Moreover, $\Omega^{\blacklozenge}(\theta, \pi^*)$ is well approximated by $\Omega^{\blacklozenge}(\theta, \pi)$ and so Ⓑ is negligible compared to Ⓐ. Again, $\hat{\Omega}(\theta, \pi)$ diverges under this case. These cases are now presented formally:

1. For all θ, π such that $\rho^{\blacksquare}(\pi) > \rho_{\text{id}}(\theta)\eta$: $\rho_D(\theta, \pi) = \eta \rho^{\blacksquare}(\pi) \prec \rho^{\blacksquare}(\pi)$. Hence Ⓒ dominates Ⓓ or more formally,

$$\sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \left\{ \mathbb{1}[\rho^{\blacksquare}(\pi) > \rho_{\text{id}}(\theta)\eta] \frac{|\text{Ⓓ}|}{\text{Ⓒ}} \right\} \prec 1.$$

Therefore, since Ⓐ + Ⓑ and Ⓒ are non-negative from steps 2 and 3,

$$\mathbb{P}\left\{ \sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \mathbb{1}[\rho^{\blacksquare}(\pi) > \rho_{\text{id}}(\theta)\eta][\hat{\Omega}(\theta, \pi) - \hat{\Omega}(\theta^*, s)] < 0 \right\} \prec 1.$$

2. For all θ, π such that $\rho^{\blacksquare}(\pi) \leq \rho_{\text{id}}(\theta)\eta$: $\rho_D(\theta, \pi) = \eta^2 \rho_{\text{id}}(\theta) \prec \rho_{\text{id}}(\theta)$. Hence Ⓐ dominates Ⓓ or more formally,

$$\sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \left\{ \mathbb{1}[\rho^{\blacksquare}(\pi) \leq \rho_{\text{id}}(\theta)\eta] \frac{|\text{Ⓓ}|}{\text{Ⓐ}} \right\} \prec 1.$$

Moreover, writing Ⓑ $= [\Phi(\theta,\pi) - \Phi(\theta,\pi^*)] + [\mathcal{L}^\blacklozenge(\theta,\pi) - \mathcal{L}^\blacklozenge(\theta,\pi^*)]$, L2(a) and L3(a) respectively show that the terms of Ⓑ are dominated by Ⓐ for $\pi \in \mathbb{\Pi}^\kappa$. As before, Ⓒ is non-negative. Combining these,

$$\mathbb{P}\Big\{\sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \mathbb{1}[\rho^\blacksquare(\pi) \leq \rho_{\text{id}}(\theta)\eta][\hat{\Omega}(\theta,\pi) - \hat{\Omega}(\theta^*,s)] < 0\Big\} \prec 1.$$

Combining cases establishes $\mathbb{P}\{\sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa}[\hat{\Omega}(\theta,\pi) - \hat{\Omega}(\theta^*,s)] < 0\} \prec 1$.

Step 6. Finally, L4 extends the above results from $\mathbb{\Pi}^\kappa$ to $\mathbb{\Pi}$. The intuition for this step focusing on a single market is illustrated in fig. 1. Fig. 1 depicts the parameter space for choice probabilities in market m, $\mathbb{\Pi}_m$; the subset $\mathbb{\Pi}_m^\kappa$ for which we have already shown consistency above, and a further subset $\mathbb{\Pi}_m^{\kappa_\pi}$. L9(b) (using G and H) shows that $\mathbb{\Pi}_m^{\kappa_\pi}$ contains π_m^* with probability approaching one. As the arrows indicate, $\mathbb{\Pi}_m^{\kappa_\pi}$ and $\mathbb{\Pi}_m^\kappa$ both increase to $\mathbb{\Pi}_m$ as $M \to \infty$. To complete this step, we must show that candidate $\pi_m \notin \mathbb{\Pi}_m^\kappa$ (the blue band) are sufficiently far from π_m^* (which is inside the green circle). L4 establishes that this distance (bounded by the maize band) shrinks sufficiently slowly such that the macro likelihood $\hat{\mathcal{L}}^\blacksquare$ dominates the other terms in $\hat{\Omega}$ and diverges faster than $\hat{\Omega}(\theta^*,\pi^*)$. The proof establishes this for the entire parameter vector π, which ensures uniformity across markets. Thus, (with probability approaching one) no $\pi \notin \mathbb{\Pi}^\kappa$ can optimize $\Omega(\theta,\pi)$ over $\Theta_\epsilon^c \times \mathbb{\Pi}$.

This completes the proof of consistency of $\hat{\theta}$. L5 uses consistency of $\hat{\theta}$ to establish consistency of the remaining parameters $\hat{\pi}, \hat{\delta}, \hat{\beta}$. □

4.5 Asymptotic normality

Having established consistency, the following theorem establishes that the CLEER of consumer heterogeneity $\hat{\theta}$ is asymptotically normal. Once this is established, normality of the remaining parameters are straightforward and we address them immediately following the proof.

Theorem 2. Under A to I, for $\hat{\Gamma}_\theta := \hat{\mathcal{Q}}_{\theta\theta}{}^{-1/2} := (\hat{\Omega}_{\theta\theta} - \hat{\Omega}_{\theta\pi}\hat{\Omega}_{\pi\pi}^{-1}\hat{\Omega}_{\pi\theta})^{-1/2}$, $\hat{\Gamma}_\theta^{-1}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}_\theta)$, for $\mathcal{V}_\theta$ given in L8. □

The formula for $\mathcal{V}_\theta$ simplifies to the identity matrix if the likelihood portion of the objective function dominates asymptotics for $\hat{\theta}$. If $(B^\triangledown B)^{-1}$ is the optimal weight matrix then $\mathcal{V}_\theta = \mathbb{I}$, also. As suggested in section 4.6, instruments can always be chosen to make $(B^\triangledown B)^{-1}$ the optimal weight matrix. The formula for $\hat{\mathcal{Q}}_{\theta\theta}$ is the familiar Hessian with respect to θ with π concentrated out.

Theorem 2 uses a self-normalizing matrix $\hat{\Gamma}_\theta$, akin to dividing by the standard error to arrive at a pivotal T-statistic in linear regression (e.g., Student, 1908), because different elements of $\hat{\theta}$ can converge at different rates and those rates depend on the data generating process. This is a feature of the conformance property. We discuss this issue in greater detail and provide rates for individual parameters across all cases in section 5.

Proof (theorem 2). The proof proceeds over five steps with supporting lemmas in app. B.2.

Step 1. Let $\hat{\Omega}^\blacklozenge = \hat{\mathcal{L}}^\blacklozenge + \hat{\Phi}$, so that $\hat{\Omega}(\theta,\pi) = \hat{\Omega}^\blacklozenge(\theta,\pi) + \hat{\mathcal{L}}^\blacksquare(\pi)$. Consider a quadratic expansion of $\hat{\mathbb{F}}(\theta,\pi,\gamma) = \hat{\Omega}^\blacklozenge(\theta,\pi) + \hat{\mathcal{L}}^\blacksquare(\gamma)$. For generic v_θ, v_π, let $t \in [0,1]$ index a convex combination between (θ^*,π^*,s) and $(\theta^* + v_\theta, \pi^* + v_\pi, \pi^* + v_\pi)$ and define,

$$\hat{\mathbb{f}}(t) = \hat{\mathbb{F}}[\theta(t),\pi(t),\gamma(t)] = \hat{\Omega}^\blacklozenge(\theta^* + tv_\theta, \pi^* + tv_\pi) - \hat{\Omega}^\blacklozenge(\theta^*,\pi^*) + \hat{\mathcal{L}}^\blacksquare[\pi^* + tv_\pi + (1-t)(s-\pi^*)].$$

Thus, $\hat{\mathbb{f}}(1) = \hat{\Omega}^\blacklozenge(\theta^* + v_\theta, \pi^* + v_\pi) - \hat{\Omega}^\blacklozenge(\theta^*,\pi^*) + \hat{\mathcal{L}}^\blacksquare(\pi^* + v_\pi)$ and $\hat{\mathbb{f}}(0) = \hat{\mathcal{L}}^\blacksquare(s) = 0$, $\hat{\mathcal{L}}_\pi^\blacksquare(s) = 0$, $\hat{\mathbb{f}}'(0) = (v_\theta^\triangledown \hat{\Omega}_\theta^\blacklozenge + v_\pi^\triangledown \hat{\Omega}_\pi^\blacklozenge)(\theta^*,\pi^*)$, and

$$\begin{aligned}\hat{\mathbb{f}}''(t) = (&v_\theta^\triangledown \hat{\Omega}_{\theta\theta} v_\theta + 2v_\pi^\triangledown \hat{\Omega}_{\pi\theta} v_\theta + v_\pi^\triangledown \hat{\Omega}^\blacklozenge_{\pi\pi} v_\pi)(\theta^* + tv_\theta, \pi^* + tv_\pi)\\ &+ [v_\pi - (s-\pi^*)]^\triangledown \hat{\mathcal{L}}^\blacksquare_{\pi\pi}[\pi^* + tv_\pi + (1-t)(s-\pi^*)][v_\pi - (s-\pi^*)].\end{aligned}$$

Applying the MVT to $\mathbb{f}$, for some $\mathring{t} \in [0,1]$, $\hat{\mathbb{f}}(1) = \hat{\mathbb{f}}(0) + \hat{\mathbb{f}}'(0) + \hat{\mathbb{f}}''(\mathring{t}) \;/\; 2 = \hat{\mathbb{f}}'(0) + \hat{\mathbb{f}}''(\mathring{t}) \;/\; 2$. Now substitute in $v_\theta = \Gamma_\theta h_\theta$ and $v_\pi = \Gamma_\pi h_\pi$ for symmetric matrices $\Gamma_\theta, \Gamma_\pi$ and vectors h_θ, h_π to obtain

$$\begin{aligned}\hat{\Omega}^\blacklozenge(\theta^* + \Gamma_\theta h_\theta, \pi^* + \Gamma_\pi h_\pi) + \hat{\mathcal{L}}^\blacksquare(\pi^* + \Gamma_\pi h_\pi) = (h_\theta^\triangledown \Gamma_\theta \hat{\Omega}_\theta + h_\pi^\triangledown \Gamma_\pi \hat{\Omega}^\blacklozenge_\pi)(\theta^*, \pi^*)\\ + \frac{1}{2}\Big((h_\theta^\triangledown \Gamma_\theta \hat{\Omega}_{\theta\theta} \Gamma_\theta h_\theta + 2\Gamma_\pi h_\pi^\triangledown \hat{\Omega}_{\pi\theta} \Gamma_\theta h_\theta + h_\pi^\triangledown \Gamma_\pi \hat{\Omega}^\blacklozenge_{\pi\pi} \Gamma_\pi h_\pi)[\theta(\mathring{t}), \pi(\mathring{t})]\\ + [\Gamma_\pi h_\pi - (s-\pi^*)]^\triangledown \hat{\mathcal{L}}^\blacksquare_{\pi\pi}[\gamma(\mathring{t})][\Gamma_\pi h_\pi - (s-\pi^*)]\Big). \quad (16)\end{aligned}$$

Based on this expansion the proof proceeds in the remaining three steps. Step 2 shows that the final term of (16) is well approximated when we replace sample objects evaluated at $(\theta, \pi, \gamma)(\mathring{t})$ with population objects evaluated at (θ^*, π^*, π^*), yielding (18). Step 3, minimizes (18) with respect to h to arrive at an asymptotic approximation of $\Gamma_\theta^{-1}(\hat{\theta} - \theta^*)$. Finally, step 4 applies a central limit theorem to establish normality.

Step 2. Defining everything at the truth, let $\mathcal{A} \coloneqq \operatorname{plim}_{M\to\infty}(B^\triangledown B \;/\; M)$,

$$\begin{aligned}\Xi_\theta &= \operatorname*{plim}_{M\to\infty}\big[\mathbb{D}_\theta^\triangledown \mathcal{P} B (B^\triangledown B)^{-1}\big] = \operatorname*{plim}_{M\to\infty}\Big[\frac{\mathbb{D}_\theta^\triangledown B}{M}\mathcal{A}^{-1}\Big(\frac{B^\triangledown \mathcal{P} B}{M}\Big)\mathcal{A}^{-1}\Big],\\ \Xi_\pi &= \operatorname*{plim}_{M\to\infty}\Big(\mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\mathbb{D}_\pi^\triangledown \mathcal{P} B (B^\triangledown B)^{-1}\Big) = \operatorname*{plim}_{M\to\infty}\Big[\frac{\mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\mathbb{D}_\pi^\triangledown B}{M}\mathcal{A}^{-1}\Big(\frac{B^\triangledown \mathcal{P} B}{M}\Big)\mathcal{A}^{-1}\Big],\end{aligned}$$

and $\Xi = \Xi_\theta - \Xi_\pi$. L6 shows that for $\Gamma_\theta = \big[\mathbb{E}(\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\mathcal{L}_{\pi\theta}) + M\Xi\mathcal{A}\Xi^\triangledown\big]^{-1/2}$ and $\Gamma_\pi = \big\{\mathbb{E}[\mathcal{L}_{\pi\pi} - \mathcal{L}_{\pi\theta}\mathcal{L}_{\theta\theta}^{+}\mathcal{L}_{\theta\pi}]\big\}^{-1/2}$, the following hold for $\|A\| = \sup_{x:\|x\|=1}\|Ax\|$,

$$\max_{\mathring{t}\in[0,1]}\Big\|\Gamma_\theta\big\{\hat{\Omega}_{\theta\theta}[\theta(\mathring{t}), \pi(\mathring{t})] - \Omega^*_{\theta\theta}\big\}\Gamma_\theta\Big\| \prec 1, \quad (17a)$$

$$\max_{\mathring{t}\in[0,1]}\Big\|\Gamma_\theta\big\{\hat{\Omega}_{\theta\pi}[\theta(\mathring{t}), \pi(\mathring{t})] - \Omega^*_{\theta\pi}\big\}\Gamma_\pi\Big\| \prec 1, \quad (17b)$$

$$\max_{\mathring{t}\in[0,1]}\Big\|\Gamma_\pi\big\{\hat{\Omega}^\blacklozenge_{\pi\pi}[\theta(\mathring{t}), \pi(\mathring{t})] - \Omega^{\blacklozenge *}_{\pi\pi}\big\}\Gamma_\pi\Big\| \prec 1, \quad (17c)$$

$$\max_{\mathring{t}\in[0,1]}\Big\|\Gamma_\pi\big\{\hat{\mathcal{L}}^\blacksquare_{\pi\pi}[\gamma(\mathring{t})] - \mathcal{L}^{\blacksquare *}_{\pi\pi}\big\}\Gamma_\pi\Big\| \prec 1. \quad (17d)$$

This allows us to replace all the hatted second derivatives in (16) with their population counterparts at the truth. This, with some algebraic manipulation, transforms the right-hand side of (16) into (evaluating everything at the truth),

$$\begin{aligned}\frac{1}{2}(s-\pi^*)^\triangledown \mathcal{L}^\blacksquare_{\pi\pi}(s-\pi^*) + \{h_\theta^\triangledown \Gamma_\theta \hat{\Omega}_\theta + h_\pi^\triangledown \Gamma_\pi[\hat{\Omega}^\blacklozenge_\pi - \mathcal{L}^\blacksquare_{\pi\pi}(s-\pi^*)]\} +\\ \frac{1}{2}\Big(h_\theta^\triangledown \Gamma_\theta \Omega_{\theta\theta} \Gamma_\theta h_\theta + 2h_\pi^\triangledown \Gamma_\pi \Omega_{\pi\theta} \Gamma_\theta h_\theta + h_\pi^\triangledown \Gamma_\pi \Omega_{\pi\pi} \Gamma_\pi h_\pi\Big). \quad (18)\end{aligned}$$

Step 3. Note that (18) is a quadratic in h_θ, h_π. Thus, $h^* = (h^*_\theta, h^*_\pi) \simeq [\Gamma_\theta^{-1}(\hat{\theta} - \theta^*), \Gamma_\pi^{-1}(\hat{\pi} - \pi^*)]$ is the solution to,

$$\min_h\left\{\frac{1}{2}(s-\pi^*)^\triangledown \mathcal{L}^\blacksquare_{\pi\pi}(s-\pi^*) + h^\triangledown\begin{bmatrix}\Gamma_\theta \hat{\Omega}_\theta\\ \Gamma_\pi[\hat{\Omega}^\blacklozenge_\pi - \mathcal{L}^\blacksquare_{\pi\pi}(s-\pi^*)]\end{bmatrix} + \frac{1}{2}h^\triangledown\begin{bmatrix}\Gamma_\theta \Omega_{\theta\theta} \Gamma_\theta & \Gamma_\theta \Omega_{\theta\pi} \Gamma_\pi\\ \Gamma_\pi \Omega_{\pi\theta} \Gamma_\theta & \Gamma_\pi \Omega_{\pi\pi} \Gamma_\pi\end{bmatrix}h\right\}.$$

Applying the partitioned inverse formula,

$$\Gamma_\theta^{-1}(\hat{\theta} - \theta^*) \simeq -\big[\Gamma_\theta\overbrace{(\Omega_{\theta\theta} - \Omega_{\theta\pi}\Omega_{\pi\pi}^{-1}\Omega_{\pi\theta})}^{=\mathcal{Q}_{\theta\theta}}\Gamma_\theta\big]^{-1}\big\{\Gamma_\theta\big[\overbrace{\hat{\Omega}_\theta - \Omega_{\theta\pi}\Omega_{\pi\pi}^{-1}(\hat{\Omega}^\blacklozenge_\pi - \mathcal{L}^\blacksquare_{\pi\pi}(s-\pi^*)}^{=:\hat{q}_\theta}\big]\big\}.$$

We apply L7 to show that the denominator term $\Gamma_\theta \mathcal{Q}_{\theta\theta} \Gamma_\theta$ converges to an identity matrix. Then, L8 shows that the numerator term $\Gamma_\theta \hat{q}_\theta \xrightarrow{d} N(0, \mathcal{V}_{\hat{\theta}})$.

Step 4. Finally, L6 implies that $\Gamma_\theta^2 \hat{\Gamma}_\theta^{-2} \xrightarrow{p} \mathbb{I}$. □

Theorem 2 can be extended to cover asymptotic normality of linear combinations of CLEER estimator of $(\beta^*, \theta^*, \delta^*)$.

Lemma 1. For any conformable matrix Λ for which $\lim_{M\to\infty}(\Lambda^\triangledown \Lambda) = \mathbb{I}$ and matrices $\hat{\mathcal{H}}, A$ presented in app. J.1,

$$(\Lambda^\triangledown \hat{\mathcal{H}}^{-1} \Lambda)^{-1/2} \Lambda^\triangledown \begin{bmatrix} \hat{\beta} - \beta^* \\ \hat{\theta} - \theta^* \\ \hat{\delta} - \delta^* \end{bmatrix} \simeq A^\triangledown \begin{bmatrix} B^\triangledown \xi \\ \hat{\mathcal{L}}_\theta \\ \hat{\mathcal{L}}_\pi \end{bmatrix} \xrightarrow{d} \mathcal{N}(0, \mathbb{I}).$$

□

The proof of L1 is similar to that of theorem 2 with the main wrinkle being that we need to use the delta method with the dimension of the argument increasing with M: see app. J.1 for the proof. L1 makes inference on linear combinations of CLEER estimates straightforward.[21]

4.6 Choice of $\hat{\mathcal{W}}$

We now return to the choice of weight matrix. In defining the sample objective function in section 4.1, we assumed $\hat{\mathcal{W}} = (B^\triangledown B)^{-1}$ which is convenient as it allows for straightforward analytical expressions when profiling out β.

As noted, the choice of $\hat{\mathcal{W}}$ is (asymptotically) immaterial for the estimation of θ^* if asymptotics come from the likelihood portion of the objective function. In other cases, it matters only for the asymptotic variance of $\hat{\theta}$, not for its consistency, provided that $\operatorname{plim}_{M\to\infty}(J\hat{\mathcal{W}})$ is positive definite.

A choice for $\hat{\mathcal{W}}$ that is made intuitive by the optimal weight matrix discussion in the GMM literature is $\hat{\mathcal{W}} = (B^\triangledown \hat{\mathcal{V}}_\xi B)^{-1}$, where $\hat{\mathcal{V}}_\xi$ is such that $B^\triangledown \hat{\mathcal{V}}_\xi B/J \xrightarrow{p} \mathbb{E}(\xi_{jm}^2 b_{jm} b_{jm}^\triangledown)$. For instance, for heteroskedasticity robust inference, $\hat{\mathcal{V}}_\xi$ could be a diagonal matrix with elements $\hat{\xi}_{jm}^2$, where $\hat{\xi}_{jm}$ is a residual from a first step estimation and $\mathcal{V}_\xi$ a matrix with diagonal elements $\mathbb{V}(\xi_{jm} \mid b_{jm})$. Such a two-step strategy can be used to achieve efficiency in the sense of either section 5.2.1 or section 5.2.2 as detailed there.

Theorem 2 already allows for this possibility since in that theorem we can redefine the instruments $\bar{B} = \mathcal{V}_\xi^{1/2} B$, which results in the PLM term $[\delta(\theta, \pi) - X\beta]^\triangledown \mathcal{V}_\xi^{-1/2} \mathcal{P}_{\bar{B}} \mathcal{V}_\xi^{-1/2} [\delta(\theta, \pi) - X\beta]/2$. With this adjustment the proof will go through provided the matrix $\mathcal{V}_\xi^{-1/2}$ is explicitly incorporated throughout, which is trivial in view of F[ii].

5 Conformance and Efficiency

5.1 Conformance

The intuition for conformance is most readily apparent by examining the Hessian of the population objective function Ω, introduced in section 4.1, evaluated at the truth,

$$H_\Omega = H_{\mathcal{L}^\blacklozenge} + H_{\mathcal{L}^\blacksquare} + H_\Phi = \begin{bmatrix} \mathcal{L}^\blacklozenge_{\theta z \theta z} & \mathcal{L}^\blacklozenge_{\theta z \theta \nu} & \mathcal{L}^\blacklozenge_{\theta z \pi} \\ \mathcal{L}^\blacklozenge_{\theta \nu \theta z} & \mathcal{L}^\blacklozenge_{\theta \nu \theta \nu} & \mathcal{L}^\blacklozenge_{\theta \nu \pi} \\ \mathcal{L}^\blacklozenge_{\pi \theta z} & \mathcal{L}^\blacklozenge_{\pi \theta \nu} & \mathcal{L}^\blacklozenge_{\pi\pi} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \mathcal{L}^\blacksquare_{\pi\pi} \end{bmatrix} + \begin{bmatrix} \Phi_{\theta z \theta z} & \Phi_{\theta z \theta \nu} & \Phi_{\theta z \pi} \\ \Phi_{\theta \nu \theta z} & \Phi_{\theta \nu \theta \nu} & \Phi_{\theta \nu \pi} \\ \Phi_{\pi \theta z} & \Phi_{\pi \theta \nu} & \Phi_{\pi\pi} \end{bmatrix}. \quad (19)$$

Up to negligible terms, $\hat{\Gamma}_\theta^2$ defined in theorem 2 is the top left two-by-two block of H_Ω^{-1}.

The entries of the block terms of (19) diverge at different rates. Moreover, any individual term can be singular. Conformance obtains because CLEER is asymptotically normal and converges at the optimal

[21] In practice, if the weight matrix is chosen as is suggested in section 4.6, standard errors can be computed from the inverse hessian of the objective function (5), otherwise the sandwich analog should be used.

rate whenever the sum of the terms is invertible.

While the middle term, $H_{\mathcal{L}^\blacksquare}$, is not a function of θ, it provides key identifying information on π^*. The first and last terms of (19) correspond to the two sources of identification for θ^* discussed earlier: micro data and product level moments.

Micro identification arises from $H_{\mathcal{L}^\blacklozenge}$. Under strong micro identification, $H_{\mathcal{L}^\blacklozenge}$ has full rank. As discussed above, a failure of micro identification occurs when $\theta^{*z} = 0$, a consequence of which is that the red entries in $H_{\mathcal{L}^\blacklozenge}$ are all 0 and $H_{\mathcal{L}^\blacklozenge} + H_{\mathcal{L}^\blacksquare}$ is singular. To see this, consider for ease of notation the case where θ^ν is a scalar and x_m^ν represents the product characteristic on which the random coefficient appears. Then,

$$\mathcal{L}_{\theta^\nu\theta^\nu m}(\theta^*, \pi_m^*) = \mathbb{E}\Big(x_m^{\nu\nabla}(\mathbb{W}_{im}^* - \mathbb{W}_m^*\mathbb{Q}_m^{*-1}\mathbb{Q}_{im}^*)\mathbb{M}_{im}^{*-1}(\mathbb{W}_{im}^* - \mathbb{W}_m^*\mathbb{Q}_m^{*-1}\mathbb{Q}_{im}^*)x_m^\nu \,\Big|\, \mathbb{A}\Big),$$

where $\mathbb{M}_{im}^* = \operatorname{diag}(\vec{\pi}_{im}^*) - \vec{\pi}_{im}^*\vec{\pi}_{im}^{*\nabla}$, $\mathbb{Q}_{im}^* = \int \mathfrak{A}_{im}^* \mathrm{d}F_\nu$, $\mathbb{W}_{im}^* = \int \mathfrak{A}_{im}^* \nu \mathrm{d}F_\nu$, with $\mathfrak{A}_{im}^* = \operatorname{diag}(\vec{\mathscr{s}}_{im}^*) - \vec{\mathscr{s}}_{im}^*\vec{\mathscr{s}}_{im}^{*\nabla}$ where $\vec{\pi}_{im}^* = \varsigma_{i\cdot m}(\theta^*, \pi_m^*)$, $\vec{\mathscr{s}}_{im}$ is a vector with elements $\mathscr{s}_{jm}(z_{im}, \nu; \theta, \delta)$. Further, $\mathbb{W}_m^*, \mathbb{Q}_m^*$ are $\mathbb{W}_{im}^*, \mathbb{Q}_{im}^*$ integrated over demographic characteristics. When $\theta^{*z} = 0$, $\vec{\pi}_{im}^* = \pi_m^*$. It follows that $\mathbb{W}_m^* = \mathbb{W}_{im}^*$, $\mathbb{Q}_m^* = \mathbb{Q}_{im}^*$ and hence $\mathcal{L}_{\theta^\nu\theta^\nu m}(\theta^*, \pi_m^*) = 0$ for all m and $H_{\mathcal{L}^\blacklozenge} + H_{\mathcal{L}^\blacksquare}$ is singular, the red case in (19) noted above. Weak micro identification corresponds to the case where θ^{*z} is close to zero and hence $H_{\mathcal{L}^\blacklozenge} + H_{\mathcal{L}^\blacksquare}$ is nearly singular.

Product identification arises from H_Φ. As with conventional GMM, identification depends on the number of instruments relative to the number of parameters to identify. Suppose the number of instruments in b is less than or equal to d_β. Then, Φ and H_Φ are both 0 and product moments do not aid in identification of θ^*. Additional instruments provide identifying power, and if $d_b \geq d_\beta + d_\theta$ then in principle all parameters can be estimated *à la* BLP95 (i.e., without relying on the micro sample). However, as in BLP95, it is necessary to supplement Φ with additional restrictions on π to achieve identification, since $d_\pi = J \gg d_b$ for all practical purposes. Most of the literature accomplishes this by constraining π to match observed market shares s following BLP95, introducing J market share constraints. In CLEER, the additional restrictions come from $\mathcal{L}^\blacksquare$, so product identification obtains when $H_{\mathcal{L}^\blacksquare} + H_\Phi$ is invertible. We further note that the instruments b may be weak, analogous to standard weak identification in GMM (see Stock et al., 2002), resulting in $H_{\mathcal{L}^\blacksquare} + H_\Phi$ being nearly singular.

CLEER's asymptotics are driven by $H_\Omega = H_{\mathcal{L}^\blacklozenge} + H_{\mathcal{L}^\blacksquare} + H_\Phi$ rather than $H_{\mathcal{L}^\blacklozenge} + H_{\mathcal{L}^\blacksquare}$ or $H_{\mathcal{L}^\blacksquare} + H_\Phi$ individually. The rates of convergence of the elements of CLEER can differ and will correspond to the *fastest* rate of divergence among the three terms (entry by entry). In the best case, both micro and product identification are strong, and these rates depend only on the number of observations used to construct $\mathcal{L}^\blacklozenge$, $\mathcal{L}_m^\blacksquare$, and Φ, which grow with the number of micro consumers I, the market sizes $\{N_m\}$, and the number of markets, M, respectively.[22] In general, the convergence rates will also depend on identification strength. The following subsection will focus on strong identification; we expand our discussion to the general case allowing for weak identification strength (drifting) parameters in section 5.1.2.

5.1.1 Strong identification. This section focuses on the case when the micro identification parameter $\lambda = \|\theta^{*z}\| > 0$ is constant and we have $d_b \geq d_\beta + d_\theta$ strong instruments. This is sufficient, but not

[22] As is standard for GMM, Φ is an inner product of moment vectors which are sums over all J products. By F, J increases at rate M, the number of markets.

necessary for strong identification.[23]

Case	$\hat{\theta}$	$\hat{\pi}_m$	$\hat{\delta}_m$	$\hat{\beta}$
$M \preceq I \preceq N_m$	$\sqrt{I}$	$\sqrt{N_m}$	$\sqrt{I}$	$\sqrt{M}$
$M \preceq N_m \preceq I$	$\sqrt{I}$	$\sqrt{N_m}$	$\sqrt{N_m}$	$\sqrt{M}$
$I \preceq M \preceq N_m$	$\sqrt{M}$	$\sqrt{N_m}$	$\sqrt{M}$	$\sqrt{M}$
$I \preceq N_m \preceq M$	$\sqrt{M}$	$\sqrt{N_m}$	$\sqrt{N_m}$	$\sqrt{M}$
$N_m \preceq M \preceq I$	$\sqrt{I}$	$\sqrt{N_m}$	$\sqrt{N_m}$	$\sqrt{M}$
$N_m \preceq I \preceq M$	$\sqrt{M}$	$\sqrt{N_m}$	$\sqrt{N_m}$	$\sqrt{M}$

(a) *Convergence rates under strong identification*

Parameter	Rate
$\hat{\theta}^z$	$\max(\sqrt{I}, \sqrt{M\phi_\theta^2})$
$\hat{\theta}^\nu$	second fastest of $\sqrt{I}, \sqrt{I\lambda^2}, \sqrt{M\phi_\nu^2}, \sqrt{M\phi_\theta^2}$
$\hat{\pi}_m$	$\sqrt{N_m}$
$\hat{\delta}_m$	$\min(\sqrt{N_m}$, rate of $\hat{\theta}^\nu)$
$\hat{\beta}$	$\min(\sqrt{M\phi_\beta^2}$, rate of $\hat{\theta}^\nu)$

(b) *General convergence rates*

Table 1: *Convergence rates comparison*

Tbl. 1a displays the convergence rates of each component of CLEER under all relative rates of divergence of M (the number of markets), I (the micro sample size), and N_m (the market population). Because of strong identification, $\hat{\theta}^z$ and $\hat{\theta}^\nu$ diverge at the same rate across all cases. The rate is the faster of $\sqrt{I}$ and $\sqrt{M}$, which correspond to (the square root of the) rates of divergence of $\hat{\mathcal{L}}^\blacklozenge$ and $\hat{\Phi}$ respectively; the term that diverges at the slower rate is asymptotically negligible. Choice probabilities, $\hat{\pi}_m$, always converge at a $\sqrt{N_m}$ rate as it can be estimated off $\hat{\mathcal{L}}_m^\blacksquare$ which diverges at rate N_m. While micro data may improve efficiency of $\hat{\pi}_m$, it does not improve the rate as π_m^* is market specific. $\hat{\delta}_m$ can be estimated from $\hat{\theta}$ and $\hat{\pi}_m$ via the delta method. Therefore, it converges at the slower rate of these two estimators. $\hat{\beta}$ always converges at the rate $\sqrt{M}$; it cannot converge faster than the infeasible IV estimator which treats δ^* as known. Although $\hat{\delta}_m$ may converge at rate slower than $\sqrt{M}$ (i.e., rows 4 through 6 of tbl. 1a), $\hat{\beta}$ maintains the $\sqrt{M}$ rate as it depends on a (weighted) average of the $\hat{\pi}_m$'s, thereby reducing its variance.

5.1.2 The general case. The convergence rates of CLEER elements in the general case are summarized in tbl. 1b. In general, weak identification or failure of identification may arise from either the micro sample or the PLMs. CLEER natively conforms to the identifying power in the data without pre-testing, which is important for applied work. Recall that weakness of micro identification is parameterized by λ drifting towards zero. For the PLMs, we consider three *concentration parameters*: one for β^*, one for $(\theta^{*\nu}, \beta^*)$, and one for (θ^*, β^*).[24] There are three because having more parameters to identify requires more and stronger moments. We denote the smallest eigenvalue of each concentration parameter as $M\phi_\beta^2$, $M\phi_\nu^2$, and $M\phi_\theta^2$, respectively.

The convergence rate of $\hat{\pi}_m$ is unaffected compared to section 5.1.1. Turning to $\hat{\theta}$, we see that in general the rates of $\hat{\theta}^z$ and $\hat{\theta}^\nu$ can differ. $\hat{\theta}^z$ converges weakly faster than $\hat{\theta}^\nu$. When CLEER estimates θ^{*z} off $\hat{\mathcal{L}}^\blacklozenge$, it uses the micro data and effectively takes deviations from 'averages,' and hence the rate of $\hat{\theta}^z$ cannot be slower than $\sqrt{I}$. However, the convergence rate of $\hat{\theta}^z$ cannot be slower than $\sqrt{M\phi_\theta^2}$, either, since with sufficiently many strong moments, CLEER estimates θ^{*z} off the PLMs.

Now $\hat{\theta}^\nu$. Note that $I \succeq I\lambda^2$ and $M\phi_\nu^2 \succeq M\phi_\theta^2$ by definition. Any of the rates for $\hat{\theta}^\nu$ in tbl. 1b can be second

[23] Strong identification may also be obtained with only $d_b \geq d_\beta + d_{\theta^\nu}$ strong instruments if I grows sufficiently fast. This case will be covered in section 5.1.2.

[24] A concentration parameter is a "measure of the strength of the instruments" (Stock et al., 2002). In our context, the concentration parameters are (up to immaterial constants) $\mathbb{E}(A^\triangledown \mid Z) Z Z^\triangledown \mathbb{E}(A \mid Z)$ for $A = X$, $A = [X \quad \mathbb{D}^*_{\theta^\nu}]$, and $A = [X \quad \mathbb{D}^*_{\theta^\nu} \quad \mathbb{D}^*_{\theta^z}]$, respectively.

fastest:[25] (1) The rate $\sqrt{I\lambda^2}$ obtains if the micro data contain sufficient identifying information for the random coefficients to make the PLMs redundant for the estimation of θ^*, i.e. if $I \geq I\lambda^2 \geq M\phi_\nu^2 \geq M\phi_\theta^2$. (2) We get the rate $\sqrt{M\phi_\nu^2}$ if $I \geq M\phi_\nu^2 \geq I\lambda^2 + M\phi_\theta^2$ because θ^{*z} is then (asymptotically) estimated off the micro data only, and $\theta^{*\nu}$ off the PLMs. (3) Third, the rate is $\sqrt{I}$ if $M\phi_\nu^2 \geq I \geq M\phi_\theta^2 + I\lambda^2$ because both θ^{*z} and $\theta^{*\nu}$ can then be estimated using a combination of micro data and product level moments, with the convergence rate no worse than the slower of the two. (4) Finally, we get $\sqrt{M\phi_\theta^2}$ if $M\phi_\nu^2 \geq M\phi_\theta^2 \geq I \geq I\lambda^2$ because the estimator then only uses the PLMs to estimate θ^*. Since it needs more of them, the smaller concentration parameter applies.

In section 5.1.1, we noted that the convergence rate for $\hat{\delta}_m$ depends on the rates of $\hat{\theta}$ and $\hat{\pi}_m$. Because in the general case $\hat{\theta}^\nu$ can converge more slowly than $\hat{\theta}^z$, the rate is now the slower of that of $\hat{\theta}^\nu$ and $\hat{\pi}_m$. Although $\hat{\beta}$ depends on both $\hat{\theta}$ and $\{\hat{\pi}_m\}$, its convergence rate is the slower of $\sqrt{M}$ and the convergence rate of $\hat{\theta}^\nu$ because it depends on a (weighted) average of the $\hat{\pi}_m$'s, thereby reducing its asymptotic variance.

5.2 Efficiency

Conformance establishes that CLEER converges at the optimal rate under general conditions. To be efficient, CLEER must further achieve the smallest possible variance at that rate. If the likelihood is the dominant term in the objective function then efficiency of $\hat{\theta}$ follows directly from maximum likelihood principles. Only if the PLMs contribute, i.e. if $M\phi_\nu^2 \geq I\lambda^2$, does the choice of instruments matter for efficiency of $\hat{\theta}$. Moment condition estimators typically operate under two distinct notions of efficiency based on the form of moment conditions. We cover both in the following two subsections.

5.2.1 Unconditional moment restrictions. In this section, we discuss efficiency for a given instrument vector b_{jm}. Since this notion of efficiency considers only unconditional moment restrictions, we relax F[ii] as follows,

Assumption J (Products: Unconditional Moments). Product characteristics and instruments in a market $(x_{jm}, \xi_{jm}, b_{jm})$ are independent across j, satisfy $\mathbb{E}(b_{jm}\xi_{jm}) = 0$. □

With this assumption, CLEER achieves efficiency when the standard optimal weight matrix from GMM is used for $\hat{\mathcal{W}}$, i.e. $\hat{\mathcal{W}}$ is chosen as suggested in section 4.6. To see this, it is convenient to first consider the gradient of the CLEER objective function (5),[26]

$$\begin{bmatrix} \partial_\beta \hat{m}^\triangledown \hat{\mathcal{W}} \hat{m} \\ -\partial_\theta \log \hat{L} \\ -\partial_\delta \log \hat{L} + \partial_\delta \hat{m}^\triangledown \hat{\mathcal{W}} \hat{m} \end{bmatrix}. \tag{20}$$

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

$$\operatorname*{arg\,min}_{\beta,\theta,\delta} \frac{1}{2} \begin{bmatrix} \hat{m}^\triangledown & \partial_{\psi^\triangledown} \log \hat{L} \end{bmatrix} \begin{bmatrix} \hat{\mathcal{W}} & 0 \\ 0 & \hat{\mathcal{W}}_L \end{bmatrix} \begin{bmatrix} \hat{m} \\ \partial_\psi \log \hat{L} \end{bmatrix}, \tag{21}$$

where $\psi = [\theta^\triangledown, \delta^\triangledown]^\triangledown$ and $\hat{\mathcal{W}}_L = \left(-\partial_{\psi\psi^\triangledown} \log \hat{L}\right)^{-1}$ evaluated at the solution $\hat{\psi}$ of (5).[27] Note that in (21) there may be more moments than parameters. Specifically, (20) has $d_\beta + d_\theta + d_\delta$ moments, whereas

[25] For the sake of ranking, ties are kept, e.g. if $\lambda = 1$ and $I > M$ then the second-fastest rate is $\sqrt{I\lambda^2} = \sqrt{I}$.

[26] While the majority of this section makes use of the objective function described in section 4.1, which is convenient for theoretical reasons, here we use (5) because δ^*, not π^* is a parameter of interest; the substantive argument is equivalent since $\delta = \delta(\theta, \pi)$ is invertible in π.

[27] We define $\hat{\mathcal{W}}_L$ in terms of (5) in case its gradient (20) is zero at multiple points. Despite their asymptotic equivalence, there are two reasons to prefer CLEER to the GMM estimators described in (20) and (21). as we detail in section 6.1.

(21) is based on $d_b + d_\theta + d_\delta$ moments. Under exact identification of (21), i.e. if $d_b = d_\beta$, both (20) and (21) 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 (21) is

$$\begin{bmatrix} \partial_\beta \hat{m}^\triangledown \hat{\mathcal{W}} \hat{m} \\ 0_{d_\theta} \\ \partial_\delta \hat{m}^\triangledown \hat{\mathcal{W}} \hat{m} \end{bmatrix} + \begin{bmatrix} 0_{d_\beta} \\ \partial_{\theta\psi^\triangledown} \log \hat{L} \hat{\mathcal{W}}_L \partial_\psi \log \hat{L} \\ \partial_{\delta\psi^\triangledown} \log \hat{L} \hat{\mathcal{W}}_L \partial_\psi \log \hat{L} \end{bmatrix}, \tag{22}$$

which yields (20) at the solution since $\hat{\mathcal{W}}_L = \left(-\partial_{\psi\psi^\triangledown} \log \hat{L}\right)^{-1}$, establishing the equivalence of these estimators.

The remainder of this subsection establishes that (21) is efficient when we weaken F[ii] to J. First, by the law of iterated expectations, at the truth, the off-diagonal blocks are zero, because

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

where the second equality follows from the the likelihood principle applied to the choice problem (without PLMs).[28] The intuition for this result follows from the fact the inner expectation is over the consumer level variables z, y, whereas z, y do not enter the PLMs. 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 PLMs. Therefore, this choice of weight matrix is optimal (when replacing F[ii] with J).

