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 $x^{(1)} = [1 \quad 0]^\nabla$, $x^{(2)} = [1 \quad 1]^\nabla$. 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 $\delta = (-.25, 25)^\nabla$ 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.

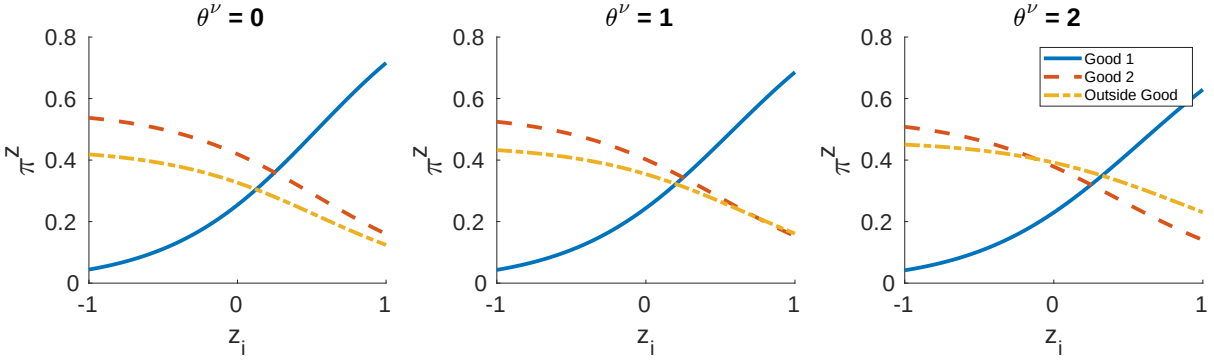

Figure 8: *Conditional shares $\tilde{\pi}^z$ are identified by the micro sample.*

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.

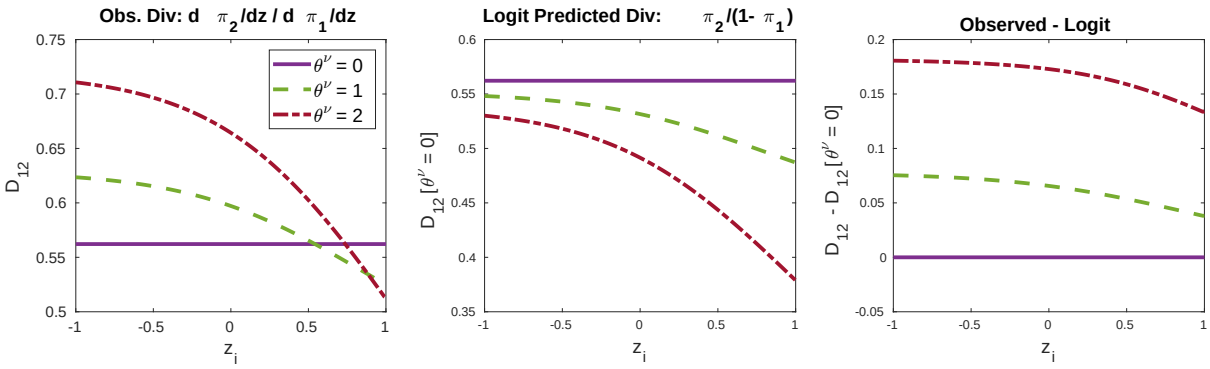

Figure 9: *Diversion and Demographics*

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,

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

In matrix notation, (50) can be written as

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

Thus, only if $X^\nabla\Theta^z$ 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 $c_m = [x_m^\nabla, b_m^\nabla, \xi_m^\nabla, N_m]^\nabla$, 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

$$0 = \sum_t q^*(\upsilon_t)H(\upsilon_t)[\delta(\theta,\upsilon_t) - X_{\upsilon t}^\nabla\beta] = \sum_t q_t^* H_t e_t(\alpha), \tag{52}$$

where $q_t^* = \mathbb{P}(c_m = \upsilon_t)$ with $\upsilon_1, \ldots, \upsilon_{\bar{t}}$ the values that c_m can take, $H_t = H(B_{\upsilon t})$ a matrix of instruments, $e_t(\alpha) = \delta_{\upsilon t}(\theta) - X_{\upsilon t}^\nabla\beta$ with $B_{\upsilon t}, X_{\upsilon t}, \delta_{\upsilon t}$ the values of δ_m (with π partialed out) if $c_m = \upsilon_t$. 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

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

Plugging q_t back into (53) yields

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

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

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

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 $\mathfrak{G}^{*\nabla}\mathfrak{V}^{*-1}\mathfrak{G}^*$ 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