5.2.2 Optimal instruments. If the PLMs do not contribute asymptotically to the estimation of θ^*, then CLEER $\hat{\theta}$ is fully efficient independent of the choice of weight matrix or instruments. Efficiency of $\hat{\beta}$ will require the use of optimal instruments, as it always relies on PLMs. *If* the PLMs contribute asymptotically to the estimation of θ^*[29] then full efficiency requires the use of optimal instruments to fully exploit the conditional moment restrictions F[ii] (Chamberlain, 1987, C87).[30] As always, the optimal instruments would have to be estimated.

A novelty of CLEER is that such instruments must incorporate both the score and the Hessian of the loglikelihood, specifically,

$$B_m^{\text{opt}} = \mathcal{V}_{\xi m}^{-1} \mathbb{E}\left(\left[\mathbb{D}^*_{\theta m} - \mathbb{D}^*_{\pi m} \mathcal{L}^{*-1}_{\pi\pi m} \mathcal{L}^*_{\pi\theta m} \quad -X_m\right] \,\middle|\, B_m\right), \tag{23}$$

where (as defined in section 4.3) $\mathbb{D}_{\theta m} = \partial_{\theta^\triangledown} \delta_m$, $\mathbb{D}_{\pi m} = \partial_{\pi_m^\triangledown} \delta_m$, and $\mathcal{V}_{\xi m} = \mathbb{V}(\xi_m \mid B_m)$.

In app. D we show that CLEER using these instruments and $\hat{\mathcal{W}} = (B^{\text{opt}\triangledown} \mathcal{V}_\xi B^{\text{opt}})^{-1}$ achieves the semiparametric efficiency bound for (θ^*, β^*). This can be implemented following section 4.6, such that $\hat{\mathcal{W}}$ is compatible with theorems 1 and 2.[31]

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

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

[29] Under strong identification this would occur if $M \succeq I$.

[30] As always, going from conditional to unconditional moment restrictions can result in a loss of identification. Optimal instruments use the local curvature of the objective function at the truth. Away from the truth, they only enforce the restriction $\mathbb{E}[g(b)a] = 0$ for one function g, which does not imply $\mathbb{E}(a|b) = 0$ globally, which would be equivalent to $\mathbb{E}[g(b)a] = 0$ for *all* functions g.

[31] Unlike in the standard case, where the choice of a weight matrix is immaterial as optimal instruments provide exact identification, CLEER must use the optimal weight matrix $(\sum_m B_m^{\text{opt}\triangledown} \mathcal{V}_{\xi m} B_m^{\text{opt}})^{-1}$ to achieve the proper weighting of

6 Comparison with Alternative Estimators

To clarify the contribution of CLEER, we now relate it to other estimators used in the discrete choice literature.

First, as noted above, with $I = N$, $\log \hat{L}$ simplifies to the mixed logit loglikelihood. If $I < N$, the only difference is that $\log \hat{L}$ exploits the market share data via the macro term. This is particularly useful when M is large relative to I, 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. (2023*b*).

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*a*).[32] The remainder of this section shows how CLEER can be converted into members of this class of estimators. There are four basic steps: using the score of CLEER to construct an asymptotically equivalent GMM estimator (section 6.1); restricting δ as a function of θ to enforce the constraint $s = \sigma(\theta, \delta)$ (section 6.2); replacing non-linear moments relating to the derivatives of $\log L$ with approximations that can be simulated without bias (section 6.3); and integrating moments over z to arrive at moment restrictions that can be employed without access to the underlying micro data (section 6.4). We provide a schematic figure of these steps in app. E.1. Since CLEER is conformant and efficient, we will point out losses of conformance and efficiency along the way. There may be a trade-off between efficiency and computational tractability that justifies using an inefficient estimator. We 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 CLEER in app. G.

6.1 Step 1: A GMM version of our estimator

In section 5.2, we presented a GMM estimator (21) which is asymptotically equivalent to our estimator, assuming that (21) does not lose identification; as we pointed out in section 5.2. 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 (21) have the norming indicated in section 5.2: 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 $\hat{\chi}$ in (5) similarly avoids distributional assumptions on ξ.

This GMM estimator has an important computational disadvantage: moving from the likelihood to a quadratic function *of the score* as the objective function replaces the computational tractability of the logit kernel in the mixed logit objective with an a more complicated loss function with respect to the underlying parameters, especially the high-dimensional δ. In practice, this increases the occurrence of

likelihood and GMM terms.

[32] An alternative class of share constrained micro likelihood estimators (e.g., Goolsbee and Petrin, 2004; Chintagunta and Dube, 2005; Train and Winston, 2007; 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 Fig. 10.

local optima and saddle points of the objective function, making the estimator difficult to compute and verify. 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 (21) one can impose share constraints, $\hat{\pi} = s$.[33] Following the intuition of Berry (1994), this means one can effectively replace the unknown π^* with data and consequently treat $\delta(\theta, \pi^*)$ as a known function of θ. Doing this allows one to replace optimization over δ with computing a fixed point of a contraction mapping to enforce the share constraints.[34]

Three issues arise when imposing the share constraints. First, because it is a one to one mapping on the interior of the probability simplex, doing so rules out the presence of zero observed shares. Second, imposing the share constraints introduces a potential loss of efficiency. Third, and most importantly, assuming $s = \pi^*$ will invalidate standard 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 provide a full discussion of these issues in app. E.2.

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

We can focus on the micro score because the macro score in (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.

$$\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{j=0}^{J_m}\frac{D_{im}y_{ijm}}{\sigma_{jm}^{z_{im}}}\int \mathscr{s}_{jm}(z_{im},\nu)\Big(x_{jm}^{k} z_{im}^{d} - \sum_{k=1}^{J_m} x_{km}^{k} z_{im}^{d}\mathscr{s}_{km}(z_{im},\nu)\Big)\mathrm{d}F(\nu), \tag{24}$$

which is a ratio of two integrals due to the presence of $\sigma_{jm}^{z_i}$ in the denominator. A commonly used approximation to the score can be found by setting $\nu = 0$ selectively as follows: Continuing from (24),

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

[33] Share constraints can also be imposed on $\log \hat{L}$ directly, see fn. 32.

[34] The underlying estimator is the same whether this mapping computed in the inner loop a nested fixed point (BLP95) or jointly when computing the estimator as in MPEC (Dubé et al., 2012), so this distinction is unrelated to our discussion.

$$= \sum_{m=1}^{M} \sum_{i=1}^{N_m} D_{im} \Big(\sum_{j=0}^{J_m} y_{ijm} x_{jm}^{k} z_{im}^{d} - \sum_{k=1}^{J_m} x_{km}^{k} z_{im}^{d} \sigma_{km}^{z_{im}} \underbrace{\sum_{j=0}^{J_m} y_{ijm}}_{=1} \Big)$$

$$= \sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} D_{im} (y_{ijm} - \sigma_{jm}^{z_{im}}) x_{jm}^{k} z_{im}^{d}, \tag{25}$$

where the final equality follows because we can reindex the final summation and $x_{0m}^{k} = 0$ by definition.

The final line of (25) 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.[35] A convenient feature of this moment is that it is linear in $\sigma_{jm}^{z_{im}}$, 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 $\delta(\theta, s)$, which is an argument to $\mathscr{s}_{jm}$, must diverge faster than I to avoid affecting efficiency and necessitating a different inference procedure,[36] and at at least the same rate as I 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 (24), replacing z_{im}^{d} with ν^{k} in the integrand, i.e.

$$\sum_{m=1}^{M} \sum_{i=1}^{N_m} \sum_{j=0}^{J_m} D_{im} \frac{y_{ijm}}{\sigma_{jm}^{z_{im}}} \int \mathscr{s}_{jm}(z_{im}, \nu) \Big(x_{jm}^{k} \nu^{k} - \sum_{k=1}^{J_m} x_{km}^{k} \nu^{k} \mathscr{s}_{km}(z_{im}, \nu) \Big) \mathrm{d}F(\nu), \tag{26}$$

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 PLMs for identification. As discussed above, doing so may slow the rate of convergence of $\hat{\theta}^{\nu}$ from $\sqrt{I}$ to $\sqrt{M}$.

A second alternative employed by e.g. Berry et al. (2004*a*) and Grieco et al. (2023*a*) is introducing second choice data based on surveys of consumer purchases to construct alternative moments. CLEER 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. If such data were available, it could easily be exploited by an extension of our estimator employing an exploded logit (Allison and Christakis, 1994). However, analysis of this estimator is outside the scope of our paper.

While we are unaware of its use in the literature, there is a third possibility that utilizes two independent ν draws per simulation r. While not efficient, such an estimator would be conformant and avoid simulation bias. We descibe this estimator in app. E.3.

6.4 Step 4: Population statistics instead of micro data

One may further alter the moment described in section 6.3.1 by integrating (25) over z,

$$\sum_{m=1}^{M} \sum_{j=0}^{J_m} \Big(\frac{1}{I_m} \sum_{i=1}^{N_m} D_{im} y_{ijm} x_{jm}^{k} z_{im}^{d} - \int \sigma_{jm}^{z} x_{jm}^{k} z^{d} \mathrm{d}G(z) \Big). \tag{27}$$

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

[35] 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.

[36] Otherwise, there would be an extra term in the moment due to the error in simulating δ, i.e. there would be one term with δ and one expansion term involving the difference between simulated and actual values of δ.

There are two possible motivations using (27) over (25). 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 minuend in (27). The subtrahend in (27) does not involve a sum over observed consumers. Therefore, (27) can be applied without direct access to the micro data. There is also a weaker motivation in terms of relaxing computational effort: Since (27) targets a population statistic, it can be approximated by an integral over the choice probabilities without simulating individual (for each i in a micro sample) objects. However, in view of PP89, the total number of simulation draws needed is the same in both cases. To simulate (25), 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 (27) one needs a number of independent draws that is at least proportional to I.

Using (27) over (25) has an additional efficiency cost. In particular, (27) 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 (27) weakly greater than (25). 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 (25) and (27) (if evaluated at the truth) have the same Jacobian in expectation. The variance contribution for observation i using (27) equals

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

which is the variance contribution of observation i in (25). These two facts combined with the sandwich formula for the asymptotic variance of the GMM estimator imply that using (25) dominates (27).

7 Monte Carlo Experiments

This section presents Monte Carlo simulations across a number of different settings to investigate the performance of CLEER and illustrate the practical importance of conformance in applied work. To do so, we compare CLEER with two non-conformant estimators: First, we consider an estimator that uses product-level information but does not fully incorporate the information in the micro data. Specifically, we use the GMM estimator described in section 6, which enforces the share constraint (section 6.2), drops moments relating to the score of $\log L$ with respect to θ^{ν} (section 6.3.2) and utilizes the correlation micro moment (27) to approximate the score of θ^{z} (section 6.4). We will refer to this estimator as “GMM Micro Moments,” or simply as GMM-M.[37] Second, we consider the MDLE introduced in section 3.1. This estimator underutilizes product level exogeneity restrictions. Recall that MDLE is a two-step estimator that first estimates (θ^*, δ^*) by minimizing $-\log \hat{L}$, and then estimates β^* by minimizing $\hat{\chi}(\beta, \hat{\delta})$. Thus, product level moment restrictions are not used in the estimation of θ^*. As we detail below, we use the same instruments to specify $\hat{\chi}$ in all three estimators.

We compare these three estimators' performance as we vary the following aspects of the data generating process (DGP): (a) The size of micro data sample available, which impacts the amount of micro information available; (b) The number of markets (and hence products) in the data, which affects the amount of product information available; (c) The underlying θ^* parameters; the magnitude of θ^{*z} directly influences the strength of micro information and the magnitude of $\theta^{*\nu}$ the strength of the PLMs; (d) The strength of the product level instrument for an endogenous characteristic, which impacts the strength of PLMs, but has no effect on the amount of micro information.

[37] We implement GMM-M using the `pyblp` package, version 1.1.0 (Conlon and Gortmaker, 2020, 2023).

To summarize, varying these settings affects the relative power of the micro observations and product level exclusion restrictions for estimation of the random coefficients $\theta^{*\nu}$, which affects the precision of all parameters of the model. Throughout, we will compare CLEER, which efficiently utilizes both these sources of identification, with the two estimators that emphasize only one. Finally, since likelihood estimators can suffer from numerical bias, we perform a final comparison of CLEER and GMM-M when $\theta^{*\nu}$ is large and this bias is likely to be most severe.

7.1 Monte Carlo Design

This section provides an abbreviated summary of our Monte Carlo design. See app. H for a comprehensive overview of the design and implementation details.

Our data generating process (DGP) includes two observable product characteristics (x^1_{jm}, x^2_{jm}); two demographic characteristics (z^1_{im}, z^2_{im}). Mean product quality is specified as $\delta^*_{jm} = \beta^*_c + \beta^*_1 x^1_{jm} + \beta^*_2 x^2_{jm} + \xi_{jm}$, The unobservable product characteristic ξ_{jm} is also distributed as a standard normal independent across j and m. One of the observable product characteristics, x^1, is correlated with unobserved characteristic ξ_{jm}, and thus endogenous; the strength of the instrument is governed by the design parameter a. (See app. H.1 for the precise specification). The remaining characteristic x^2 is exogenous. Consumers have observed (if in the micro sample) characteristics, $z_{im} = (z^1_{im}, z^2_{im})$ and unobserved characteristics $\nu_{im} = (\nu^1_{im}, \nu^2_{im})$ which are independent and drawn from the standard normal distribution. Preference heterogeneity is parameterized according to $\mu^{z_{im}}_{jm} = \theta^{*z}_1 z^1_{im} x^1_{jm} + \theta^{*z}_2 z^2_{im} x^2_{jm}$, and $\mu^{\nu_{im}}_{jm} = \theta^{*\nu}_1 \nu^1_{im} x^1_{jm} + \theta^{*\nu}_2 \nu^2_{im} x^2_{jm}$..

In addition to the instrument b^1 for x^1 as well as a constant and the exogenous characteristic x^2, we utilize three additional "BLP instruments" in $\hat{\chi}$ constructed from product characteristics: We construct two differentiation IVs for x^1, x^2 following GH20 allowing for the fact that x^1 is endogenous (see app. H.1); finally, we include the number of products in the market (which varies across markets). Consequently, $d_b = 6 > d_\beta = 3$, so $\hat{\chi}$ is overidentified for β^* and the extra exclusion restrictions are potentially useful to identify θ^*.

We organize our experiments around a baseline specification of the data generating process, described in app. H.2. Except where they are explicitly varied, these parameters are held at the baseline throughout of this section. Implementation details for the estimators are provided in app. H.3.

Before turning to our results we briefly summarize the role of each estimator in our study. While GMM-M utilizes PLMs for the identification of $\theta^{*\nu}$, it fails to incorporate all the information in the likelihood of the consumer sample. MDLE does the opposite: fully utilizing micro data for the estimation of $\theta^{*\nu}$ while not leveraging the information in the PLMs. CLEER fully exploits all available information from the data, making it conformant.

7.2 Sample size results

This subsection illustrates the practical implication of conformance with regard to sample sizes. These experiments vary the size of the micro sample and the number of markets observed in a setting of strong identification. The following subsection will consider variations in the DGP that alter the strength of identification while holding sample sizes fixed.

7.2.1 Varying micro sample size. The first experiment varies the size of the micro sample. Growing I_m while holding M fixed increases the amount of information in the micro sample relative to the PLMs. While increasing I_m should improve the precision of all estimators, GMM-M only benefits via greater precision in the estimation of the demographic micro-moment whereas MDLE and CLEER fully exploit the consumer data via the micro-likelihood though the score with respect to θ^z, θ^ν, and δ. Fig. 2 presents the distribution of $\hat{\theta}_1^z$, $\hat{\theta}_1^\nu$ and $\hat{\beta}_1$ from this experiment. Each plot compares the distribution of the three estimators for a specific consumer sample size (rows) and a given parameter (columns). CLEER is a solid blue line, GMM-M is a dashed green line, MDLE is a dotted black line. The central row, $I_m = 1\,000$, represents our baseline DGP.

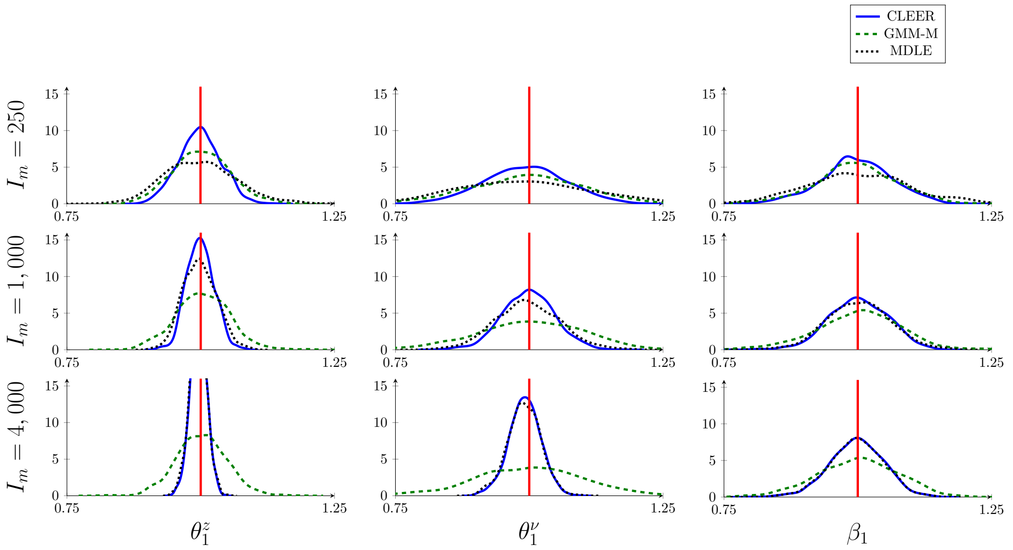


Figure 2: *Distribution of parameters for different sizes of consumer sample.*

Beginning with the smallest micro sample $I_m = 250$, CLEER dominates both GMM-M and MDLE for $\hat{\theta}^z$. At this small micro-sample size, CLEER and GMM-M perform similarly for $\hat{\theta}^\nu$ and $\hat{\beta}$, outperforming MDLE. As I_m increases, there is significant improvement in the precision of both $\hat{\theta}^z$ and $\hat{\theta}^\nu$ for CLEER and MDLE, both of which utilize the score of the likelihood with respect to θ^ν. In contrast, GMM-M has a smaller improvement as I_m increases. This is intuitive given its inefficient use of the micro data. At $I_m = 4\,000$ the MDLE and CLEER almost coincide and outperform GMM-M. The similarity of MDLE and CLEER for large I_m is also an implication of conformance.

The left panel of Tbl. 2 presents distribution and inference statistics relating to $\hat{\theta}_1^\nu$ for this DGP. Summary statistics for the other parameters are presented in app. K, which covers all experiments presented in this section. The median absolute error (MAE) numbers in the left panel of tbl. 2 reflect that increasing I_m improves the precision of CLEER and MDLE, but has minimal impact on GMM-M for the reasons discussed above. In this experiment, none of the estimators suffers from significant bias. The final two panels of the table consider inference. While all three estimators are close to the targeted 0.95 acceptance probability, we see that, as expected, the standard errors of CLEER and MDLE shrink fast as I_m grows. Moreover, when $I_m = 250$ or $I_m = 1\,000$, CLEER is able to generate more precise estimates.

7.2.2 Varying the number of markets. We now reverse the experiment and consider increasing M while holding I fixed. Note that consequently the size of the consumer sample for each market, I_m, decreases with M. Intuitively, increasing the number of markets increases the amount of information available from PLMs relative to the micro sample.

As with increasing I_m, increasing M improves the precision of all estimators. However, whereas GMM-M and CLEER fully exploit the information in the PLMs, MDLE benefits only in the second step

		Varying Consumer Sample				Varying Number of Markets		
Criteria	I_m	CLEER	GMM-M	MDLE	M	CLEER	GMM-M	MDLE
Median absolute error	250	0.051	0.069	0.087	10	0.039	0.149	0.041
	1 000	0.032	0.068	0.041	50	0.032	0.068	0.041
	4 000	0.020	0.067	0.021	1 000	0.015	0.018	0.042
Bias	250	-0.007	-0.004	-0.007	10	-0.008	-0.010	-0.008
	1 000	-0.002	0.003	-0.003	50	-0.002	0.003	-0.003
	4 000	-0.008	-0.000	-0.009	1 000	0.001	0.001	-0.008
Acceptance probability	250	0.951	0.967	0.945	10	0.957	0.974	0.957
	1 000	0.960	0.953	0.957	50	0.960	0.953	0.957
	4 000	0.936	0.957	0.935	1 000	0.949	0.945	0.952
Median S.E.	250	0.078	0.103	0.123	10	0.060	0.218	0.063
	1 000	0.051	0.101	0.061	50	0.051	0.101	0.061
	4 000	0.029	0.101	0.031	1 000	0.022	0.026	0.061

This Table shows the results from two separate experiments. The left panel shows result for varying the size of the consumer sample (I_m), while holding other parameters fixed. The right panel shows results for varying the number of markets while holding other parameters fixed (including the total consumer sample size).

Table 2: *Monte Carlo Results for θ_1^ν*

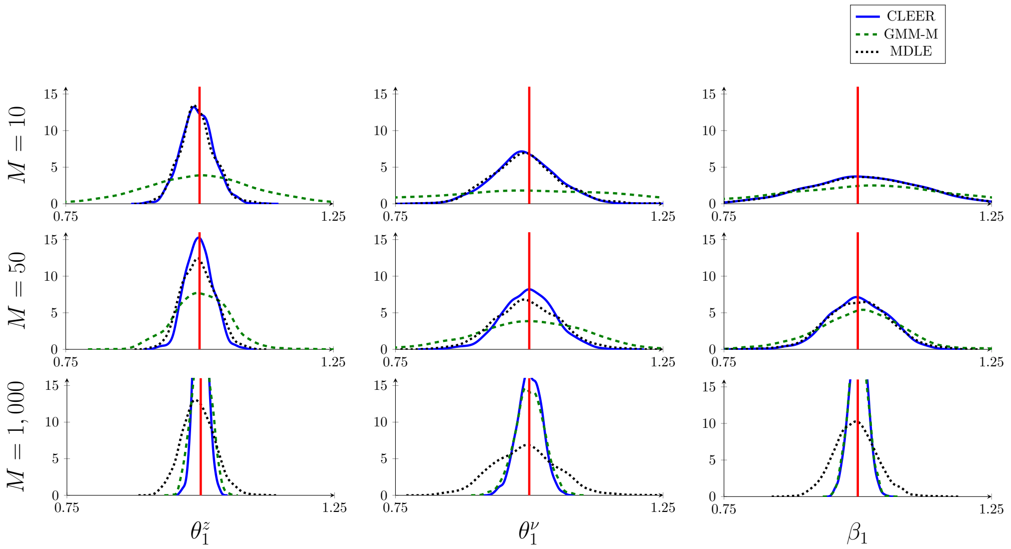


Figure 3: *Distribution of parameters for different numbers of markets*

(estimating β^*). Fig. 3 presents results for this experiment. Again, each plot compares the distribution of the three estimators for a specific overall number of markets (rows) and a given parameter (columns). For reference, the central row in this figure corresponds to the baseline DGP, $M = 50$, which is also the central DGP of fig. 2.

Visually, both CLEER and MDLE dominate GMM-M when $M = 10$. With little exploitable information in the PLMs, CLEER and MDLE almost coincide for all parameters. Both methods outperform GMM-M, which relies more heavily on the relatively sparse information in the PLMs. As we increase M (and thus J), there is significant improvement in the precision of both $\hat{\theta}^z$ and $\hat{\theta}^\nu$ for CLEER and GMM-M, both of which are able to fully exploit the information in $\hat{\chi}$. In contrast, MDLE improves only for β^*, as it relies on $\hat{\chi}$ exclusively in the second stage. At $M = 1\,000$ GMM-M and CLEER outperform MDLE for β^*, θ^{*z} and $\theta^{*\nu}$. Both estimators perform better than MDLE because the information of the PLMs on θ^* dominates that from the likelihood. Although MDLE also benefits from more identifying information in $\hat{\chi}$ for β^*, it performs worse than GMM-M and CLEER even for β^* because less precision in estimating θ^* (and thus δ^*) causes a loss of precision in the second stage. While GMM-M and CLEER perform similarly for β^* when M is large, CLEER still outperforms GMM-M for θ^{*z} and $\theta^{*\nu}$ because it efficiently combines identifying information from the PLMs with information from the likelihood. However, this distinction would disappear were we to raise M even further.

We present MAE, bias, coverage probabilities, and median standard errors for $\theta_1^{*\nu}$ in the right panel of tbl. 2. The results for MAE are intuitive, CLEER and MDLE outperform GMM-M when M is small; CLEER and GMM-M outperform MDLE when M is large. There is little evidence of bias in any estimator, and coverage probabilities perform similarly well. As in right panel, CLEER is able to deliver more precise standard errors while maintaining coverage probabilities.

Finally, while computational speed is not our focus. It is worth pointing out that CLEER is computationally tractable. The largest problem presented in these Monte Carlos, with $M = 1\,000$ and $d_\delta = 19\,000$, takes about 4 hours to compute using a single standard processing core on the Penn State ICDS ROAR computing cluster. Moreover, implementation of CLEER is able to compute estimates for a version of our DGP with $M = 500$ and $J_m \approx 190$ such that $d_\delta = 95\,000$ in under four hours on a laptop with a (16-core) Intel Core Ultra 7 155H processor and 64Gb of RAM.[38]

7.3 Parameterization Results

We next consider the estimators' performance for different parameterization of the DGP while fixing I_m and M at their baseline values. These exercises allow us to examine the impact of changes in identification strength while holding the size of the data sample fixed.

7.3.1 Varying θ^*. In our first exercise, we consider the estimators' performance as we vary θ^*. Specifically, we conduct nine experiments where we alter the values of $\theta^{*\nu}$ and θ^{*z}.[39] We focus on the distribution of the random coefficient $\hat{\theta}_1^{\nu}$, which is plotted in fig. 4. Moving along a row of this figure varies θ^{*z} while holding $\theta^{*\nu}$ fixed; moving down a column varies $\theta^{*\nu}$ while holding θ^{*z} fixed. Our baseline parameter values are in the central panel in fig. 4.[40]

As discussed in detail in section 4 and app. C, the identifying power of the consumer sample for $\theta^{*\nu}$ becomes weak as $\theta^{*z} \to 0$. Intuitively, if changes in observable demographics do not affect utility across products, then comparisons between consumers are not useful in measuring substitution. This explains the poor performance of MDLE for small values of θ^{*z} (first column). In contrast, GMM-M and CLEER perform similarly when θ^{*z} is small. Both rely on the variation from the PLMs that compose $\hat{\chi}$ to estimate $\theta^{*\nu}$. While CLEER also incorporates information from the micro sample, this is negligible when θ^{*z} is small.

Fig. 4 also documents cases where GMM-M performs relatively poorly but MDLE is comparable to CLEER. In particular, this occurs when θ^{*z} is large relative to $\theta^{*\nu}$ (i.e., panels below the 45 degree line). As is well known, when random coefficients are normally distributed, the objective function of GMM-M is symmetric at 0. Consequently, the Hessian is singular, leading to weak identification of $\theta^{\nu *}$

[38] This timing result uses full version of CLEER, a implementation of the less computationally intensive but asymptotically equivalent version described in app. G computes the same problem in under two hours.

[39] For parsimony of presenting the results, we let $\theta_1^{*z} = \theta_2^{*z}$ and $\theta_1^{*\nu} = \theta_2^{*\nu}$ throughout this experiment.

[40] A small number of replications for MDLE estimator fail when $\theta^z = 0.3$ due to the gradient becoming numerically unstable after several iterations. We do not include these replications in the figures or reported statistics. This issue does not occur for the CLEER or GMM-M estimators or for MDLE in any other specifications.

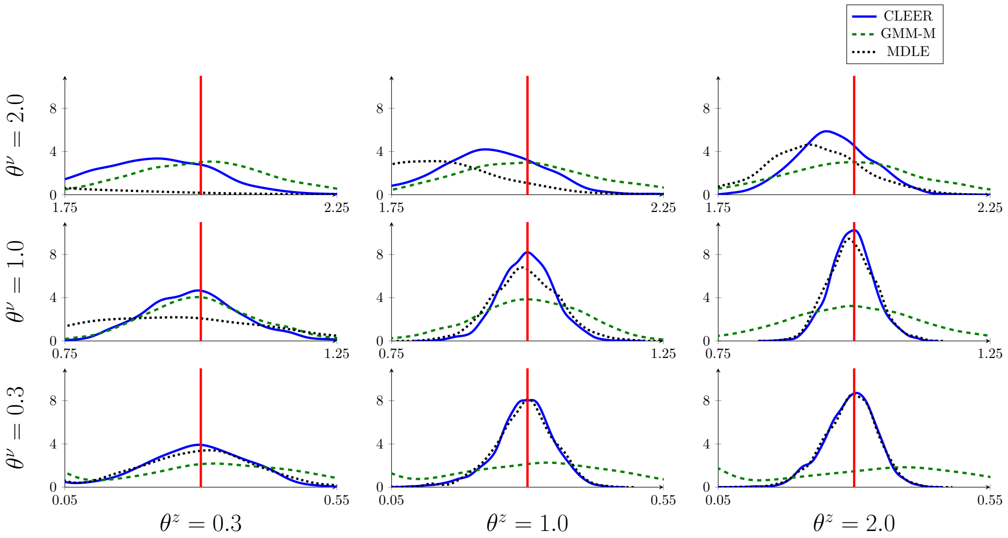


Figure 4: *Distribution of $\hat{\theta}^\nu$ across three estimators as we vary $\theta^{*\nu}$ (rows) and θ^{*z} (columns).*

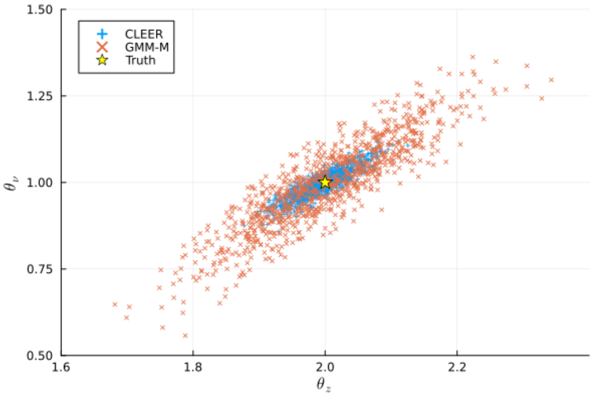


Figure 5: *Joint distribution of $(\hat{\theta}^z, \hat{\theta}^\nu)$ for CLEER and GMM-M when $\theta^{*z} = 2.0$ and $\theta^{*\nu} = 1.0$.*

near 0. The effects of this are visible in the bottom row of fig. 4. While a similar issue arises for CLEER and MDLE, it is mitigated by the use of the likelihood score.

To illustrate this phenomenon further, fig. 5 plots the joint distribution of $(\hat{\theta}^z_1, \hat{\theta}^\nu_1)$ for CLEER (blue cross) and GMM-M (red x) when $\theta^{*\nu} = 1$ and $\theta^{*z} = 2$.[41] As we see, the joint distributions of both CLEER and GMM-M are centered at the true parameter values and exhibit strong correlation between $\hat{\theta}^\nu_1$ and $\hat{\theta}^z_1$. Intuitively, this correlation is driven by the need to solve the "correlation" micro moment for GMM-M and the θ^z-micro likelihood score for CLEER. Of course, the θ^z-score contains more information than the correlation moment as we explain in section 6.3.1. However, these factors alone can only pin down a curve through the $(\hat{\theta}^z, \hat{\theta}^\nu)$ plane. To identify these parameters individually, GMM-M relies primarily on the differentiation IVs in the PLMs, whereas CLEER combines the same PLMs with additional information from the θ^ν-micro likelihood score; see section 6. Both of these sources of identification become weak as $\theta^{*\nu} \to 0$. However, the added information means that CLEER will (asymptotically) outperform GMM-M along this path.

As we look across all panels of fig. 4, CLEER performs well.[42] When only one source of identification is useful, it roughly matches the performance of the estimator that exploits that source. When both sources

[41] We choose to focus on this combination of parameter values because when $\theta^{*\nu} = 0.3$ the GMM-M micro moments estimator exhibits a mode at the lower bound of parameter space for θ^ν at 0 which is visible in the lower panels of fig. 4.

[42] CLEER does exhibit some bias in the top row when $\theta^{*\nu} = 2$, we explore this further in section 7.4 below.

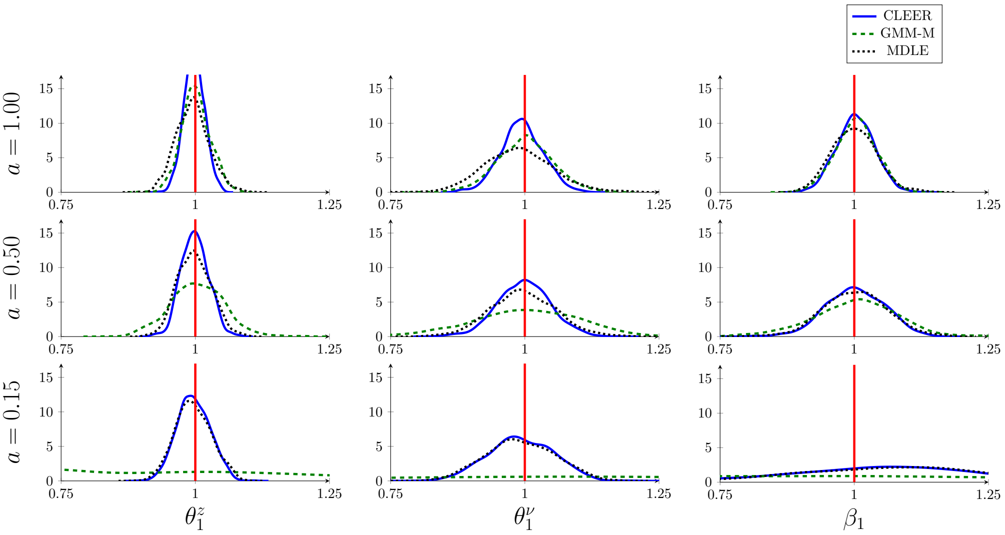


Figure 6: *Distribution of parameter estimates for* $(\theta_1^{*z}, \theta_1^{*\nu}, \beta_1^*)$ *varying the strength of the instrument for* x^1*. When* $a = 1$*,* $x_1 = b^1$*, the correlation between* x^1 *and* b^1 *increases with* a*.*

are useful, it efficiently weights the two. It does so with no pre-testing, tuning, or other adjustments on the part of the researcher. This exercise provides a finite sample illustration of the value of the conformance property in practice.

7.3.2 Varying instrument strength. So far, we have assumed to have access to a strong instrument, which is a strong data requirement in many empirical applications. In this experiment, we assess the sensitivity of the three estimators to varying the strength of the instrument b^1 for the endogenous variable x^1. As explained in app. H.1, our DGP parameter $a \in [0, 1]$ governs the strength of this instrument. Fig. 6 plots the distribution of estimates for $(\theta_1^{*z}, \theta_1^{*\nu}, \beta_1)$ (columns) varying a (rows).

We start with $a = 1$, such that x^1 is exogenous and $b^1 = x^1$. In this case, all three estimators perform well, but CLEER has a slightly tighter distribution around the truth. When $a = 0.5$—our baseline case where the mean F-statistic of the first stage regression of instruments on x^1 over all replications is 190.71 (s.d. 18.05) the instrument can be considered strong. For the θ parameters, this has no effect on the MDLE two-step, which does not use the instrument to identify θ^*. CLEER and GMM-M both become less precise than when $a = 1$. The biggest decline in performance comes from GMM-M, which ignores the micro data variation. CLEER, which was always the most precise estimator, remains visibly more precise than MDLE for $\hat{\theta}^z$, although its advantage for $\hat{\theta}_1^\nu$ and $\hat{\beta}_1$ is smaller. Finally, when $a = 0.15$, the mean first stage F-stat is 6.74 (s.d. 2.21), so the instrument is considered weak. As expected, GMM-M performs poorly for all three parameters. This weakness will carry over to the differentiation IV which is constructed using the predicted values from this regression. However, the distributions of CLEER and MDLE are essentially identical and remain precise (and approximately normal) for θ^*. MDLE suffers essentially no loss of precision for the estimate of θ^* from the $a = 0.5$ case. CLEER is no longer more precise than MDLE owing to the fact that the PLMs are no longer adding useful information for θ^*, but it matches MDLE's performance. There is also a difference for β^* between GMM-M and the two likelihood estimators. Since MDLE and CLEER identify θ^* and δ^* from the micro data, all the useful variation in b^1 is preserved for the estimation of β^*.

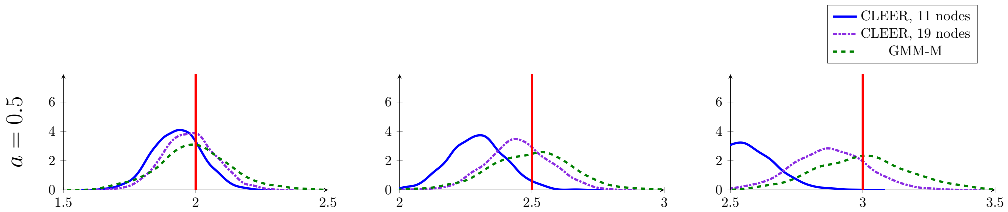


Figure 7: *Distribution of $\hat{\theta}_1^{*\nu}$ for different values of $\theta^{*\nu}$ comparing the CLEER estimator with 11 node and 19 node quadrature integration with GMM-M.*

7.4 Numerical bias

As discussed in app. G.2, in practice likelihood-based estimators are subject to bias due to the use of numerical integration over ν. This bias will grow more severe as $\theta^{*\nu}$ rises. All of our experiments so far have used 11 point Gaussian quadrature (121 nodes over two dimensions of ν) to approximate the likelihood. We now compare the performance of CLEER using 19 point quadrature (361 nodes) and the GMM-M estimator—which is unbiased—when $\theta^{*\nu}$ is large. Specifically we consider $\theta^{*\nu} \in \{2.0, 2.5, 3\}$. Fig. 7 displays the results of this experiment. As before, CLEER using 11 point quadrature is displayed in blue (solid), and GMM-M is in green (dashed). We introduce a 19 point quadrature implementation of CLEER displayed in purple (dashed dot).

The first panel, $\theta^{*\nu} = 2$, corresponds to the top-center panel of fig. 4. A slight bias is visible in the 11-point quadrature while the 19 point quadrature CLEER and GMM-M have similar means (GMM-M has wider dispersion). In the center panel, $\theta^{*\nu} = 2.5$, bias for the 11 point quadrature estimator is more apparent; it is largely but not completely eliminated by moving to the 19 point quadrature. The GMM-M estimator is more disperse but relatively unbiased. These trends are extended further in the right panel, when $\theta^{*\nu} = 3$.

We have also considered the same experiment when x is exogenous, i.e., $a = 1$ (see Grieco et al., 2023*b*, fig. 7). Across all values of $\theta^{*\nu}$, there is significantly smaller bias for both the 11 and 19 point approximations.This interaction between the strength of the instrument and the degree of bias is intuitive; a stronger instrument leads CLEER to rely more heavily on the PLMs in estimation.

The degree of approximation bias is under the control of the researcher and can be alleviated at the expense of more computational resources. Of course, computational demands will rise with the dimension of $\theta^{*\nu}$. However, stipulating that the variation exists to identify a high dimensional $\theta^{*\nu}$, 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). These results suggest that the bias of CLEER can be contained to acceptable levels given modern computing resources.[43]

7.5 Diversion

We conclude this section by examining how our three estimators perform in estimating substitution patterns across the sampling and DGP specifications we have considered.

Specifically, we present results on diversion with respect to unobserved quality, defined at the truth as, $\mathcal{D}_{jkm}^* = -\partial_{\xi_{jm}}\sigma_{km}^* \,/\, \partial_{\xi_{jm}}\sigma_{jm}^*$, with diversion as a function of model parameters $\psi = (\theta, \delta, \beta)$ denoted $\mathcal{D}_{jkm}(\psi)$.[44]

[43] App. E.3 proposes a conformant, but not efficient, GMM estimator that does not suffer from integration bias.

[44] We highlight diversion with respect to ξ_{jm} rather than x_{jm} because in our DGP consumers can have both positive and negative preferences for both observed characteristics. While this degree of horizontal differentiation is reasonable for many empirically relevant characteristics, an implication is that $\partial_{x_{jm}}\pi_{jm}$ may be zero resulting in diversion being

Experiment	Specification	CLEER 11 nodes	GMM-M	MDLE	CLEER 19 nodes	Logit
Baseline		1.039	1.498	1.304		27.636
		[1.651]	[2.919]	[2.032]		[31.093]
Vary I_m	$I_m = 250$	1.286	1.557	2.292		27.663
		[2.291]	[2.897]	[4.035]		[31.003]
	$I_m = 4{,}000$	0.829	1.517	0.863		27.606
		[1.172]	[2.907]	[1.238]		[30.820]
Vary M	$M = 10$	1.208	3.079	1.298		27.353
		[2.098]	[7.018]	[2.175]		[35.567]
	$M = 1{,}000$	0.725	0.737	1.231		27.081
		[0.999]	[1.021]	[2.105]		[34.814]
Vary (θ^z, θ^ν)	$\theta^z = 0.3, \theta^\nu = 0.3$	0.644	0.953	0.789		2.712
		[1.133]	[1.739]	[1.359]		[3.428]
	$\theta^z = 1.0, \theta^\nu = 0.3$	0.607	1.112	0.652		14.854
		[0.900]	[1.998]	[0.985]		[17.039]
	$\theta^z = 2.0, \theta^\nu = 0.3$	1.343	2.075	1.358		60.232
		[1.671]	[3.744]	[1.721]		[65.964]
	$\theta^z = 0.3, \theta^\nu = 1.0$	1.073	1.197	2.614		14.874
		[2.118]	[2.408]	[4.682]		[17.399]
	$\theta^z = 2.0, \theta^\nu = 1.0$	1.768	2.863	1.878		76.397
		[2.404]	[5.228]	[2.577]		[83.200]
	$\theta^z = 0.3, \theta^\nu = 2.0$	2.640	2.427	15.302		60.433
		[5.038]	[4.522]	[22.650]		[66.293]
	$\theta^z = 1.0, \theta^\nu = 2.0$	2.645	2.863	5.338		76.330
		[4.569]	[5.367]	[8.483]		[82.870]
	$\theta^z = 2.0, \theta^\nu = 2.0$	3.297	4.257	4.227		132.494
		[4.800]	[7.605]	[6.512]		[143.081]
Vary a	$a = 0.15$	1.200	7.959	1.415		29.214
		[1.969]	[14.937]	[2.299]		[32.918]
	$a = 1.00$	0.808	0.949	1.087		23.023
		[1.206]	[1.506]	[1.778]		[25.670]
Integration	$a = 0.50, \theta^\nu = 2.0$	2.631	2.856		2.497	76.109
Bias		[4.524]	[5.578]		[4.186]	[82.823]
	$a = 0.50, \theta^\nu = 2.5$	5.492	3.871		3.596	118.481
		[8.533]	[6.914]		[5.869]	[127.526]
	$a = 0.50, \theta^\nu = 3.0$	12.901	5.272		5.365	172.998
		[17.299]	[9.231]		[8.978]	[185.756]

Note: 90th percentile is presented in brackets.

Table 3: *Median and 90th percentile of $\mathfrak{D}(\hat{\psi})$ across estimators and experiments*

We compare estimators based on the following MAE-based summary statistic,

$$\mathfrak{D}(\hat{\psi}) = 100 \cdot M^{-1} \sum_{m=1}^{M} \sum_{j=1}^{J_m} \frac{s_{jm}}{1 - s_{0m}} \sum_{k=0}^{J_m} |\mathcal{D}^*_{jkm} - \mathcal{D}_{jkm}(\hat{\psi})|. \tag{28}$$

In words, $\mathfrak{D}(\hat{\psi})$ is the aggregate absolute error in diversion from inside good j to all other good (including the outside good) averaged over all products j weighting by their inside share. $\mathfrak{D}(\hat{\psi}) \in [0, 200)$ due to the properties of $\mathcal{D}_{jkm}(\psi)$. We weight by inside share so that the more popular products (which tend to be more important in an antitrust setting) receive more weight.[45]

undefined for some products.Moreover, diversion will be numerically unstable when $\partial_{x^{\cdot}_{jm}} \pi_{jm}$ is small. In contrast, utility is monotone in ξ_{jm} across all consumers—like price in most scenarios—resulting in well behaved diversion ratios between 0 and 1.

[45] In practice, diversion must be numerically approximated. Since we know the true distribution of z is normal for our DGP, we compute $\mathfrak{D}(\hat{\psi})$ using precise quadrature rules to integrate over both z and ν. Hence, integration computing $\mathfrak{D}(\hat{\psi})$ is more accurate than that used in estimation. We use the same integration method to compute $\mathfrak{D}(\hat{\psi})$ for all estimators. As a

Tbl. 3 presents the median and 90th percentile of the distribution of $\mathfrak{D}(\hat{\psi})$ over all of the experiments presented in this section. For reference, the final column also provides $\mathfrak{D}(\hat{\psi})$ for the simple logit model, which is straightforwardly calculated from observed market shares since $\hat{\mathcal{D}}^{\text{logit}}_{jkm} = s_{km}/(1 - s_{jm})$. We view this as an index of the degree of unobserved heterogeneity in each underlying DGP.

All estimators considered dramatically out-perform the logit model across all specifications. This is no surprise given the logit model lacks the flexibility to capture the heterogeneity in our DGPs. Our estimators tend to capture substitution patterns well; this is intuitive given they are correctly specified.

Overall, CLEER tends to outperform GMM-M and MDLE. Indeed, CLEER outperforms MDLE across all specifications, although the difference is small in cases where micro data is the dominant source of identification (i.e., when the micro sample is large, there is a significant degree of micro sample variation, and the product level instruments are weak). Comparing CLEER to GMM-M, the only case in the top panel where GMM-M outperforms CLEER is when $\theta^{*z} = 0.3, \theta^{*\nu} = 2$. Here, the integration bias from the likelihood when $\theta^{*\nu}$ is large outweighs the advantage from leveraging the micro sample when θ^{*z} is small. The bottom panel shows that integration bias continues to be exacerbated as we set $\theta^{*\nu}$ even higher. However, increasing the quadrature precision from 11 to 19 nodes largely alleviates the issue. Indeed, at the 90th percentile, CLEER with 19 node quadrature outperforms GMM-M for across all experiments. Finally, it is worth noting that the fact that logit model performs so poorly in the bottom panel suggests that the degree of heterogeneity may be higher than is empirically relevant.

8 Conclusion

Random coefficients discrete choice demand models are a workhorse of applied industrial organization. In this paper, we propose CLEER, which optimally combines the likelihood for purchase data with product level exogeneity restrictions. This estimator does not require additional parametric assumptions relative to a GMM estimator. CLEER is a unified estimator that conforms to a wide variety of data environments and achieves efficiency in each. It has an additional advantage that inference is straightforward and is correct under more general assumptions than the standard approach. Finally, CLEER is computationally tractable, suggesting that it will be directly useful for applied work.

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.

ALLISON, P. D., AND N. A. CHRISTAKIS. 1994. “Logit models for sets of ranked items.” *Sociological methodology*, 199–228.

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.

result, variation across diversion estimators we report in tbl. 3 is entirely due to variation in the distribution of $\hat{\psi}$.

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. 2024. “Nonparametric identification of differentated products demand using micro data.” *Econometrica*, 92(4): 1135–1162.

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

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*.

Conlon, C. T., and J. H. Mortimer. 2013. “Demand estimation under incomplete product availability.” *American Economic Journal: Microeconomics*, 5(4): 1–30.

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.

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

Dubé, J.-P., J. T. Fox, and C.-L. Su. 2012. “Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation.” *Econometrica*, 80(5): 2231–2267.

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*a*. “The Evolution of Market Power in the U.S. Automobile Industry.” *working paper*.

Grieco, P., C. Murry, J. Pinkse, and S. Sagl. 2023*b*. “Conformant and efficient estimation of discrete choice demand models.” Penn State working paper.

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.

Hausman, J. A., and D. A. Wise. 1978. “A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences.” *Econometrica: Journal of the econometric society*, 403–426.

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.

McFadden, D. 1974. “Conditional logit analysis of qualitative choice behavior.” In *Frontiers in Econometrics*. Chapter 4, 105–142.

McFadden, D. 1989. “A method of simulated moments for estimation of discrete response models without numerical integration.” *Econometrica: Journal of the Econometric Society*, 995–1026.

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

Moon, H. R., M. Shum, and M. Weidner. 2018. “Estimation of random coefficients logit demand models with interactive fixed effects.” *Journal of Econometrics*, 206(2): 613–644.

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.

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

STOCK, J. H., J. H. WRIGHT, AND M. YOGO. 2002. "A survey of weak instruments and weak identification in generalized method of moments." *Journal of Business & Economic Statistics*, 20(4): 518–529.

STUDENT. 1908. "The probable error of a mean." *Biometrika*, 1–25.

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.

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

A Consumer demographics

The paper and proof assume that the distribution of consumer demographics G is observed and constant across markets. It moreover assumes that the micro sample is a random sample from the population. In this appendix, we show how both of these assumptions can be relaxed.

A.1 Estimation of G or G_m

If we maintain that G does not vary across markets, an estimator of G using a sample size that grows at rate faster than M and I can be used in its place without affecting the asymptotics. We view this as a typical case when one uses a population census (such as the Decennial Census) or large demographic survey (such as the Current Population Survey or American Community Survey) to estimate the superpopulation. One may also want to allow the superpopulation G_m to vary by market. In this case, it is reasonable to assume that the fastest possible estimator of G_m converges at rate N_m since this is the size of the observed population.

Regardless of how G_m is (consistently) estimated, our estimator remains consistent, but in some cases inference would need to be adjusted to account for estimation error in $\hat{G}_m$. If I_m and M are small relative to N_m—a common case—estimation of G_m will again be asymptotically negligible. If I_m is large relative to N_m and a market-specific G_m is estimated at rate $\sqrt{N_m}$, our estimator's inference procedure would need to be adjusted. Alternatively, one may consider an alternative estimator of the (finite) population analog of G_m, in which one accounts for inclusion in the micro sample when integrating out z_{im} for individuals outside the micro sample. However, this is beyond the scope of this paper.

A.2 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

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

This allows for a variety of forms of selection. Clearly, random selection poses no difficulty as in this case $\pi_{jm}^{D=0} = \mathbb{P}(D_{im} = 0)\pi_{jm}$, leading to the loglikelihood presented in (6) (up to a constant).

Interestingly, deterministic selection based on y_{im} 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, $\mathbb{P}(D_{im} = 1 \cap y_{ijm} = 1 \mid z_{im}) = \mathbb{P}(D_{im}^{*} = 1)\pi_{jm}^{z_{im}}\mathbb{1}(j \in \mathbb{J})$, so we have

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

Moreover, in both of the above cases, because only logarithms of the choice probabilities appear in the loglikelihood, the $\mathbb{P}(D_{im}^{*} = 0)$ 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,

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

More general forms would have to be explicitly modeled and are outside the scope of this paper. For example, we are unable to write down a likelihood that incorporates selection into the micro sample based on ν (e.g., taste for sugar) without further assumptions. However, these assumptions would *also* be needed to form micro moments that would address to this form of selection.

Selection into the micro sample is closely related to the issue of compatibility described in CG23 which is visible when there are discrepancies between population and micro sample distributions. They propose assuming compatibility of specific moments for use in estimation when other moments indicate that the micro sample and population distributions are incompatible. See CG23 for examples.

B Lemmas for proofs of consistency and asymptotic normality

This appendix contains the primary supporing lemmas used in theorems 1 and 2. It concludes with L5 which establishes consistency of $\hat{\pi}$, $\hat{\delta}$, and $\hat{\beta}$. L9, 10 and 12 to 15, which are referenced here, are relegated to apps. J.2 and J.3.

B.1 Consistency

Lemma 2 ($\hat{\Phi}$, Φ approximations)**.** Given our assumptions, (a) $\sup_{\Theta_\epsilon^c \times \Pi^\kappa} |\Phi(\theta,\pi) - \Phi(\theta,\pi^*)|\mathbb{1}[\rho^{\blacksquare}(\pi) \le \rho_{\mathrm{id}}(\theta)\eta] \;/\; \rho_{\mathrm{id}}(\theta) \prec 1$; (b) $\sup_{\Theta_\epsilon^c \times \Pi^\kappa} |\Delta\hat{\Phi}(\theta,\pi)| \;/\; \rho_D(\theta,\pi) \prec 1$; (c) $\Phi(\theta^*, s) \preceq 1$, $\hat{\Phi}(\theta^*, s) \preceq 1$, and $\hat{\Phi}(\theta^*, \pi^*) \preceq 1$.

Proof. First recall that $\hat{\Phi}(\theta,\pi) = \|\mathcal{P}\delta(\theta,\pi)\|^2 \,/\, 2$ and $\Phi(\theta,\pi) = \|\mathcal{P}[\delta(\theta,\pi) - \delta(\theta^*,\pi^*)]\|^2 \,/\, 2$. For (a), rearranging terms,[46] we apply the triangle and Schwarz inequalities to obtain,

$$|\Phi(\theta,\pi) - \Phi(\theta,\pi^*)| \leq \|\mathcal{P}[\delta(\theta,\pi) - \delta(\theta,\pi^*)]\|^2 + \|\mathcal{P}[\delta(\theta,\pi) - \delta(\theta,\pi^*)]\| \cdot \|\mathcal{P}[\delta(\theta,\pi^*) - \delta(\theta^*,\pi^*)]\|. \quad (29)$$

Let $\tilde{B} = B(B^{\triangledown}B)^{-1/2}$. By the MVT, $\|\mathcal{P}[\delta(\theta,\pi) - \delta(\theta,\pi^*)]\| = \|\mathcal{P}\mathbb{D}_\pi(\theta,\mathring{\pi})(\pi - \pi^*)\|$ for some $\mathring{\pi}$. Now, since $\mathcal{P}\tilde{B}\tilde{B}^{\triangledown} = \mathcal{P}$, $\tilde{B}^{\triangledown}\mathcal{P}\tilde{B} \leq \tilde{B}^{\triangledown}\tilde{B} = \mathbb{I}$, and hence $\mathcal{P}^{\triangledown}\mathcal{P} = \tilde{B}(\tilde{B}^{\triangledown}\mathcal{P}\tilde{B})\tilde{B}^{\triangledown} \leq \tilde{B}\tilde{B}^{\triangledown}$,[47]

$$\begin{aligned}\|\mathcal{P}[\delta(\theta,\pi) - \delta(\theta,\pi^*)]\| \leq \|\tilde{B}^{\triangledown}\mathbb{D}_\pi(\theta,\mathring{\pi})(\pi - \pi^*)\| &\overset{\text{triangle}}{\leq} \|\tilde{B}^{\triangledown}\mathbb{D}_\pi(\theta,\mathring{\pi})(\pi - s)\| + \|\tilde{B}^{\triangledown}\mathbb{D}_\pi(\theta,\mathring{\pi})(s - \pi^*)\| \\ &\overset{\text{Schwarz}}{\leq} \sqrt{\rho^{\blacksquare}(\pi)}\|\tilde{B}^{\triangledown}\mathbb{D}_\pi(\theta,\mathring{\pi})\|_N + \|s - \pi^*\| \cdot \|\tilde{B}^{\triangledown}\mathbb{D}_\pi(\theta,\mathring{\pi})\|, \quad (30)\end{aligned}$$

where $\|x\|_N = \sqrt{\sum_m \|x_m\|^2 \,/\, N_m}$ and $\rho^{\blacksquare}(\pi)$ as defined in theorem 1, step *3*. Now, by F[iii]

$$\sup_{\Theta_\epsilon^c \times \sqcap^\kappa} \|\tilde{B}^{\triangledown}\mathbb{D}_\pi(\theta,\pi)\| \leq \boldsymbol{\lambda}_{\max}\Big[\Big(\frac{B^{\triangledown}B}{M}\Big)^{-1/2}\Big] \sup_{\Theta_\epsilon^c \times \sqcap^\kappa} \sqrt{\frac{1}{M}\sum_m \|B_m\|^2\|\mathbb{D}_{\pi m}(\theta,\pi_m)\|^2} \overset{\text{L9(d)}}{\preceq} \kappa^{-3}. \quad (31)$$

Moreover, $\sup_{\Theta_\epsilon^c \times \sqcap^\kappa} \|\tilde{B}^{\triangledown}\mathbb{D}_\pi(\theta,\pi)\|_N \leq \rho_u \sup_{\Theta_\epsilon^c \times \sqcap^\kappa} \|\tilde{B}^{\triangledown}\mathbb{D}_\pi(\theta,\pi)\| \overset{\text{L9(d)}}{\preceq} \rho_u\kappa^{-3}$, where $\rho_u = 1/\min_m \sqrt{N_m}$. Consequently, since $\mathbb{E}\|s - \pi^*\|^2 \preceq \sum_m N_m^{-1} = \rho_N$, substituting into (30),

$$\|\mathcal{P}[\delta(\theta,\pi) - \delta(\theta,\pi^*)]\| \leq \sqrt{\rho^{\blacksquare}(\pi)}O_p(\rho_u\kappa^{-3}) + O_p(\sqrt{\rho_N}\kappa^{-3}). \quad (32)$$

Thus,

$$\begin{aligned}\sup_{\Theta_\epsilon^c \times \sqcap^\kappa} \frac{\|\mathcal{P}[\delta(\theta,\pi) - \delta(\theta,\pi^*)]\|\mathbb{1}[\rho^{\blacksquare}(\pi) \leq \rho_{\text{id}}(\theta)\eta]}{\sqrt{\rho_{\text{id}}(\theta)}} &\preceq \sup_{\Theta_\epsilon^c \times \sqcap^\kappa} \frac{\rho_u\sqrt{\rho^{\blacksquare}(\pi)}\mathbb{1}[\rho^{\blacksquare}(\pi) \leq \rho_{\text{id}}(\theta)\eta] + \sqrt{\rho_N}}{\kappa^3\sqrt{\rho_{\text{id}}(\theta)}} \\ &\leq \frac{\rho_u\sqrt{\eta}}{\kappa^3} + \sqrt{\frac{\rho_N}{\kappa^6\breve{\rho}_{\text{id}}(\epsilon)}} \overset{\text{C,I}}{\prec} 1. \quad (33)\end{aligned}$$

Finally, from (29) we obtain

$$\sup_{\Theta_\epsilon^c \times \sqcap^\kappa} \frac{|\Phi(\theta,\pi) - \Phi(\theta,\pi^*)|\mathbb{1}[\rho^{\blacksquare}(\pi) \leq \rho_{\text{id}}(\theta)\eta]}{\rho_{\text{id}}(\theta)} \overset{\text{(33),B}}{\leq} o_p(1) + o_p(1)\sup_{\Theta_\epsilon^c} \underbrace{\sqrt{\frac{\rho^{\Phi}(\theta)}{\rho_{\text{id}}(\theta)}}}_{\leq 1} \prec 1,$$

which concludes the proof of (a).

Now (b). Let $\delta = \delta(\theta,\pi)$ and $\delta^* = \delta(\theta^*,\pi^*)$. First, $\|\mathcal{P}\delta^*\| = \|\mathcal{P}\xi\| \preceq 1$. Reusing fn. 46 with $x = \mathcal{P}(\delta + \delta^*)$, $y = \mathcal{P}\delta$, and $z = \mathcal{P}\delta^*$ and applying the triangle and Schwarz inequalities,

$$\begin{aligned}|\hat{\Phi}(\theta,\pi) - \Phi(\theta,\pi)| = \big|\|\mathcal{P}\delta\|^2 - \|\mathcal{P}(\delta - \delta^*)\|^2\big| \,/\, 2 \leq \\ \|\mathcal{P}\delta^*\|^2 + \|\mathcal{P}\delta^*\| \cdot \|\mathcal{P}(\delta - \delta^*)\| = O_p(1)[O_p(1) + \|\mathcal{P}(\delta - \delta^*)\|]. \quad (34)\end{aligned}$$

By the triangle inequality $\|\mathcal{P}(\delta - \delta^*)\| \leq \|\mathcal{P}[\delta(\theta,\pi) - \delta(\theta,\pi^*)]\| + \|\mathcal{P}[\delta(\theta,\pi^*) - \delta^*]\|$. Working with the first term, by (32) and because $\rho_D(\theta,\pi) \geq \sqrt{\eta^3\rho_{\text{id}}(\theta)\rho^{\blacksquare}(\pi)}$ and $\rho_D(\theta,\pi) \geq \eta^2\rho_{\text{id}}(\theta) = \kappa^6\rho_{\text{id}}(\theta)$,

[46] We use the equality $\|x - z\|^2 - \|y - z\|^2 = \|x - y\|^2 + 2(x - y)^{\triangledown}(y - z)$ with $x = \mathcal{P}\delta(\theta,\pi)$, $y = \mathcal{P}\delta(\theta,\pi^*)$, and $z = \mathcal{P}\delta(\theta^*,\pi^*)$.

[47] Using $\|\mathcal{P}\tilde{B}x\|^2 = x^{\triangledown}\tilde{B}^{\triangledown}\mathcal{P}\tilde{B}x \leq \|x\|^2$.

$$\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\frac{\|\mathcal{P}[\delta(\theta,\pi)-\delta(\theta,\pi^*)]\|}{\rho_D(\theta,\pi)}\leq\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\frac{\sqrt{\rho^{\blacksquare}(\pi)}O_p(\rho_u)+O_p(\sqrt{\rho_N})}{\rho_D(\theta,\pi)\kappa^3}\overset{\text{C,I}}{\leq}\sup_{\Theta_\epsilon^c}\frac{O_p(\rho_u)}{\kappa^3\sqrt{\eta^3\rho_{\mathrm{id}}(\theta)}}+o_p(1)\overset{\text{C}}{\prec}1.$$

For the second term,

$$\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\frac{\|\mathcal{P}[\delta(\theta,\pi^*)-\delta^*]\|}{\rho_D(\theta,\pi)}\lesssim\sup_{\Theta_\epsilon^c}\frac{\sqrt{\rho^{\Phi}(\theta)}}{\eta^2\rho_{\mathrm{id}}(\theta)}\leq\frac{1}{\eta^2\sqrt{\breve{\rho}_{\mathrm{id}}(\epsilon)}}\overset{\text{C}}{\prec}1.$$

Combining terms, we have the asserted result. Finally, (c):

$$\begin{aligned}2\Phi(\theta^*,s)=\|\mathcal{P}[\delta(\theta^*,s)-\delta^*]\|^2&\leq\|\tilde{B}^{\nabla}[\delta(\theta^*,s)-\delta^*]\|^2\\&\overset{\text{MVT}}{\leq}\max_{0\leq t\leq 1}\|\tilde{B}^{\nabla}\mathbb{D}_\pi[\theta^*,\pi^*+t(s-\pi^*)](s-\pi^*)\|^2\overset{\text{L9(d)}}{\lesssim}\kappa^{-6}\rho_N\overset{\text{I}}{\lesssim}1,\end{aligned}$$

so $\Phi(\theta^*,s)\lesssim 1$. For $\hat{\Phi}(\theta^*,s)$, we reuse (34) to obtain $\hat{\Phi}(\theta^*,s)\leq|\hat{\Phi}(\theta^*,s)-\Phi(\theta^*,s)|+\Phi(\theta^*,s)\lesssim\sqrt{\Phi(\theta^*,s)+1}+\Phi(\theta^*,s)\lesssim 1$. Finally, because $\Phi(\theta^*,\pi^*)=0$ we similarly get $\hat{\Phi}(\theta^*,\pi^*)\lesssim 1$. □

Lemma 3 ($\hat{\mathcal{L}}^{\blacklozenge},\mathcal{L}^{\blacklozenge}$ **approximations).** Given our assumptions, (a) $\sup_{\Theta_\epsilon^c\times\Pi^\kappa}|\mathcal{L}^{\blacklozenge}(\theta,\pi)-\mathcal{L}^{\blacklozenge}(\theta,\pi^*)|\mathbb{1}[\rho^{\blacksquare}(\pi)\leq\rho_{\mathrm{id}}(\theta)\eta]\;/\;\rho_{\mathrm{id}}(\theta)\prec 1$; (b) $\sup_{\Theta_\epsilon^c\times\Pi^\kappa}|\Delta\hat{\mathcal{L}}^{\blacklozenge}(\theta,\pi)|\;/\;\rho_D(\theta,\pi)\prec 1$; (c) $\sup_{\Theta_\epsilon^c\times\Pi^\kappa}|\hat{\mathcal{L}}^{\blacklozenge}(\theta^*,s)|\;/\;\rho_D(\theta,\pi)\lesssim|\hat{\mathcal{L}}^{\blacklozenge}(\theta^*,s)|\;/\;[\eta^2\breve{\rho}_{\mathrm{id}}(\epsilon)]\prec 1$.

Proof. For (a), we break up the problem into two smaller ones. Let $\tau_m(\pi)=\mathbb{1}(\|\pi_m-\pi_m^*\|\leq\kappa)$ and $\alpha(\theta,\pi)=\mathbb{1}[\rho^{\blacksquare}(\pi)\leq\rho_{\mathrm{id}}(\theta)\eta]$. We first show that $\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\sum_m(1-\tau_m(\pi))\mathcal{L}_m^{\blacklozenge}(\theta,\pi_m)\alpha(\theta,\pi)\;/\;\rho_{\mathrm{id}}(\theta)\prec 1$. Noting that $\max_m\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\mathcal{L}_m^{\blacklozenge}(\theta,\pi)/N_m\overset{\text{L13}}{\lesssim}1$ and dropping the arguments from α,τ_m, we have uniformly over $\Theta_\epsilon^c\times\Pi^\kappa$,[48]

$$\begin{aligned}\underbrace{\frac{\alpha\sum_m N_m(1-\tau_m)}{\rho_{\mathrm{id}}(\theta)}}_{\text{①}}\leq\frac{\alpha\sum_m N_m(1-\tau_m)\|\pi_m-\pi_m^*\|^2}{\kappa^2\rho_{\mathrm{id}}(\theta)}&\leq\frac{2\alpha\sum_m N_m(1-\tau_m)\big(\|\pi_m-s_m\|^2+\|s_m-\pi_m^*\|^2\big)}{\kappa^2\rho_{\mathrm{id}}(\theta)}\\&\overset{\text{L12(d)}}{\leq}\underbrace{\frac{2\alpha\rho^{\blacksquare}(\pi)}{\kappa^2\rho_{\mathrm{id}}(\theta)}}_{\text{②}}+o_p(1)\underbrace{\alpha\frac{(\log M)^2}{\kappa^2}\sum_m\frac{(1-\tau_m)}{\rho_{\mathrm{id}}(\theta)}}_{\text{③}},\end{aligned}$$

recalling that $\rho^{\blacksquare}=\sum_m N_m\|\pi_m-s_m\|^2$. Now, use the definition of α and recall $\eta=\kappa^3$ such that ② $\leq 2\kappa\prec 1$. Further, ③ is negligible (relative to ①) because $(\log M)^2/\kappa^2\overset{\text{I}}{\prec}\min_m N_m$. So $\sup_{\Theta_\epsilon^c\times\Pi^\kappa}$ ① $\prec 1$. That leaves us with

$$\begin{aligned}\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\frac{\alpha\sum_m\tau_m|\mathcal{L}_m^{\blacklozenge}(\theta,\pi_m)-\mathcal{L}_m^{\blacklozenge}(\theta,\pi_m^*)|}{\rho_{\mathrm{id}}(\theta)}&\overset{\substack{\text{MVT}\\\text{Scwz.}}}{\leq}\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\frac{\alpha\sum_m\tau_m\|\pi_m-\pi_m^*\|\cdot\|\mathcal{L}_{\pi m}^{\blacklozenge}(\theta,\mathring{\pi}_m)\|}{\rho_{\mathrm{id}}(\theta)}\\&\overset{\text{Schwarz}}{\leq}\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\sqrt{\frac{\alpha\sum_m N_m\|\pi_m-\pi_m^*\|^2}{\rho_{\mathrm{id}}(\theta)}}\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\sqrt{\sum_m\frac{\alpha\tau_m\|\mathcal{L}_{\pi m}^{\blacklozenge}(\theta,\pi_m)\|^2}{N_m\rho_{\mathrm{id}}(\theta)}}\lesssim\sqrt{\eta}\times\frac{1}{\sqrt{\breve{\rho}_{\mathrm{id}}(\epsilon)}}\overset{\text{C}}{\prec}1,\end{aligned}$$

where the penultimate inequality follows from L12(d), L15(e), and L9(f). This completes (a).

Now (b). We can replace $\Delta\hat{\mathcal{L}}^{\blacklozenge}(\theta,\pi)$ with $\Delta\hat{\mathcal{L}}^{\blacklozenge}(\theta,\pi)-\Delta\hat{\mathcal{L}}^{\blacklozenge}(\theta,\pi^*)$ because

$$\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\frac{|\Delta\hat{\mathcal{L}}^{\blacklozenge}(\theta,\pi^*)|}{\rho_D(\theta,\pi)}\overset{\text{L10(a)}}{\prec}\sup_{\Theta_\epsilon^c\times\Pi^\kappa}\frac{\sqrt{\rho_{\mathrm{id}}(\theta)}}{\rho_D(\theta,\pi)}\log^2 I_+\lesssim\frac{\log^2 I_+}{\eta^2\breve{\rho}_{\mathrm{id}}(\epsilon)}\overset{\text{C}}{\prec}1.$$

[48] For the second inequality, note that $\|a+b\|^2=2(\|a\|^2+\|b\|^2)-\|a-b\|^2$ for $a=\pi_m-s_m, b=s_m-\pi_m^*$.

Thus, since $\|\theta^z\| \leq C_\epsilon \|\theta - \theta^\ast\|_\lambda$ by L12(b), we have

$$\begin{aligned}\sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \frac{|\varDelta\hat{\mathcal{L}}^{\blacklozenge}(\theta,\pi) - \varDelta\hat{\mathcal{L}}^{\blacklozenge}(\theta,\pi^\ast)|}{\rho_D(\theta,\pi)} &\overset{\substack{\text{MVT}\\ \text{Scwz.}}}{\leq} C_\epsilon \sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \frac{\|\theta-\theta^\ast\|_\lambda \cdot \|\varDelta\hat{\mathcal{L}}^{\blacklozenge}_{\theta^z\pi}(\mathring{\theta}^z, \theta^\nu, \mathring{\pi})(\pi - \pi^\ast)\|}{\rho_D(\theta,\pi)} \\ &\overset{\text{Schwarz}}{\leq} C_\epsilon \sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \frac{\sqrt{\sum_m \|\theta-\theta^\ast\|_\lambda^2 \cdot \|\varDelta\hat{\mathcal{L}}^{\blacklozenge}_{\theta^z\pi m}(\mathring{\theta}^z, \theta^\nu, \mathring{\pi})\|^2 \cdot \|\pi_m - \pi_m^\ast\|^2}}{\rho_D(\theta,\pi)} \\ &\overset{\text{L10(b)}}{\prec} \kappa^{-6}(\log I_+)^3 \sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \frac{\sqrt{\sum_m I_m \|\theta-\theta^\ast\|_\lambda^2 \cdot \|\pi_m - \pi_m^\ast\|^2}}{\rho_D(\theta,\pi)} =: \textcircled{4}.\end{aligned}$$

Applying fn. 48, we can thus break up ④ into two terms. The term corresponding to $s_m - \pi_m^\ast$, is by L12(d) bounded above by

$$\rho_u \kappa^{-6}(\log I_+)^3(\log M) \sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \frac{\sqrt{\rho_{\text{id}}(\theta)}}{\eta^2 \rho_{\text{id}}(\theta)} \overset{\eta=\kappa^3}{\preceq} \frac{\rho_u (\log I_+)^3 \log M}{\kappa^{12}\sqrt{\breve{\rho}_{\text{id}}(\epsilon)}} \overset{\text{C,I[i]}}{\prec} 1.$$

Now the $\pi_m - s_m$ component,

$$\begin{aligned}(\log I_+)^3 \sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \frac{\sqrt{\sum_m I_m \|\theta-\theta^\ast\|_\lambda^2 \cdot \|\pi_m - s_m\|^2}}{\kappa^6 \rho_D(\theta,\pi)} &\overset{\text{I[ii]}}{\preceq} (\log I_+)^3 \rho_u \sup_{\Theta_\epsilon^c \times \mathbb{\Pi}^\kappa} \frac{\sqrt{\rho_{\text{id}}(\theta)\rho^{\blacksquare}(\pi)}}{\kappa^6 \sqrt{\eta^3 \rho_{\text{id}}(\theta)\rho^{\blacksquare}(\pi)}} \\ &\overset{\eta=\kappa^3}{=} \frac{(\log I_+)^3 \rho_u}{\kappa^{10.5}} \overset{\text{C,I[i]}}{\prec} 1.\end{aligned}$$

Hence, (b) holds.

Finally, (c). Recall that $\rho_D(\theta,\pi) = \eta \max(\eta\rho_{\text{id}}(\theta), \rho^{\blacksquare}(\pi)) \geq \eta^2 \breve{\rho}_{\text{id}}(\epsilon)$ for all $\theta \in \Theta_\epsilon^c$, so we focus on the final inequality of the statement. For $\mathring{\pi}$ between s and $\pi^\ast$,

$$\frac{|\hat{\mathcal{L}}^{\blacklozenge}(\theta^\ast, s)|}{\eta^2 \breve{\rho}_{\text{id}}(\epsilon)} \overset{\substack{\text{MVT}\\ \text{tri.}}}{\leq} \overbrace{\left|(s-\pi^\ast)^{\triangledown} \frac{\hat{\mathcal{L}}^{\blacklozenge}_\pi(\theta^\ast,\pi^\ast)}{\eta^2 \breve{\rho}_{\text{id}}(\epsilon)}\right|}^{\textcircled{x}} + \overbrace{\frac{|(s-\pi^\ast)^{\triangledown} \hat{\mathcal{L}}^{\blacklozenge}_{\pi\pi}(\theta^\ast, \mathring{\pi})(s-\pi^\ast)|}{\eta^2 \breve{\rho}_{\text{id}}(\epsilon)}}^{\textcircled{y}}.$$

We show separately that ⓧ and ⓨ vanish. First, by Schwarz, ⓧ is bounded above by $\|s-\pi^\ast\| \cdot \|\hat{\mathcal{L}}^{\blacklozenge\ast}_\pi\| \,/\, [\eta^2 \breve{\rho}_{\text{id}}(\epsilon)]$. Now, since $s_{jm} = N_m^{-1} \sum_i y_{ijm}$ and $\mathbb{V} y_{ijm} = \pi_{jm}(1-\pi_{jm})$, $\mathbb{E}\|s-\pi^\ast\|^2 = \sum_{mj} \mathbb{E}|s_{jm} - \pi_{jm}^\ast|^2 = \sum_{mj} N_m^{-2} \sum_i \mathbb{E}[\pi_{jm}^\ast(1-\pi_{jm}^\ast)] \leq \rho_N \overset{\text{I}}{\prec} 1$. Further, by the information matrix equality, we show ⓧ $\prec 1$ since,

$$\frac{\mathbb{E}\big(\|\hat{\mathcal{L}}^{\blacklozenge\ast}_\pi\|^2 \mid \mathbb{A}\big)}{\eta^4 \breve{\rho}^2_{\text{id}}(\epsilon)} = \frac{\text{tr}(\mathcal{L}^{\blacklozenge\ast}_{\pi\pi})}{\eta^4 \breve{\rho}^2_{\text{id}}(\epsilon)} \overset{\text{L15(f)}}{\preceq} \frac{I\lambda^2}{\kappa^{12} \breve{\rho}^2_{\text{id}}(\epsilon)} \overset{\text{C}}{\preceq} \frac{1}{\kappa^{12} \breve{\rho}_{\text{id}}(\epsilon)} \overset{\text{C}}{\prec} 1.$$

That leaves ⓨ. Let $\mathbb{\Pi}_m^{\text{nbh}} = \{\pi_m \colon \sqrt{N_m}\|\pi_m - \pi_m^\ast\| \leq \log M\}$. Noting that $\forall m \colon \|\mathring{\pi}_m - \pi_m^\ast\| \leq \|s_m - \pi_m^\ast\|$, such that by L12(d), $\mathbb{P}(\exists m \colon \mathring{\pi}_m \notin \mathbb{\Pi}_m^{\text{nbh}}) \prec 1$. By Schwarz, recalling $\lambda = \|\theta^{\ast z}\|$, $\hat{\mathcal{L}} = \varDelta\hat{\mathcal{L}} + \mathcal{L}$,

$$\textcircled{y} \preceq \|s-\pi^\ast\|^2 \max_m \max_{\mathbb{\Pi}_m^{\text{nbh}}} \frac{\|\hat{\mathcal{L}}^{\blacklozenge}_{\pi\pi m}(\theta^\ast, \pi_m)\|}{\eta^2 \breve{\rho}_{\text{id}}(\epsilon)} \overset{\text{MVT:}\theta^z}{\leq} \frac{\rho_N}{\eta^2 \breve{\rho}_{\text{id}}(\epsilon)} \max_m \max_{\mathbb{\Pi}_m^{\text{nbh}}} \big(\textcircled{a}_m + \textcircled{b}_m + \textcircled{c}_m\big), \tag{35}$$

where $\textcircled{a}_m = \|\hat{\mathcal{L}}^{\blacklozenge}_{\pi\pi m}(0, \theta^{\nu\ast}, \pi_m)\| \overset{\text{L15(a)}}{=} 0$, $\textcircled{b}_m = \lambda \max_{0 \leq t \leq 1} \|\varDelta\hat{\mathcal{L}}^{\blacklozenge}_{\pi\pi\theta^z m}(t\theta^{\ast z}, \theta^{\nu\ast}, \pi_m)\|$, and $\textcircled{c}_m = \lambda \max_{0 \leq t \leq 1} \|\mathcal{L}^{\blacklozenge}_{\pi\pi\theta^z m}(t\theta^{\ast z}, \theta^{\nu\ast}, \pi_m)\|$. Finally,

$$\frac{\rho_N}{\eta^2 \breve{\rho}_{\mathrm{id}}(\epsilon)} \max_m \max_{\mathbb{\Pi}_m^{\mathrm{nbh}}} \textcircled{b}_m \overset{\mathrm{L10(b)}}{\prec} \max_m \frac{\sqrt{I_m \lambda^2}}{\sqrt{\breve{\rho}_{\mathrm{id}}(\epsilon)}} \frac{\rho_N (\log I_+)^3}{\eta^2 \kappa^9 \sqrt{\breve{\rho}_{\mathrm{id}}(\epsilon)}} \overset{\substack{\mathrm{C,I} \\ \eta = \kappa^3}}{\prec} 1;$$

$$\frac{\rho_N}{\eta^2 \breve{\rho}_{\mathrm{id}}(\epsilon)} \max_m \max_{\mathbb{\Pi}_m^{\mathrm{nbh}}} \textcircled{c}_m \overset{\mathrm{L15(f)}}{\prec} \frac{\rho_N}{\eta^2 \breve{\rho}_{\mathrm{id}}(\epsilon)} \max_m (I_m \lambda^2) \overset{\eta = \kappa^3}{\prec} \frac{\rho_N}{\kappa^6} \overset{\mathrm{I[i]}}{\prec} 1,$$

So $\textcircled{y} \prec 1$, which completes the proof of (c). □

The following lemma establishes step *6* of theorem 1, showing that the restriction of steps *1* to *5* to $\mathbb{\Pi}^\kappa \subset \mathbb{\Pi}$ is without loss of generality.

Lemma 4 (step 6). With probability approaching one, $(\hat{\theta}, \hat{\pi}) \in \Theta \times \mathbb{\Pi}^\kappa$.

Proof. It is sufficient to show that $\lim_{M \to \infty} \mathbb{P}\big[\inf_{\Theta \times \mathbb{\Pi}^{\kappa c}} \hat{\Omega}(\theta, \pi) - \hat{\Omega}(\theta^*, \pi^*) \leq 0\big] = 0$. Since $\hat{\Phi}(\theta^*, \pi^*) \preceq 1$ by L2(c) and $\hat{\Phi}(\theta, \pi) \geq 0$ by definition, we only need to establish that the likelihood difference $\inf_{\Theta \times \mathbb{\Pi}^{\kappa c}} \hat{\mathcal{L}}(\theta, \pi) - \hat{\mathcal{L}}(\theta^*, \pi^*)$ diverges with probability approaching one. For notational parsimony, we tackle the most challenging case, i.e. where $I = N$, which implies,

$$\hat{\mathcal{L}}_m(\theta, \pi_m) = \sum_{ij} y_{ijm} \log \frac{\varsigma^*_{ijm} \pi_{jm}}{\varsigma_{ijm} \pi^*_{jm}} + N_m \sum_j s_{jm} \log \frac{s_{jm}}{\pi_{jm}} = \sum_{ij} y_{ijm} \log \frac{\varsigma^*_{ijm}}{\varsigma_{ijm}} + N_m \sum_j s_{jm} \log \frac{s_{jm}}{\pi^*_{jm}}$$
$$= \sum_{ij} y_{ijm} \log \frac{\varsigma^*_{ijm}}{\varsigma_{ijm}} + \hat{\mathcal{L}}_m(\theta^*, \pi^*).$$

Define $\mathcal{E}_m = \{j \colon \pi_{jm} \geq \kappa\}$. We will split $\hat{\mathcal{L}}_m(\theta, \pi) - \hat{\mathcal{L}}_m(\theta^*, \pi^*_m)$ into terms involving $\mathcal{E}_m$ and terms involving its complement $\mathcal{E}^c_m$ and bound these terms individually.