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

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

$$\mathfrak{G}^{*\nabla}\mathfrak{V}^{*-1}\mathfrak{G}^* = \sum_k q^*_{(k)} A_k H^\nabla_{(k)} \Big(\sum_k q^*_{(k)} H_{(k)} V_k H^\nabla_{(k)}\Big)^{-1} \sum_k q^*_{(k)} H_{(k)} A_k^\nabla$$
$$\leq \sum_k q^*_{(k)} A_k V_k^{-1} A_k^\nabla = \mathbb{E}(B_m^{\text{opt}\nabla} \mathcal{V}_{\xi m} B_m^{\text{opt}}). \tag{55}$$

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

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

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 $S^\nabla = [\sqrt{q^*_{(1)}} A_1 V_1^{-1/2}, \ldots, \sqrt{q^*_{(\bar{k})}} A_{\bar{k}} V_{\bar{k}}^{-1/2}] \in \mathbb{R}^{d_\alpha \times (\bar{k} J_m)}$ 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 $\|\mathbb{D}_{\pi m}\| = \|\partial_{\pi_m^\nabla}\delta_m\|$ 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 $j = 1, \dots, J_m$,

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

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

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

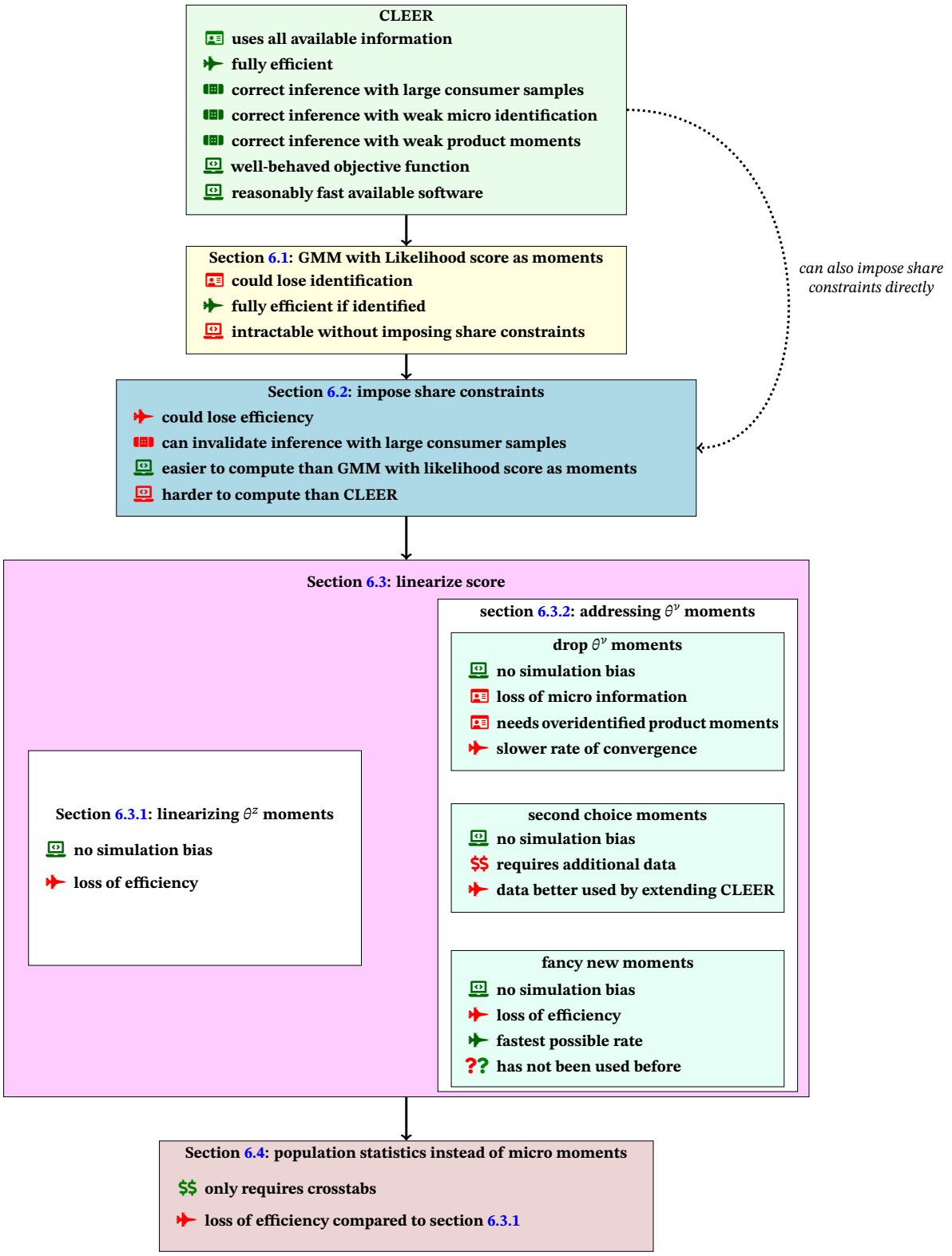

Figure 10: *Schematic comparison of our estimator to alternatives. See text for details.*