First, the $\mathcal{E}^c_m$ term. If $\mathcal{E}^c_m$ is empty, this term is zero, so suppose $\mathcal{E}^c_m$ is non-empty. Let $\ell_{ijm}(\theta, \pi_m) = \log \varsigma_{ijm}(\theta, \pi_m)$.By L13, for some fixed $C < \infty$,

$$\sum_i \sum_{\mathcal{E}^c_m} y_{ijm} \log \frac{\varsigma^*_{ijm}}{\varsigma_{ijm}} \overset{\mathrm{L13}}{\geq} \sum_i \sum_{\mathcal{E}^c_m} y_{ijm} \Big[\log \frac{\pi^*_{jm}}{\pi_{jm}} - C\|z_{im}\| - C\Big] = \overbrace{N_m \sum_{\mathcal{E}^c_m} s_{jm} \log \frac{\pi^*_{jm}}{\pi_{jm}}}^{\textcircled{1}_m} - \overbrace{C \sum_i y_{i\mathcal{E}^c m} (\|z_{im}\| + 1)}^{\textcircled{2}_m},$$

with $y_{i\mathcal{E}^c m} = \sum_{\mathcal{E}^c_m} y_{ijm}$. Recalling $\kappa = \kappa_\pi^{4/3} = \exp(-4\kappa_\delta^\uparrow)$, $\min_{jm} \pi^*_{jm}/\kappa = \exp(\kappa_\delta^\uparrow) \min_{jm} \pi^*_{jm}/\kappa_\pi \overset{\mathrm{L9(b)}}{\succ} \exp(\kappa_\delta^\uparrow) \succ 1$. Hence,

$$\textcircled{1}_m \geq N_m \sum_{\mathcal{E}^c_m} s_{jm} \log \frac{\pi^*_{jm}}{\kappa} \geq N_m \sum_{\mathcal{E}^c_m} s_{jm} \kappa_\delta^\uparrow = N_m \Big(\kappa_\delta^\uparrow \sum_{\mathcal{E}^c_m} \pi^*_{jm} - \kappa_\delta^\uparrow \sum_{\mathcal{E}^c_m} (s_{jm} - \pi^*_{jm})\Big) =: N_m \Big(\kappa_\delta^\uparrow \pi^*_{\mathcal{E}^c_m} - \mathcal{T}_{1m}\Big).$$

Recall $\rho_u = 1 / \min_m \sqrt{N_m}$, so $\max_m \mathcal{T}_{1m} \prec \kappa$ since $\max_m \|s_m - \pi^*_m\| \leq \max_m \rho_u \sqrt{N_m} \|s_m - \pi^*_m\| \overset{\mathrm{L12(d)}}{\prec} \rho_u \log M \overset{\mathrm{C,I}}{\prec} \kappa / \kappa_\delta^\uparrow$. Thus, we can treat $N_m \pi^*_{\mathcal{E}^c m} \kappa_\delta^\uparrow$as a lower bound to $\textcircled{1}_m$.

We now turn to $\textcircled{2}_m$,[49]

$$\mathbb{P}\big[\exists m \colon |\textcircled{2}_m - \mathbb{E}(\textcircled{2}_m \mid \mathbb{A})| > N_m \pi^*_{\mathcal{E}^c m}\big] \overset{\mathrm{Bonferroni}}{\leq} \sum_m \mathbb{P}\big[|\textcircled{2}_m - \mathbb{E}(\textcircled{2}_m \mid \mathbb{A})| > N_m \pi^*_{\mathcal{E}^c m}\big] \overset{\mathrm{G[iii],\ subg}}{\leq} 2 \sum_m \exp\Big(-\frac{N_m^2 \pi^{*2}_{\mathcal{E}^c m}}{2 N_m c^*_z}\Big) \prec 1.$$

Further,

$$-\mathbb{E}(\textcircled{2}_m \mid \mathbb{A}) = -C N_m \mathbb{E}\big(y_{i\mathcal{E}^c m} \|z_{im}\| \mid \mathbb{A}\big)$$

[49] Because the sum of i.i.d. subgaussians is subgaussian with OVP that is N_m times as large, $\textcircled{1}_m - \mathbb{E}(\textcircled{2}_m \mid \mathbb{A})$ is subgaussian (conditional on $\mathbb{A}$) with OVP no greater than $N_m c^*_z$.

$$
\begin{aligned}
&= -CN_m\mathbb{E}\big(y_{i\mathcal{E}^c m}\|z_{im}\|\mathbb{1}[\|z_{im}\| \le \kappa_\delta^{\uparrow}/(2C)] \mid \mathbb{A}\big) - CN_m\mathbb{E}\big(y_{i\mathcal{E}^c m}\|z_{im}\|\mathbb{1}[\|z_{im}\| > \kappa_\delta^{\uparrow}/(2C)] \mid \mathbb{A}\big) \\
&\overset{\text{Hölder}}{\ge} -N_m\kappa_\delta^{\uparrow}\pi^*_{\mathcal{E}^c m}/2 - \underbrace{CN_m(\pi^*_{\mathcal{E}^c m})^{1/p_1}[\mathbb{E}(\|z_{im}\|^{p_2})]^{1/p_2}\big[\mathbb{P}(\|z_{im}\| > \kappa_\delta^{\uparrow}/(2C) \mid \mathbb{A})\big]^{1/p_3}}_{\text{will show to be negligible}}, \qquad (36)
\end{aligned}
$$

for any $p_1, p_2, p_3 > 0$ whose reciprocals sum to one. We now show that the last right-hand side term in (36) is negligible compared to the first. We have, $\mathbb{P}(\|z_{im}\| > \kappa_\delta^{\uparrow}/(2C) \mid \mathbb{A}) \overset{\text{G[iii]}}{\le} 2\exp\big[-\kappa_\delta^{\uparrow 2}/(8C^2c_z^*)\big] \overset{\text{C}}{=} 2M^{-c_\xi^*/(C^2c_z^*)}$, whereas $\pi^*_{\mathcal{E}^c m} \overset{\text{L9(b)}}{\succeq} \kappa$, which by C decreases more slowly than any power of M.

So up to negligible terms (uniformly in m), $N_m\kappa_\delta^{\uparrow}\pi^*_{\mathcal{E}^c m}/2$ is an upper bound to $|②_m|$ and hence a lower bound to $①_m - ②_m$ (i.e., the $\mathcal{E}^c_m$ term).

Finally, the $\mathcal{E}_m$ term of $\hat{\mathcal{L}}_m$, i.e. $\sum_i \sum_{\mathcal{E}_m} y_{ijm} \log(\varsigma^*_{ijm}/\varsigma_{ijm})$. We first show that left hand of (37) is negligible, so we can simply consider $\sum_i \sum_{\mathcal{E}_m} \varsigma_{ijm} \log(\varsigma^*_{ijm}/\varsigma_{ijm})$.

$$
\max_m \sup_{\Theta\times\sqcap}\left|\frac{2}{N_m\kappa_\delta^{\uparrow}\pi^*_{\mathcal{E}^c m}}\sum_i\sum_{\mathcal{E}_m}(y_{ijm} - \varsigma^*_{ijm})\log\frac{\varsigma^*_{ijm}}{\varsigma_{ijm}}\right| \overset{\text{L10(b)}}{\preceq} \max_m \sup_{\Theta\times\sqcap}\left|\frac{\sqrt{N_m}(\log N)^3}{N_m\kappa_\delta^{\uparrow}\pi^*_{\mathcal{E}^c m}}\right| \overset{\text{L9(b)}}{\preceq} \frac{\rho_u(\log N)^3}{\kappa_\delta^{\uparrow}\kappa_\pi} \overset{\text{I[i]}}{\prec} 1. \qquad (37)
$$

Now, for each i, fixing $\varsigma_{i\mathcal{E}^c m}$ and minimizing $\sum_{\mathcal{E}_m} \varsigma^*_{ijm} \log(\varsigma^*_{ijm}/\varsigma_{ijm})$ with respect to the ς_{ijm}, $j \in \mathcal{E}_m$, subject to the adding up constraint $\sum_{\mathcal{E}_m} \varsigma_{ijm} = 1 - \varsigma_{i\mathcal{E}^c m}$ yields $\varsigma^\circ_{ijm} = \varsigma^*_{ijm}(1 - \varsigma_{i\mathcal{E}^c m})/\varsigma^*_{i\mathcal{E}m}$. Hence,

$$
\begin{aligned}
\sum_i\sum_{\mathcal{E}_m}\varsigma^*_{ijm}\log\frac{\varsigma^*_{ijm}}{\varsigma_{ijm}} \ge \sum_i\sum_{\mathcal{E}_m}\varsigma^*_{ijm}\log\frac{\varsigma^*_{ijm}}{\varsigma^\circ_{ijm}} &= \sum_i \varsigma^*_{i\mathcal{E}m}\log\frac{\varsigma^*_{i\mathcal{E}m}}{1 - \varsigma_{i\mathcal{E}^c m}} \\
&\ge \sum_i \varsigma^*_{i\mathcal{E}m}\log\varsigma^*_{i\mathcal{E}m} \overset{\text{Jensen: } t\log t \text{ convex}}{\ge} N_m\bar{\pi}^*_{\mathcal{E}m}\log\bar{\pi}^*_{\mathcal{E}m} \overset{\text{MVT}}{\ge} -N_m\bar{\pi}^*_{\mathcal{E}^c m},
\end{aligned}
$$

where $\bar{\pi}^*_{jm} = N_m^{-1}\sum_i \varsigma^*_{ijm}$.[50] So $\mathcal{E}_m$ term is bounded below by $-N_m\bar{\pi}^*_{\mathcal{E}^c m}$ up to negligible terms.

Finally, we show the $\mathcal{E}_m$ term is negligible relative to the $\mathcal{E}^c_m$ term by taking the ratio of their bounds:

$$
\begin{aligned}
\max_m \frac{N_m\bar{\pi}^*_{\mathcal{E}^c m}}{N_m\pi^*_{\mathcal{E}^c m}\kappa_\delta^{\uparrow}} = \frac{1}{\kappa_\delta^{\uparrow}}\Big(1 + \max_m \frac{\bar{\pi}^*_{\mathcal{E}^c m} - \pi^*_{\mathcal{E}^c m}}{\pi^*_{\mathcal{E}^c m}}\Big) &\overset{\text{L9(b)}}{\preceq} \\
\frac{1}{\kappa_\delta^{\uparrow}}\Big(1 + \max_m \frac{|\bar{\pi}^*_{\mathcal{E}^c m} - \pi^*_{\mathcal{E}^c m}|}{\kappa_\pi}\Big) \overset{\text{L12(d)}}{\preceq} \frac{1}{\kappa_\delta^{\uparrow}}\Big(1 + \max_m \frac{\log M}{\sqrt{N_m}\kappa_\pi}\Big) &\overset{\text{I[i]}}{\preceq} \frac{1}{\kappa_\delta^{\uparrow}} \prec 1,
\end{aligned}
$$

where the rate inequality marked L12(d) uses L12(d) with s_{jm}'s replaced with $\bar{\pi}^*_{jm}$'s.

Sum $\mathcal{E}^c_m$ and $\mathcal{E}_m$ terms to show $\lim_{M\to\infty}\mathbb{P}\big[\inf_{\Theta\times\sqcap^{\kappa c}} \hat{\mathcal{L}}(\theta,\pi) - \hat{\mathcal{L}}(\theta^*,\pi^*) \le c\big] = 0$ for any $c > 0$. □

The following lemma establishes consistency of $\hat{\pi}, \hat{\delta}, \hat{\beta}$. The rate for $\hat{\pi}$ is intermediate in that it is superseded by theorem 2 which shows asymptotic normality at a $\sqrt{N_m}$ rate.

Lemma 5 (Consistency of $\hat{\pi}, \hat{\delta}, \hat{\beta}$). For $r_m = \sqrt[4]{I_m}\sqrt{(\log I_+)^3} + \log M$, (a) $\max_m[N_m r_m^{-2}\|\hat{\pi}_m - \pi^*_m\|^2] \prec 1$; (b) $\hat{\beta} \xrightarrow{p} \beta^*$; (c) $\forall m$: $\hat{\delta}_m - \delta^*_m \prec 1$.

Proof. First (a). Let $\hat{\pi}_{\Delta m}(\pi_m)$ denote the vector $\hat{\pi}$ with its m-th element replaced with π_m, so $\hat{\pi}_{\Delta m}(\hat{\pi}_m) = \hat{\pi}$. By definition, $\hat{\pi}_m$ is the minimizer of $\hat{\Omega}[\hat{\theta}, \hat{\pi}_{\Delta m}(\pi_m)] = \sum_{m'\neq m} \hat{\mathcal{L}}_{\tilde{m}}(\hat{\theta}, \hat{\pi}_{m'}) + \hat{\mathcal{L}}_m(\hat{\theta}, \pi_m) + \hat{\Phi}[\hat{\theta}, \hat{\pi}_{\Delta m}(\pi_m)]$. The first term does not depend on the choice of π_m, and can be ignored.

We will use $\hat{\pi}_{\Delta m}(\hat{\pi}_m)$ to show that $\hat{\pi}_m$ is close to π^*_m by contradiction. Fix any constant $c > 0$. Suppose that for some m, $\sqrt{N_m}\|\hat{\pi}_m - \pi^*_m\| > cr_m$. We show that $\hat{\Omega}[\hat{\theta}, \hat{\pi}_{\Delta m}(\hat{\pi}_m)] - \hat{\Omega}[\hat{\theta}, \hat{\pi}_{\Delta m}(\pi^*_m)] > 0$ with probability approaching one, providing the contradiction with $\hat{\pi}_m$ being an optimum.

[50] For the MVT, $0 < t \le 1$, $\log t = (t-1)/\mathring{t} = -(1-t)/\mathring{t}$ for $t \le \mathring{t} \le 1$. So $t\log t \ge -(1-t)$.

First, noting that in (15) $\min(\pi^*_m, s_m) \leq \mathring{\pi}_m$, we have the following bounds,

$$\begin{aligned} 2\min_m \hat{\mathcal{L}}^{\blacksquare}_m(\hat{\pi}_m) &\overset{(15)}{\geq} \min_m N_m \|\hat{\pi}_m - s_m\|^2 > c^2 \min_m r_m^2 \geq c^2(\log M)^2, \\ 2\max_m \hat{\mathcal{L}}^{\blacksquare}_m(\pi^*_m) &\overset{(15)}{\leq} \max_m N_m \frac{\|s_m - \pi^*_m\|^2}{\min(\pi^*_m, s_m)} \overset{\text{L12(d)}}{\prec} (\log M)^2. \end{aligned} \tag{38}$$

So $\min_m[\hat{\mathcal{L}}^{\blacksquare}_m(\hat{\pi}_m) - \hat{\mathcal{L}}^{\blacksquare}_m(\pi^*_m)] \geq c^2(\log M)^2 \ / \ 4$ with probability approaching one.

Second, $\hat{\mathcal{L}}^{\blacklozenge}_m(\theta^*, \pi^*_m) = 0$ by construction. Since $\hat{\mathcal{L}}^{\blacklozenge}_m = \mathcal{L}^{\blacklozenge}_m + \Delta\hat{\mathcal{L}}^{\blacklozenge}_m$ and $\mathcal{L}^{\blacklozenge}_m(\hat{\theta}, \hat{\pi}_m) \geq \mathcal{L}^{\blacklozenge}_m(\theta^*, \pi^*_m) = 0$,

$$\min_m \frac{\hat{\mathcal{L}}^{\blacklozenge}_m(\hat{\theta}, \hat{\pi}_m)}{r_m^2} \geq \min_m \frac{\mathcal{L}^{\blacklozenge}_m(\hat{\theta}, \hat{\pi}_m)}{r_m^2} - \max_m \frac{|\Delta\hat{\mathcal{L}}^{\blacklozenge}_m(\hat{\theta}, \hat{\pi}_m)|}{r_m^2} \overset{\text{L10(b)}}{\geq} 0 - o_p(1).$$

Third, we provide a lower bound on the contribution $\hat{\Phi}(\hat{\theta}, \hat{\pi}) - \hat{\Phi}[\hat{\theta}, \hat{\pi}_{\Delta m}(\pi^*_m)]$. Recall that $\mathcal{P} \leq \mathcal{P}_B = B(B^\triangledown B)^{-1}B^\triangledown$, so expanding $\delta(\hat{\theta}, \hat{\pi})$ around $\hat{\pi}_{\Delta m}(\pi^*_m)$,

$$\begin{aligned} \|\mathcal{P}\{\delta(\hat{\theta}, \hat{\pi}) - \delta[\hat{\theta}, \hat{\pi}_{\Delta m}(\pi^*_m)]\}\| &\overset{\text{MVT}}{\leq} \|(B^\triangledown B)^{-1/2}B^\triangledown \mathbb{D}_\pi(\hat{\theta}, \mathring{\pi})[\hat{\pi} - \hat{\pi}_{\Delta m}(\pi^*_m)]\| \\ &= \|(B^\triangledown B)^{-1/2}B^\triangledown_m \mathbb{D}_{\pi m}(\hat{\theta}, \mathring{\pi}_m)(\hat{\pi}_m - \pi^*_m)\| \overset{\text{F[iii],Schwarz}}{\leq} O_p(M^{-1/2})\|B^\triangledown_m \mathbb{D}_{\pi m}(\hat{\theta}, \mathring{\pi}_m)\| \cdot \|\hat{\pi}_m - \pi^*_m\|. \end{aligned}$$

Recall $\mathcal{P}\mathcal{P} = \mathcal{P}$. Since for any $\delta, \bar{\delta}$, $|\delta^\triangledown \mathcal{P}\delta - \bar{\delta}^\triangledown \mathcal{P}\bar{\delta}| \leq \|\mathcal{P}(\delta - \bar{\delta})\|^2 + 2\|\mathcal{P}(\delta - \bar{\delta})\| \cdot \|\mathcal{P}\delta\|$,[51]

$$\begin{aligned} &2|\hat{\Phi}(\hat{\theta}, \hat{\pi}) - \hat{\Phi}[\hat{\theta}, \hat{\pi}_{\Delta m}(\pi^*_m)]| \\ &\quad \leq O_p(M^{-1})\|B^\triangledown_m \mathbb{D}_{\pi m}(\hat{\theta}, \mathring{\pi}_m)\|^2 \cdot \|\hat{\pi}_m - \pi^*_m\|^2 + 2O_p(M^{-1/2})\|B^\triangledown_m \mathbb{D}_{\pi m}(\hat{\theta}, \mathring{\pi}_m)\| \cdot \|\hat{\pi}_m - \pi^*_m\|. \end{aligned} \tag{39}$$

Now, $\max_m \|B^\triangledown_m \mathbb{D}_{\pi m}(\hat{\theta}, \mathring{\pi}_m)\|^2 \overset{\text{L9(d)}}{\precsim} \kappa^{-3} \max_m \|B_m\| \overset{\text{G[iv]}}{\precsim} \kappa^{-3}\sqrt{M}$, which implies that the right-hand side in (39) is bounded above by $O_p(N_m^{-1}\kappa^{-6})N_m\|\hat{\pi}_m - \pi^*_m\|^2 + O_p(N_m^{-1/2}\kappa^{-3})\sqrt{N_m}\|\hat{\pi}_m - \pi^*_m\| \overset{(38)}{\prec} \hat{\mathcal{L}}^{\blacksquare}_m(\hat{\pi}_m)$, uniformly in m because $\max_m N_m^{-1}\kappa^{-6} \overset{\text{C,I}}{\prec} 1$.

Combining the three terms, noting that $\hat{\mathcal{L}}^{\blacksquare}_m(\hat{\pi}_m) - \hat{\mathcal{L}}^{\blacksquare}_m(\pi^*_m)$ is dominant, we have $\hat{\Omega}[\hat{\theta}, \hat{\pi}_{\Delta m}(\hat{\pi}_m)] - \hat{\Omega}[\hat{\theta}, \hat{\pi}_{\Delta m}(\pi^*_m)] > 0$ with probability one, showing the contradiction and completing (a).

Now (b). Recall that $\tilde{B} = B(B^\triangledown B)^{-1/2}$ such that $\hat{\beta} = (X^\triangledown \mathcal{P}_B X)^{-1}X^\triangledown \mathcal{P}_B \delta(\hat{\theta}, \hat{\pi}) = (\tilde{B}^\triangledown X)^+\tilde{B}^\triangledown \delta(\hat{\theta}, \hat{\pi})$, and

$$\begin{aligned} \|\hat{\beta} - \beta^*\| &= \|(\tilde{B}^\triangledown X)^+\tilde{B}^\triangledown \delta(\hat{\theta}, \hat{\pi}) - \beta^*\| = \|(\tilde{B}^\triangledown X)^+\tilde{B}^\triangledown \xi + (\tilde{B}^\triangledown X)^+\tilde{B}^\triangledown[\delta(\hat{\theta}, \hat{\pi}) - \delta(\theta^*, \pi^*)]\| \\ &\overset{\text{MVT}}{=} \| \underbrace{(\tilde{B}^\triangledown X)^+\tilde{B}^\triangledown \xi}_{\textcircled{1}} + \underbrace{(\tilde{B}^\triangledown X)^+\tilde{B}^\triangledown \mathbb{D}_\theta(\mathring{\theta}, \mathring{\pi})}_{\textcircled{2}} \underbrace{(\hat{\theta} - \theta^*)}_{\textcircled{3}} + \underbrace{(\tilde{B}^\triangledown X)^+}_{\textcircled{4}} \underbrace{\tilde{B}^\triangledown \mathbb{D}_\pi(\mathring{\theta}, \mathring{\pi})(\hat{\pi} - \pi^*)}_{\textcircled{5}} \| \prec 1, \end{aligned}$$

where the last $\prec$ holds because ① $\precsim M^{-1/2}$ (standard linear IV error); ② $\precsim 1$ (multiple applications of WLLN); ③ $\prec 1$ (shown above); ④ $\precsim M^{-1/2}$ (multiple applications of WLLN); and,

$$\begin{aligned} \textcircled{5} &= (B^\triangledown B)^{-1/2}\sum_m B^\triangledown_m \mathbb{D}_{\pi m}(\mathring{\theta}, \mathring{\pi}_m)(\hat{\pi}_m - \pi^*_m) \\ &\overset{\text{(a)}}{\precsim} \|(B^\triangledown B)^{-1/2}\sum_m B^\triangledown_m \mathbb{D}_{\pi m}(\mathring{\theta}, \mathring{\pi}_m)\| \frac{r_m}{\sqrt{N_m}} \precsim \sqrt{M} \max_m \frac{r_m}{\sqrt{N_m}} \prec M^{1/2}, \end{aligned}$$

so ④ × ⑤ $\prec 1$. Finally, (c) follows from Slutsky's theorem, $\hat{\delta}_m = \delta_m(\hat{\theta}, \hat{\pi}_m) \xrightarrow{p} \delta_m(\theta^*, \pi^*_m) = \delta^*_m$. □

B.2 Asymptotic normality

Lemma 6 (Negligibility of higher order terms in (16)). Statements (17a) to (17d) hold.

[51] Take $a = \mathcal{P}\delta$, $b = \mathcal{P}\bar{\delta}$ and note $\left|\|a\|^2 - \|b\|^2\right| = \left|\|a - b\|^2 - 2(a - b)^\triangledown a\right|$. Apply triangle and Schwarz inequalities.

Proof. First (17d). We first develop a lower bound for Γ_π^{-2}. since $\mathcal{L}^{\blacklozenge\ast}_{\pi\pi} - \mathcal{L}^{\ast}_{\pi\theta}\mathcal{L}^{\ast+}_{\theta\theta}\mathcal{L}^{\ast}_{\theta\pi} \geq 0$,[52]

$$\Gamma_\pi^{-2} = \mathbb{E}(\mathcal{L}^{\ast}_{\pi\pi} - \mathcal{L}^{\ast}_{\pi\theta}\mathcal{L}^{\ast+}_{\theta\theta}\mathcal{L}^{\ast}_{\theta\pi}) = \mathbb{E}(\mathcal{L}^{\blacksquare\ast}_{\pi\pi} + \mathcal{L}^{\blacklozenge\ast}_{\pi\pi} - \mathcal{L}^{\ast}_{\pi\theta}\mathcal{L}^{\ast+}_{\theta\theta}\mathcal{L}^{\ast}_{\theta\pi}) \geq \mathbb{E}\mathcal{L}^{\blacksquare\ast}_{\pi\pi}.$$

Dividing $\mathcal{L}^{\blacksquare}_{\pi\pi}$ into its diagonal blocks, $\mathcal{L}^{\blacksquare\ast}_{\pi\pi m} = N_m(\Pi^{\ast-1}_m + \pi^{\ast-1}_{0m}u^{\triangledown}) \geq N_m\mathbb{I}$. By L5(a), we may consider only $\pi_m \in \sqcap^r_m = \{\pi_m\colon \|\pi_m - \pi^{\ast}_m\| \leq r_m/\sqrt{N_m}\}$. Noting the (k,t) elements of $\hat{\mathcal{L}}^{\blacksquare}_{\pi\pi m}$ and $\mathcal{L}^{\blacksquare\ast}_{\pi\pi m}$ are $N_m(s_{km}\pi^{-2}_{km}\mathbb{1}[k=t] + s_{0m}\pi^{-2}_{0m})$ and $N_m(\pi^{\ast-1}_{km}\mathbb{1}[k=t] + \pi^{\ast-1}_{0m})$ respectively,

$$\begin{aligned}\max_m \sup_{\sqcap^r_m}\big[N_m^{-1}\|\hat{\mathcal{L}}^{\blacksquare}_{\pi\pi m}(\pi_m) - \mathcal{L}^{\blacksquare}_{\pi\pi m}(\pi^{\ast}_m)\|\big] &\leq 2\max_{m,j}\sup_{\sqcap^r_m}\left|\frac{s_{jm}}{\pi^2_{jm}} - \frac{1}{\pi^{\ast}_{jm}}\right| \\ &= 2\max_{m,j}\sup_{\sqcap^r_m}\left|\frac{(s_{jm}-\pi^{\ast}_{jm})\pi^{\ast}_{jm} - (\pi_{jm}-\pi^{\ast}_{jm})\pi^{\ast}_{jm} - (\pi_{jm}-\pi^{\ast}_{jm})\pi_{jm}}{\pi^2_{jm}\pi^{\ast}_{jm}}\right| \prec 1.\end{aligned}$$

So $\max_{\mathring{t}\in[0,1]}\|\Gamma_\pi\{\hat{\mathcal{L}}^{\blacksquare}_{\pi\pi}[\gamma(\mathring{t})] - \mathcal{L}^{\blacksquare\ast}_{\pi\pi}\}\Gamma_\pi\| \leq \max_{\mathring{t}\in[0,1]}\|(\mathbb{E}\mathcal{L}^{\blacksquare\ast}_{\pi\pi})^{-1/2}\{\hat{\mathcal{L}}^{\blacksquare}_{\pi\pi}[\gamma(\mathring{t})] - \mathcal{L}^{\blacksquare\ast}_{\pi\pi}\}(\mathbb{E}\mathcal{L}^{\blacksquare\ast}_{\pi\pi})^{-1/2}\|$ which is the norm of a block diagonal matrix with blocks $\prec \max_m[N_m^{-1/2}\|\mathbb{I}\| \times N_m \times N_m^{-1/2}\|\mathbb{I}\|] \preceq 1$, establishing (17d).

The remaining three results are similar to each other and are shown in L11. □

Lemma 7 (Denominator approximation). At the truth, $\Gamma_\theta\mathcal{Q}_{\theta\theta}\Gamma_\theta \xrightarrow{p} \mathbb{I}$.

Proof. We will show the equivalent result $\Gamma_\theta(\mathcal{Q}_{\theta\theta} - \Gamma_\theta^{-2})\Gamma_\theta \prec 1$. Recall that $\mathcal{Q}_{\theta\theta} = \overbrace{\Omega_{\theta\theta}}^{①} - \overbrace{\Omega_{\theta\pi}\Omega^{-1}_{\pi\pi}\Omega_{\pi\theta}}^{②}$, We will separate the Ω's into their constituent parts $\mathcal{L}, \Phi$ and then drop negligible terms. We begin with ②, defining $\mathcal{R} = \mathcal{L}^{-1}_{\pi\pi}\mathcal{L}_{\pi\theta}$,

1. Establish that ② $\simeq \Omega_{\theta\pi}\mathcal{L}^{-1}_{\pi\pi}\Omega_{\pi\theta} - \Omega_{\theta\pi}\mathcal{L}^{-1}_{\pi\pi}\Phi_{\pi\pi}\mathcal{L}^{-1}_{\pi\pi}\Omega_{\pi\theta} =:$ ⑥ $-$ ⑦;

2. Show that ⑥ $\simeq \mathcal{L}_{\theta\pi}\mathcal{L}^{-1}_{\pi\pi}\mathcal{L}_{\pi\theta} + \mathcal{R}^{\triangledown}\Phi_{\pi\theta} + \Phi_{\theta\pi}\mathcal{R}$;

3. Show that ⑦ $\simeq \mathcal{R}^{\triangledown}\Phi_{\pi\pi}\mathcal{R}$

Now define $\mathcal{Q}^{\mathcal{L}}_{\theta\theta} = \mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\pi}\mathcal{L}^{-1}_{\pi\pi}\mathcal{L}_{\pi\theta}$. Since ① $= \mathcal{L}_{\theta\theta} + \Phi_{\theta\theta}$, we have $\mathcal{Q}_{\theta\theta} \simeq \mathcal{Q}^{\mathcal{L}}_{\theta\theta} + \Phi_{\theta\theta} - \mathcal{R}^{\triangledown}\Phi_{\pi\theta} - \Phi_{\pi\theta}\mathcal{R} + \mathcal{R}^{\triangledown}\Phi_{\pi\pi}\mathcal{R}$. We now proceed with the terms of the right hand side:

4. Show that $\mathcal{Q}^{\mathcal{L}}_{\theta\theta} \simeq \mathbb{E}\mathcal{Q}^{\mathcal{L}}_{\theta\theta}$;

5. Show that $\Phi_{\theta\theta} \simeq M\Xi_\theta\mathcal{A}\Xi^{\triangledown}_\theta$, $\Phi_{\theta\pi}\mathcal{R} \simeq M\Xi_\theta\mathcal{A}\Xi^{\triangledown}_\pi$, and $\mathcal{R}^{\triangledown}\Phi_{\pi\pi}\mathcal{R} \simeq M\Xi_\pi\mathcal{A}\Xi^{\triangledown}_\pi$;

6. That produces $\mathcal{Q}_{\theta\theta} \simeq \mathbb{E}(\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\pi}\mathcal{L}^{-1}_{\pi\pi}\mathcal{L}_{\pi\theta}) + M\Xi\mathcal{A}\Xi^{\triangledown} = \Gamma_\theta^{-2}$, as promised.

The most complicated part is step 1. Let $A = \mathbb{D}^{\triangledown}_\pi\tilde{B}(\tilde{B}^{\triangledown}\mathcal{P}\tilde{B})^{1/2}$ such that $\Phi_{\pi\pi} = \mathbb{D}^{\triangledown}_\pi\mathcal{P}\mathbb{D}_\pi = \mathbb{D}^{\triangledown}_\pi\mathcal{P}_B\mathcal{P}\mathcal{P}_B\mathbb{D}_\pi = AA^{\triangledown}$. Applying the Woodbury matrix identity to $(\mathcal{L}_{\pi\pi} + A\mathbb{I}A^{\triangledown})^{-1}$,

$$\begin{aligned}\Omega^{-1}_{\pi\pi} = (\mathcal{L}_{\pi\pi} + \Phi_{\pi\pi})^{-1} &= \mathcal{L}^{-1}_{\pi\pi} - \mathcal{L}^{-1}_{\pi\pi}A[\mathbb{I} + \overbrace{A^{\triangledown}\mathcal{L}^{-1}_{\pi\pi}A}^{③}]^{-1}A^{\triangledown}\mathcal{L}^{-1}_{\pi\pi} \\ &\simeq \mathcal{L}^{-1}_{\pi\pi} - \mathcal{L}^{-1}_{\pi\pi}AA^{\triangledown}\mathcal{L}^{-1}_{\pi\pi} = \mathcal{L}^{-1}_{\pi\pi} - \mathcal{L}^{-1}_{\pi\pi}\Phi_{\pi\pi}\mathcal{L}^{-1}_{\pi\pi}, \qquad (40)\end{aligned}$$

because ③ $\prec 1$ as we now show,

[52] This is easy to see for the case when $H^{\ast}_{\mathcal{L}\bullet}$, defined in (19), is full rank because $(\mathcal{L}^{\blacklozenge\ast}_{\pi\pi} - \mathcal{L}^{\ast}_{\pi\theta}\mathcal{L}^{\ast+}_{\theta\theta}\mathcal{L}^{\ast}_{\theta\pi})^{-1}$ is a main diagonal block of $H^{\ast-1}_{\mathcal{L}\bullet} > 0$. Using the Moore Penrose inverse covers the case when $H^{\ast}_{\mathcal{L}\bullet}$ is not full rank.

$$\textcircled{3} = (\underbrace{\tilde{B}^\triangledown \mathcal{P} \tilde{B}}_{\leq \tilde{B}^\triangledown \tilde{B} = \mathbb{I}})^{1/2} \overbrace{\tilde{B}^\triangledown}^{(B^\triangledown B)^{-1/2} B^\triangledown} \mathbb{D}_\pi \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown \tilde{B} (\tilde{B}^\triangledown \mathcal{P} \tilde{B})^{1/2} = O_p(1)(B^\triangledown B)^{-1/2} \overbrace{B^\triangledown \mathbb{D}_\pi \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown B}^{\textcircled{4}} (B^\triangledown B)^{-1/2} O_p(1)$$

$$= O_p(M^{-1/2}) \times \textcircled{4} \times O_p(M^{-1/2}). \quad (41)$$

Noting that $\mathcal{L}_{\pi\pi m}^{-1} \leq \mathcal{L}_{\pi\pi m}^{\blacksquare -1} = N_m^{-1}(\varPi_m^* - \pi_m^* \pi_m^{*\triangledown})$ because $\mathcal{L}_{\pi\pi m}^{\blacklozenge} \geq 0$,

$$\textcircled{4} = \sum_m B_m^\triangledown \mathbb{D}_{\pi m} \mathcal{L}_{\pi\pi m}^{-1} \mathbb{D}_{\pi m}^\triangledown B_m \preceq \sum_m \frac{1}{N_m} B_m^\triangledown \mathbb{D}_{\pi m} \underbrace{(\varPi_m^* - \pi_m^* \pi_m^{*\triangledown})}_{\leq \mathbb{I}} \mathbb{D}_{\pi m}^\triangledown B_m \preceq \sum_m \frac{1}{N_m} B_m^\triangledown \mathbb{D}_{\pi m} \mathbb{D}_{\pi m}^\triangledown B_m$$

$$\simeq \sum_m \frac{1}{N_m} \mathbb{E}(B_m^\triangledown \mathbb{D}_{\pi m} \mathbb{D}_{\pi m}^\triangledown B_m) \sim \sum_m \frac{1}{N_m} \prec 1, \quad (42)$$

it follows that $\textcircled{3} \prec M^{-1}$, which establishes the $\simeq$ in (40). Substituting (40) into $\textcircled{2}$ completes step 1.

For step 2, substitute $\varOmega_{\theta\pi} = \mathcal{L}_{\theta\pi} + \varPhi_{\theta\pi}$ and multiply out,

$$\textcircled{6} = \mathcal{L}_{\theta\pi} \mathcal{L}_{\pi\pi}^{-1} \mathcal{L}_{\pi\theta} + \mathcal{R}^\triangledown \varPhi_{\pi\theta} + \varPhi_{\theta\pi} \mathcal{R} + \overbrace{\varPhi_{\theta\pi} \mathcal{L}_{\pi\pi}^{-1} \varPhi_{\pi\theta}}^{\textcircled{8}},$$

so it suffices to show that $\textcircled{8}$ is negligible. We have,

$$\textcircled{8} = \mathbb{D}_\theta^\triangledown \mathcal{P} \mathbb{D}_\pi \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown \mathcal{P} \mathcal{D}_\theta = \overbrace{\mathbb{D}_\theta^\triangledown \mathcal{P} B (B^\triangledown B)^{-1}}^{\preceq 1} B^\triangledown \mathbb{D}_\pi \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown B \overbrace{(B^\triangledown B)^{-1} B^\triangledown \mathcal{P} \mathbb{D}_\theta}^{\preceq 1} \prec 1, \quad (43)$$

which follows from $B^\triangledown \mathbb{D}_\pi \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown B = \textcircled{4} \prec 1$ as shown in (42).

Moving on to step 3, recall $\textcircled{7} = \varOmega_{\theta\pi} \mathcal{L}_{\pi\pi}^{-1} \varPhi_{\pi\pi} \mathcal{L}_{\pi\pi}^{-1} \varOmega_{\pi\theta}$, we must show that

$$\textcircled{7} - \mathcal{R}^\triangledown \varPhi_{\pi\pi} \mathcal{R} = \overbrace{\mathcal{R}^\triangledown \varPhi_{\pi\pi} \mathcal{L}_{\pi\pi}^{-1} \varPhi_{\pi\theta}}^{\textcircled{9}} + \varPhi_{\theta\pi} \mathcal{L}_{\pi\pi}^{-1} \varPhi_{\pi\pi} \mathcal{R} + \varPhi_{\theta\pi} \mathcal{L}_{\pi\pi}^{-1} \varPhi_{\pi\pi} \mathcal{L}_{\pi\pi}^{-1} \varPhi_{\pi\theta}, \quad (44)$$

is negligible. Starting with $\textcircled{9}$, we have, $\varPhi_{\pi\pi} = \mathbb{D}_\pi^\triangledown \mathcal{P} \mathbb{D}_\pi = \mathbb{D}_\pi^\triangledown \mathcal{P}_B \mathcal{P} \mathcal{P}_B \mathbb{D}_\pi$. Recalling $\mathcal{P}_B = B(B^\triangledown B)^{-1} B^\triangledown$,

$$\textcircled{9} = \overbrace{\frac{\mathcal{L}_{\theta\pi} \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown B}{M} \left(\frac{B^\triangledown B}{M}\right)^{-1} \frac{B^\triangledown \mathcal{P} B}{M} \left(\frac{B^\triangledown B}{M}\right)^{-1}}^{\xrightarrow{p} \varXi_\pi} [B^\triangledown \mathbb{D}_\pi \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown B] \overbrace{\left(\frac{B^\triangledown B}{M}\right)^{-1} \frac{B^\triangledown \mathcal{P} B}{M} \left(\frac{B^\triangledown B}{M}\right)^{-1} \frac{B^\triangledown \mathbb{D}_\theta}{M}}^{\xrightarrow{p} \varXi_\theta^\triangledown}. \quad (45)$$

Hence, $\textcircled{9}$ is negligible since like before, $B^\triangledown \mathbb{D}_\pi \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown B = \textcircled{4} \prec 1$. The second term of (44) is the transpose of $\textcircled{9}$. Following analogous steps, the final term of (44) $\simeq \varXi_\theta^\triangledown \textcircled{4} \varXi_\theta \prec 1$, concluding step 3.

Collecting terms from above, we now have, $\mathcal{Q}_{\theta\theta} \simeq \mathcal{Q}_{\theta\theta}^{\mathcal{L}} + \varPhi_{\theta\theta} - \mathcal{R}^\triangledown \varPhi_{\pi\theta} - \varPhi_{\pi\theta} \mathcal{R} + \mathcal{R}^\triangledown \varPhi_{\pi\pi} \mathcal{R}$. For step 4, let $\mathbb{U}_m := \varGamma_\theta [\mathcal{Q}_{\theta\theta m}^{\mathcal{L}} - \mathbb{E}(\mathcal{Q}_{\theta\theta m}^{\mathcal{L}} \mid I_m)] \varGamma_\theta$, where $\mathcal{Q}_{\theta\theta m}^{\mathcal{L}} = \mathcal{L}_{\theta\theta m} - \mathcal{L}_{\theta\pi m} \mathcal{L}_{\pi\pi m}^{-1} \mathcal{L}_{\pi\theta m}$ and $\mathbb{U} := \sum_m \mathbb{U}_m$, the result follows from the fact that for any fixed vector v, $\mathbb{E}\|\mathbb{U}v\|^2 \overset{\mathrm{D}}{\prec} 1$. This completes step 4.

Now, step 5. We will show the second result, $\mathcal{R}^\triangledown \varPhi_{\pi\theta} \simeq M \varXi_\pi \mathcal{A} \varXi_\theta^\triangledown$. We have,

$$\mathcal{R}^\triangledown \varPhi_{\pi\theta} = \mathcal{L}_{\theta\pi} \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown \mathcal{P} \mathbb{D}_\theta = \mathcal{L}_{\theta\pi} \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown \overbrace{\mathcal{P}_B \mathcal{P} \mathcal{P}_B \mathcal{P} \mathcal{P}_B}^{= \mathcal{P}} \mathbb{D}_\theta =$$

$$M \overbrace{\frac{\mathcal{L}_{\theta\pi} \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown B}{M} \left(\frac{B^\triangledown B}{M}\right)^{-1} \frac{B^\triangledown \mathcal{P} B}{M} \left(\frac{B^\triangledown B}{M}\right)^{-1}}^{\xrightarrow{p} \varXi_\pi} \overbrace{\frac{B^\triangledown B}{M}}^{\xrightarrow{p} \mathcal{A}} \overbrace{\left(\frac{B^\triangledown B}{M}\right)^{-1} \frac{B^\triangledown \mathcal{P} B}{M} \left(\frac{B^\triangledown B}{M}\right)^{-1} \frac{B^\triangledown \mathbb{D}_\theta}{M}}^{\xrightarrow{p} \varXi_\theta^\triangledown} \simeq M \varXi_\pi \mathcal{A} \varXi_\theta^\triangledown.$$

The other two results follow by analogously applying $\mathcal{P} = \mathcal{P}_B \mathcal{P} \mathcal{P}_B \mathcal{P} \mathcal{P}_B$ and are left to the reader.

Finally, step 6, collecting the above results we have,

$$\mathcal{Q}_{\theta\theta} \simeq \mathbb{E}\mathcal{Q}^{\mathcal{L}}_{\theta\theta} + \Xi_\theta \mathcal{A} \Xi_\theta^{\triangledown} - \Xi_\theta \mathcal{A} \Xi_\pi^{\triangledown} - \Xi_\pi \mathcal{A} \Xi_\theta^{\triangledown} + \Xi_\pi \mathcal{A} \Xi_\pi^{\triangledown} = \mathbb{E}\mathcal{Q}^{\mathcal{L}}_{\theta\theta} + \Xi \mathcal{A} \Xi^{\triangledown} = \Gamma_\theta^{-2}. \qquad \square$$

Lemma 8 (Numerator normality)**.** Evaluated at the truth, for $\hat{q}_\theta = \overbrace{\hat{\Omega}_\theta}^{\textcircled{a}} - \overbrace{\underbrace{\Omega_{\theta\pi}\Omega_{\pi\pi}^{-1}}_{\mathcal{R}^{\Omega\triangledown}}[\underbrace{\hat{\Omega}^{\blacklozenge}_\pi - \mathcal{L}^{\blacksquare}_{\pi\pi}(s - \pi^*)}_{\hat{r}^{\Omega}}]}^{\textcircled{b}}$, $\Gamma_\theta \hat{q}_\theta \xrightarrow{d} N(0, \mathcal{V}_\theta)$.

Proof. We will first expand the Ω's and $\hat{\Omega}$'s into their constituent parts and then drop negligible terms before applying a central limit theorem. The steps of this proof are as follows, we begin with ⓑ:

a. Show that ⓑ $\simeq \Omega_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\hat{r}^{\Omega} - \Omega_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\pi}\mathcal{L}_{\pi\pi}^{-1}\hat{r}^{\Omega} =:$ ⓒ − ⓓ;

b. Recalling that $\mathcal{R} = \mathcal{L}_{\pi\pi}^{-1}\mathcal{L}_{\pi\theta}$ and defining $\hat{r} = \hat{r}^{\Omega} - \hat{\Phi}_\pi = \hat{\mathcal{L}}^{\blacklozenge}_\pi - \mathcal{L}^{\blacksquare}_{\pi\pi}(s - \pi^*)$, show that ⓒ $\simeq \mathcal{R}^{\triangledown}\hat{r} + \mathcal{R}^{\triangledown}\hat{\Phi}_\pi$;

c. Show that $\Gamma_\theta \times$ ⓓ $\prec 1$.

Use the above steps with ⓐ $= \hat{\Omega}_\theta = \hat{\mathcal{L}}_\theta + \hat{\Phi}_\theta$ to rearrange $\hat{q}_\theta \simeq (\hat{\mathcal{L}}_\theta - \mathcal{R}^{\triangledown}\hat{r}) + (\hat{\Phi}_\theta - \mathcal{R}^{\triangledown}\hat{\Phi}_\pi)$. Continuing,

d. Show $\hat{\Phi}_\theta \simeq \Xi_\theta B^{\triangledown}\xi, \mathcal{R}^{\triangledown}\hat{\Phi}_\pi \simeq \Xi_\pi B^{\triangledown}\xi$; so $\hat{q}_\theta \simeq \sum_m (\hat{\mathcal{L}}_{\theta m} - \mathcal{L}_{\theta\pi m}\mathcal{L}_{\pi\pi m}^{-1}\hat{r}_m + \Xi B_m^{\triangledown}\xi_m) =: \sum_m \mathbb{T}_m$;

e. Use a central limit theorem to establish that $\Gamma_\theta \hat{q}_\theta \xrightarrow{d} \mathcal{N}[0, \lim_{M\to\infty} \sum_m (\Gamma_\theta \mathcal{V}_{\mathbb{T}m}\Gamma_\theta)] = \mathcal{N}(0, \mathcal{V}_{\hat{\theta}})$.

We will establish these results in order, reusing results from the proof of L7 where helpful. For instance, in step a we re-use (40), $\Omega_{\pi\pi}^{-1} \simeq \mathcal{L}_{\pi\pi}^{-1} - \mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\pi}\mathcal{L}_{\pi\pi}^{-1}$.

For step b, we have ⓒ $= \mathcal{R}^{\triangledown}\hat{r}^{\Omega} + \Phi_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\hat{r}^{\Omega} =: \mathcal{R}^{\triangledown}\hat{r}^{\Omega} +$ ⓔ. To show that ⓔ vanishes, we first show that for any $\mathbb{A}$-measurable K,

$$\|K^{\triangledown}\hat{r}^{\Omega}\| \preceq \|K^{\triangledown}\Omega_{\pi\pi}K\|^{1/2}. \tag{46}$$

To see this, note $K^{\triangledown}\hat{r}^{\Omega} = K^{\triangledown}\hat{r} + K^{\triangledown}\hat{\Phi}_\pi$. Proceeding term by term, $\mathbb{E}(\|K^{\triangledown}\hat{r}\|^2 \mid \mathbb{A}) = K^{\triangledown}\mathcal{L}_{\pi\pi}K$ because

$$\mathbb{E}(\hat{r}\hat{r}^{\triangledown} \mid \mathbb{A}) = E[\hat{\mathcal{L}}^{\blacklozenge}_\pi \hat{\mathcal{L}}^{\blacklozenge\triangledown}_\pi + \mathcal{L}^{\blacksquare}_{\pi\pi}(s - \pi^*)(s - \pi^*)^{\triangledown}\mathcal{L}^{\blacksquare}_{\pi\pi} \mid \mathbb{A}] = \mathcal{L}^{\blacklozenge}_{\pi\pi} + \mathcal{L}^{\blacksquare}_{\pi\pi} = \mathcal{L}_{\pi\pi},$$

where the first equality follows from $\mathbb{E}[(s - \pi^*)\hat{\mathcal{L}}^{\blacklozenge\triangledown}_\pi \mid \mathbb{A}] = 0$[53] and the second uses the information matrix equalities for $\mathcal{L}^{\blacklozenge}$ and $\mathcal{L}^{\blacksquare}$. For the second term, $K^{\triangledown}\hat{\Phi}_\pi = K^{\triangledown}\mathbb{D}_\pi^{\triangledown}\mathcal{P}\xi = K^{\triangledown}\mathbb{D}_\pi^{\triangledown}\mathcal{P}B(B^{\triangledown}B)^{-1}B^{\triangledown}\xi$. Since $\mathbb{V}[B^{\triangledown}\xi] \preceq B^{\triangledown}B$ by F[ii] and G[iv], $\|K^{\triangledown}\hat{\Phi}_\pi\|^2 \preceq K^{\triangledown}\mathbb{D}_\pi^{\triangledown}\mathcal{P}\mathbb{D}_\pi K = K^{\triangledown}\Phi_{\pi\pi}K$, which establishes (46). Applying this result is sufficient for ⓔ to vanish since setting $K = \mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\theta}$ we have,

$$K^{\triangledown}\Omega_{\pi\pi}K = \Phi_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}(\mathcal{L}_{\pi\pi} + \Phi_{\pi\pi})\mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\theta} = \overbrace{\Phi_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\theta}}^{\textcircled{8}\text{ in (43)}} + \overbrace{\Phi_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\pi}\mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\theta}}^{\text{last term in (44)}} \prec 1,$$

which completes step b.

Now step c. Take $K = \mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\pi}\mathcal{L}_{\pi\pi}^{-1}\Omega_{\pi\theta}$ in (46) and note that ⓓ $= K^{\triangledown}\hat{r}^{\Omega}$. First,

$$\boldsymbol{\lambda}_{\max}(\mathcal{L}_{\pi\pi}^{-1/2}\Phi_{\pi\pi}\mathcal{L}_{\pi\pi}^{-1/2}) = \boldsymbol{\lambda}_{\max}(\mathcal{L}_{\pi\pi}^{-1/2}AA^{\triangledown}\mathcal{L}_{\pi\pi}^{-1/2}) = \boldsymbol{\lambda}_{\max}(A^{\triangledown}\mathcal{L}_{\pi\pi}^{-1}A) = \boldsymbol{\lambda}_{\max}(\textcircled{3}) \overset{(41)}{\prec} M^{-1}, \tag{47}$$

where A is defined in L7, step 1. Now, write $K^{\triangledown}\Omega_{\pi\pi}K = K^{\triangledown}\mathcal{L}_{\pi\pi}K + K^{\triangledown}\Phi_{\pi\pi}K$ and note that

$$K^{\triangledown}\Phi_{\pi\pi}K = K^{\triangledown}\mathcal{L}_{\pi\pi}^{1/2}(\mathcal{L}_{\pi\pi}^{-1/2}\Phi_{\pi\pi}\mathcal{L}_{\pi\pi}^{-1/2})\mathcal{L}_{\pi\pi}^{1/2}K \overset{(47)}{\prec} \frac{1}{M}K^{\triangledown}\mathcal{L}_{\pi\pi}K,$$

[53] For any k, m, $\mathbb{E}[(s_{km} - \pi^*_{km})\hat{\mathcal{L}}^{\blacklozenge\triangledown}_{\pi m}(\theta^*, \pi^*_m) \mid \mathbb{A}] = I_m \sum_j \mathbb{E}[(y_{ikm} - \pi^*_{km})y_{ijm}\Delta\ell^*_{\pi ijm} \mid \mathbb{A}] = I_m\{\mathbb{E}[\varsigma^*_{ikm}\Delta\ell^*_{\pi ikm} \mid \mathbb{A}] - \pi^*_{km}\sum_j \mathbb{E}[\varsigma^*_{ijm}\Delta\ell^*_{\pi ijm} \mid \mathbb{A}]\} = 0$, noting that $\varsigma^*_{ijm}\Delta\ell^*_{\pi ijm} = \varsigma^*_{\pi ijm} - \varsigma^*_{ijm}\varsigma^*_{\pi jm}/\varsigma^*_{jm}$.

such that $K^{\triangledown}\Phi_{\pi\pi}K$ is negligible relative to $K^{\triangledown}\mathcal{L}_{\pi\pi}K$.

$$K^{\triangledown}\mathcal{L}_{\pi\pi}K = \Omega_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\pi}\mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\pi}\mathcal{L}_{\pi\pi}^{-1}\Omega_{\pi\theta} \overset{(47)}{\prec} \frac{1}{M}\Omega_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\Phi_{\pi\pi}\mathcal{L}_{\pi\pi}^{-1}\Omega_{\pi\theta} \overset{\text{L7 step 3}}{\simeq} \frac{1}{M}\mathcal{R}^{\triangledown}\Phi_{\pi\pi}\mathcal{R} \overset{\text{L7 step 5}}{\preceq} \Xi_{\pi}\mathcal{A}\Xi_{\pi}^{\triangledown} \prec 1,$$

and hence $\textcircled{d} \prec 1$. Since $\Gamma_{\theta} \prec 1$ by definition, this completes step c.

We now have $\hat{q}_{\theta} \simeq (\hat{\mathcal{L}}_{\theta} - \mathcal{R}^{\triangledown}\hat{r}) + (\hat{\Phi}_{\theta} - \mathcal{R}^{\triangledown}\hat{\Phi}_{\pi})$. We continue with step d. First, using Ξ_{θ} and Ξ_{π} defined in theorem 2, step *2*, $\hat{\Phi}_{\theta} = \mathbb{D}_{\theta}^{\triangledown}\mathcal{P}\xi = \mathbb{D}_{\theta}^{\triangledown}\mathcal{P}_{B}\mathcal{P}B(B^{\triangledown}B)^{-1}B^{\triangledown}\xi \simeq \Xi_{\theta}B^{\triangledown}\xi$, and $\mathcal{R}^{\triangledown}\hat{\Phi}_{\pi} = \mathcal{R}^{\triangledown}\mathbb{D}_{\pi}^{\triangledown}\mathcal{P}\xi = \mathcal{R}^{\triangledown}\mathbb{D}_{\pi}^{\triangledown}\mathcal{P}_{B}\mathcal{P}B(B^{\triangledown}B)^{-1}B^{\triangledown}\xi \simeq \Xi_{\pi}B^{\triangledown}\xi$, analogous to (45). This completes step d. So we can write,

$$\hat{q}_{\theta} \simeq \sum_{m}\left(\hat{\mathcal{L}}_{\theta m} - \mathcal{L}_{\theta\pi m}\mathcal{L}_{\pi\pi m}^{-1}\hat{r}_{m} + \Xi B_{m}^{\triangledown}\xi_{m}\right) = \sum_{m}\mathbb{T}_{m}. \tag{48}$$

Finally, step e. Let $\tilde{\mathcal{V}}_{\theta} = \Gamma_{\theta}\sum_{m}\mathbb{V}(\mathbb{T}_{m})\Gamma_{\theta}$. By D, $\{\mathbb{T}_{m}\}$ are independent and so are $\{\mathcal{T}_{m}\} = \{\tilde{\mathcal{V}}_{\theta}^{-1/2}\Gamma_{\theta}\mathbb{T}_{m}\}$. We only need to verify the Lindeberg condition (Davidson, 1994, 23.6), which in view of our assumptions, notably I[ii] is not in doubt. Hence, $\sum_{m}\mathcal{T}_{m} \overset{d}{\to} \mathcal{N}(0,\mathbb{I})$. Thus, by Cramér's theorem, $\Gamma_{\theta}\hat{q}_{\theta} = \tilde{\mathcal{V}}_{\theta}^{1/2}\Gamma_{\theta}\sum_{m}\mathbb{T}_{m} \overset{d}{\to} N(0,\mathcal{V}_{\theta})$, where

$$\mathcal{V}_{\theta} = \lim_{M\to\infty}(\Gamma_{\theta}\sum_{m}\mathbb{V}\mathbb{T}_{m}\Gamma_{\theta}) = \lim_{M\to\infty}\left[\Gamma_{\theta}\left(\mathbb{E}\mathcal{Q}_{\theta\theta}^{\mathcal{L}} + \Xi\mathbb{E}(B^{\triangledown}V_{\xi}B)\Xi^{\triangledown}\right)\Gamma_{\theta}\right]. \qquad \square$$

Remaining Material
for
Online Publication

C Identifying unobserved heterogeneity from micro data

In this appendix we discuss a specific example to illustrate the underlying variation in the micro sample that provides identification of θ^ν in our parametric model. This example represents a special case of the nonparametric arguments in Berry and Haile (2024).

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.[54] 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 $x^{(1)} = [1 \quad 0]^\triangledown$, $x^{(2)} = [1 \quad 1]^\triangledown$. 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 $\tilde{\pi}^z = \mathbb{P}(y_{i\cdot} = 1 \,|\, z, x)$. Fig. 8 plots this function over $z \in [-1, 1]$ for three different parametrization of the model, namely $\theta^\nu = \{0, 1, 2\}$ with $\delta = (-.25, 25)^\triangledown$ 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.

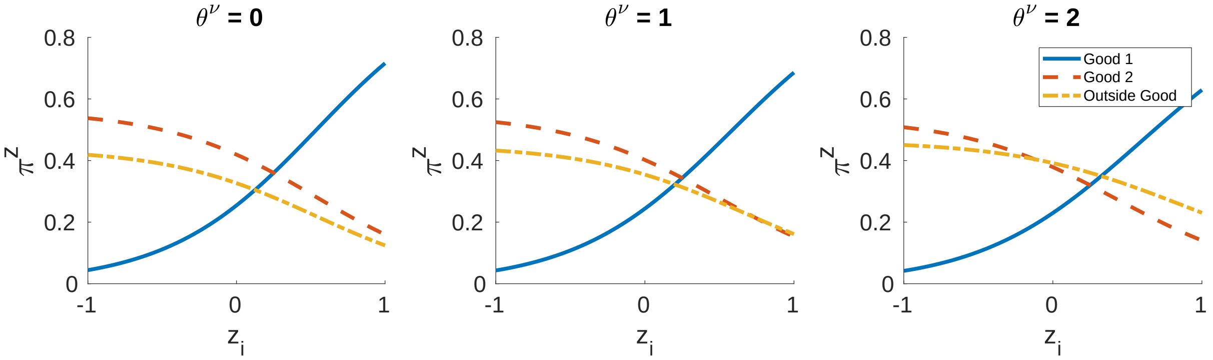


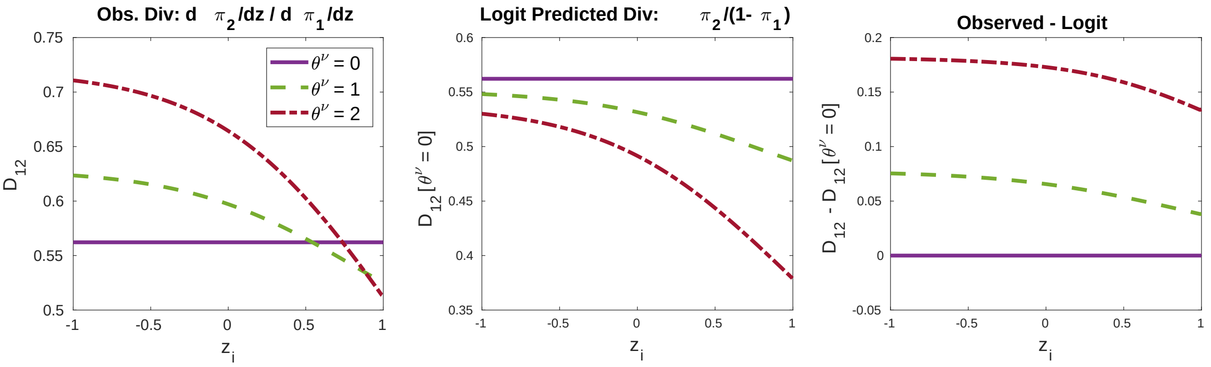
Figure 8: *Conditional shares $\tilde{\pi}^z$ are identified by the micro sample.*

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 $j = \{0, 2\}$,

$$\frac{\mathrm{d}_z \tilde{\pi}_j^z}{\mathrm{d}_z \tilde{\pi}_1^z} = \frac{\partial_{u_1} \pi_j^z}{\partial_{u_1} \pi_1^z} = D_{1j}^z. \tag{49}$$

Equation (49) 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.

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


Obs. Div: d π2/dz / d π1/dz

Logit Predicted Div: π2/(1- π1)

Observed - Logit

Figure 9: *Diversion and Demographics*

Fig. 9 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^z_{12}. 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 fig. 8: 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 fig. 9 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 fig. 9 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 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 $\hat{\chi}$ term of our estimator would be redundant. To see a case where $\hat{\chi}$ 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 demographics. μ^z is typically specified as a linear combination of interactions between product characteristics and consumer demographics, e.g.,

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

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

$$\mathrm{d}_{z^d}\tilde{\pi}^z_j = \sum_{k=1}^{K}\sum_{\ell=1}^{J}\theta^{z(k,d)}x^k_\ell\partial_{u_\ell}\pi^z_j. \tag{50}$$

In matrix notation, (50) can be written as

$$\mathrm{d}_{z^\triangledown}\tilde{\pi}^z = \partial_{u^\triangledown}\pi^z\partial_{z^\triangledown}u = \partial_{u^\triangledown}\pi^z\partial_{z^\triangledown}\mu^z = \partial_{u^\triangledown}\pi^z X^\triangledown\Theta^z. \tag{51}$$

Thus, only if $X^\triangledown\Theta^z$ has maximum column rank, does there exist a unique $\partial_{u^\triangledown}\pi^z$ that solves (51). 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.

D Optimal instruments for CLEER

This appendix shows that CLEER $\hat{\alpha} = (\hat{\theta}, \hat{\beta})$ achieves the semiparametric efficiency bound for the model presented in section 2.1.

Following C87, we show the result for all multinomial submodels[55] and rely on the arguments in C87 to take us to the general case. The derivation below differs from C87 only because CLEER combines moments with a likelihood.

We will work with the superpopulation likelihood of the model after concentrating out π. Specifically, we show that if the distribution of product-level variables is multinomial, then the Hessian of the superpopulation loglikelihood constrained to satisfy the moments with respect to $\alpha = (\theta, \beta)$ coincides (up to asymptotically negligible terms) with the Hessian of CLEER if instruments are chosen according to (23). As the Hessians are equivalent, CLEER attains the Cramér Rao lower bound for any multinomial submodel.

We first write the moment conditions for an arbitrary multinomial submodel. Treating N_m as random and making the notational simplification of identical J_m across markets, let $c_m = [x^\triangledown_m, b^\triangledown_m, \xi^\triangledown_m, N_m]^\triangledown$, where x_m, b_m are vectorized-versions of X_m, B_m. In view of the multinomial assumption, we follow C87 and express the (population) PLMs as

$$0 = \sum_t q^*(v_t)H(v_t)[\delta(\theta, v_t) - X^\triangledown_{vt}\beta] = \sum_t q^*_t H_t e_t(\alpha), \tag{52}$$

where $q^*_t = \mathbb{P}(c_m = v_t)$ with $v_1, \ldots, v_{\bar{t}}$ the values that c_m can take, $H_t = H(B_{vt})$ a matrix of instruments, $e_t(\alpha) = \delta_{vt}(\theta) - X^\triangledown_{vt}\beta$ with $B_{vt}, X_{vt}, \delta_{vt}$ the values of δ_m (with π partialed out) if $c_m = v_t$. Equation (52) is an unconditional moment condition since H incorporates all possible combinations of instrument values.

We now construct the parametric likelihood of the submodel. Since we do not know the values of the q^*_t, the objective function will now have q_t's in them as an auxiliary parameter. Let a_m be a vector containing all (y_{im}, z_{im}, D_{im})'s in a given market, where the value of z_{im} is only observed if $D_{im} = 1$. The superpopulation loglikelihood incorporating the multinomial distribution of the product-level variables is

$$\mathfrak{A}(\alpha, q) := M\sum_t q^*_t[\mathbb{L}_t(\alpha) + \log q_t], \tag{53}$$

[55] A parametric submodel is any given parametric model that satisfies the imposed conditions. A multinomial submodel is a parametric submodel in which certain variables are assumed to have a multinomial distribution.

where $\mathbb{L}_t$is the expected value of the loglikelihood for a single market conditional on $c_m = v_t$, after concentrating out π_m. Without $\mathbb{L}_t$ in (53), the optimal instruments defined below would exactly mirror C87.

Next, we derive the Hessian of (53) at its optimum. For given value of α, maximizing (53) with respect to q subject to $\sum_t q_t = 1$ and $\sum_t q_t H_t e_t(\alpha) = 0$, yields the solution

$$q_t = \frac{q_t^*}{1 + \mathfrak{l}^\triangledown(\theta,\beta) H_t e_t(\theta,\beta)}, \text{where} \mathfrak{l} = \Big(\sum_t q_t^* H_t e_t e_t^\triangledown H_t^\triangledown\Big)^{-1} \sum_t q_t^* H_t e_t =: \mathfrak{B}^{-1}(\theta,\beta) \sum_t q_t^* H_t e_t(\alpha).$$

Plugging q_t back into (53) yields

$$\bar{\mathfrak{A}}(\alpha) := M \sum_t q_t^* \{\mathbb{L}_t(\alpha) - \log[1 + \mathfrak{l}^\triangledown(\alpha) H_t e_t(\alpha)]\} + M \sum_t q_t^* \log q_t^*.$$

Letting $\mathfrak{G}^* = \sum_t q_t^* H_t \partial_{\alpha^\triangledown} e_t(\alpha^*)$, and noting that $\mathfrak{l}(\alpha) = 0$ for all α, the Hessian of $\bar{\mathfrak{A}}$ at the truth is

$$M\Big(\mathbb{L}^*_{\alpha\alpha} - \frac{1}{2} \mathfrak{G}^{*\triangledown} \mathfrak{B}^{*-1} \mathfrak{G}^*\Big). \tag{54}$$

Taking the inverse of the (minus) Hessian yields the Cramér Rao lower bound.

Finally, we show that minus (54) coincides with the Hessian of the CLEER superpopulation objective function if c_m has a multinomial distribution and the instruments are chosen according to (23). To see this, we first note that $\mathbb{L}^*_{\theta\theta} = \mathbb{E}[\mathcal{L}_{\theta\theta m} - \mathcal{L}_{\theta\pi m} \mathcal{L}^{-1}_{\pi\pi m} \mathcal{L}_{\pi\theta m}]$, and that all other elements of $\mathbb{L}^*_{\alpha\alpha}$ are zero since β does not enter the likelihood. That leaves us with the $\mathfrak{G}^{*\triangledown} \mathfrak{B}^{*-1} \mathfrak{G}^*$ component. Let $\{B_{(k)}\}$ be the values that B_m can take, $q^*_{(k)} := \sum_t q_t^* \mathbb{1}(B_{vt} = B_{(k)}) = \mathbb{P}(B_m = B_{(k)})$, $H_{(k)} = H(B_{(k)})$, $V_k = \mathbb{V}(\xi_m \mid B_m = B_{(k)})$, and

$$A_k = \sum_t \frac{q_t^*}{q_{(k)}} \mathbb{1}(B_{vt} = B_{(k)}) \partial_\alpha e_t^{*\triangledown} = \mathbb{E}\left(\begin{bmatrix} \mathbb{D}^{*\triangledown}_{\theta m} - \mathcal{L}_{\theta\pi m} \mathcal{L}^{-1}_{\pi\pi m} \mathbb{D}^{*\triangledown}_{\pi m} \\ -X_m^\triangledown \end{bmatrix} \,\Big|\, B_m = B_{(k)}\right).$$

Now, since $S^\triangledown \mathcal{P}_R S \leq S^\triangledown S$ for any matrices R, S,[56]

$$\begin{aligned} \mathfrak{G}^{*\triangledown} \mathfrak{B}^{*-1} \mathfrak{G}^* &= \sum_k q^*_{(k)} A_k H^\triangledown_{(k)} \Big(\sum_k q^*_{(k)} H_{(k)} V_k H^\triangledown_{(k)}\Big)^{-1} \sum_k q^*_{(k)} H_{(k)} A_k^\triangledown \\ &\leq \sum_k q^*_{(k)} A_k V_k^{-1} A_k^\triangledown = \mathbb{E}(B_m^{\text{opt}\triangledown} \mathcal{V}_{\xi m} B_m^{\text{opt}}). \end{aligned} \tag{55}$$

Now consider the Hessian of the PLM portion of the CLEER objective function at the truth divided by M using our proposed instruments,

$$\frac{1}{M} \begin{bmatrix} (\mathbb{D}_\theta - \partial_\theta \sigma^\triangledown \mathbb{D}_\pi) B^{\text{opt}} \\ -X^\triangledown B^{\text{opt}} \end{bmatrix} (B^{\text{opt}\triangledown} \mathcal{V}_\xi B^{\text{opt}})^{-1} \begin{bmatrix} (\mathbb{D}_\theta^\triangledown - \partial_\theta \sigma^\triangledown \mathbb{D}_\pi^\triangledown) B^{\text{opt}} & -X^\triangledown B^{\text{opt}} \end{bmatrix}^\triangledown \simeq \mathbb{E}(B_m^{\text{opt}\triangledown} \mathcal{V}_{\xi m}^{-1} B_m^{\text{opt}}),$$

i.e. up to negligible terms it is the right-hand side in (55). To conclude the argument, the left-hand side of (55) cannot be less than the right-hand side since that would make our estimator more efficient than the maximum likelihood estimator in the parametric submodel.[57] So the left-hand side and right-hand side in (55) must be equal. Consequently, the Hessian of CLEER using optimal instruments at the truth

[56] Make $S^\triangledown = [\sqrt{q^*_{(1)}} A_1 V_1^{-1/2}, \dots, \sqrt{q^*_{(\bar{k})}} A_{\bar{k}} V_{\bar{k}}^{-1/2}] \in \mathbb{R}^{d_\alpha \times (\bar{k} J_m)}$ where $\bar{k}$ is the number of values B_m can take.

[57] Recall that the Cramér Rao lower bound is the inverse of (minus) the Hessian of a loglikelihood function.

is (54). So CLEER achieves the Cramér Rao bound in every multinomial submodel.

E Estimator Comparison

In this appendix, we present additional details on the comparison between CLEER and estimators employed in the applied literatuere.

E.1 Schematic Sketch of Section section 6

Fig. 10 provides a summary of the steps presented in section 6. The top node in the tree represents CLEER. Each node below represents an alteration to arrive at an alternative estimator. The large pink box representing section 6.3 proposes several alternative estimators which we will rationalize as modifications of the score. One can stop the process at any node in the tree, so in total the figure describes nine alternative estimators (including share-constrained likelihood, see fn. 32). At each node, we briefly list the primary costs (red) and benefits (green) of the step relating to identification (🪪), econometric efficiency (both rate and variance, ✈), 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.

E.2 Share Constraint

In section 6.2 we listed three drawbacks to the imposition of share constraints on a likelihood or GMM estimator relating to robustness to zero shares, efficiency and inference. This section discusses each of these issues in turn.

First, because it is a one to one mapping on the interior of the probability simplex, doing so rules out the presence of zero observed shares. Moreover, the contraction can become unstable as observed shares tend towards zero and $\|\mathbb{D}_{\pi m}\| = \|\partial_{\pi_m^{\triangledown}} \delta_m\|$ tends to infinity. While this is reasonable for conditional choice probabilities, applied cases have arisen where zero shares are observed in data due to finite market sizes N_m and small choice probabilities. In this case, even when shares are non-zero, they will be imprecisely estimated. CLEER offers some robustness to zero or small shares because it does not enforce unconditional choice probabilities equal market shares.

Second, imposing the share constraints introduces a potential loss of efficiency. Suppose that $\theta^{*z} \neq 0$ and I is large relative to M such that contribution of the PLMs to the estimation of θ^{ν} are asymptotically negligible (as discussed in section 5). Then this efficiency loss occurs unless the population in the *smallest* market diverges faster than both I and M. Examples 1 and 2 in Grieco et al. (2023*b*) illustrate that this efficiency loss can be substantial.

For intuition, we now show that imposing share constraints is equivalent to placing infinite weight on the macro likelihood in CLEER. To see this, separate out the micro and macro terms of $\log \hat{L}$ as specified in (7) and consider the derivative of the macro loglikelihood with respect to δ, i.e. for all $m = 1, \dots, M$ and all $j = 1, \dots, J_m$,

$$\sum_{k=0}^{J_m} \frac{S_{km}}{\sigma_{km}} \int \mathscr{s}_{km}(z,\nu)\Big(\mathbb{1}(k=j) - \mathscr{s}_{jm}(z,\nu)\Big) \mathrm{d}F(\nu)\mathrm{d}G(z) = 0, \tag{56}$$

where $\mathscr{s}$ was defined in (2). Setting $\delta = \delta(\theta, s)$ such that $\sigma(\theta,\delta) = s$ solves (56) as the left hand side becomes

$$\int \mathscr{s}_{jm}(z,\nu)\mathrm{d}F(\nu)\mathrm{d}G(z) - \int \mathscr{s}_{jm}(z,\nu) \underbrace{\sum_{k=0}^{J_m} \mathscr{s}_{km}(z,\nu)}_{=1} \mathrm{d}F(\nu)\mathrm{d}G(z).$$

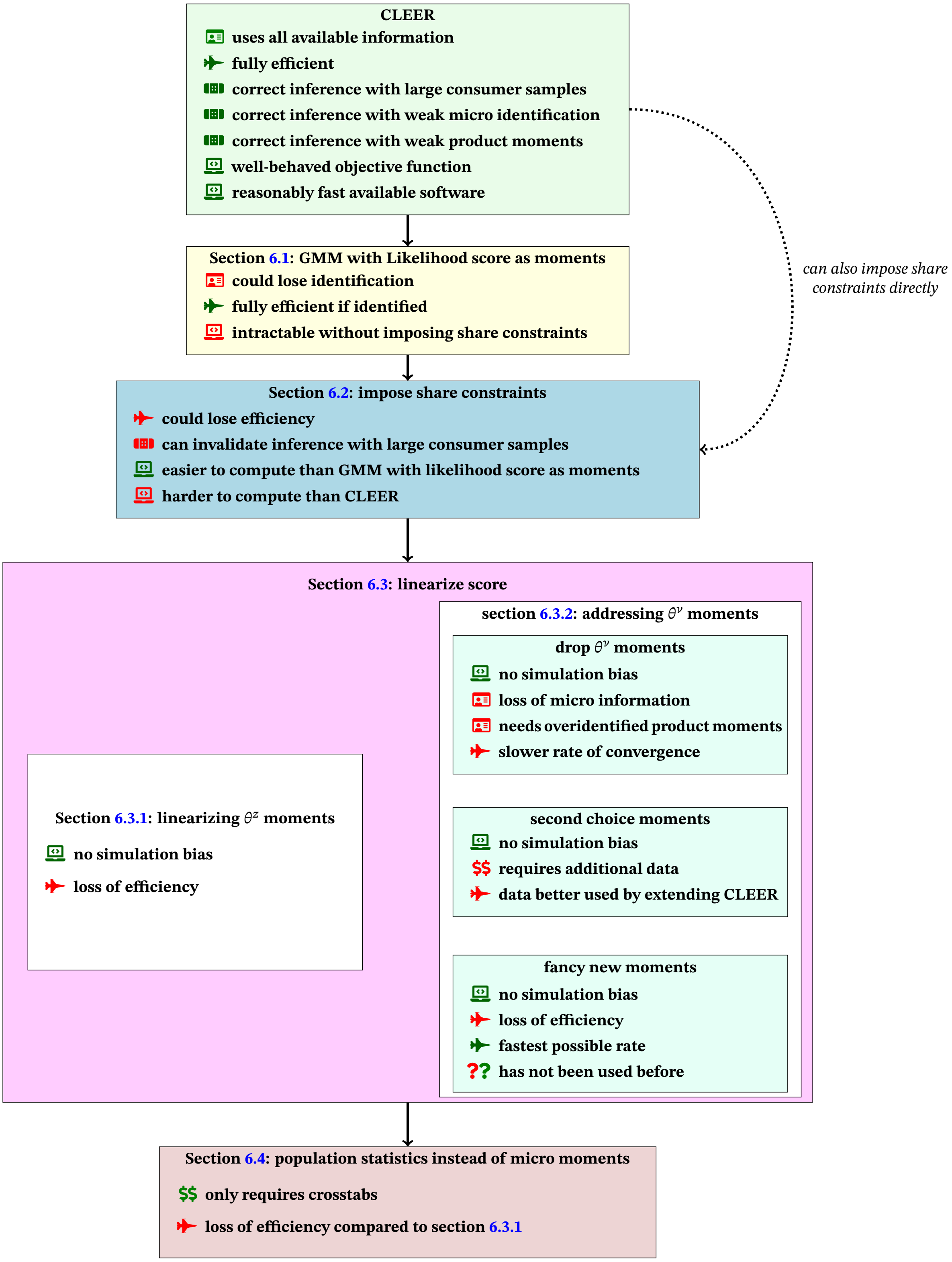

CLEER
uses all available information
fully efficient
correct inference with large consumer samples
correct inference with weak micro identification
correct inference with weak product moments
well-behaved objective function
reasonably fast available software
Section 6.1: GMM with Likelihood score as moments
could lose identification
fully efficient if identified
intractable without imposing share constraints
can also impose share constraints directly
Section 6.2: impose share constraints
could lose efficiency
can invalidate inference with large consumer samples
easier to compute than GMM with likelihood score as moments
harder to compute than CLEER
Section 6.3: linearize score
Section 6.3.1: linearizing θ^z moments
no simulation bias
loss of efficiency
section 6.3.2: addressing θ^ν moments
drop θ^ν moments
no simulation bias
loss of micro information
needs overidentified product moments
slower rate of convergence
second choice moments
no simulation bias
$$ requires additional data
data better used by extending CLEER
fancy new moments
no simulation bias
loss of efficiency
fastest possible rate
?? has not been used before
Section 6.4: population statistics instead of micro moments
$$ only requires crosstabs
loss of efficiency compared to section 6.3.1

Figure 10: *Schematic comparison of our estimator to alternatives. See text for details.*

By Berry (1994), this solution is unique for every θ. Therefore, imposing share constraints effectively places infinite weight on this moment.[58] 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 I and M were negligibly small compared to N_m because then the macro score runs over more terms than the other moments.

Third, and most importantly, assuming $s = \pi^*$ will invalidate standard inference unless the total

[58] If one places more weight on a moment in GMM estimation (without changing the rest of the weight matrix) then that moment at the estimate gets closer to zero. If one increases the weight on a moment to infinity, then that moment evaluated at the estimate must go to zero.

number of consumers in all markets is negligible compared to the square root of the population in the smallest market. If one treats $\delta(\theta, \pi^*)$ as a known deterministic function of θ, one ignores the uncertainty arising from approximating π^* with 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 I is negligibly small compared to $\min_m \sqrt{N_m}$.

To illustrate, consider inference on a linear combination of δ^*_m. Imposing share constraints, it would be tempting to use the delta method to conclude that for any vector $\upsilon \neq 0$,

$$\frac{\sqrt{I}\upsilon^\triangledown(\hat{\delta}_m - \delta^*_m)}{\sqrt{\upsilon^\triangledown \partial_{\theta^\triangledown} \mathbb{D}_{\theta m}(\hat{\theta}, s_m)\hat{\mathcal{V}}_\theta \mathbb{D}^\triangledown_{\theta m}(\hat{\theta}, s_m)\upsilon}} \xrightarrow{d} \mathcal{N}(0,1), \tag{57}$$

where $\mathbb{D}_{\theta m}$ is the derivative of δ_m with respect to θ 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 υ for which $\upsilon^\triangledown \mathbb{D}_{\theta m}(\theta^*, \pi^*) = 0$,[59] where the left-hand side of (57) diverges. The space of such vectors υ is of dimension no less than $J_m - d_\theta > 0$ since $\delta_m(\cdot, \pi^*)\colon \mathbb{R}^{d_\theta} \to \mathbb{R}^{J_m}$. Using the bootstrap the way it is typically used does not solve this problem.[60] We provide the correct variance formulas for the GMM estimators under strong micro identification when $I \succ M$ in app. F. Grieco et al. (2023*b*) 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. This issue extends to any estimator in which the share constraints are imposed to hold. In contrast, inference using CLEER can be done using standard extremum estimation techniques.[61]

E.3 Conformant GMM

This subsection presents the approximation to the derivative of $\log L$ with respect to θ^ν that can be employed to avoid simulation bias in a GMM estimator.

First, note that[62] $\sum_{j=0}^{J_m} \mathscr{s}_{jm}\big(x^k_{jm}\nu^k - \sum_{k=0}^{J_m} \mathscr{s}_{km} x^k_{km}\nu^k\big) = 0$, such that (26) can be expressed as

$$\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{j=0}^{J_m} D_{im} \frac{y_{ijm} - \sigma^{z_{im}}_{jm}}{\sigma^{z_{im}}_{jm}} \int \mathscr{s}_{jm}(z_{im}, \nu)\Big(x^k_{jm}\nu^k - \sum_{k=0}^{J_m} x^k_{km}\nu^k \mathscr{s}_{km}(z_{im}, \nu)\Big)\mathrm{d}F(\nu),$$

because summing the integrand over j equals zero and $\sigma^{z_{im}}_{jm}/\sigma^{z_{im}}_{jm} = 1$. 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

$$\sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{j=0}^{J_m} \int D_{im}\{y_{ijm} - \mathscr{s}_{jm}(z_{im}, \nu^*)\}\mathrm{d}F(\nu^*) \int \mathscr{s}_{jm}(z_{im}, \nu)\Big(x^k_{jm}\nu^k - \sum_{k=0}^{J_m} \mathscr{s}_{km}(z_{im}, \nu)x^k_{km}\nu^k\Big)\mathrm{d}F(\nu). \tag{58}$$

[59] Indeed, then by a Taylor expansion,

$$\upsilon^\triangledown[\delta_m(\hat{\theta}, s_m) - \delta_m(\theta, \pi^*_m)] \simeq \underbrace{\upsilon^\triangledown[\delta_m(\hat{\theta}, s_m) - \delta_m(\hat{\theta}, \pi^*_m)]}_{\preceq N_m^{-1/2}} + \underbrace{\upsilon^\triangledown \mathbb{D}^*_{\theta m}}_{=0}(\hat{\theta} - \theta^*) + \frac{1}{2}\sum_j \upsilon_j \underbrace{(\hat{\theta} - \theta^*)^\triangledown \partial_{\theta\theta^\triangledown}\delta_{jm}(\theta^*, \pi^*_m)(\hat{\theta} - \theta^*)}_{\preceq I^{-1}},$$

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

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

[61] We are implicitly assuming that the integrals can be computed sufficiently accurately so as not to affect the asymptotics.

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

Thus, approximating the integrals with sums using mutually independent Monte Carlo draws results in a simulated moment that has mean zero because simulation error enters linearly.[63] 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.

F Variance comparison under strong identification

This appendix provides a variance comparison between the CLEER $(\hat{\theta}, \hat{\delta})$ and the corresponding share constrained estimator $(\hat{\theta}^{\text{SHCON}}, \hat{\delta}^{\text{SHCON}})$ that maximizes the mixed logit objective function subject to the share constraints. It then demonstrates that for the share constrained estimator, ignoring the contribution of the estimation of π^* often results in incorrect inference. Throughout, we focus on the strong micro identification case and $I \succ M$ so that we can ignore $\hat{\Phi}$ which is asymptotically negligible for the estimation of (θ^*, δ^*).