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 $v \neq 0$,

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

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 v for which $v^\triangledown \mathbb{D}_{\theta m}(\theta^*, \pi^*) = 0$,[59] where the left-hand side of (57) diverges. The space of such vectors v 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] $\sum_{j=0}^{J_m} s_{jm}\big(x_{jm}^k \nu^k - \sum_{k=0}^{J_m} s_{km} x_{km}^k \nu^k\big) = 0$, such that (26) can be expressed as

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

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

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

[59] Indeed, then by a Taylor expansion,

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

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 $\mathcal{V}_{\theta\delta}^{\text{CLEER}}$ be such that for any conformable C with a fixed number of columns,

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

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

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

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

Next, consider $\mathcal{V}_{\theta\delta}^{\text{SHCON}}$. This estimator is equivalent to placing infinite weight on $\mathcal{L}^\blacksquare$, however, since $\mathcal{L}_\theta^\blacksquare = 0$, 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 $\mathcal{L}_\theta^\blacklozenge = \mathcal{L}_\theta$ etc. So for $\bar{\mathcal{A}} = \mathcal{L}_{\pi\pi}^{\blacksquare -1}\mathcal{L}_{\pi\theta}, \bar{\mathcal{L}}_{\pi\pi} = \mathcal{L}_{\pi\pi}^\blacksquare + \mathcal{L}_{\pi\theta}\mathcal{L}_{\theta\theta}^{-1}\mathcal{L}_{\theta\pi}$, and $\bar{\mathcal{Q}}_{\theta\theta}^{\mathcal{L}} = \mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\pi}\bar{\mathcal{L}}_{\pi\pi}^{-1}\mathcal{L}_{\pi\theta}$, and $\bar{\mathfrak{G}} = \mathbb{D}_\theta - \mathbb{D}_\pi\bar{\mathcal{A}}\mathcal{L}_{\theta\theta}^{-1}\bar{\mathcal{Q}}_{\theta\theta}^{\mathcal{L}}$, we have,

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

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

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 $\mathcal{V}_{\theta\delta}^{\text{SHCON}}$ but often neglected in practice. If π^* were known, one could approximate $\mathcal{V}_{\theta\delta}^{\text{SHCON}}$ by an oracle equivalent,

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

[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., 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, $\mathbb{D}_\theta \mathcal{L}_{\theta\theta}^{-1} \mathbb{D}_\theta^\nabla$ has many eigenvalues equal to zero and so $(C^\nabla \mathcal{V}_{\theta\delta}^{\text{CLEER}} C)^{-1/2}$ is undefined. For the corresponding eigenvector-directions, the term $\mathbb{D}_\pi \mathcal{L}_{\pi\pi}^{-1} \mathbb{D}_\pi^\nabla$ is first order and hence needed to avoid this degeneracy. So an acceptable substitute under the assumption $N_m \succ I_m$ would be,

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

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.

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

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 $\partial_{\theta^\nu} \delta^{\text{sol}}(\theta) = -f_{\delta\delta}^{-1} f_{\delta\theta}$. 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 $\mathcal{K} = \mathcal{U}_B \mathcal{M}_{\mathcal{U}_{\bar{B}}^\triangledown \mathcal{U}_{\bar{X}}}$ 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 $\mathcal{P}_B - \mathcal{P}_{\mathcal{P}_B X} = (\mathcal{P}_C + \mathcal{P}_{\bar{B}}) - (\mathcal{P}_C + \mathcal{P}_{\mathcal{P}_B \bar{X}}) = \mathcal{P}_{\bar{B}} - \mathcal{P}_{\mathcal{P}_B \bar{X}}$ 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 $(\theta^{*z}_1, \theta^{*z}_2)$; and two random coefficients $(\theta^{*\nu}_1, \theta^{*\nu}_2)$.

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 $\mu^{z_{im}}_{jm} = \theta^{*z}_1 z^1_{im} x^1_{jm} + \theta^{*z}_2 z^2_{im} x^2_{jm}$,

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 $\mu^{\nu_{im}}_{jm} = \theta^{*\nu}_1 \nu^1_{im} x^1_{jm} + \theta^{*\nu}_2 \nu^2_{im} x^2_{jm}$.

In addition to the instrument b^1 for x^1 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 $\beta^* = [-6, 1, 1]^\nabla$, $\theta^{*z} = [1, 1]^\nabla$, and $\theta^{*\nu} = [1, 1]^\nabla$ 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 $(B^\nabla B)^{-1}$. 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.