First, we compare the asymptotic variance of linear combinations of the estimators $(\hat{\theta}, \hat{\delta})$ and $(\hat{\theta}^{\text{SHCON}}, \hat{\delta}^{\text{SHCON}})$.[64] Specifically, let the matrix $\mathcal{V}^{\text{CLEER}}_{\theta\delta}$ be such that for any conformable C with a fixed number of columns,

$$(C^{\triangledown} \mathcal{V}^{\text{CLEER}}_{\theta\delta} C)^{-1/2} C^{\triangledown} \begin{bmatrix} \hat{\theta} - \theta^* \\ \hat{\delta} - \delta^* \end{bmatrix} \xrightarrow{d} \mathcal{N}(0, \mathbb{I}). \tag{59}$$

Analogously, the matrix $\mathcal{V}^{\text{SHCON}}_{\theta\delta}$ does the same for the share constrained estimator with the identical C.

We can ascertain relative efficiency by comparing the elements of $\mathcal{V}^{\text{CLEER}}_{\theta\delta}$ and $\mathcal{V}^{\text{SHCON}}_{\theta\delta}$. For $\mathcal{Q}^{\mathcal{L}}_{\theta\theta} = \mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\pi} \mathcal{L}^{-1}_{\pi\pi} \mathcal{L}_{\pi\theta}$, $\mathcal{A} = \mathcal{L}^{-1}_{\pi\pi} \mathcal{L}_{\pi\theta}$, $\mathfrak{G} = \mathbb{D}_\theta - \mathbb{D}_\pi \mathcal{A}$, we have

$$\mathcal{V}^{\text{CLEER}}_{\theta\delta} = \begin{bmatrix} (\mathcal{Q}^{\mathcal{L}}_{\theta\theta})^{-1} & (\mathcal{Q}^{\mathcal{L}}_{\theta\theta})^{-1} \mathfrak{G}^{\triangledown} \\ \mathfrak{G} (\mathcal{Q}^{\mathcal{L}}_{\theta\theta})^{-1} & \mathfrak{G} (\mathcal{Q}^{\mathcal{L}}_{\theta\theta})^{-1} \mathfrak{G}^{\triangledown} + \mathbb{D}_\pi \mathcal{L}^{-1}_{\pi\pi} \mathbb{D}^{\triangledown}_\pi \end{bmatrix}$$

Next, consider $\mathcal{V}^{\text{SHCON}}_{\theta\delta}$. This estimator is equivalent to placing infinite weight on $\mathcal{L}^{\blacksquare}$, however, since $\mathcal{L}^{\blacksquare}_\theta = 0$, the other terms ($\mathcal{L}^{\blacklozenge}$ and in general Φ, though not for this example) will still appear in the score and Hessian. Indeed, note that $\mathcal{L}^{\blacklozenge}_\theta = \mathcal{L}_\theta$ etc. So for $\bar{\mathcal{A}} = \mathcal{L}^{\blacksquare -1}_{\pi\pi} \mathcal{L}_{\pi\theta}$, $\bar{\mathcal{L}}_{\pi\pi} = \mathcal{L}^{\blacksquare}_{\pi\pi} + \mathcal{L}_{\pi\theta} \mathcal{L}^{-1}_{\theta\theta} \mathcal{L}_{\theta\pi}$, and $\bar{\mathcal{Q}}^{\mathcal{L}}_{\theta\theta} = \mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\pi} \bar{\mathcal{L}}^{-1}_{\pi\pi} \mathcal{L}_{\pi\theta}$, and $\bar{\mathfrak{G}} = \mathbb{D}_\theta - \mathbb{D}_\pi \bar{\mathcal{A}} \mathcal{L}^{-1}_{\theta\theta} \bar{\mathcal{Q}}^{\mathcal{L}}_{\theta\theta}$, we have,

$$\mathcal{V}^{\text{SHCON}}_{\theta\delta} = \begin{bmatrix} (\bar{\mathcal{Q}}^{\mathcal{L}}_{\theta\theta})^{-1} & (\bar{\mathcal{Q}}^{\mathcal{L}}_{\theta\theta})^{-1} \bar{\mathfrak{G}}^{\triangledown} \\ \bar{\mathfrak{G}} (\bar{\mathcal{Q}}^{\mathcal{L}}_{\theta\theta})^{-1} & \bar{\mathfrak{G}} (\bar{\mathcal{Q}}^{\mathcal{L}}_{\theta\theta})^{-1} \bar{\mathfrak{G}}^{\triangledown} + \mathbb{D}_\pi \bar{\mathcal{L}}^{-1}_{\pi\pi} \mathbb{D}^{\triangledown}_\pi \end{bmatrix}.$$

To see directly that CLEER is at least as efficient for θ^* as SHCON, note first that $\mathcal{L}_{\pi\pi} - \bar{\mathcal{L}}_{\pi\pi} = \mathcal{L}^{\blacklozenge}_{\pi\pi} - \mathcal{L}_{\pi\theta} \mathcal{L}^{-1}_{\theta\theta} \mathcal{L}_{\theta\pi} = \mathcal{L}^{\blacklozenge}_{\pi\pi} - \mathcal{L}^{\blacklozenge}_{\pi\theta} \mathcal{L}^{\blacklozenge -1}_{\theta\theta} \mathcal{L}^{\blacklozenge}_{\theta\pi} \geq 0$ and then that $\mathcal{Q}^{\mathcal{L}}_{\theta\theta} - \bar{\mathcal{Q}}^{\mathcal{L}}_{\theta\theta} = \mathcal{L}_{\theta\pi} (\bar{\mathcal{L}}^{-1}_{\pi\pi} - \mathcal{L}^{-1}_{\pi\pi}) \mathcal{L}_{\pi\theta} \geq 0$.

Next, we discuss the potential hazards of conducting inference on the share constrained estimator. The fundamental issue is that π^* is estimated by s, which is accounted for in $\mathcal{V}^{\text{SHCON}}_{\theta\delta}$ but often neglected in practice. If π^* were known, one could approximate $\mathcal{V}^{\text{SHCON}}_{\theta\delta}$ by an oracle equivalent,

$$\mathcal{V}^{\text{ORACLE}}_{\theta\delta} = \begin{bmatrix} \mathcal{L}^{-1}_{\theta\theta} & \mathcal{L}^{-1}_{\theta\theta} \mathbb{D}^{\triangledown}_\theta \\ \mathbb{D}_\theta \mathcal{L}^{-1}_{\theta\theta} & \mathbb{D}_\theta \mathcal{L}^{-1}_{\theta\theta} \mathbb{D}^{\triangledown}_\theta \end{bmatrix}$$

[63] This is necessary to satisfy condition (iii) of Theorem 3 in PP89. Many of the other assumptions in PP89 hold trivially because our simulated moment (58) is infinitely differentiable in θ and also infinitely differentiable in the simulation draws due to the properties of the mixed logit demand specification (i.e., $\mathscr{s}$ is infinitely differentiable with respect to ν).

[64] Since the dimension of δ grows with M, we focus on linear combinations of fixed length. That is, C has a fixed number of columns while its number of rows grows with M.

However, substituting $\mathcal{V}_{\theta\delta}^{\text{ORACLE}}$ for $\mathcal{V}_{\theta\delta}^{\text{CLEER}}$ in (59) does not result in an asymptotically normal distribution for many choices of C when π^* is estimated by s. To see why, note that since $\mathbb{D}_\theta$ has many more rows than it has columns, $\mathbb{D}_\theta \mathcal{L}_{\theta\theta}^{-1} \mathbb{D}_\theta^\triangledown$ has many eigenvalues equal to zero and so $(C^\triangledown \mathcal{V}_{\theta\delta}^{\text{CLEER}} C)^{-1/2}$ is undefined. For the corresponding eigenvector-directions, the term $\mathbb{D}_\pi \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown$ is first order and hence needed to avoid this degeneracy. So an acceptable substitute under the assumption $N_m \succ I_m$ would be,

$$\mathcal{V}_{\theta\delta}^{\text{CORRECTED}} = \begin{bmatrix} \mathcal{L}_{\theta\theta}^{-1} & \mathcal{L}_{\theta\theta}^{-1} \mathbb{D}_\theta^\triangledown \\ \mathbb{D}_\theta \mathcal{L}_{\theta\theta}^{-1} & \mathbb{D}_\theta \mathcal{L}_{\theta\theta}^{-1} \mathbb{D}_\theta^\triangledown + \underbrace{\mathbb{D}_\pi \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\triangledown}_{\text{needed}} \end{bmatrix}$$

However, to our knowledge, this method of inference has never been applied in any estimator employing share constraints.

G Computation

While CLEER is of theoretical interest, it must also be computationally tractable in order to be appropriate for applied use. This appendix discusses two critical computational aspects of our estimator. First, CLEER 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.[65] The choice of integration method will impact that accuracy of the estimator. We discuss several approaches in app. G.2.

G.1 Dimensionality

We now describe two feasible algorithms for the computation of CLEER which make use of Newton's method with Trust Regions.[66] Recall from (5) that our optimization problem is

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

Like BLP95, we start by concentrating out β which leaves

$$(\hat{\theta}, \hat{\delta}) = \arg\min_{\theta,\delta} \Big(-\log \hat{L}(\theta, \delta) + \hat{\chi}\{\hat{\beta}(\delta), \delta\} \Big). \tag{60}$$

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.

$$\hat{\theta} = \arg\min_{\theta} \Big(-\log \hat{L}(\theta, \delta(\theta)) + \hat{\chi}\{\hat{\beta}(\delta(\theta)), \delta(\theta)\} \Big) \tag{61}$$
$$\text{s.t.: } \delta(\theta) = \arg\min_{\delta} \Big(-\log \hat{L}(\theta, \delta) + \hat{\chi}\{\hat{\beta}(\delta), \delta\} \Big)$$

This approach is similar to that in BLP95 with the important exception that the inner loop objective is to optimize (5)—the same as the outer loop objective—rather than satisfying the share constraint $\pi^* = s$.

The outer loop is over a low dimensional parameter vector, albeit computations of the derivatives

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

[66] As noted below, one of these algorithms computes an estimator that is asymptotically equivalent to CLEER but less computationally intensive.

involves application of the chain rule to account for inner loop optimization.[67]

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, $\hat{\chi}[\hat{\beta}(\delta), \delta] = 0$ for all values of δ, in which case we only need to minimize $-\log \hat{L}$ in the inner loop. Conveniently, $-\log \hat{L}$ is additively separable across markets in δ_m in each δ_m. So we can parallelize the computation of $\hat{\delta}_m(\theta)$ market by market, and each computation is highly tractable.

The overidentified case is more complicated. Since $\hat{\chi}[\hat{\beta}(\delta), \delta] > 0$ and is not additively separable in π_m. However, there are several convenient features which make the inner loop of (61) tractable, even in this case. To simplify exposition but without loss of generality, we will take $\hat{\mathcal{W}}$ in the definition of $\hat{\chi}$ in (9) to be $(B^\triangledown B)^{-1}$ where B is a $J \times d_b$ matrix with rows $b_{jm}^\triangledown$, the instruments introduced in (10).

The first such feature is that $\hat{\beta}(\delta)$ is simply a linear IV estimator, i.e. $\hat{\beta}(\delta) = (X^\triangledown \mathcal{P}_B X)^{-1} X^\triangledown \mathcal{P}_B \delta$, with $\mathcal{P}_B = B(B^\triangledown B)^{-1} B^\triangledown$ an orthogonal projection matrix. Second, $\hat{\chi}$ is quadratic in δ. Thus, writing $\mathcal{P}_{\mathcal{P}_B X} = \mathcal{P}_B X (X^\triangledown \mathcal{P}_B X)^{-1} X^\triangledown \mathcal{P}_B$, the minimand of (60) of becomes

$$-\log \hat{L}(\theta, \delta) + \frac{1}{2} \delta^\triangledown (\mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X}) \delta \tag{62}$$

Third, (62) 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 $\mathcal{K}\mathcal{K}^\triangledown$ 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.[68]

We now turn to the primary complication of applying Newton's method to optimize (62) 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 (62), $H + \mathcal{K}\mathcal{K}^\triangledown$, 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 (62) can by the Woodbury matrix identity be written as $H^{-1} - H^{-1}\mathcal{K}(\mathbb{I} + \mathcal{K}^\triangledown H^{-1} \mathcal{K})^{-1} \mathcal{K}^\triangledown H^{-1}$,

Since $\log \hat{L}$ is additively separable in the δ_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 $H + \mathcal{K}\mathcal{K}^\triangledown$ 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

[67] Note that, as in any nested optimization problem, the outer loop of an optimization problem with objective function of the form $f(\theta, \delta)$ has gradient $f_\theta[\theta, \delta^{\text{sol}}(\theta)]$ since the inner loop solution $\delta^{\text{sol}}(\theta)$ has made $f_\delta[\theta, \delta^{\text{sol}}(\theta)] = 0$ which, by the implicit function theorem, implies that $\partial_{\theta^\triangledown} \delta^{\text{sol}}(\theta) = -f_{\delta\delta}^{-1} f_{\delta\theta}$. Hence, the Hessian becomes $f_{\theta\theta} - f_{\theta\delta} f_{\delta\delta}^{-1} f_{\delta\theta}$. In practice, we do use a change of variables on the θ^ν's in that we optimize over their logarithms to allow for an unconstrained optimization.

[68] To obtain an explicit form for $\mathcal{K}$, let C denote the columns that B and X have in common and $\bar{B}, \bar{X}$ the columns that are unique to each matrix. Then, an explicit form is $\mathcal{K} = \mathcal{U}_B \mathcal{M}_{\mathcal{U}_B^\triangledown \mathcal{U}_X}$ with $\mathcal{U}_B, \mathcal{U}_X$ matrices with orthonormal columns spanning the column spaces of $\mathcal{M}_C \bar{B}, \mathcal{M}_C \bar{X}$, respectively, and $\mathcal{M}$ denoting an annihilator matrix. This follows by expressing $\mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X} = (\mathcal{P}_C + \mathcal{P}_{\bar{B}}) - (\mathcal{P}_C + \mathcal{P}_{\mathcal{P}_{\bar{B}} \bar{X}}) = \mathcal{P}_{\bar{B}} - \mathcal{P}_{\mathcal{P}_{\bar{B}} \bar{X}}$ and applying the singular value decomposition to $\bar{B}$ and $\bar{X}$.

makes the optimization step of the inner loop practical for many products.

For even larger problems, one may consider an alternative approach which is implemented in the Grumps package as the "cheap" estimator. Here, we alter the inner loop dropping $\hat{\chi}$, so the full problem becomes,

$$\hat{\theta} = \arg\min_{\theta}\Big(-\log \hat{L}(\theta, \delta(\theta)) + \hat{\chi}\{\hat{\beta}(\delta(\theta)), \delta(\theta)\}\Big) \tag{63}$$
$$\text{s.t.: } \delta(\theta) = \arg\min_{\delta} -\log \hat{L}(\theta, \delta)$$

This yields a different, but asymptotically equivalent, estimator to CLEER. However, this estimator is not robust to zero shares. We further note that the "cheap" estiamator may be useful as a warm start for CLEER in some cases.

G.2 Numerical integration

As we have pointed out, the largest disadvantage of our estimator is that a computable version relies on numerical integration. This is costly since to avoid affecting the asymptotic behavior, numerical integration error must be negligible. Of course, as always, we can arbitrarily reduce the numerical approximation error by incurring a higher computational cost. In contrast, GMM estimators can be computed via the method of simulated moments (MSM). MSM can achieve the same convergence rate as its theoretical counterpart by averaging over noisy approximations of these integrals. However, as discussed 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.

CLEER evaluates two types of integrals, those over ν (e.g., π^z_m) and those over both ν and z (e.g., π_m). 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 usually of small dimension, so the curse of dimensionality associated with tensor product quadrature methods is less binding.[69] We examine the sensitivity of CLEER's numerical performance to the number of nodes used for numerical integration in section 7.4.

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 ($\sum_m J_m I_m$). 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. In our implementation for section 7, we use 10,000 quasi-monte carlo draws to approximate these integral for all estimators.

H Monte Carlo Design

In this appendix we present the full details of our Monte Carlo design and implementation. While some of this material is redundant with the summary presented in section 7, it is also included here in order to provide for a single, comprehensive overview.

[69] 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.

H.1 Data Generating Process

Our empirical design includes two observable 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 $(\theta_1^{*z}, \theta_2^{*z})$; and two random coefficients $(\theta_1^{*\nu}, \theta_2^{*\nu})$.

Mean product quality is specified as $\delta^*_{jm} = \beta^*_c + \beta^*_1 x^1_{jm} + \beta^*_2 x^2_{jm} + \xi_{jm}$, The unobservable product characteristic ξ_{jm} is also distributed as a standard normal independent across j and m.[70]

We specify that one of the observable product characteristics, x^1, is correlated with unobserved characteristics ξ_{jm}, and thus endogenous. Specifically, so that x^1 is normally distributed, let a vector of instruments b^1 and random noise u both be vectors drawn from a standard normal distribution independent of ξ and each other. Then construct x^1 according to,

$$x^1_{jm} = w_a b^1_{jm} + \sqrt{1 - w_a^2}\,(w_c u_{jm} + \sqrt{1 - w_c^2}\xi_{jm}) \tag{64}$$

where $w_a = w(a) = a/\sqrt{a^2 + (1-a)^2}$ for $a \in [0,1]$ governs the strength of the instrument b^1 and $w_c = w(c)$ for $c \in [0,1]$ governs the degree to which the remaining variation in x^1 is due to random noise versus the product's unobserved quality. In estimation, we use b^1 as an observed instrument for x^1. The remaining characteristic x^2_{jm} is exogenous and drawn from a standard normal independent of all other variables.

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 $\mu^{z_{im}}_{jm} = \theta_1^{*z} z^1_{im} x^1_{jm} + \theta_2^{*z} z^2_{im} x^2_{jm}$,

As in section 7, altering θ^{*z} affects the strength of identification of $\theta^{*\nu}$ 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 independent and drawn from the standard normal distribution. Following the model as well as standard practice, this distribution is assumed to be known to the researcher. The unobserved heterogeneity term in utility is $\mu^{\nu_{im}}_{jm} = \theta_1^{*\nu}\nu^1_{im}x^1_{jm} + \theta_2^{*\nu}\nu^2_{im}x^2_{jm}$.

In addition to the instrument b^1 for x^1 described above as well as a constant and the exogenous characteristic x^2, we utilize three additional "BLP instruments" constructed from product characteristics for the PLMs (4). We construct a differentiation IV for x^2 following GH20. Specifically, for b^2 we use,

$$b^2_{jm} = \sum_{j' \in J_m \setminus j} (x^2_{jm} - x^2_{j'm})^2, \tag{65}$$

This instrument is valid since it depends entirely on the exogenous vector x^2. We also construct the differentiation instrument for x^1. Here, we must make use of 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 b^3 analogous to (65). The final instrument is simply the number of products in each market m. This varies across markets but not within market. Since $d_b = 6 > d_\beta = 3$, $\hat{\chi}$ is overidentified for β^* and the extra exclusion restrictions are potentially useful to identify θ^*. Note that since $d_\theta = 4$, the score of the likelihood for CLEER and MDLE, and the two covariance moments for GMM-M are necessary to identify the full parameter vector.

[70] In a previous version of this paper (Grieco et al., 2023*b*), we have used a Pareto distribution for ξ_{jm}. The Pareto distribution more closely mimics the "80/20" rule commonly observed in market share data. However, the Pareto distribution has thicker tails than allowed by G. This choice results in a bias in the PLMs which is visible for some simulations. In practice CLEER still outperforms the other estimators.

We include the same instruments in all three of the estimators we consider.

H.2 Baseline Parameterization

We organize all of our experiments around a baseline specification of the data generating process, which we now describe. Except where they are explicitly varied, these specifications remain constant throughout section 7.

The parameters $\beta^* = [-6, 1, 1]^\triangledown$, $\theta^{*z} = [1, 1]^\triangledown$, and $\theta^{*\nu} = [1, 1]^\triangledown$ were chosen so that in the baseline specification, average share of the inside products is .0206; and the first decile of shares is 0.0006 on average. The average share of the outside good is .6095, with a standard deviation of .1326. We let $a = 0.5$ and $c = 0.5$; which results in a the mean F-stat for the first stage regression of x^1 on the instruments of 190.71 with a standard deviation of 18.05 across our 1000 simulations.

We draw data for $M = 50$ 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} products.[71] There are $N_m = 100\,000$ consumers in each market. We take a random sample (without replacement) of size $I_m = 1\,000$ for the micro dataset of each market.

All three estimators must integrate over both ν and z to compute the function 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 7.4.

H.3 Implementation

For all experiments, we estimate the model for each of 1 000 replications of the data generating process. In rare instances, we draw a dataset where some product has a share of zero, in which case we discard the draw and sample again. Because GMM-M requires $s_m > 0$, it is unable to handle these cases, our other estimators may also be affected as we describe in app. I. In practice, most practitioners drop products when no sales are observed, since it is difficult to determine whether they were actually available for purchase. In, app. I, we investigate performance of all three estimators following this practice. For CLEER and MDLE we use a single, arbitrary, starting point. For GMM-M, which is known to have local optima, we multi-start from three values, including the truth. From the three runs, we use the one generating the smallest minimum.

Finally, we must choose weight matrices for all three estimators. For CLEER and MDLE two step, we use the standard initial choice of $(B^\triangledown B)^{-1}$. Hence, our results do not take advantage of optimal instruments. For GMM-M, we follow the `pyblp` default, which constructs a weight matrix for both PLMs and micro-moments that would be optimal if the initial parameter were the truth. Note that since we perform a modest multistart for GMM-M with one starting point being the truth, this means that one of the GMM-M implementations utilizes the *true* optimal weight matrix (as opposed to a consistent estimate thereof).

For these reasons, if one wishes to view our results as a comparison between the implementations of the three estimators—which is *not* our goal—one should view results in favor of CLEER or MDLE as conservative. However, our primary purpose with these experiments is to straightforwardly illustrate the theoretical properties of CLEER and the alternative estimators across a variety of designs.

[71] For the experiment varying the number of markets, we similarly vary the number of products in each market with one market of {10, 12, 14, 16, 18, 20, 22, 24, 26, 28} products for $M = 10$, and 100 markets of {10, 12, 14, 16, 18, 20, 22, 24, 26, 28} products for $M = 1\,000$.

I Small market population

In this appendix we discuss performance of CLEER and the other estimators presented in section 7 when the population size is small. We also present Monte Carlo results for this situation.

When N_m is low, two issues arise. First, sampling error in s increases, which makes imposing the share constraint more costly in terms of efficiency. This impacts GMM-M but not CLEER or MDLE.

Second, when N_m is small, the probability that some offered products are not purchased, so $s_{jm} = 0$, increases. When $s_{jm} = 0$, the share constraint cannot be solved, making it impossible to compute an estimate for our GMM-M estimator. CLEER and the MDLE two step face similar but less severe issues when a product is not purchased.

The MDLE objective function $\log \hat{L}$ is well defined when $s_{jm} = 0$, however its score with respect to δ_{jm} is negative for all finite δ_{jm}.[72] In our view, the first step of MDLE is quite robust, as one can simply drop δ_{jm} from the parameter set when $s_{jm} = 0$ without affecting the likelihood to recover θ and the remaining elements of δ.[73] However, dropping δ_{jm} does impact the PLMs, so the second step of MDLE will suffer from selection bias in the estimation of β.

In principle, CLEER can address this issue when $d_b > d_\beta$, since once the PLMs are added to the objective function, it is no longer optimal to let $\delta_{jm} \to -\infty$, as this will cause $\hat{\chi}$ to diverge, see (10). However, once $d_b - d_\beta$ shares are zero in the data, $\hat{\chi}$ can be set to 0 for any θ, so $\hat{\theta}$ must be estimated from the micro data. If the number of zero shares is larger than $d_b - d_\beta$, then the PLMs can be satisfied with equality using only a subset of δ_{jm} for zero share products and the remainder are free diverge as above. Consequently, CLEER can only be computed provided that the number of products with zero shares is no greater than than $d_b - d_\beta$. This means that while CLEER can be estimated for allowing the presence of a small number of zeros, it will eventually break down for markets with very low N_m as the number of zero share products increases.

For empirical applications, practical considerations also arise when $s_{jm} = 0$ is observed in data. Foremost among them is that the researcher is usually uncertain as to whether or not product j was actually available to consumers in market m as it may have been out of stock or simply not offered. The issue of stock outs is broader than simply observing zero shares, but has been typically ignored in the applied literature.[74]

In practice, applied researchers have commonly dropped products with zero market share from the choice set of market m while assuming all other products were available to all consumers. We now examine the impact of this practice when, following our model, all products are available but some were not purchased by any consumer in the market population.

Specifically, we consider our baseline DGP from section 7, but lower the market population size from 100,000 in the baseline to $N_m = \{10,000; 5,000; 1,000\}$. This reduction in N_m makes the probability of drawing a product with a market share of 0 increase from being negligible in the baseline to 0.22, 0.90, and 7.79 percent respectively. Consequently, the probability that a market contains a product with zero share for these experiments is, 7.9, 28.22, and 92.18 percent. Thus we consider the three cases presented to be examples of small, moderate, and extreme zero shares problems.

[72] One can this immediately for $L^{\blacksquare}$ by examining (56). For $L^{\blacklozenge}$, it is intuitive since $\sigma^z_{jm} > 0$ for any finite δ_{jm} and $\sum_{\ell=0}^{J_m} \sigma^z_{\ell m} = 1$.

[73] Note that since δ are location normalized against the outside goods, the remaining δ will be unbiased provided $s_{0m} > 0$.

[74] An important exception is Conlon and Mortimer (2013), which leverages periodic observations of product availability to estimate a demand model with endogenous stock outs. We do not consider availability of such data in our analysis.

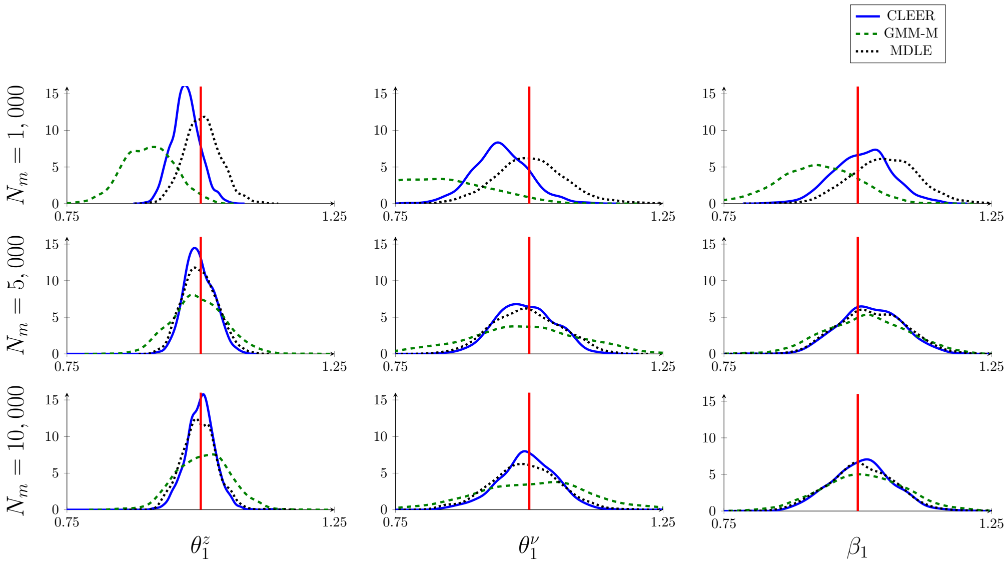


Figure 11: *Distribution of parameters for different population sizes N_m.*

The distributions of our estimators are presented in fig. 11. To reiterate, when a product has a market share of zero, it is dropped from the dataset prior to estimation, enabling all three estimators to be computed. When $N_m = 10,000$, there is relatively little difference between these results and our baseline for any estimator. When $N_m = 5,000$, the variance of GMM-M appears to increase slightly for $\hat{\theta}$, and some bias for $\hat{\beta}$ appears for all three estimators. However, overall performance is acceptable. In the extreme case where $N_m = 1,000$, the GMM-M estimator is severely biased for all parameters, this is a direct result of its reliance on the product level moments which now suffer from significant selection bias. CLEER is also biased for the same reason, but to a lesser extent as it combines information from the biased PLMs and the likelihood. On the other hand, MDLE, which ignores the PLMs when estimating $\hat{\theta}$, remains unbiased for $\hat{\theta}$ and performs well.

All three estimators exhibit bias for $\hat{\beta}$. Interesting, the bias for MDLE and GMM-M are in opposite directions. It is intuitive that the distribution of CLEER is between the other two estimators, however there is no reason to expect CLEER will be unbiased for $\hat{\beta}$ in general.

To summarize these results, the standard practice of dropping zero or small share products, while inducing bias, did not substantially affect any of the estimators. Bias did become apparent in our extreme case once the share of products dropped rose to over 7 percent. Because this bias is entirely due to selection affecting the PLMs, the first step of MDLE remains consistent even in the extreme case.

In cases where the share of zero share products significant and it is known these products were available to consumers, our results indicate it may be fruitful to consider adjusting the second stage of MDLE to account for selection. We leave such a possibility for future research.

J Technical Lemmas

In this appendix, we cover technical lemmas we refer to in our paper, which are relegated to the online appendix due to space constraints.

J.1 Asymptotic normality of other parameters

Proof (of L1). The statement of L1 says that for $\hat{\upsilon} = [(B^\triangledown \xi)^\triangledown, \mathcal{L}_\theta^\triangledown, \mathcal{L}_\pi^\triangledown]^\triangledown$ and $\omega = [\beta^\triangledown, \theta^\triangledown, \delta^\triangledown]^\triangledown$, $(A^\triangledown \hat{\mathcal{H}}^{-1} A)^{-1/2} A^\triangledown (\hat{\omega} - \omega^*) \simeq A^\triangledown \hat{\upsilon} \xrightarrow{d} \mathcal{N}(0, \mathbb{I})$, for matrices $A, \mathcal{H}$.

Here we establish asymptotic normality of linear combinations of $(\hat{\beta}, \hat{\theta}, \hat{\delta})$. Theorem 2 is a special case for $\varLambda^{\triangledown} = \begin{bmatrix} \mathbb{0} & \mathbb{I} & \mathbb{0} \end{bmatrix}$. In theorem 2, we showed asymptotic normality of $\hat{\theta}$ by writing it as a linear combination of $\hat{\upsilon}$. Specifically, L8 showed that $\mathcal{V}_{\theta}^{-1/2}\varGamma_{\theta}\hat{q}_{\theta} \xrightarrow{d} \mathcal{N}(0, \mathbb{I})$. Recalling (48) and substituting $\hat{r} = \hat{\mathcal{L}}_{\pi}^{\blacklozenge} - \mathcal{L}_{\pi\pi}^{\blacksquare-1}(s - \pi^*) \simeq \hat{\mathcal{L}}_{\pi}$,

$$\hat{q}_{\theta} \simeq \Big(\sum_{m}(\varXi B_m^{\triangledown}\xi_m + \hat{\mathcal{L}}_{\theta m} - \mathcal{L}_{\theta\pi m}\mathcal{L}_{\pi\pi m}^{-1}\hat{\mathcal{L}}_{\pi m})\Big) = \begin{bmatrix} \varXi & \mathbb{I} & -\mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1} \end{bmatrix} \hat{\upsilon},$$

which implies for $A^{\triangledown} = \mathcal{V}_{\theta}^{-1/2}\varGamma_{\theta}\begin{bmatrix} \varXi & \mathbb{I} & -\mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1} \end{bmatrix}$ we have $A^{\triangledown}\hat{\upsilon} \xrightarrow{d} \mathcal{N}(0, \mathbb{I})$. To show asymptotic normality of linear combinations of $(\hat{\beta}, \hat{\theta}, \hat{\delta})$, we reuse the same argument for a general $\varLambda$.

Before providing the general form of A, note that the initial steps are identical: The quadratic approximation for $(\hat{\theta}, \hat{\pi})$ obtained in theorem 2 (steps *1* and *2*) can be reused verbatim.

In the general case, $A^{\triangledown}$ will be of the form, $(\varLambda^{\triangledown}\mathcal{H}^{-1}\varLambda)^{-1/2}\varLambda^{\triangledown}\varUpsilon$, where $\hat{\mathcal{H}}^{-1}$ is a sample analog of,

$$\mathcal{H}^{-1} \coloneqq \mathbb{V}(\varUpsilon\hat{\upsilon} \mid \mathbb{A}) = \varUpsilon \begin{bmatrix} B^{\triangledown}\mathcal{V}_{\xi}B & \mathbb{0} & \mathbb{0} \\ \mathbb{0} & \mathcal{L}_{\theta\theta} & \mathcal{L}_{\theta\pi} \\ \mathbb{0} & \mathcal{L}_{\pi\theta} & \mathcal{L}_{\pi\pi} \end{bmatrix} \varUpsilon^{\triangledown}.$$

For $\varUpsilon$,

$$\varLambda^{\triangledown}(\hat{\omega} - \omega^*) \simeq \varLambda^{\triangledown}\varUpsilon_1 \begin{bmatrix} B^{\triangledown}\xi \\ \hat{\theta} - \theta^* \\ \hat{\delta} - \delta^* \end{bmatrix} \simeq \varLambda^{\triangledown}\varUpsilon_1\varUpsilon_2 \begin{bmatrix} B^{\triangledown}\xi \\ \hat{\theta} - \theta^* \\ \hat{\pi} - \pi^* \end{bmatrix} \simeq \varLambda^{\triangledown}\varUpsilon_1\varUpsilon_2\varUpsilon_3\upsilon,$$

where $\varUpsilon_1, \varUpsilon_2, \varUpsilon_3$ are respectively given by,

$$\begin{bmatrix} (\mathcal{P}_B X)^{+}B^{+\triangledown} & \mathbb{0} & (\mathcal{P}_B X)^{+} \\ \mathbb{0} & \mathbb{I} & \mathbb{0} \\ \mathbb{0} & \mathbb{0} & \mathbb{I} \end{bmatrix}, \begin{bmatrix} \mathbb{I} & \mathbb{0} & \mathbb{0} \\ \mathbb{0} & \mathbb{I} & \mathbb{0} \\ \mathbb{0} & \mathbb{D}_{\theta} & \mathbb{D}_{\pi} \end{bmatrix}, \begin{bmatrix} \mathbb{I} & \mathbb{0} & \mathbb{0} \\ \mathbb{0} & -\mathcal{Q}_{\theta\theta}^{-1} & \mathbb{0} \\ \mathbb{0} & \mathbb{0} & -\mathcal{Q}_{\pi\pi}^{-1} \end{bmatrix} \begin{bmatrix} \mathbb{I} & \mathbb{0} & \mathbb{0} \\ \mathbb{D}_{\theta}^{\triangledown}\mathcal{P}B^{+\triangledown} & \mathbb{I} & -\mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1} \\ \mathbb{D}_{\pi}^{\triangledown}\mathcal{P}B^{+\triangledown} & -\mathcal{L}_{\pi\theta}\mathcal{L}_{\theta\theta}^{+} & \mathbb{I} \end{bmatrix}.$$

The formula for $\varUpsilon_1$ is due to the fact that $\hat{\beta} - \beta^* = (\mathcal{P}_B X)^{+}\hat{\delta} - \beta^* = (\mathcal{P}_B X)^{+}[(\hat{\delta} - \delta^*) + B^{+\triangledown}B^{\triangledown}\xi]$, $\varUpsilon_2$ is essentially applying the delta method to the transformation from π to δ, and $\varUpsilon_3$ amounts to a linearization of $B^{\triangledown}\xi$, $\hat{\theta} - \theta^*$, $\hat{\pi} - \pi^*$. The product $\varUpsilon = \varUpsilon_1\varUpsilon_2\varUpsilon_3$ is for $\mathcal{C} = (\mathcal{P}_B X)^{+}(\mathbb{I} - \mathbb{D}_{\theta}\mathcal{Q}_{\theta\theta}^{-1}\mathcal{P} - \mathbb{D}_{\pi}\mathcal{Q}_{\pi\pi}^{-1}\mathbb{D}_{\pi}^{\triangledown}\mathcal{P})B^{+\triangledown}$ given by,

$$\begin{bmatrix} \mathcal{C} & -(\mathcal{P}_B X)^{+}(\mathbb{D}_{\theta}\mathcal{Q}_{\theta\theta}^{-1} - \mathbb{D}_{\pi}\mathcal{Q}_{\pi\pi}^{-1}\mathcal{L}_{\pi\theta}\mathcal{L}_{\theta\theta}^{+}) & -(\mathcal{P}_B X)^{+}(\mathbb{D}_{\pi}\mathcal{Q}_{\pi\pi}^{-1} - \mathbb{D}_{\theta}\mathcal{Q}_{\theta\theta}^{-1}\mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}) \\ -\mathcal{Q}_{\theta\theta}^{-1}\mathbb{D}_{\theta}^{\triangledown}\mathcal{P}B^{+\triangledown} & -\mathcal{Q}_{\theta\theta}^{-1} & \mathcal{Q}_{\theta\theta}^{-1}\mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1} \\ \mathbb{0} & -(\mathbb{D}_{\theta}\mathcal{Q}_{\theta\theta}^{-1} - \mathbb{D}_{\pi}\mathcal{Q}_{\pi\pi}^{-1}\mathcal{L}_{\pi\theta}\mathcal{L}_{\theta\theta}^{+}) & -(\mathbb{D}_{\pi}\mathcal{Q}_{\pi\pi}^{-1} - \mathbb{D}_{\theta}\mathcal{Q}_{\theta\theta}^{-1}\mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}) \end{bmatrix}.$$

So $\varUpsilon$ transforms $\hat{\upsilon}$ into $(\hat{\beta}, \hat{\theta}, \hat{\delta})$ up to negligible terms and $\mathcal{H}^{-1}$ is the variance of $\varUpsilon\hat{\upsilon} \simeq (\hat{\beta}, \hat{\theta}, \hat{\delta})$. □

J.2 Other lemmas referred to in the main text and app. B

The model implies that π^* is in the interior of $\mathbb{\Pi}$. The following lemma establishes a bound for π^* and related objects following our assumptions, especially G.

Lemma 9 (Bounds for δ^*, π^*, and related objects). Recall that $\kappa_{\delta}^{\uparrow} = 2\sqrt{2c_{\xi}^{*}\log M}$, $\kappa = \exp(-4\kappa_{\delta}^{\uparrow})$, and let $\kappa_{\pi} = \kappa^{3/4} = \exp(-3\kappa_{\delta}^{\uparrow})$, so $\kappa_{\pi} \succ \kappa$. Then, (a) $\mathbb{P}\big(\max_{m,j}|\delta_{jm}^{*}| > \kappa_{\delta}^{\uparrow}\big) \prec 1$; (b) $\max_{m,j}\kappa_{\pi}/\pi_{jm}^{*} \prec 1$ and $\max_{m,j}\kappa_{\pi}/s_{jm} \prec 1$; (c) $\min_m \inf_{\mathbb{\Pi}_m^{\kappa c}}[\hat{\mathcal{L}}_m^{\blacksquare}(\pi_m) \ / \ N_m] \succ \kappa_{\pi}\log(\kappa_{\pi}/\kappa)$; (d) For a constant C, $\mathbb{P}\big\{\max_m \sup_{\varTheta\times\mathbb{\Pi}_m^{\kappa}} \boldsymbol{\lambda}_{\max}[\mathbb{D}_{\pi m}(\theta, \pi_m)] \le C\kappa^{-3}\big\} = 1$; (e) For any $0 < p < \infty$ and some constant C only depending on p, $\max_m \mathbb{E}\big\{\sup_{\varTheta}\boldsymbol{\lambda}_{\max}^{p}[\mathbb{D}_{\pi m}(\theta, \pi_m^{*})]\big\} \le C$; (f) For any $0 < p < \infty$ and some constant

C only depending on p, $\max_m \mathbb{E} \sup_{\Theta \times \mathbb{\Pi}_m^\kappa} \mathbb{1}(\|\pi_m - \pi_m^*\| \le \kappa) \boldsymbol{\lambda}_{\max}^p[\mathbb{D}_{\pi m}(\theta, \pi_m)] \le C$.

Proof. First (a). By the triangle inequality, for and some fixed $C < \infty$,

$$\mathbb{P}(\max_{m,j} |\delta_{jm}^*| > \kappa_\delta^\uparrow) \le \mathbb{P}(\exists m, j\colon |\xi_{jm}^*| > \kappa_\delta^\uparrow - |x_{jm}^\triangledown \beta^*|) \overset{\text{G[i],H}}{\le} \mathbb{P}(\exists m, j\colon |\xi_{jm}^*| > \kappa_\delta^\uparrow - C) \overset{\text{Bonferroni}}{\le}$$
$$\sum_{m,j} \mathbb{P}(|\xi_{jm}^*| > \kappa_\delta^\uparrow - C) \overset{\text{G[ii]}}{\le} 2M\bar{J} \exp\big[-(\kappa_\delta^\uparrow - C)^2/(2c_\xi^*)\big] \prec 1.$$

For (b), for some fixed $c > 0$,

$$\pi_{jm}^* = \int_{z,\nu} \frac{\exp\big[\mu(\theta^*, z, \nu, x_{jm})\big] \exp(\delta_{jm}^*)}{\sum_t \exp\big[\mu(\theta^*, z, \nu, x_{tm})\big] \exp(\delta_{tm}^*)} \ge \frac{\exp(\delta_{jm}^*)}{\max_t \exp(\delta_{tm}^*)} \int_{z,\nu} \frac{\exp\big[\mu(\theta^*, z, \nu, x_{jm})\big]}{\sum_t \exp\big[\mu(\theta^*, z, \nu, x_{tm})\big]}$$
$$\overset{\text{G,H}}{\ge} c \exp\big(\delta_{jm}^* - \max_t \delta_{tm}^*\big). \qquad (66)$$

Hence, for any $C < \infty$,

$$\mathbb{P}\Big(\min_{m,j} \pi_{jm}^* < C\kappa_\pi\Big) \le \mathbb{P}\Big[c \exp\Big(\min_{m,j} \delta_{jm}^* - \max_{m,j} \delta_{jm}^*\Big) < C\kappa_\pi\Big] \le \mathbb{P}\Big(-2 \max_{m,j} |\delta_{jm}^*| < \log \frac{\kappa_\pi C}{c}\Big)$$
$$= \mathbb{P}\Big(\max_{m,j} |\delta_{jm}^*| > -\frac{1}{2} \log \frac{\kappa_\pi C}{c}\Big) = \mathbb{P}\Big(\max_{m,j} |\delta_{jm}^*| > \frac{3}{2}\kappa_\delta^\uparrow - \frac{1}{2} \log \frac{C}{c}\Big) \overset{\text{(a)}}{\prec} 1,$$

which establishes the first half of the assertion. The other half then follows from L12(d).