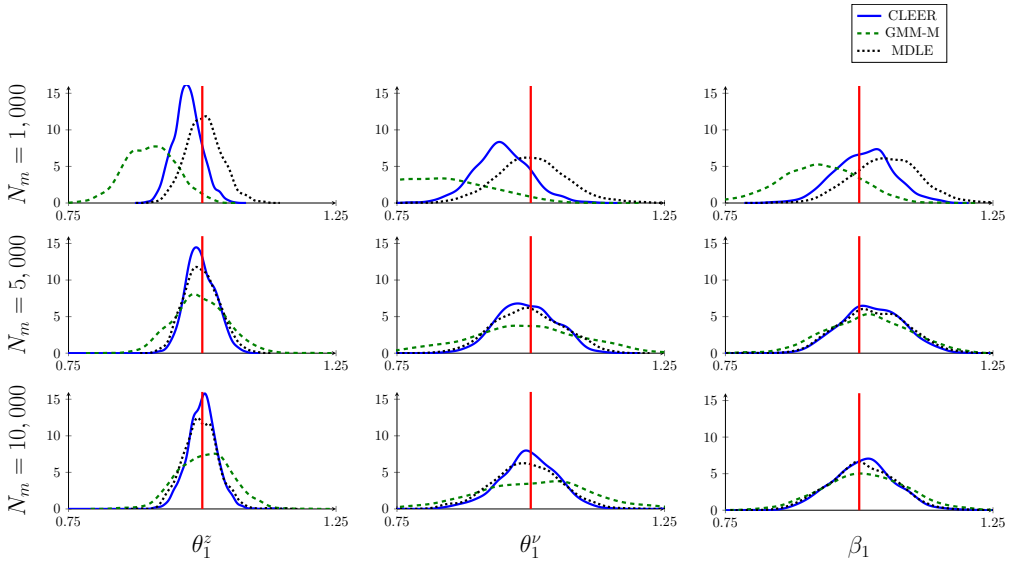

Figure 11: *Distribution of parameters for different population sizes N_m.*

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 $\hat{\upsilon} = [(B^\nabla \xi)^\nabla, \hat{\mathcal{L}}_\theta^\nabla, \hat{\mathcal{L}}_\pi^\nabla]^\nabla$ and $\omega = [\beta^\nabla, \theta^\nabla, \delta^\nabla]^\nabla$, $(\Lambda^\nabla \hat{\mathcal{H}}^{-1} \Lambda)^{-1/2} \Lambda^\nabla (\hat{\omega} - \omega^*) \simeq A^\nabla \hat{\upsilon} \xrightarrow{d} \mathcal{N}(0, \mathbb{I})$, 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 $\Lambda^\triangledown = \begin{bmatrix}\mathbb{0} & \mathbb{I} & \mathbb{0}\end{bmatrix}$. In theorem 2, we showed asymptotic normality of $\hat\theta$ by writing it as a linear combination of $\hat\upsilon$. Specifically, L8 showed that $\mathcal{V}_\theta^{-1/2}\Gamma_\theta\hat q_\theta \xrightarrow{d} \mathcal{N}(0,\mathbb{I})$. Recalling (48) and substituting $\hat r = \hat{\mathcal{L}}^\blacklozenge_\pi - \mathcal{L}^{\blacksquare-1}_{\pi\pi}(s-\pi^*) \simeq \hat{\mathcal{L}}_\pi$,

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

which implies for $A^\triangledown = \mathcal{V}_\theta^{-1/2}\Gamma_\theta\begin{bmatrix}\Xi & \mathbb{I} & -\mathcal{L}_{\theta\pi}\mathcal{L}^{-1}_{\pi\pi}\end{bmatrix}$ 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 Λ.

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, $(\Lambda^\triangledown\mathcal{H}^{-1}\Lambda)^{-1/2}\Lambda^\triangledown\Upsilon$, where $\hat{\mathcal{H}}^{-1}$ is a sample analog of,

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

For Υ,

$$\Lambda^\triangledown(\hat\omega-\omega^*) \simeq \Lambda^\triangledown\Upsilon_1\begin{bmatrix} B^\triangledown\xi \\ \hat\theta-\theta^* \\ \hat\delta-\delta^*\end{bmatrix} \simeq \Lambda^\triangledown\Upsilon_1\Upsilon_2\begin{bmatrix} B^\triangledown\xi \\ \hat\theta-\theta^* \\ \hat\pi-\pi^*\end{bmatrix} \simeq \Lambda^\triangledown\Upsilon_1\Upsilon_2\Upsilon_3\upsilon,$$

where $\Upsilon_1,\Upsilon_2,\Upsilon_3$ are respectively given by,

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

The formula for Υ_1 is due to the fact that $\hat\beta-\beta^* = (\mathcal{P}_B X)^+\hat\delta - \beta^* = (\mathcal{P}_B X)^+[(\hat\delta-\delta^*) + B^{+\triangledown}B^\triangledown\xi]$, Υ_2 is essentially applying the delta method to the transformation from π to δ, and Υ_3 amounts to a linearization of $B^\triangledown\xi$, $\hat\theta-\theta^*$, $\hat\pi-\pi^*$. The product $\Upsilon = \Upsilon_1\Upsilon_2\Upsilon_3$ is for $\mathcal{C} = (\mathcal{P}_B X)^+(\mathbb{I} - \mathbb{D}_\theta\mathcal{Q}^{-1}_{\theta\theta}\mathcal{P} - \mathbb{D}_\pi\mathcal{Q}^{-1}_{\pi\pi}\mathbb{D}^\triangledown_\pi\mathcal{P})B^{+\triangledown}$ given by,