Now (c). Suppose without loss of generality that within a market products are such that $\pi_{1m} = \min_j \pi_{jm}$. Then,

$$\inf_{\mathbb{\Pi}_m^{\kappa c}} \frac{\hat{\mathcal{L}}_m^\blacksquare(\pi_m)}{N_m} = \inf_{\mathbb{\Pi}_m^{\kappa c}} \sum_j s_{jm} \log \frac{s_{jm}}{\pi_{jm}} \ge \inf_{\pi_{1m} \le \kappa} \Big(s_{1m} \log \frac{s_{1m}}{\pi_{1m}} + (1 - s_{1m}) \log \frac{1 - s_{1m}}{1 - \pi_{1m}}\Big).$$

By L12(d), the infimum is (for all m simultaneously) attained at $\pi_{1m} = \kappa$, such that the infimum is bounded below by $s_{1m} \log(s_{1m}/\kappa) + (1 - s_{1m}) \log[(1 - s_{1m})/(1 - \kappa)]$. The stated result then follows from (b).

Next, (d). Note first that $\mathbb{D}_{\pi m}(\theta, \pi_m) = \mathbb{Q}_m^{-1}(\theta, \pi_m)$, where for $\mathcal{S}_m = \operatorname{diag}(\mathscr{s}_m)$,

$$\mathbb{Q}_m = \int_{z,\nu} (\mathcal{S}_m - \mathscr{s}_m \mathscr{s}_m^\triangledown) = \int_{z,\nu} \mathcal{S}_m^{1/2}[\mathbb{I} - \mathcal{S}_m^{-1/2} \mathscr{s}_m \mathscr{s}_m^\triangledown \mathcal{S}_m^{-1/2}]\mathcal{S}_m^{1/2} \ge \int_{z,\nu} \mathcal{S}_m \mathscr{s}_{0m} \ge \min_j \int_{z,\nu} \mathscr{s}_{jm} \mathscr{s}_{0m} \mathbb{I},$$

where the penultimate inequality follows from the fact that $\mathbb{I} - \mathcal{S}_m^{-1/2} \mathscr{s}_m \mathscr{s}_m^\triangledown \mathcal{S}_m^{-1/2}$ has eigenvalues that are bounded below by $\mathscr{s}_{0m}$.[75] Analogous to the proof of (b), we have $\min_j \int_{z,\nu} \mathscr{s}_{jm} \mathscr{s}_{0m} \ge C_3 \exp(-3 \max_j |\delta_{jm}|)$ for some fixed $C_3 > 0$. Consequently,

$$\boldsymbol{\lambda}_{\max}[\mathbb{D}_{\pi m}(\theta, \pi_m)] = \frac{1}{\boldsymbol{\lambda}_{\min}[\mathbb{Q}_m(\theta, \pi_m)]} \le \exp[3 \max_j |\delta_{jm}(\theta, \pi_m)|] \,/\, C_3. \qquad (67)$$

For uniformity, it remains to be shown that $\max_\Theta \max_m \max_{\mathbb{\Pi}_m^\kappa} \exp[\max_j |\delta_{jm}(\theta, \pi_m)|] \le C_\gamma/\kappa$. By the definition of $\mathbb{\Pi}_m^\kappa$, $\kappa \le \pi_{0m} = \textstyle\int [\sum \exp(\delta_{jm} + \mu_{jm})]^{-1} \le C_\alpha \exp(-\max_j \delta_{jm})$. Moreover, $\kappa \le \min_j \pi_{jm} = \min_j \textstyle\int \exp(\delta_{jm} + \mu_{jm}) \,/\, [\sum \exp(\delta_{tm} + \mu_{tm})] \le C_\beta \exp(\min_j \delta_{jm})$. Combining these, we have for all $\delta_{jm}\colon \kappa \,/\, C_\beta \le \exp(\delta_{jm}) \le C_\alpha \,/\, \kappa$, and so $\exp(|\delta_{jm}|) \le \max(C_\alpha, C_\beta) \,/\, \kappa$, which establishes (d).

[75] We use the fact that the smallest eigenvalue of $\mathbb{I} - \upsilon\upsilon^\triangledown$ corresponds to the eigenvector υ and is equal to $1 - \|\upsilon\|^2$, which for $\upsilon = \mathcal{S}_m^{-1/2} \mathscr{s}_m$ is equal to $1 - \sum_{j>0} \mathscr{s}_{jm} = \mathscr{s}_{0m}$.

Penultimately, (e). Apply (67) for $\pi_m = \pi^*_m$. Using the analogous argument as follows (67), for some fixed C_1 and all $\theta \in \Theta, m, j$: $\exp[|\delta_{jm}(\theta, \pi^*_m)|] \leq C_1/\pi^*_{jm}$. Now, for some fixed C_2 and any $p > 0$,

$$\mathbb{E}[(\pi^*_{jm})^{-p}] \leq C_2^p \mathbb{E}[\exp(p|\delta^*_{jm}|)] \overset{\text{triangle}}{\leq} C_2^p \mathbb{E}[\exp(p|x^{\triangledown}_{jm}\beta|)\exp(p|\xi_{jm}|)] \overset{\text{G[i],[ii], H[ii]}}{\leq} \infty,$$

Finally, (f) follows from (e) and (b). □

In the remainder, we (as earlier) use the symbols $\ell_{ijm} = \log \varsigma_{ijm}$, $\ell_{jm} = \log \varsigma_{jm} = \mathbb{E}(\ell_{ijm} \mid \mathbb{A})$, and $\Delta\ell_{ijm} = \ell_{ijm} - \ell_{jm}$, and use a tilde to indicate when ℓ is used as a function of δ instead of π, e.g. $\tilde{\ell}_{ijm}$.

Lemma 10 (Uniform convergence of $\Delta\hat{\mathcal{L}}^{\blacklozenge}$ and its derivatives). (a) Let $r_\epsilon(\theta) = \sqrt{\max[\rho_{\mathrm{id}}(\theta), \breve{\rho}_{\mathrm{id}}(\epsilon)]} \log^2 I_+$. Then, $\sup_\Theta |\Delta\hat{\mathcal{L}}^{\blacklozenge}(\theta, \pi^*)/r_\epsilon(\theta)| \prec 1$; (b) Let u be a vector of nonnegative integers indicating derivative order with respect to each element of $\psi_m = (\theta, \pi_m)$, let $|u|$ denote the sum of the elements in u, and let u_z denote the number of derivatives with respect to elements of θ^z. Let further, $r^u_m = \sqrt{I_m}\kappa^{-3|u|}(\log I_+)^{2+\max(u_z,1)} + \exp(-N)$, where the $\exp(-N)$ term serves to ensure that we are not dividing by zero. Then, for $|u| > 0$, $\max_m \sup_{\Theta\times\mathbb{\Pi}^\kappa_m}\big[|\partial^u \Delta\hat{\mathcal{L}}^{\blacklozenge}_m(\theta, \pi_m)|/r^u_m\big] \prec 1$; and if $|u| = 0$ then $\max_m \sup_{\Theta\times\mathbb{\Pi}_m}\big[|\Delta\hat{\mathcal{L}}^{\blacklozenge}_m(\theta, \pi_m)|/r^u_m\big] \prec 1$.

Proof. $\Delta\hat{\mathcal{L}}^{\blacklozenge}$ is a sum over I terms, so if I does not grow then the results are trivial. So, suppose that $I \succ 1$.

We first show (a). We use L14(a) conditional on the $\mathcal{I}_m$'s. Since we do not need a result for each market separately, we have no use for the g subscript in L14. Using the superscript $^{\mathrm{L14}}$ to distinguish objects in L14 from objects here, take $\psi^{\mathrm{L14}} = \theta$, $n^{\mathrm{L14}} = I$, and write $\hat{\mathbb{f}}^{\mathrm{L14}}(\theta) = \Delta\hat{\mathcal{L}}^{\blacklozenge}(\theta, \pi^*)/r_\epsilon(\theta)$ as $\sum_m \sum_{\mathcal{I}_m} \zeta_{im}(\theta)$ for $\zeta_{im} = \sum_j (y_{ijm} - \varsigma^*_{ijm})[\Delta\ell_{ijm}(\theta, \pi_m) - \Delta\ell_{ijm}(\theta^*, \pi^*_m)]/r_\epsilon(\theta)$, where each $(\zeta^{\mathrm{L14}}_i, z^{\mathrm{L14}}_i)$ corresponds to (ζ_{im}, z_{im}) for one $(m \in \{1, \ldots, M\}, i \in \mathcal{I}_m)$ combination. We now verify the conditions of L14.

First, L14[i] is satisfied if we make $\boldsymbol{\delta}^{\mathrm{L14}}$ decrease at a sufficiently fast polynomial rate of I because $\hat{\mathbb{f}}^{\mathrm{L14}}$ is differentiable and by L15(c). We now establish condition L14[ii] using L14(b), for which we need to check L14[iii],[iv],[v]. L14[iii] holds by G[iii]. For [iv] and [v], take $\boldsymbol{\beta}^{\mathrm{L14}} = \log I$, such that [iv] is satisfied. Finally, [v]. Note that $\mathbb{h}^{\mathrm{L14}} \preceq \breve{\rho}^{-1}_{\mathrm{id}}(\epsilon)\log I$ by L15(c). By L15(d), $\bar{\boldsymbol{\sigma}}^{\mathrm{L14}2} \preceq (\log I)^{-4}$. To verify [v], first note that $\exp(-c/\mathbb{h}^{\mathrm{L14}})$ and $\exp(-c/\bar{\boldsymbol{\sigma}}^{\mathrm{L14}2})$ decrease faster than any power of I. Now, due to the compactness of Θ, we can choose $\boldsymbol{T}^{\mathrm{L14}}$ to increase at a (sufficiently fast) polynomial rate of I (that depends on our choice of $\boldsymbol{\delta}^{\mathrm{L14}}$) to make the requirements on $\boldsymbol{T}^{\mathrm{L14}}, \boldsymbol{\delta}^{\mathrm{L14}}$ hold, showing L14[v]. This completes (a).

The proof of (b) follows the same steps as that of (a), except that we now do use the g subscript in L14. First the case $|u| > 0$. Take $\mathrm{g}^{\mathrm{L14}} = m$, $z^{\mathrm{L14}}_{i\mathrm{g}} = z_{im}$, $\psi^{\mathrm{L14}}_{\mathrm{g}} = (\theta, \pi_m)$, $n^{\mathrm{L14}}_{\mathrm{g}} = I_m$. Now $\hat{\mathbb{f}}^{\mathrm{L14}}_m(\theta, \pi_m) = \partial^u \Delta\hat{\mathcal{L}}^{\blacklozenge}_m(\theta, \pi_m)/r^u_m = \sum_{\mathcal{I}_m} \zeta_{im}(\theta, \pi_m)$ with $\zeta_{im}(\theta, \pi_m) = \sum_j (y_{ijm} - \varsigma^*_{ijm})\partial^u \Delta\ell_{ijm}(\theta, \pi_m)/r^u_m$. First, L14[i] is satisfied if we make $\boldsymbol{\delta}^{\mathrm{L14}}_{n\mathrm{g}}$ decrease at a sufficiently fast polynomial rate of I because $\hat{\mathbb{f}}^{\mathrm{L14}}_m$ is differentiable and by L15(c). We now establish condition L14[ii] using L14(b), for which we need to check L14[iii],[iv],[v]. L14[iii] holds by G[iii]. For [iv] and [v], take $\boldsymbol{\beta}^{\mathrm{L14}}_m = \log I$, such that [iv] is satisfied. Finally, [v]. By L15(c), $\max_m \mathbb{h}^{\mathrm{L14}}_m \preceq (\log I)^{-2}$.

Further, noting that for implicit $\{a_{ijm}\}$, $\mathbb{V}[\sum_{\mathcal{I}_m} \zeta_{im} \mid \mathcal{I}_m] = I_m \mathbb{V}\zeta_{im} = I_m \mathbb{V}[\sum_j a_{ijm}] \leq I_m \bar{J} \sum_j \mathbb{E}a^2_{ijm}$,

$$\max_m \bar{\boldsymbol{\sigma}}^{\mathrm{L14}2}_m \leq \bar{J} \max_m \Big\{ I_m \sup_{\Theta\times\mathbb{\Pi}^\kappa_m} \sum_j \mathbb{E}\Big[\varsigma^*_{ijm}\Big(\frac{\partial^u \Delta\ell_{ijm}(\theta, \pi_m)}{r^u_m}\Big)^2\Big]\Big\} \overset{\text{L15(c)}}{\preceq} (\log I)^{-4}.$$

To verify [v], note that $\max_m \exp(-c/\mathbb{h}^{\mathrm{L14}}_m)$ and $\max_m \exp(-c/\bar{\boldsymbol{\sigma}}^{\mathrm{L14}2}_m)$ decrease faster than any power of I.

Now, due to the compactness of Θ, we can choose $\boldsymbol{T}_{ng}^{\text{L14}}$ to increase at a (sufficiently fast) polynomial rate of I (that depends on our choice of $\boldsymbol{\delta}_{ng}^{\text{L14}}$) to make the requirements on $\boldsymbol{T}_{ng}^{\text{L14}}, \boldsymbol{\delta}_{ng}^{\text{L14}}$ hold, showing L14[v].

Finally, if $|u| = 0$ then we do not have to take derivatives of δ and by L13 we have an upper bound on ℓ_{ijm} that applies to all of $\mathbb{\Pi}_m$. The remainder of the proof is identical. This completes (b).

□

J.3 Lemmas referred to in app. J.2

Lemma 11 (Ω approximations)**.** Statements (17a) to (17c),

$$\max_{\mathring{t} \in [0,1]} \left\| \Gamma_\theta \{ \hat{\Omega}_{\theta\theta}[\theta(\mathring{t}), \pi(\mathring{t})] - \Omega_{\theta\theta}^{*} \} \Gamma_\theta \right\| \prec 1,$$
$$\max_{\mathring{t} \in [0,1]} \left\| \Gamma_\theta \{ \hat{\Omega}_{\theta\pi}[\theta(\mathring{t}), \pi(\mathring{t})] - \Omega_{\theta\pi}^{*} \} \Gamma_\pi \right\| \prec 1,$$
$$\max_{\mathring{t} \in [0,1]} \left\| \Gamma_\pi \{ \hat{\Omega}_{\pi\pi}^{\blacklozenge}[\theta(\mathring{t}), \pi(\mathring{t})] - \Omega_{\pi\pi}^{\blacklozenge *} \} \Gamma_\pi \right\| \prec 1,$$

hold.

Proof. Let $(\tilde{\theta}, \tilde{\pi}) = (\theta(\mathring{t}), \pi(\mathring{t}))$, which by the MVT lies between $(\hat{\theta}, \hat{\pi})$ and (θ^*, π^*). We will first show that $\Gamma_\theta[\hat{\mathcal{L}}_{\theta\theta}(\tilde{\theta}^z, \tilde{\pi}^z) - \mathcal{L}_{\theta\theta}^{*}]\Gamma_\theta \prec 1$, which is more challenging than the PLM component. We will assume $I \to \infty$, since if it is fixed the result is trivial. For a convenient scaling, let R be block diagonal with blocks $\mathbb{I}$ and $\mathbb{I}/\lambda$ then all elements of $R\Gamma_\theta^2 R$ converge at rate $1/I$ (if micro identification dominates) or faster (if identification comes from PLM).

Let $\hat{K}(\theta, \pi), K(\theta, \pi)$ be the (r, c) element of $\Gamma_\theta \hat{\mathcal{L}}_{\theta\theta}(\theta, \pi)\Gamma_\theta$, and $\Gamma_\theta \mathcal{L}_{\theta\theta}(\theta, \pi)\Gamma_\theta$, respectively (to avoid three-dimensional arrays of derivatives). Then, by adding and subtracting,

$$\hat{K}(\tilde{\theta}, \tilde{\pi}) - K^* = [\Delta\hat{K}(\tilde{\theta}, \tilde{\pi}) - \Delta\hat{K}^*] + \Delta\hat{K}^* + [K(\tilde{\theta}, \tilde{\pi}) - K^*] =: \textcircled{1} + \textcircled{2} + \textcircled{3}.$$

First,

$$|\textcircled{3}| \overset{\text{MVT}}{=} |(\tilde{\theta} - \theta^*)^{\triangledown} K_\theta(\mathring{\theta}, \mathring{\pi}) + (\tilde{\pi} - \pi^*)^{\triangledown} K_\pi(\mathring{\theta}, \mathring{\pi})|$$
$$\overset{\text{tri., Schwarz}}{\leq} \underbrace{\|\hat{\theta} - \theta^*\|}_{\text{Thm.1}} \cdot \| \underbrace{K_\theta(\mathring{\theta}, \mathring{\pi})}_{\text{L15(f)}} \| + \underbrace{\max_m \|\hat{\pi}_m - \pi_m^*\|}_{\text{Thm.1}} \sum_m \| \underbrace{K_{\pi m}(\mathring{\theta}, \mathring{\pi}_m)}_{\text{L15(f)}} \| \prec 1.$$

Further, since $\mathbb{E}\textcircled{2}^2$ is the variance of a sample mean,

$$\mathbb{E}\textcircled{2}^2 = \sum_m \mathbb{E}\textcircled{2}_m^2 \lesssim \frac{1}{I^2} \sum_m I_m \prec 1.$$

Finally, by MVT, triangle, and Schwarz inequalities,

$$|\textcircled{1}| \leq \underbrace{\|\hat{\theta} - \theta^*\|}_{\text{Thm.1}} \cdot \| \underbrace{\Delta\hat{K}_\theta(\mathring{\theta}, \mathring{\pi})}_{\text{L10(b)}} \| + \underbrace{\max_m \|\hat{\pi}_m - \pi_m^*\|}_{\text{Thm.1}} \cdot \sum_m \| \underbrace{\Delta\hat{K}_{\pi m}(\mathring{\theta}, \mathring{\pi}_m)}_{\text{L10(b)}} \| \prec 1.$$

Now the PLM component, $\Gamma_\theta[\hat{\Phi}_{\theta\theta}(\tilde{\theta}^z, \tilde{\pi}^z) - \Phi_{\theta\theta}^{*}]\Gamma_\theta$. Note that Φ is quadratic in $\delta(\theta, \pi)$ and $\mathbb{D}_{\theta m} = -\mathbb{D}_{\pi m}\partial_{\theta^{\triangledown}}\sigma_m$ and $\mathbb{D}_{\pi m}$ are bounded near the truth by L9(f), so $\Gamma_\theta[\hat{\Phi}_{\theta\theta}(\tilde{\theta}^z, \tilde{\pi}^z) - \Phi_{\theta\theta}^{*}]\Gamma_\theta \prec 1$.

Summing the two components completes the proof. The remaining results follow analogously by redefining K for the likelihood term and using the same argument for the PLM term. □

The next lemma contains some simple results, several of which are well-known, albeit typically presented less conveniently (for us).

Lemma 12 (Trivial technical results). Recall that $I_+ = \max(I, \mathrm{e})$. (a) If $\{a_i\}$ are subgaussian with common OVP $c_a < \infty$ then $\max_i |a_i| \prec \sqrt{c_a \log I_+}$; (b) for a fixed $C < \infty$, $\forall \theta \in \Theta_\epsilon^c$: $\|\theta^z\|^2 \le C\|\theta - \theta^*\|_\lambda^2$; (c) for a fixed $C < \infty$, $\forall \theta \in \Theta_\epsilon^c$: $\lambda^2 \le C\|\theta - \theta^*\|_\lambda^2$; (d) $\max_{m,j}(\sqrt{N_m}|s_{jm} - \pi_{jm}^*|/\sqrt{\pi_{jm}^*}) \prec \log M$.

Proof. First, (a). Suppose without loss of generality that $\forall i$: $\mathbb{E}a_i = 0$. Take $K > 1$ to obtain

$$\mathbb{P}(\max_i |a_i| > K\sqrt{2c_a \log I^+}) \overset{\text{Bonferroni}}{\le} \sum_i \mathbb{P}(|a_i| > K\sqrt{2c_a \log I^+}) \overset{\text{fn. 20}}{\le} 2I \exp(-K^2 \log I_+).$$

Let $I \to \infty$ followed by $K \to \infty$.

Now (b). Since $\theta \in \Theta_\epsilon^c$, $\|\theta - \theta^*\|^2 \ge \epsilon^2$. If $\|\theta^z - \theta^{*z}\|^2 \ge \epsilon^2/2$ then $\|\theta^z\|^2 \le \theta_\varnothing^{z2} \le (2\theta_\varnothing^{z2} \;/\; \epsilon^2)\|\theta^z - \theta^{*z}\|^2 \le C_1\|\theta - \theta^*\|_\lambda^2$, where $C_1 = 2\theta_\varnothing^{z2} \;/\; \epsilon^2$. Now suppose that $\|\theta^z - \theta^{*z}\|^2 < \epsilon^2/2$. Then by the triangle inequality, $\|\theta^\nu - \theta^{*\nu}\|^2 \ge \epsilon^2/2$. Take $C_2 = 2 + 4/\epsilon^2$. Then, $\|\theta^z\|^2 \le 2(\|\theta^z - \theta^{*z}\|^2 + \lambda^2) \le 2[\|\theta^z - \theta^{*z}\|^2 + \lambda^2\|\theta^\nu - \theta^{*\nu}\|^2/(\epsilon^2/2)] \le C_2\|\theta - \theta^*\|_\lambda^2$. Take $C = \max(C_1, C_2)$.

For (c), note that for $\lambda > 0$ and any $\theta \in \Theta_\epsilon^c$, $\|\theta - \theta^*\|_\lambda^2 = \|\theta^z - \theta^{*z}\|^2 + \lambda^2\|\theta^\nu - \theta^{*\nu}\|^2 \ge \|\theta - \theta^*\|^2 \min(\lambda^2, 1) \ge \epsilon^2 \min(\lambda^2, 1)$. Take $C = \epsilon^2$.

Finally, (d). We have,

$$\mathbb{P}\Big(\max_{m,j}(\sqrt{N_m}|s_{jm} - \pi_{jm}^*|/\sqrt{\pi_{jm}^*}) > \log M \Bigm| \mathbb{A}\Big) \overset{\text{Bonferroni}}{\le} \sum_{m,j} \mathbb{P}\Big(\sqrt{N_m}|s_{jm} - \pi_{jm}^*| > \sqrt{\pi_{jm}^*} \log M \Bigm| \mathbb{A}\Big)$$

$$\overset{\text{Bernstein}}{\le} 2\sum_{m,j} \exp\Big(-\frac{3N_m\pi_{jm}^* \log^2 M}{6N_m\pi_{jm}^*(1 - \pi_{jm}^*) + 2\sqrt{N_m\pi_{jm}^*} \log M}\Big) \le 2\sum_{m,j} \exp\Big(-\frac{\min(\log^2 M, \sqrt{N_m\pi_{jm}^*} \log M)}{6}\Big) \overset{\text{L9(b),C}}{\prec} 1. \qquad \square$$

Lemma 13 (Uniform upper bounds on contributions to the micro likelihood). $|\Delta\ell_{ijm}(\theta, \pi_m)| \le C\|\theta^z\| \cdot (\|z_{im}\| + C)$.

Proof. Recall that $\ell_{ijm}(\theta, \pi_m) = \log \varsigma_{ijm}(\theta, \pi_m)$ and $\ell_{jm}(\theta, \pi_m) = \log \varsigma_{jm}(\theta, \pi_m)$. Let $r_{jm}(\nu) = \sum_k \theta_k^\nu x_{jm(k)}^\nu \nu_k + \delta_{jm}$ where x_{jm}^ν represents the elements of x_{jm} associated with θ^ν (i.e., the elements of x_{jm} with random coefficients). Because of G[i], H[i], and the Schwarz inequality, $|\sum_k \theta_k^z x_{jm(k)} z_{im(k)}| \le C_1\|z_{im}\| \cdot \|\theta^z\|$. So, for all θ, π_m:

$$\begin{aligned}\varsigma_{ijm}(\theta, \pi_m) &= \int_\nu \frac{\exp\big[\sum_k \theta_k^z x_{jm(k)} z_{im(k)} + r_{jm}\big]}{\sum_t \exp\big[\sum_k \theta_k^z x_{tm(k)} z_{im(k)} + r_{tm}\big]} \le \int_\nu \frac{\exp\big[C_1\|z_{im}\| \cdot \|\theta^z\|\big] \exp\big[r_{jm}\big]}{\sum_t \exp\big[-C_1\|z_{im}\| \cdot \|\theta^z\|\big] \exp\big[r_{tm}\big]} \\ &\le \exp(2C_1\|z_{im}\| \cdot \|\theta^z\|) \int_\nu \frac{\exp(r_{jm})}{\sum_t \exp(r_{tm})};\end{aligned}$$

$$\pi_{jm} = \varsigma_{jm}(\theta, \pi_m) = \int_z \int_\nu \frac{\exp\big[\sum_k \theta_k^z x_{jm(k)} z + r_{jm}\big]}{\sum_t \exp\big[\sum_k \theta_k^z x_{tm(k)} z + r_{tm}\big]} \ge \Big(\int_z \overbrace{\exp(-2C_1\|\theta^z\| \cdot \|z\|)}^{=:\mathbb{f}(\theta^z, z)}\Big)\Big(\int_\nu \frac{\exp(r_{jm})}{\sum_t \exp(r_{tm})}\Big),$$

where the final line applies the inequality oppositely on numerator and denominator respectively. Then for $C_2 = 2C_1 \sup_\Theta \int_z \|z\|\mathbb{f}(\theta^z, z) \;/\; \int_z \mathbb{f}(\theta^z, z)$,[76] we have $\ell_{ijm}(\theta, \pi_m) - \ell_{jm}(\theta, \pi_m) \le \log \exp\big[\|\theta^z\|(2C_1\|z_{im}\| + C_2)\big] = \|\theta^z\|(2C_1\|z_{im}\| + C_2)$. This establishes an upper bound. A lower bound can be obtained analogously. $\square$

L14 below shows a general uniform convergence result for growing vectors of functions. We need L14 in

[76] This definition of C_2 is motivated the MVT expansion around $\theta^z = 0$, $|\log \int_z \mathbb{f}(\theta^z, z)| \le |\log 1| + 2C_1 \int_z \|z\|\mathbb{f}(\mathring{\theta}^z, z) \;/\; \int_z \mathbb{f}(\mathring{\theta}^z, z)$.

our proof, because as M grows, we need M different random functions of (θ, π_m) to converge uniformly in both the arguments and in m. For example, in L10(b) we need $\max_{m=1,\ldots,M} \sup_{\Theta_\epsilon^c \times \sqcap_m^\kappa} \|\Delta\hat{\mathcal{L}}^\blacklozenge_m(\theta, \pi_m)/r_m^u\|$ to converge.

Although there are many uniform convergence results in the literature, we have failed to find one that covers our case. Specifically, the fact that we have an increasing number M of functions and an increasing number of parameter vectors. Nevertheless, L14 uses a familiar method of proof.

Lemma 14 (Uniform convergence over growing vectors of functions)**.** For each of a possibly growing number of $\boldsymbol{G}$ groups indexed by $g = 1, \ldots, \boldsymbol{G}$, we define a parameter space Ψ_g, a true parameter vector ψ_g^*, and a sequence of independent random vectors $\{v_{ig}\}$, with $i \in \{1, \ldots, n_g\}$. We partition each space Ψ_g into $\boldsymbol{T}_{ng}$ sets $\Psi_{g1}, \ldots, \Psi_{g\boldsymbol{T}_{ng}}$ and let $\boldsymbol{\delta}_{ng} = \max_{t=1,\ldots,\boldsymbol{T}_{ng}} \sup_{\psi_g, \psi_g^\circ \in \Psi_{gt}} \|\psi_g - \psi_g^\circ\|$ denote the greatest distance possible between two points in the same Ψ_{gt}. Let $\bar{\psi}_{gt}$ be an arbitrary point in each Ψ_{gt}. Define functions $\hat{\mathbb{f}}_{ng}\colon \Psi_g \to \mathbb{R}^{d_{\psi_g}}$, where $\hat{\mathbb{f}}_{ng}(\psi_g) = \hat{\mathbb{f}}_g[\psi_g, \{v_{ig}\}]$.

(a) If [i] $\max_g \sup_{\psi_g, \psi_g^\circ \in \Psi_g : \|\psi_g - \psi_g^\circ\| \le \delta_{ng}} \|\hat{\mathbb{f}}_{ng}(\psi_g) - \hat{\mathbb{f}}_{ng}(\psi_g^\circ)\| \prec 1$; and [ii] $\max_{g,t} \|\hat{\mathbb{f}}_{ng}(\bar{\psi}_{gt})\| \prec 1$; then $\max_g \sup_{\Psi_g} \|\hat{\mathbb{f}}_{ng}(\psi_g)\| \prec 1$.

(b) Suppose that we can write $\hat{\mathbb{f}}_{ng}(\psi_g) = \sum_{i=1}^{n_g} \zeta_g(\psi_g, v_{ig})$. Let z_{ig} be a subvector of v_{ig} for which $\forall g, i, \psi_g \colon \mathbb{E}[\zeta_g(\psi_g, v_{ig}) \mid z_{ig}] = 0$. Define $\bar{\boldsymbol{\sigma}}_g^2 = \sup_{\Psi_g} \sum_i \|\mathbb{V}\zeta_g(\psi_g, v_{ig})\|$. Let for some $\boldsymbol{\beta}_g$, $\mathbb{h}_g := \mathbb{h}(\boldsymbol{\beta}_g) =$ esssup $\sup_{\Psi_g}[\|\zeta_g(\psi_g, v_{ig})\| \mathbb{1}(\|z_{ig}\| \le \boldsymbol{\beta}_g)]$. Then [ii] is satisfied if [iii] z_{ig} is subgaussian with OVP c_z^*; [iv] $\sum_g n_g \exp\big[-\boldsymbol{\beta}_g^2/(2c_z^*)\big] \prec 1$; [v] for any fixed $\boldsymbol{\varepsilon} > 0$, $\sum_g \boldsymbol{T}_{ng} \exp[-3\boldsymbol{\varepsilon}^2/(6\bar{\boldsymbol{\sigma}}_g^2 + 2\mathbb{h}_g \boldsymbol{\varepsilon})] \prec 1$.

Proof. Consider (a). We have by the triangle inequality,

$$\max_g \sup_{\Psi_g} \|\hat{\mathbb{f}}_{ng}(\psi_g)\| = \max_{g,t} \sup_{\Psi_{gt}} \|\hat{\mathbb{f}}_{ng}(\psi_g)\| \le \max_{g,t} \sup_{\psi_g \in \Psi_{gt}} \|\hat{\mathbb{f}}_{ng}(\psi_g) - \hat{\mathbb{f}}_{ng}(\bar{\psi}_{gt})\| + \max_{g,t} \|\hat{\mathbb{f}}_{ng}(\bar{\psi}_{gt})\| \overset{[i],[ii]}{\prec} 1.$$

Now (b). For $\boldsymbol{\varepsilon} > 0$, write

$$\begin{aligned}
\mathbb{P}[\max_{g,t} \|\hat{\mathbb{f}}_{ng}(\bar{\psi}_{gt})\| > \boldsymbol{\varepsilon}] &= \mathbb{P}\Big(\max_{g,t} \|\sum_i \zeta_g(\bar{\psi}_{gt}, v_{ig})\| > \boldsymbol{\varepsilon}\Big) \\
&\le \mathbb{P}\Big(\max_{g,t} \Big\|\sum_i \zeta_g(\bar{\psi}_{gt}, v_{ig}) \mathbb{1}(\|z_{ig}\| \le \boldsymbol{\beta}_g)\Big\| > \boldsymbol{\varepsilon}\Big) + \mathbb{P}(\max_{g,i} \|z_{ig}\| > \boldsymbol{\beta}_g) \\
&\overset{\text{Bonferroni},[iii]}{\le} \sum_{g,t} \mathbb{P}\Big(\|\sum_i \zeta_g(\bar{\psi}_{gt}, v_{ig}) \mathbb{1}(\|z_{ig}\| \le \boldsymbol{\beta}_g)\| > \boldsymbol{\varepsilon}\Big) + \sum_{g,i} \exp[-\boldsymbol{\beta}_g^2/(2c_z^*)] \\
&\overset{\text{Bernstein},[iv]}{\le} 2\sum_{g,t} \exp\Big(-\frac{3\boldsymbol{\varepsilon}^2}{6\bar{\boldsymbol{\sigma}}_g^2 + 2\mathbb{h}_g \boldsymbol{\varepsilon}}\Big) + o(1) \overset{[v]}{=} o(1) + o(1) \prec 1. \quad \square
\end{aligned}$$

The following lemma provides bounds on derivatives of (contributions to) the micro log likelihood, $\hat{\mathcal{L}}^\blacklozenge$, and its expectation when π is in the interior of its parameter space, i.e., $\pi \in \sqcap^\kappa$. It is used, for example, in step *4* of theorem 1 and in the proof of theorem 2, after consistency of $\hat{\pi}$ has been shown which implies $\hat{\pi} \in \sqcap^\kappa$ with probability approaching one. For notational convenience, recall $\ell_{ijm} = \log \varsigma_{ijm}$, $\ell_{jm} = \log \varsigma_{jm}$, and $\Delta\ell_{ijm} = \ell_{ijm} - \ell_{jm}$. Now, since $\varsigma_{jm}(\theta, \pi) = \pi$, we have $\hat{\mathcal{L}}^\blacklozenge(\theta, \pi) = -\sum_{ijm} D_{ijm} y_{ijm}[\Delta\ell_{ijm}(\theta, \pi_m) - \Delta\ell_{ijm}(\theta^*, \pi_m^*)]$.

Lemma 15. Let u be a vector of nonnegative integers indicating the number of partial derivatives with respect to elements of $\psi_m = (\theta, \pi_m)$, let u_z denote the total number of partial derivatives with respect to elements of θ^z, and $|u|$ the total number of partial derivatives. Recall that $\Delta\ell_{ijm} = \ell_{ijm} - \ell_{jm}$. (a) If $u_z = 0$ then $\forall \theta, \pi_m \in \Theta \times \sqcap_m^\kappa \colon \partial^u \Delta\ell_{ijm}(0, \theta^\nu, \pi_m) = 0$; (b) If $u_z = 1$ then $\forall \theta, \pi_m \in \Theta \times \sqcap_m^\kappa \colon \mathbb{E}[\partial^u \Delta\ell_{ijm}(0, \theta^\nu, \pi_m) \mid \mathbb{A}] = 0$; (c) For a constant $C < \infty$: $\mathbb{P}\big[\max_m \max_{\Theta\times\Pi^\kappa_m} |\partial^u \ell_{ijm}(\theta,\pi_m)| > C\|\mathbb{D}_{\pi m}(\theta,\pi_m)\|^{|u|}\|z_{im}\|^{\max(u_z,1)} \mid \mathbb{A}\big] = 0$ and $\mathbb{P}\big[\max_m \max_{\Theta\times\Pi^\kappa_m} |\partial^u \ell_{ijm}(\theta,\pi_m)| > C\kappa^{-3|u|}\|z_{im}\|^{\max(u_z,1)} \mid \mathbb{A}\big] = 0$; (d) $\sup_\theta \sum_m \sum_{\mathcal{J}_m} \mathbb{V}\{\sum_j (y_{ijm} - \varsigma^*_{ijm})[\Delta\ell_{ijm}(\theta,\pi^*_m) - \Delta\ell^*_{ijm}]\} \;/\; \|\theta - \theta^*\|^2_\lambda \le 1$; (e) For some fixed $C > 0$, $\max_m \sup_{\Theta^c_\epsilon\times\Pi^\kappa}\big(|\partial^u \mathcal{L}^\blacklozenge_m(\theta,\pi_m)|/\{\max(I_m,1)\boldsymbol{\lambda}^{|u|}_{\max}[\mathbb{D}_{\pi m}(\theta,\pi_m)]\}\big) \le C$ and $\max_m \sup_{\Theta^c_\epsilon\times\Pi^\kappa}\big(|\partial^u \mathcal{L}^\blacklozenge_m(\theta,\pi_m)|/\max(I_m,1) \le \kappa^{-3|u|}$; (f) Let $\tau_m = \mathbb{1}(\|\pi_m - \pi^*_m\| \le \kappa)$. If $u_z \le 1$, then for some fixed C, $\max_m \mathbb{E}|\partial^u \mathcal{L}^\blacklozenge_m(\theta,\pi_m)\tau_m|/\max(I_m,1) \le C(\|\theta^z\| + \lambda)^{2-u_z}$.

Proof. (a) and (b) are trivial since $\forall\theta^\nu,\pi_m$: $\Delta\ell_{ijm}(0,\theta^\nu,\pi_m) = 0$ and $\forall\theta^\nu,\pi_m$: $\mathbb{E}[\Delta\ell_{\theta^z ijm}(0,\theta^\nu,\pi_m) \mid \mathbb{A}_m] = 0$, and hence so are all their derivatives with respect to θ^ν,π_m.

Now (c). Write $\ell_{ijm}(\theta,\pi_m) = \tilde\ell_{ijm}[\theta,\delta_m(\theta,\pi_m)]$. The partial derivatives of $\tilde\ell_{ijm}$ with respect to $\theta^z_k,\theta^\nu_k,\delta_{km}$ are given by

$$\begin{aligned}
\tilde\ell_{\delta_k ijm} &= \mathbb{1}(j=k) - \frac{\int s_{ijm}s_{ikm}}{\sigma_{ijm}} \Rightarrow |\tilde\ell_{\delta_k ijm}| \le 1;\\
\tilde\ell_{\theta^z_k ijm} &= z_{im(k)}\Big(x_{jm(k)} - \sum_t x_{tm(k)}\frac{\int s_{ijm}s_{ikm}}{\sigma_{ijm}}\Big) \Rightarrow |\tilde\ell_{\theta^z_k ijm}| \overset{\text{G[i][iii]}}{\le} C|z_{im(k)}|;\\
\tilde\ell_{\theta^\nu_k ijm} &= \frac{1}{\sigma_{ijm}}\int s_{ijm}\nu_k\Big(x^\nu_{jm(k)} - \sum_t s_{itm}x^\nu_{tm(k)}\Big) \Rightarrow |\tilde\ell_{\theta^\nu_k ijm}| \overset{\text{G[i]}}{\le} C.
\end{aligned}\tag{68}$$

Now, by the chain rule, $\ell_{\pi ijm} = \mathbb{D}^\triangledown_{\pi m}\tilde\ell_{\delta ijm}$ and $\ell_{\theta ijm} = \tilde\ell_{\theta ijm} + \mathbb{D}^\triangledown_{\theta m}\tilde\ell_{\delta ijm} = \tilde\ell_{\theta ijm} - \partial_\theta\sigma^\triangledown_m\mathbb{D}^\triangledown_{\pi m}\tilde\ell_{\delta ijm}$.[77] Each partial derivative of ℓ_{ijm} with respect to any element of θ or π_m adds a factor of (a norm of) $\mathbb{D}_{\pi m}$. For $|u| > 0$, result (c) then follows from the bounds in (68). If $|u| = 0$ then (c) follows from L13.

For, (d), define $a_{ijm}(\theta) = \Delta\ell_{ijm}(\theta,\pi^*_m) - \Delta\ell^*_{ijm}$ and let C be a constant,

$$\begin{aligned}
\mathbb{V}\Big(\sum_j (y_{ijm} - \varsigma^*_{ijm})a_{ijm}(\theta)\Big) &\le \bar J\sum_j \mathbb{E}[\varsigma^*_{ijm}a^2_{ijm}(\theta)]\\
&\overset{\text{MVT}}{=} \bar J\sum_j \mathbb{E}\{\varsigma^*_{ijm}\big[(\theta^z - \theta^{*z})^\triangledown a_{\theta^z ijm}(\mathring\theta) + (\theta^\nu - \theta^{*\nu})^\triangledown a_{\theta^\nu ijm}(\mathring\theta)\big]^2\}\\
&\overset{\text{MVT,(a)}}{=} \bar J\sum_j \mathbb{E}\{\varsigma^*_{ijm}\big[(\theta^z - \theta^{*z})^\triangledown a_{\theta^z ijm}(\mathring\theta) + (\theta^\nu - \theta^{*\nu})^\triangledown a_{\theta^\nu\theta^z ijm}(\ddot\theta^z,\mathring\theta^\nu)\mathring\theta^z\big]^2\}\\
&\overset{\text{(c)}}{\le} C\big(\|\theta^z - \theta^{*z}\|^2 + \|\theta^\nu - \theta^{*\nu}\|^2\cdot\|\mathring\theta^z\|^2\big)\max_\theta\max_{1\le p\le 4}\mathbb{E}\|\mathbb{D}_{\pi m}(\theta,\pi^*_m)\|^p \le C^2\|\theta - \theta^*\|^2_\lambda,
\end{aligned}$$

where the last inequality follows from L9(d), and $\|\mathring\theta^z\| \le \|\theta^z - \theta^{*z}\| + \lambda$ by the triangle inequality.