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

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

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

The model implies that π^* is in the interior of Π. 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 $\kappa^\uparrow_\delta = 2\sqrt{2c^*_\xi\log M}$, $\kappa = \exp(-4\kappa^\uparrow_\delta)$, and let $\kappa_\pi = \kappa^{3/4} = \exp(-3\kappa^\uparrow_\delta)$, so $\kappa_\pi \succ \kappa$. Then, (a) $\mathbb{P}\big(\max_{m,j}|\delta^*_{jm}| > \kappa^\uparrow_\delta\big) \prec 1$; (b) $\max_{m,j}\kappa_\pi/\pi^*_{jm} \prec 1$ and $\max_{m,j}\kappa_\pi/s_{jm} \prec 1$; (c) $\min_m \inf_{\Pi^{\kappa c}_m}[\hat{\mathcal{L}}^\blacksquare_m(\pi_m) \;/\; N_m] \succ \kappa_\pi\log(\kappa_\pi/\kappa)$; (d) For a constant C, $\mathbb{P}\big\{\max_m \sup_{\Theta\times\Pi^\kappa_m}\boldsymbol{\lambda}_{\max}[\mathbb{D}_{\pi m}(\theta,\pi_m)] \le C\kappa^{-3}\big\} = 1$; (e) For any $0<p<\infty$ and some constant C only depending on p, $\max_m \mathbb{E}\big\{\sup_\Theta \boldsymbol{\lambda}^p_{\max}[\mathbb{D}_{\pi m}(\theta,\pi^*_m)]\big\} \le C$; (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 $c > 0$,

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

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 $\mathcal{S}_m = \operatorname{diag}(s_m)$,

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

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

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

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 $\mathbb{\Pi}_m^\kappa$, $\kappa \le \pi_{0m} = \int[\sum\exp(\delta_{jm}+\mu_{jm})]^{-1} \le C_\alpha\exp(-\max_j\delta_{jm})$. Moreover, $\kappa \le \min_j\pi_{jm} = \min_j\int\exp(\delta_{jm}+\mu_{jm}) \,/\, [\sum\exp(\delta_{tm}+\mu_{tm})] \le C_\beta\exp(\min_j\delta_{jm})$. Combining these, we have for all δ_{jm}: $\kappa \,/\, C_\beta \le \exp(\delta_{jm}) \le C_\alpha \,/\, \kappa$, 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 $\mathbb{I} - vv^\triangledown$ corresponds to the eigenvector v and is equal to $1 - \|v\|^2$, which for $v = \mathcal{S}_m^{-1/2}s_m$ is equal to $1 - \sum_{j>0}s_{jm} = s_{0m}$.

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 $p > 0$,

$$\mathbb{E}[(\pi_{jm}^*)^{-p}] \leq C_2^p \mathbb{E}[\exp(p|\delta_{jm}^*|)] \overset{\text{triangle}}{\leq} C_2^p \mathbb{E}[\exp(p|x_{jm}^\nabla \beta|)\exp(p|\xi_{jm}|)] \overset{\text{G[i],[ii], H[ii]}}{\leq} \infty,$$

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 $r_\epsilon(\theta) = \sqrt{\max[\rho_{\text{id}}(\theta), \check{\rho}_{\text{id}}(\epsilon)]}\log^2 I_+$. 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, $r_m^u = \sqrt{I_m}\kappa^{-3|u|}(\log I_+)^{2+\max(u_z,1)} + \exp(-N)$, where the $\exp(-N)$ term serves to ensure that we are not dividing by zero. Then, for $|u| > 0$, $\max_m \sup_{\Theta\times\mathbb{\Pi}_m^\kappa}\left[|\partial^u\Delta\hat{\mathcal{L}}_m^\blacklozenge(\theta, \pi_m)|/r_m^u\right] \prec 1$; and if $|u| = 0$ then $\max_m \sup_{\Theta\times\mathbb{\Pi}_m}\left[|\Delta\hat{\mathcal{L}}_m^\blacklozenge(\theta, \pi_m)|/r_m^u\right] \prec 1$.

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 $^{\text{L14}}$ to distinguish objects in L14 from objects here, take $\psi^{\text{L14}} = \theta$, $n^{\text{L14}} = I$, and write $\hat{\mathbb{P}}^{\text{L14}}(\theta) = \Delta\hat{\mathcal{L}}^\blacklozenge(\theta, \pi^*)/r_\epsilon(\theta)$ 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 $(\zeta_i^{\text{L14}}, z_i^{\text{L14}})$ corresponds to (ζ_{im}, z_{im}) for one $(m \in \{1, ..., M\}, i \in \mathcal{I}_m)$ combination. We now verify the conditions of L14.

First, L14[i] is satisfied if we make $\boldsymbol{\delta}^{\text{L14}}$ decrease at a sufficiently fast polynomial rate of I because $\hat{\mathbb{P}}^{\text{L14}}$ 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 $\boldsymbol{\beta}^{\text{L14}} = \log I$, such that [iv] is satisfied. Finally, [v]. Note that $\mathbb{h}^{\text{L14}} \preceq \check{\rho}_{\text{id}}^{-1}(\epsilon)\log I$ by L15(c). By L15(d), $\bar{\boldsymbol{\sigma}}^{\text{L14}2} \preceq (\log I)^{-4}$. To verify [v], first note that $\exp(-c/\mathbb{h}^{\text{L14}})$ and $\exp(-c/\bar{\boldsymbol{\sigma}}^{\text{L14}2})$ decrease faster than any power of I. Now, due to the compactness of Θ, we can choose $\boldsymbol{T}^{\text{L14}}$ to increase at a (sufficiently fast) polynomial rate of I (that depends on our choice of $\boldsymbol{\delta}^{\text{L14}}$) to make the requirements on $\boldsymbol{T}^{\text{L14}}$, $\boldsymbol{\delta}^{\text{L14}}$ 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 $g^{\text{L14}} = m$, $z_{ig}^{\text{L14}} = z_{im}$, $\psi_g^{\text{L14}} = (\theta, \pi_m)$, $n_g^{\text{L14}} = I_m$. Now $\hat{\mathbb{P}}_m^{\text{L14}}(\theta, \pi_m) = \partial^u\Delta\hat{\mathcal{L}}_m^\blacklozenge(\theta, \pi_m)/r_m^u = \sum_{\mathcal{I}_m} \zeta_{im}(\theta, \pi_m)$ with $\zeta_{im}(\theta, \pi_m) = \sum_j (y_{ijm} - \varsigma_{ijm}^*)\partial^u\Delta\ell_{ijm}(\theta, \pi_m)/r_m^u$. First, L14[i] is satisfied if we make $\boldsymbol{\delta}_{ng}^{\text{L14}}$ decrease at a sufficiently fast polynomial rate of I because $\hat{\mathbb{P}}_m^{\text{L14}}$ 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 $\boldsymbol{\beta}_m^{\text{L14}} = \log I$, such that [iv] is satisfied. Finally, [v]. By L15(c), $\max_m \mathbb{h}_m^{\text{L14}} \preceq (\log I)^{-2}$.

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

$$\max_m \bar{\boldsymbol{\sigma}}_m^{\text{L14}2} \leq \bar{J}\max_m\left\{I_m \sup_{\Theta\times\mathbb{\Pi}_m^\kappa} \sum_j \mathbb{E}\left[\varsigma_{ijm}^*\left(\frac{\partial^u\Delta\ell_{ijm}(\theta, \pi_m)}{r_m^u}\right)^2\right]\right\} \overset{\text{L15(c)}}{\preceq} (\log I)^{-4}.$$

To verify [v], note that $\max_m \exp(-c/\mathbb{h}_m^{\text{L14}})$ and $\max_m \exp(-c/\bar{\boldsymbol{\sigma}}_m^{\text{L14}2})$ 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 $\mathbb{M}_m$. 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),

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

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,

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

First,

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

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

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

Finally, by MVT, triangle, and Schwarz inequalities,

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

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 $C < \infty$, $\forall \theta \in \Theta_\epsilon^c$: $\|\theta^z\|^2 \leq C\|\theta - \theta^*\|_\lambda^2$; (c) for a fixed $C < \infty$, $\forall \theta \in \Theta_\epsilon^c$: $\lambda^2 \leq C\|\theta - \theta^*\|_\lambda^2$; (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

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

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

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

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

Finally, (d). We have,

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

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

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

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 $r_{jm}(\nu) = \sum_k \theta^\nu_k x^\nu_{jm(k)}\nu_k + \delta_{jm}$ where x^ν_{jm} 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, $|\sum_k \theta^z_k x_{jm(k)} z_{im(k)}| \leq C_1\|z_{im}\| \cdot \|\theta^z\|$. So, for all θ, π_m:

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

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

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 $\ell_{ijm}(\theta, \pi_m) - \ell_{jm}(\theta, \pi_m) \leq \log\exp\big[\|\theta^z\|(2C_1\|z_{im}\| + C_2)\big] = \|\theta^z\|(2C_1\|z_{im}\| + C_2)$. 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 $\theta^z = 0$, $|\log \int_z \mathbb{f}(\theta^z, z)| \leq |\log 1| + 2C_1 \int_z \|z\|\mathbb{f}(\mathring{\theta}^z, z) \,/\, \int_z \mathbb{f}(\mathring{\theta}^z, z)$.