The first half of (e) follows trivially from (c) and the second half from L9(d).

Finally (f). First, suppose $u_z = 1$. For some $0 \le t \le 1$,

$$\begin{aligned}
\partial^u\mathcal{L}^\blacklozenge_m(\theta,\pi_m)\tau_m &= I_m\sum_j \mathbb{E}\Big(\varsigma_{ijm}(\theta^*,\pi^*_m)\partial^u\Delta\ell_{ijm}(\theta,\pi_m)\tau_m \mid \mathbb{A}\Big)\\
&\overset{\text{MVT}}{=} I_m\sum_j \mathbb{E}\Big(\big[\varsigma_{jm}(0,\theta^{*\nu},\pi^*_m) + \theta^{*z\triangledown}\partial_{\theta^z}\varsigma^*_{ijm}(t\theta^{*z},\theta^{*\nu},\pi^*_m)\big]\times\\
&\qquad \big[\partial^u\Delta\ell_{ijm}(0,\theta^\nu,\pi_m) + \theta^{z\triangledown}\partial_{\theta^z}\partial^u\Delta\ell_{ijm}(t\theta^z,\theta^\nu,\pi_m)\big]\tau_m \;\Big|\; \mathbb{A}\Big).
\end{aligned}\tag{69}$$

Multiplying out the right-hand side of (69) yields four terms, one of which is zero by (b). The other three are by L9(f) bounded in absolute value by a constant times I_m times one of $\|\theta^z\|, \lambda, \lambda\|\theta^z\|$.

Second, if $u_z = 0$, take the mean value expansion around θ^z to the second order, which yields

[77] By the implicit function theorem on $\delta[\theta,\sigma(\theta,\bar\delta)] = \bar\delta$ which yields $\mathbb{D}_\theta + \mathbb{D}_\pi\partial_{\theta^\triangledown}\sigma = 0$.

nine terms all of which are bounded by a constant times I_m times one of $\|\theta^z\|^2$, λ^2, and higher powers thereof. □

K Additional Monte Carlo Results

In this appendix, we display results from the Monte Carlo experiments in table format. Each table is a different statistic and each row is a different experiment. The major columns denote parameters, (θ_1^z, θ_2^z, θ_1^ν, θ_2^ν, and β_1) and the sub-columns denote the three methods, CLEER, GMM-M, and MDLE.

The first set of five tables (for the five statistics) displays results for all of the experiments, except the integration bias experiments. The sixth table displays the results for the integration bias experiments, where we combine all of the statistics into a single table.

The five statistics we display are

1. median absolute error;
2. bias;
3. acceptance probability as a percentage;
4. median standard error;
5. the percentage of runs where the estimate was at the zero boundary.

The results generally confirm the plots and surrounding discussion in Section 7 in the main text. We observe that for small true values of θ^ν, some runs of the estimator converge to the zero boundary, for example see experiments `[6]`, `[9]`, and `[11]`. This happens most frequently for GMM-M for all three specifications (recall that we even start GMM-M from the truth), and the problem get smaller for CLEER and MDLE as θ^z grows, but the problem persists for GMM-M regardles of θ^z. We exclude these cases from the calculations in acceptance probability and standard error tables.

	θ_1^z			θ_2^z			θ_1^ν			θ_2^ν			β_1		
	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE
[1] baseline*	0.017	0.034	0.022	0.014	0.020	0.022	0.032	0.068	0.041	0.026	0.034	0.043	0.038	0.051	0.040
Vary S_m															
[2] $S_m = 250$	0.026	0.036	0.046	0.020	0.025	0.045	0.051	0.069	0.087	0.033	0.038	0.083	0.041	0.047	0.063
[3] $S_m = 4,000$	0.011	0.031	0.011	0.010	0.016	0.011	0.020	0.067	0.021	0.019	0.032	0.021	0.034	0.052	0.034
Vary M															
[4] $M = 10$	0.020	0.067	0.021	0.019	0.040	0.022	0.039	0.149	0.041	0.037	0.080	0.043	0.072	0.110	0.073
[5] $M = 1,000$	0.009	0.012	0.021	0.007	0.011	0.021	0.015	0.018	0.042	0.009	0.012	0.042	0.012	0.012	0.026
Vary (θ^z, θ^ν)															
[6] $(0.3, 0.3)$	0.009	0.013	0.009	0.007	0.010	0.010	0.070	0.138	0.079	0.051	0.078	0.091	0.034	0.041	0.036
[7] $(0.3, 1.0)$	0.011	0.013	0.020	0.008	0.010	0.019	0.061	0.068	0.129	0.033	0.033	0.124	0.045	0.050	0.072
[8] $(0.3, 2.0)$	0.014	0.015	0.070	0.010	0.012	0.068	0.102	0.086	0.594	0.049	0.043	0.578	0.060	0.056	0.296
[9] $(1.0, 0.3)$	0.012	0.030	0.013	0.011	0.020	0.013	0.032	0.127	0.034	0.034	0.075	0.041	0.031	0.045	0.031
[10] $(1.0, 2.0)$	0.030	0.036	0.071	0.019	0.021	0.072	0.081	0.090	0.186	0.048	0.045	0.181	0.055	0.061	0.098
[11] $(2.0, 0.3)$	0.018	0.054	0.019	0.016	0.036	0.018	0.030	0.183	0.030	0.026	0.110	0.029	0.033	0.046	0.033
[12] $(2.0, 1.0)$	0.025	0.069	0.028	0.022	0.036	0.026	0.026	0.081	0.029	0.023	0.045	0.027	0.036	0.055	0.038
[13] $(2.0, 2.0)$	0.043	0.067	0.064	0.033	0.038	0.063	0.056	0.087	0.085	0.043	0.051	0.080	0.040	0.056	0.049
Vary 1st Stage															
[14] $a = 0.15$	0.021	0.217	0.024	0.014	0.044	0.022	0.042	0.482	0.045	0.027	0.042	0.043	0.131	0.296	0.137
[15] $a = 1.00$	0.013	0.017	0.021	0.014	0.018	0.021	0.027	0.034	0.043	0.026	0.037	0.042	0.024	0.025	0.028

This table displays the Median Absolute Error for our Monte Carlo analysis across different experiments for the four main parameters of interest.

* Baseline is $M = 50$, $S_m = 1,000$, $\theta_1^z = 1.0$, and $a = 0.5$.

Table 4: *Monte Carlo Results: Median Absolute Error*

	θ^z_1			θ^z_2			θ^ν_1			θ^ν_2			β_1		
	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE
[1] baseline*	-0.001	+0.004	-0.003	-0.001	+0.003	-0.004	-0.002	+0.003	-0.005	-0.002	+0.004	-0.011	+0.001	+0.001	+0.001
Vary S_m															
[2] $S_m = 250$	-0.001	+0.002	-0.002	-0.002	+0.001	-0.005	-0.007	-0.004	-0.011	-0.002	-0.001	-0.007	-0.002	-0.003	-0.002
[3] $S_m = 4,000$	-0.004	+0.001	-0.005	-0.003	+0.002	-0.004	-0.008	-0.000	-0.009	-0.005	+0.002	-0.007	-0.002	-0.001	-0.001
Vary M															
[4] $M = 10$	-0.004	+0.005	-0.006	-0.003	+0.006	-0.002	-0.008	-0.010	-0.009	-0.005	+0.006	-0.005	+0.009	+0.004	+0.013
[5] $M = 1,000$	+0.000	+0.001	-0.006	+0.001	+0.001	-0.007	+0.001	+0.001	-0.008	+0.001	+0.001	-0.017	+0.001	+0.001	-0.005
Vary (θ^z, θ^ν)															
[6] $(0.3, 0.3)$	-0.001	+0.002	-0.002	-0.000	-0.000	+0.001	-0.036	-0.031	-0.017	-0.010	-0.034	-0.000	-0.005	-0.001	-0.002
[7] $(0.3, 1.0)$	-0.001	+0.000	-0.007	-0.000	+0.001	-0.007	-0.013	+0.001	-0.055	-0.005	+0.001	-0.056	-0.002	+0.002	-0.023
[8] $(0.3, 2.0)$	-0.010	+0.000	-0.070	-0.002	+0.001	-0.068	-0.092	+0.002	-0.594	-0.037	+0.004	-0.577	-0.040	+0.003	-0.295
[9] $(1.0, 0.3)$	-0.001	+0.007	-0.002	-0.001	+0.001	-0.001	-0.007	-0.026	-0.003	-0.007	-0.034	-0.005	-0.000	+0.004	+0.000
[10] $(1.0, 2.0)$	-0.025	+0.001	-0.070	-0.015	+0.000	-0.071	-0.069	-0.000	-0.185	-0.038	-0.002	-0.180	-0.033	-0.004	-0.094
[11] $(2.0, 0.3)$	-0.003	+0.016	-0.004	-0.002	+0.007	-0.001	-0.007	-0.034	-0.004	-0.004	-0.044	-0.001	+0.002	+0.013	+0.002
[12] $(2.0, 1.0)$	-0.004	+0.002	-0.006	-0.004	+0.003	-0.007	-0.004	-0.004	-0.006	-0.004	+0.001	-0.007	-0.001	-0.003	-0.001
[13] $(2.0, 2.0)$	-0.029	+0.002	-0.059	-0.023	-0.001	-0.057	-0.039	+0.001	-0.076	-0.029	-0.001	-0.075	-0.015	-0.001	-0.033
Vary 1st Stage															
[14] $a = 0.15$	-0.003	+0.027	-0.005	-0.003	+0.005	-0.006	-0.008	-0.055	-0.010	-0.005	+0.005	-0.007	+0.026	+0.042	+0.062
[15] $a = 1.00$	-0.002	+0.001	-0.006	-0.001	+0.001	-0.004	-0.005	+0.001	-0.012	-0.003	-0.000	-0.008	+0.002	+0.003	-0.001

This table displays the Bias for our Monte Carlo analysis across different experiments for the four main parameters of interest.
[*] Baseline is $M = 50$, $S_m = 1,000$, $\theta^z_1 = 1.0$, and $a = 0.5$.

Table 5: *Monte Carlo Results: Bias*

	θ_1^z			θ_2^z			θ_1^ν			θ_2^ν			β_1		
	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE
[1] baseline*	96.7	94.2	95.0	94.7	94.8	95.0	96.0	95.3	95.7	95.0	95.0	94.7	95.8	94.0	95.2
Vary S_m															
[2] $S_m = 250$	95.0	95.7	93.9	94.1	94.3	92.8	95.1	96.7	94.5	94.5	93.6	94.6	94.0	94.6	93.8
[3] $S_m = 4,000$	93.3	95.3	92.8	94.7	94.8	94.3	93.6	95.7	93.5	95.1	96.1	93.9	95.5	95.4	95.2
Vary M															
[4] $M = 10$	94.6	94.4	95.2	94.3	94.1	93.9	95.7	97.4	95.7	93.5	93.2	94.3	95.6	94.3	96.0
[5] $M = 1,000$	95.9	95.7	94.9	94.9	95.2	94.4	94.9	94.5	95.2	95.5	94.8	94.1	94.9	94.6	95.2
Vary (θ^z, θ^ν)															
[6] $(0.3, 0.3)$	91.1	96.9	91.4	94.5	97.2	91.8	90.5	95.0	89.8	94.4	95.0	91.0	94.9	96.4	94.2
[7] $(0.3, 1.0)$	93.1	93.8	90.4	95.2	94.8	90.5	94.0	94.6	90.8	96.1	95.3	92.3	94.9	94.7	91.9
[8] $(0.3, 2.0)$	89.4	94.1	40.2	95.3	95.1	42.8	86.4	94.6	41.3	90.7	94.7	41.6	92.1	94.1	45.4
[9] $(1.0, 0.3)$	94.7	96.0	94.4	94.7	94.9	93.8	96.6	95.2	97.0	96.2	97.3	96.5	95.3	95.7	95.6
[10] $(1.0, 2.0)$	85.9	93.1	69.9	90.9	94.4	71.4	86.1	93.8	70.3	89.8	94.1	72.3	92.1	94.2	78.7
[11] $(2.0, 0.3)$	93.7	96.1	94.6	93.7	95.7	93.9	96.8	95.0	97.0	96.9	96.8	97.3	94.6	96.5	95.2
[12] $(2.0, 1.0)$	93.9	95.1	94.9	95.8	95.5	95.0	93.5	95.0	94.7	95.2	94.3	94.7	94.3	93.7	94.7
[13] $(2.0, 2.0)$	92.1	94.6	86.0	91.9	93.5	86.3	91.7	94.2	87.5	90.7	94.3	86.6	95.0	94.6	93.1
Vary 1st Stage															
[14] $a = 0.15$	94.1	88.4	93.2	95.5	95.1	92.8	94.3	99.8	94.4	94.3	96.3	92.7	92.1	96.2	91.8
[15] $a = 1.00$	95.7	95.3	94.4	94.6	93.7	93.4	94.5	95.0	93.8	95.0	94.4	93.7	96.3	95.7	95.8

This table displays the Acceptance Probability (%) for our Monte Carlo analysis across different experiments for the four main parameters of interest.

* Baseline is $M = 50$, $S_m = 1,000$, $\theta_1^z = 1.0$, and $a = 0.5$.

Table 6: *Monte Carlo Results: Acceptance Probability (%)*

	θ_1^z			θ_2^z			θ_1^ν			θ_2^ν			β_1		
	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE
[1] baseline*	0.026	0.048	0.032	0.021	0.028	0.031	0.051	0.101	0.061	0.040	0.052	0.062	0.056	0.076	0.060
Vary S_m															
[2] $S_m = 250$	0.040	0.053	0.063	0.028	0.037	0.062	0.078	0.103	0.123	0.049	0.057	0.124	0.066	0.075	0.089
[3] $S_m = 4,000$	0.015	0.047	0.016	0.013	0.026	0.016	0.029	0.101	0.031	0.026	0.051	0.031	0.050	0.075	0.050
Vary M															
[4] $M = 10$	0.030	0.098	0.032	0.028	0.055	0.032	0.060	0.218	0.063	0.055	0.110	0.064	0.108	0.161	0.109
[5] $M = 1,000$	0.013	0.017	0.031	0.010	0.015	0.031	0.022	0.026	0.061	0.014	0.018	0.061	0.016	0.017	0.039
Vary (θ^z, θ^ν)															
[6] $(0.3, 0.3)$	0.013	0.024	0.014	0.011	0.015	0.015	0.096	0.180	0.112	0.075	0.095	0.129	0.051	0.061	0.053
[7] $(0.3, 1.0)$	0.015	0.019	0.026	0.012	0.014	0.026	0.084	0.096	0.164	0.047	0.048	0.167	0.066	0.070	0.102
[8] $(0.3, 2.0)$	0.018	0.021	0.030	0.015	0.017	0.030	0.114	0.125	0.262	0.063	0.064	0.254	0.075	0.079	0.140
[9] $(1.0, 0.3)$	0.018	0.050	0.019	0.016	0.029	0.019	0.051	0.177	0.054	0.048	0.105	0.056	0.050	0.070	0.050
[10] $(1.0, 2.0)$	0.036	0.050	0.051	0.026	0.031	0.051	0.091	0.128	0.132	0.060	0.066	0.131	0.066	0.081	0.083
[11] $(2.0, 0.3)$	0.026	0.096	0.027	0.024	0.055	0.027	0.045	0.234	0.046	0.042	0.168	0.045	0.049	0.080	0.049
[12] $(2.0, 1.0)$	0.037	0.096	0.041	0.033	0.054	0.041	0.039	0.122	0.042	0.034	0.067	0.042	0.052	0.081	0.054
[13] $(2.0, 2.0)$	0.053	0.096	0.068	0.043	0.056	0.068	0.071	0.135	0.090	0.055	0.071	0.089	0.059	0.082	0.066
Vary 1st Stage															
[14] $a = 0.15$	0.030	0.311	0.032	0.021	0.073	0.031	0.059	0.787	0.061	0.040	0.062	0.062	0.183	0.486	0.186
[15] $a = 1.00$	0.020	0.027	0.031	0.020	0.027	0.031	0.040	0.052	0.062	0.039	0.052	0.062	0.037	0.039	0.042

This table displays the Median Standard Error for our Monte Carlo analysis across different experiments for the four main parameters of interest.

* Baseline is $M = 50$, $S_m = 1,000$, $\theta_1^z = 1.0$, and $a = 0.5$.

Table 7: *Monte Carlo Results: Median Standard Error*

	θ_1^z			θ_2^z			θ_1^ν			θ_2^ν			β_1		
	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE	CLEER	GMM-M	MDLE
[1] baseline*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Vary S_m															
[2] $S_m = 250$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[3] $S_m = 4,000$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Vary M															
[4] $M = 10$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[5] $M = 1,000$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Vary (θ^z, θ^ν)															
[6] $(0.3, 0.3)$	0.0	0.0	0.0	0.0	0.0	0.0	8.7	15.4	10.0	2.6	11.4	11.2	0.0	0.0	0.0
[7] $(0.3, 1.0)$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
[8] $(0.3, 2.0)$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0
[9] $(1.0, 0.3)$	0.0	0.0	0.0	0.0	0.0	0.0	0.3	11.1	0.5	0.3	10.0	0.5	0.0	0.0	0.0
[10] $(1.0, 2.0)$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[11] $(2.0, 0.3)$	0.0	0.0	0.0	0.0	0.0	0.0	0.1	18.4	0.1	0.0	12.2	0.0	0.0	0.0	0.0
[12] $(2.0, 1.0)$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[13] $(2.0, 2.0)$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Vary 1st Stage															
[14] $a = 0.15$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
[15] $a = 1.00$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

This table displays the Boundary Proportion (%) for our Monte Carlo analysis across different experiments for the four main parameters of interest.

* Baseline is $M = 50$, $S_m = 1,000$, $\theta_1^z = 1.0$, and $a = 0.5$.

Table 8: *Monte Carlo Results: Boundary Proportion (%)*

	θ_1^z			θ_2^z			θ_1^ν			θ_2^ν			β_1		
	CLEER	GMM-M	CLEER (19)	CLEER	GMM-M	CLEER (19)	CLEER	GMM-M	CLEER (19)	CLEER	GMM-M	CLEER (19)	CLEER	GMM-M	CLEER (19)
Panel A: Median Absolute Error															
[16] $a = 1.0, \theta\nu = 2$	0.019	0.019	0.017	0.020	0.020	0.018	0.047	0.044	0.042	0.045	0.047	0.040	0.029	0.027	0.027
[17] $a = 1.0, \theta\nu = 2.5$	0.030	0.020	0.018	0.030	0.022	0.019	0.094	0.053	0.051	0.093	0.053	0.050	0.032	0.028	0.028
[18] $a = 1.0, \theta\nu = 3$	0.061	0.021	0.021	0.061	0.022	0.023	0.215	0.057	0.067	0.211	0.059	0.067	0.058	0.028	0.028
[19] $a = 0.5, \theta\nu = 2$	0.032	0.035	0.027	0.021	0.022	0.018	0.081	0.087	0.071	0.049	0.047	0.042	0.049	0.054	0.047
[20] $a = 0.5, \theta\nu = 2.5$	0.066	0.034	0.029	0.038	0.022	0.020	0.206	0.104	0.089	0.114	0.053	0.054	0.091	0.058	0.051
[21] $a = 0.5, \theta\nu = 3$	0.133	0.036	0.039	0.078	0.024	0.026	0.465	0.117	0.136	0.257	0.062	0.074	0.177	0.059	0.056
Panel B: Bias															
[16] $a = 1.0, \theta\nu = 2$	-0.012	+0.000	-0.006	-0.011	+0.001	-0.005	-0.031	-0.000	-0.013	-0.029	+0.002	-0.012	-0.007	+0.000	+0.002
[17] $a = 1.0, \theta\nu = 2.5$	-0.030	+0.000	-0.006	-0.029	+0.002	-0.007	-0.093	+0.001	-0.019	-0.091	+0.005	-0.020	-0.025	+0.001	-0.001
[18] $a = 1.0, \theta\nu = 3$	-0.061	+0.002	-0.014	-0.061	+0.002	-0.015	-0.216	+0.003	-0.049	-0.214	+0.004	-0.051	-0.056	+0.002	-0.008
[19] $a = 0.5, \theta\nu = 2$	-0.026	-0.001	-0.009	-0.015	+0.000	-0.005	-0.069	-0.004	-0.024	-0.037	+0.002	-0.012	-0.032	-0.004	-0.007
[20] $a = 0.5, \theta\nu = 2.5$	-0.067	-0.001	-0.017	-0.038	+0.003	-0.009	-0.207	-0.004	-0.054	-0.112	+0.004	-0.030	-0.090	-0.003	-0.018
[21] $a = 0.5, \theta\nu = 3$	-0.134	+0.001	-0.035	-0.079	+0.001	-0.021	-0.469	+0.002	-0.127	-0.259	+0.003	-0.063	-0.182	+0.000	-0.042
Panel C: Acceptance Probability (%)															
[16] $a = 1.0, \theta\nu = 2$	90.6	93.5	93.2	93.1	93.7	94.7	90.9	94.3	94.3	94.0	95.5	95.2	93.2	93.7	93.8
[17] $a = 1.0, \theta\nu = 2.5$	78.6	95.2	94.2	79.3	93.7	93.8	74.2	93.5	93.6	73.2	94.0	93.8	90.9	93.6	94.5
[18] $a = 1.0, \theta\nu = 3$	38.1	93.3	92.1	37.7	92.9	90.0	20.5	94.3	90.2	20.8	93.7	90.7	70.0	94.7	94.6
[19] $a = 0.5, \theta\nu = 2$	87.8	93.6	93.8	91.3	94.6	94.7	87.4	93.5	93.1	91.9	94.6	95.1	92.1	93.1	94.0
[20] $a = 0.5, \theta\nu = 2.5$	56.7	94.5	92.5	70.9	94.7	93.0	51.1	94.5	92.6	62.6	94.0	94.1	74.7	93.8	93.5
[21] $a = 0.5, \theta\nu = 3$	5.3	94.8	85.3	19.5	92.3	87.5	3.6	94.4	84.4	11.0	93.9	88.0	22.2	94.6	91.8
Panel D: Median Standard Error															
[16] $a = 1.0, \theta\nu = 2$	0.025	0.029	0.025	0.025	0.029	0.025	0.059	0.065	0.060	0.059	0.065	0.060	0.039	0.040	0.040
[17] $a = 1.0, \theta\nu = 2.5$	0.026	0.030	0.027	0.026	0.030	0.027	0.069	0.076	0.072	0.069	0.076	0.072	0.039	0.040	0.040
[18] $a = 1.0, \theta\nu = 3$	0.027	0.032	0.028	0.027	0.032	0.028	0.077	0.087	0.083	0.077	0.087	0.083	0.039	0.041	0.040
[19] $a = 0.5, \theta\nu = 2$	0.036	0.050	0.037	0.026	0.031	0.026	0.092	0.129	0.096	0.060	0.066	0.062	0.066	0.081	0.068
[20] $a = 0.5, \theta\nu = 2.5$	0.037	0.051	0.041	0.027	0.032	0.028	0.107	0.150	0.120	0.070	0.078	0.074	0.067	0.083	0.072
[21] $a = 0.5, \theta\nu = 3$	0.036	0.052	0.042	0.027	0.034	0.029	0.119	0.170	0.140	0.079	0.090	0.085	0.066	0.082	0.073

Notes: This table presents Monte Carlo results for integration bias experiments.

Table 9: *Monte Carlo Results: Integration Bias*

L Glossaries of Common Results and Notation

This appendix includes a listing of common results used throughout the paper and referenced by name and a glossary of some of the notation used.

L.1 Common results referenced by name

annihilator matrix For given matrix A, $\mathcal{M}_A = \mathbb{I} - \mathcal{P}_A$ with $\mathcal{P}_A$ a projection matrix

Bernstein inequality If $\{x_i\}$ are independent with variances σ_i^2 and common upper bound $\bar{x}$ then $\mathbb{P}(|\sum_i x_i| > C) \leq 2\exp[-3C^2 \,/\, (6\sum_i \sigma_i^2 + 2C\bar{x})]$

Bonferroni inequality $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$

concentration parameter In a single regressor linear model $y = x\beta + u$, where $x = Z\pi + v$ for instruments Z, the number $\|Z^\triangledown \pi\|^2/\sigma_v^2$; generalizations thereof in more complicated models.

Cramér's theorem If $x_n \overset{p}{\to} x$ and $y_n \overset{d}{\to} y$ then $x_n y_n \overset{d}{\to} xy$

esssup essential supremum (in this context the top of the support of the random variable)

Hoeffding inequality If $\{x_i\}$ are independent with upper and lower bounds u_i, ℓ_i then $\mathbb{P}(\|\sum_i x_i\| > C) \leq 2\exp(-2C^2 \,/\, \sum_i (u_i - \ell_i)^2)$

Hölder inequality $\mathbb{E}(\|x\| \; \|y\|) \leq (\mathbb{E}\|x\|^p)^{1/p}(\mathbb{E}\|y\|^{p/(p-1)})^{1-1/p}$ for any $1 < p < \infty$ for which the expectations exist (special case of Jensen inequality)

information matrix equality For likelihood estimators, the expectation of the outer product of the gradients equals minus that of the Hessian

Jensen inequality If g is convex then $g(\mathbb{E}x) \leq \mathbb{E}g(x)$ provided that both expectations exist

Lindeberg condition For a triangular independent array $\{x_{in}\}$, with $\sum_i \mathbb{V}x_{in} = 1$, $\forall \epsilon > 0$: $\sum_i \mathbb{E}[x_{in}^2 \mathbb{1}(|x_{in}^2 \geq \epsilon)] \prec 1$

Markov inequality $\mathbb{P}(\|x\| \geq t) \leq t^{-r}\mathbb{E}\|x\|^r$ for any $t > 0$ and $r > 0$ for which the moment exists

Moore Penrose inverse For an arbitrary matrix A, the unique matrix A^+ for which $AA^+ = (AA^+)^\triangledown$, $A^+A = (A^+A)^\triangledown$, $AA^+A = A$, $A^+AA^+ = A^+$, i.e. $VD^{-1}U^\triangledown$ when the singular value decomposition is used with D a matrix with only the nonzero singular values

mean value theorem $f(t) = f(0) + f'(\lambda t)t$ for some $0 \leq \lambda \leq 1$ (or a higher order analog thereof)

norm of a matrix We use $\|A\| = \max_{\|x\|=1} \|Ax\|$, i.e. the square root of the largest eigenvalue of $A^\triangledown A$.

partitioned inverse Assuming the existence of the inverses,

$$\begin{bmatrix} A & B^\triangledown \\ B & C \end{bmatrix}^{-1} = \begin{bmatrix} (A - B^\triangledown C^{-1}B)^{-1} & -(A - B^\triangledown CB)^{-1}B^\triangledown C^{-1} \\ \cdot & (C - BA^{-1}B^\triangledown)^{-1} \end{bmatrix}.$$

projection matrix For given matrix A, the matrix $\mathcal{P}_A = A(A^\triangledown A)^{-1}A^\triangledown$ (or more generally AA^+)

Schwarz inequality Hölder inequality for $p = 2$

Slutsky $x_n \overset{p}{\to} x \Rightarrow g(x_n) \overset{p}{\to} g(x)$ if g is continuous

sigma algebra information set

slowly varying function a function for which $\lim_{x\to\infty} f(tx) \,/\, f(x) = 1$ for all $t > 0$; logarithms are an example

singular value decomposition Any real matrix A can be written as $UDV^\triangledown$, where U and V have orthonormal columns ($U^\triangledown U = \mathbb{I}$ and $V^\triangledown V = \mathbb{I}$) and D is a diagonal matrix.

triangle inequality $\|x + y\| \leq \|x\| + \|y\|$

weak law of large numbers (WLLN) any of a number of results showing convergence of a sample mean to its expectation

Woodbury matrix identity $(A + BC^{-1}B^{\triangledown})^{-1} = A^{-1} - A^{-1}B(C + B^{\triangledown}A^{-1}B)^{-1}B^{\triangledown}A^{-1}$

L.2 Notation (incomplete list)

A

$\mathcal{A}$ $\operatorname{plim}_{M\to\infty}(B^{\triangledown}B \,/\, M)$. 15

$\mathbb{A}$ the sigma algebra generated by product characteristics and the D_{im}'s. 9

B

B matrix of instruments. 7

b_{jm} vector of instruments. 6

$B_m^{\mathbf{opt}}$ optimal instruments for market m. 20

$\hat{\beta}$ CLEER estimator of β^*. 6

$\mathcal{B}$ parameter space of β^*. 11

β^* (true value of) product level regression coefficients. 5

C

c^*_ξ optimal variance proxy (OVP) for ξ_{jm}, see G. 10

c^*_z OVP for z_{im}, see G. 11

$\hat{\chi}$ product level moments part of the objective function defined in terms of δ, β. 6

$\to$ converges (or diverges) to. 6

$\xrightarrow{p}$ converges in probability to. 6

D

d_b number of instruments. 6

d_β dimension of β^*. 6

D_{im} dummy to indicate whether consumer i is included in the micro sample. 6

d_ν number of random coefficients. 5

d_θ dimension of θ. 8

d_x number of observed product characteristics. 4

d_z number of demographic characteristics. 5

∂ partial derivative(s) with respect to its subscript(s). 5

$\hat{\delta}$ CLEER estimator of δ^*. 6

δ^*_{jm} (true) 'mean' utility. 5

$\Delta\hat{\mathcal{L}}$ $\hat{\mathcal{L}} - \mathcal{L}$ (analogously when endowed with a ◆ superscript). 13

$\Delta\ell_{ijm}$ $\log\varsigma_{ijm} - \log\varsigma_{jm}$. 19

δ_m Berry inversion (when used as a function). 8

$\Delta\hat{\Omega}$ $\hat{\Omega} - \Omega$ (analogously when endowed with a ◆ superscript). 13

$\Delta\hat{\Phi}$ $\hat{\Phi} - \Phi$. 13

$\mathbb{D}_\pi$ derivative of Berry (1994) inversion with respect to π, $\mathbb{D}_\theta = \partial_{\pi^\triangledown}\delta$. 10

$\mathbb{D}_\theta$ derivative of Berry (1994) inversion with respect to θ, $\mathbb{D}_\theta = \partial_{\theta^\triangledown}\delta$. 10

d differential used in integration. 5

d_b number of (product level) instruments. 11

$\mathcal{D}^*_{jkm}$ diversion ratio from good j to good j with respect to unobserved quality, defined at the truth. 31

$\mathfrak{D}$ mean absolute error based diversion statistic. 32

E

$\mathbb{E}$ expectation. 6

ℓ_{ijm} $\log \varsigma_{ijm}$. 19, 40

ϵ used as a distance in the consistency proof. 9

ε_{im} idiosyncratic product specific taste shocks. 5

η convenience rate $\eta = \varkappa^3$. 13

F

F_m distribution of unobservable demographics. 5

G

G_m distribution of observable demographics. 5

$\mathit{\Gamma}_\pi = \left\{\mathbb{E}[\mathcal{L}_{\pi\pi} - \mathcal{L}_{\pi\theta}\mathcal{L}^+_{\theta\theta}\mathcal{L}_{\theta\pi}]\right\}^{-1/2}$. 15

$\mathit{\Gamma}_\theta = \left[\mathbb{E}(\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\pi}\mathcal{L}^{-1}_{\pi\pi}\mathcal{L}_{\pi\theta}) + M\mathit{\Xi}\mathit{\Lambda}\mathit{\Xi}^{\triangledown}\right]^{-1/2}$; basically a population analog to $\hat{\mathit{\Gamma}}_\theta$. 15

$\hat{\mathit{\Gamma}}_\theta$ square root of the $\theta\theta$ block of the inverse Hessian of $\mathit{\Omega}$, $\hat{\mathit{\Gamma}}_\theta = \hat{\mathcal{Q}}^{-1/2}_{\theta\theta}$. 14

$\succeq$ left hand side is element-wise of greater or equal order than the right-hand side. 10

$\succ$ indicates that the right-hand side is element-wise negligible to the left-hand side. 10

H

H Hessian matrix of subscript function evaluated at the truth, e.g., $H_{\mathit{\Omega}}$. 16

I

I total number of consumers in the micro sample (across all markets). 8

i consumer index. 5

I_m number of consumers in the micro sample in market m. 6

∞ infinity. 6

J

J total number of products across all markets. 7

j product index. 4

J_m number of products in market m. 4

K

$\varkappa$ rate used in C, $\varkappa = \exp(-4\varkappa^{\uparrow}_{\delta})$. 10

$\varkappa^{\uparrow}_{\delta}$ rate used in C, $\varkappa^{\uparrow}_{\delta} = 2\sqrt{2c^*_\xi \log M}$. 10

κ_π we show that $\min_{m,j} \pi^*_{jm} \geq \kappa_\pi$. 35

L

$\hat{L}$ mixed data likelihood defined in terms of θ, δ. 6

$\hat{\mathcal{L}}^{\blacksquare}$ minus macro loglikelihood defined in terms of π. 8

$\tilde{\ell}_{ijm}$ $\log \sigma_{ijm}$. 19

ℓ_{jm} $\log \varsigma_{jm}$. 19

$\hat{L}^{\blacksquare}$ macro likelihood as a function of θ, δ. 7

$\hat{L}^{\blacklozenge}$ micro likelihood as a function of θ, δ. 7

$\mathcal{L}^{\blacklozenge}$ (minus) micro loglikelihood. 9

$\hat{\mathcal{L}}$ (minus) sample loglikelihood defined in terms of θ, π. 8

$\hat{\mathcal{L}}^{\blacklozenge}$ minus the micro loglikelihood. 8

λ index of micro identification strength, $\|\theta^{*z}\|$. 10

$\|\theta - \theta^*\|^2_\lambda$ norm used in definition of $\rho^{\blacklozenge}$, $\|\theta - \theta^*\|^2_\lambda = \|\theta^z - \theta^{*z}\|^2 + \lambda^2 \|\theta^\nu - \theta^{*\nu}\|^2$. 10

$\prec$ indicates that the left-hand side is (element by element) of smaller order than (negligible compared to) the right-hand side. 12

M

M number of markets. 4

m market index. 4

$\hat{m}$ sample product level moment. 7

$M\phi^2_\beta$ smallest eigenvalue of the concentration parameter for β^*. 18

$M\phi^2_\nu$ smallest eigenvalue of the concentration parameter for $(\theta^{\nu *}, \beta^*)$. 18

$M\phi^2_\theta$ smallest eigenvalue of the concentration parameter for (θ^*, β^*). 18

$\mu^{\nu im}_{jm}$ deviation due to taste shock. 5

μ^{zim}_{jm} deviation from mean utility due to observed demographic variables. 5

N

N_m population size. 5

ν_{im} unobserved demographics. 5

O

Ω population objective function. 9

$\hat{\Omega}$ sample objective function. 8

P

$\mathbb{P}$ probability. 5

$\mathcal{P}_B$ orthogonal projection matrix. 9

$\mathcal{P}$ projection matrix that arises after β has been profiled out. 9

Φ population product level moments objective function. 9

$\hat{\Phi}$ product level moments objective function defined in terms of θ, π. 8

$\hat{\pi}$ CLEER estimator of π^*. 9

$\mathbb{\Pi}^\kappa$ a subset of $\mathbb{\Pi}$, $\prod_m \mathbb{\Pi}_m^\kappa$ where $\mathbb{\Pi}_m^\kappa = \{\pi_m \colon \min_j \pi_{jm} \geq \kappa\}$. 12

$\mathbb{\Pi}_m^{\kappa_\pi}$ a subset of $\mathbb{\Pi}^\kappa$ replacing κ with $\kappa_\pi = \kappa^{3/4}$. 14

π_m vector of π_{jm}'s (excluding π_{0m}). 8

π_m^* true product level choice probabilities. 6

$\mathbb{\Pi}$ parameter space of π^*. 11

$\mathscr{p}_{jm}$ endogenous product characteristics. 5

$\cdot^+$ Moore Penrose inverse. 9

Q

$\hat{\mathcal{Q}}_{\theta\theta}$ the inverse of the θ, θ block of the inverse Hessian of $\hat{\Omega}$, $\mathcal{Q}_{\theta\theta} = \hat{\Omega}_{\theta\theta} - \hat{\Omega}_{\theta\pi}\hat{\Omega}_{\pi\pi}^{-1}\hat{\Omega}_{\pi\theta}$. 14

$\hat{\mathscr{q}}_\theta$ component of numerator term in quadratic expansion, $\hat{\mathscr{q}}_\theta = \hat{\Omega}_\theta - \Omega_{\theta\pi}\Omega_{\pi\pi}^{-1}(\hat{\Omega}_\pi^{\blacklozenge} - \mathcal{L}_{\pi\pi}^{\blacksquare}(s - \pi^*)$, see L8. 15

R

ρ_D rate governing $\Delta\hat{\Omega}^{\blacklozenge}(\theta, \pi) - \hat{\Omega}^{\blacklozenge}(\theta^*, s)(\theta, \pi)$, $\rho_D(\theta, \pi) = \eta \max\{\eta\rho_{\mathrm{id}}(\theta), \rho^{\blacksquare}(\pi)\}$. 13

$\rho_{\mathbf{id}}$ identification strength (as a function of θ) $\rho^\Phi(\theta) = \|\mathcal{P}\mathbb{D}_\theta(\theta^*, \pi^*)(\theta - \theta^*)\|^2$, see C. 9

$\rho^{\blacksquare}$ rate (function of π) governing convergence of market shares to choice probabilities π, $\rho^{\blacksquare}(\pi) = \sum_m \rho_m^{\blacksquare}(\pi_m) = \sum_m N_m \|s_m - \pi_m\|^2$. 13

$\rho^{\blacklozenge}$ micro identification strength, $\rho^{\blacklozenge}(\theta) = I\|\theta - \theta^*\|_\lambda^2$, see A. 9

ρ_N rate governing total population increase, $\rho_N = \sum_m N_m^{-1}$. 12

ρ^Φ product level moment identification strength, see B. 9

ρ_u rate governing smallest market population increase, $\rho_u = 1 \,/\, \min_m \sqrt{N_m}$. 12

S

$\mathscr{s}_{ijm}(\nu; \theta, \delta)$ choice probability before integrating out random coefficients. 5

s_{jm} observed market share. 6

$\mathscr{s}_{jm}$ choice probability before integrating out random coefficients. 5

ς_{ijm} micro choice probability function defined in terms of θ, π_m. 8

σ_{jm} unconditional choice probability function. 5

$\sigma_{jm}^{z_{im}}$ micro choice probability function in terms of θ, δ. 5

$*$ when used as a superscript to a parameter it indicates the true value of that parameter; if used as a superscript to a function it indicates that the function is evaluated at the true values. 5, 32

T

Θ parameter space. 8

Θ_ϵ ϵ neighborhood of θ^*. 9

$\hat{\theta}$ CLEER estimator of θ^*. 6

$\boldsymbol{\Theta}^{*\nu}$ (true) matrix of utility coefficients on $\nu \times x$. 5

$\theta^{*\nu}$ vector of free utility coefficients on unobservable demographics. 5

$\boldsymbol{\Theta}^{*z}$ (true) matrix of utility coefficients on $z \times x$. 5

θ^{*z} vector of free utility coefficients on observable demographics. 5

$\triangledown$ transposition. 5

U

u_{ijm} utility. 5

V

π vector of π_{jm}'s across all markets. 8

$\mathcal{V}_\theta$ variance of the asymptotic distribution of $\hat{\Gamma}_\theta^{-1}(\hat{\theta} - \theta^*)$.. 14

$\mathbb{V}$ variance function. 11

W

$\hat{\mathcal{W}}$ weight matrix. 7

X

$\tilde{x}_{jm}$ exogenous product characteristics. 5

X_m observed matrix of product characteristics for market m. 5

$\mathcal{X}_m$ support of x_{jm}. 11

Ξ difference of Ξ_θ and Ξ_π. 15

Ξ_π at the truth, $\text{plim}_{M\to\infty}\left(\mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\mathbb{D}_\pi^\triangledown \mathcal{P}B(B^\triangledown B)^{-1}\right)$. 15

Ξ_θ at the truth, $\text{plim}_{M\to\infty}\left[\mathbb{D}_\theta^\triangledown \mathcal{P}B(B^\triangledown B)^{-1}\right]$. 15

ξ_{jm} unobserved product attribute. 4

ξ_m unobserved product characteristics. 5

x_{jm} vector of observed product characteristics. 4

Y

y_{ijm} consumer choice dummy. 5

y_{im} vector of y_{ijm}'s for all inside goods. 6

Z

z_{im} demographic characteristics. 5

$\mathcal{Z}$ support of consumer characteristics, z_{im}. 10