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 $\max_{m=1,\dots,M} \sup_{\Theta^c_\varepsilon \times \mathbb{\Pi}^\kappa_m} \|\Delta\hat{\mathcal{L}}^\blacklozenge_m(\theta,\pi_m)/r^u_m\|$ 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 $g = 1, \dots, \boldsymbol{G}$, we define a parameter space Ψ_g, a true parameter vector ψ^*_g, and a sequence of independent random vectors $\{\upsilon_{ig}\}$, with $i \in \{1, \dots, n_g\}$. We partition each space Ψ_g into $\boldsymbol{T}_{ng}$ sets $\Psi_{g1}, \dots, \Psi_{g\boldsymbol{T}_{ng}}$ and let $\boldsymbol{\delta}_{ng} = \max_{t=1,\dots,\boldsymbol{T}_{ng}} \sup_{\psi_g,\psi^\circ_g \in \Psi_{gt}} \|\psi_g - \psi^\circ_g\|$ 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 $\hat{\mathbb{f}}_{ng}(\psi_g) = \hat{\mathbb{f}}_g[\psi_g, \{\upsilon_{ig}\}]$.

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

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 $\varepsilon > 0$, write

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

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., $\pi \in \mathbb{\Pi}^\kappa$. 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 $\hat{\pi} \in \mathbb{\Pi}^\kappa$ 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 $\forall \theta, \pi_m \in \Theta \times \mathbb{\Pi}^\kappa_m\colon \partial^u \Delta\ell_{ijm}(0, \theta^\nu, \pi_m) = 0$; (b) If $u_z = 1$ then $\forall \theta, \pi_m \in \Theta \times \mathbb{\Pi}^\kappa_m\colon \mathbb{E}[\partial^u \Delta\ell_{ijm}(0, \theta^\nu, \pi_m) \mid \mathbb{A}] = 0$; (c) For

a constant $C < \infty$: $\mathbb{P}\big[\max_m \max_{\Theta\times\mathbb{\Pi}^\kappa_m} |\partial^u \ell_{ijm}(\theta,\pi_m)| > C\|\mathbb{D}_{\pi m}(\theta,\pi_m)\|^{|u|}\|z_{im}\|^{\max(u_z,1)} \mid \mathbb{A}\big] = 0$ and $\mathbb{P}\big[\max_m \max_{\Theta\times\mathbb{\Pi}^\kappa_m} |\partial^u \ell_{ijm}(\theta,\pi_m)| > C\kappa^{-3|u|}\|z_{im}\|^{\max(u_z,1)} \mid \mathbb{A}\big] = 0$; (d) $\sup_\theta \sum_m \sum_{\mathcal{I}_m} \mathbb{V}\{\sum_j (y_{ijm} - \varsigma^*_{ijm})[\Delta\ell_{ijm}(\theta,\pi^*_m) - \Delta\ell^*_{ijm}]\} \;/\; \|\theta - \theta^*\|^2_\lambda \preceq 1$; (e) For some fixed $C > 0$, $\max_m \sup_{\Theta^c_\xi\times\mathbb{\Pi}^\kappa}(|\partial^u \mathcal{L}^\blacklozenge_m(\theta,\pi_m)|/\{\max(I_m,1)\boldsymbol{\lambda}^{|u|}_{\max}[\mathbb{D}_{\pi m}(\theta,\pi_m)]\}) \leq C$ and $\max_m \sup_{\Theta^c_\xi\times\mathbb{\Pi}^\kappa}(|\partial^u \mathcal{L}^\blacklozenge_m(\theta,\pi_m)|/\max(I_m,1) \preceq \kappa^{-3|u|}$; (f) Let $\tau_m = \mathbb{1}(\|\pi_m - \pi^*_m\| \leq \kappa)$. If $u_z \leq 1$, then for some fixed C, $\max_m \mathbb{E}|\partial^u \mathcal{L}^\blacklozenge_m(\theta,\pi_m)\tau_m|/\max(I_m,1) \leq C(\|\theta^z\| + \lambda)^{2-u_z}$.

Proof. (a) and (b) are trivial since $\forall\theta^\nu,\pi_m\colon \Delta\ell_{ijm}(0,\theta^\nu,\pi_m) = 0$ and $\forall\theta^\nu,\pi_m\colon \mathbb{E}[\Delta\ell_{\theta^z ijm}(0,\theta^\nu,\pi_m) \mid \mathbb{A}_m] = 0$, 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

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

Now, by the chain rule, $\ell_{\pi ijm} = \mathbb{D}^\nabla_{\pi m}\tilde{\ell}_{\delta ijm}$ and $\ell_{\theta ijm} = \tilde{\ell}_{\theta ijm} + \mathbb{D}^\nabla_{\theta m}\tilde{\ell}_{\delta ijm} = \tilde{\ell}_{\theta ijm} - \partial_\theta\sigma^\nabla_m\mathbb{D}^\nabla_{\pi m}\tilde{\ell}_{\delta ijm}$.[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,

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

where the last inequality follows from L9(d), and $\|\mathring{\theta}^z\| \leq \|\theta^z - \theta^{*z}\| + \lambda$ 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 $0 \leq t \leq 1$,

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

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 $\mathbb{D}_\theta + \mathbb{D}_\pi\partial_{\theta^\nabla}\sigma = 0$.

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^y			θ_2^y			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^y)															
[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_1^z			θ_2^z			θ_1^y			θ_2^y			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^y)															
[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 $M = 50$, $S_m = 1,000$, $\theta_1^z = 1.0$, and $a = 0.5$.

Table 5: *Monte Carlo Results: Bias*

	θ_1^z			θ_2^z			θ_1^y			θ_2^y			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^y)															
[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^y			θ_2^y			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^y)															
[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^y			θ_2^y			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^y)															
[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^v			θ_2^v			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] $a = 1.0, \theta v = 2$	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] $a = 1.0, \theta v = 2.5$	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] $a = 1.0, \theta v = 3$	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] $a = 0.5, \theta v = 2$	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] $a = 0.5, \theta v = 2.5$	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] $a = 0.5, \theta v = 3$	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] $a = 1.0, \theta v = 2$	-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] $a = 1.0, \theta v = 2.5$	-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] $a = 1.0, \theta v = 3$	-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] $a = 0.5, \theta v = 2$	-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] $a = 0.5, \theta v = 2.5$	-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] $a = 0.5, \theta v = 3$	-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] $a = 1.0, \theta v = 2$	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] $a = 1.0, \theta v = 2.5$	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] $a = 1.0, \theta v = 3$	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] $a = 0.5, \theta v = 2$	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] $a = 0.5, \theta v = 2.5$	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] $a = 0.5, \theta v = 3$	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] $a = 1.0, \theta v = 2$	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] $a = 1.0, \theta v = 2.5$	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] $a = 1.0, \theta v = 3$	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] $a = 0.5, \theta v = 2$	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] $a = 0.5, \theta v = 2.5$	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] $a = 0.5, \theta v = 3$	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 $x_n \xrightarrow{p} x$ and $y_n \xrightarrow{d} y$ then $x_n y_n \xrightarrow{d} xy$

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 $x_n \xrightarrow{p} x \Rightarrow g(x_n) \xrightarrow{p} g(x)$ 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

$\mathcal{A}\lim_{M\to\infty}(B^{\triangledown}B \,/\, M)$. 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 $\blacklozenge$ superscript). 13

$\Delta\ell_{ijm}$ $\log s_{ijm} - \log s_{jm}$. 19

δ_m Berry inversion (when used as a function). 8

$\Delta\hat{\Omega}$ $\hat{\Omega} - \Omega$ (analogously when endowed with a $\blacklozenge$ 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 $\eta = \kappa^3$. 13

F

F_m distribution of unobservable demographics. 5

G

G_m distribution of observable demographics. 5

$\Gamma_\pi = \left\{\mathbb{E}[\mathcal{L}_{\pi\pi} - \mathcal{L}_{\pi\theta}\mathcal{L}^+_{\theta\theta}\mathcal{L}_{\theta\pi}]\right\}^{-1/2}$. 15

$\Gamma_\theta = \left[\mathbb{E}(\mathcal{L}_{\theta\theta} - \mathcal{L}_{\theta\pi}\mathcal{L}^{-1}_{\pi\pi}\mathcal{L}_{\pi\theta}) + M\Xi\mathcal{A}\Xi^\nabla\right]^{-1/2}$; basically a population analog to $\hat{\Gamma}_\theta$. 15

$\hat{\Gamma}_\theta$ square root of the $\theta\theta$ block of the inverse Hessian of Ω, $\hat{\Gamma}_\theta = \hat{\mathcal{Q}}^{-1/2}_{\theta\theta}$. 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_Ω. 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

κ rate used in C, $\kappa = \exp(-4\kappa^\uparrow_\delta)$. 10

$\kappa^\uparrow_\delta$ rate used in C, $\kappa^\uparrow_\delta = 2\sqrt{2c^*_\xi \log M}$. 10

κ_π we show that $\min_{m,j} \pi^*_{jm} \succeq \kappa_\pi$. 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
$\ell_{jm} \log \varsigma_{jm}$. 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
$\mu^{z_{im}}_{jm}$ 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 $\mathbb{\Pi}_m^\kappa = \{\pi_m : \min_j \pi_{jm} \geq \kappa\}$. 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
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
$\hat{q}_\theta$ component of numerator term in quadratic expansion, $\hat{q}_\theta = \hat{\Omega}_\theta - \Omega_{\theta\pi}\Omega_{\pi\pi}^{-1}(\hat{\Omega}_\pi^\blacklozenge - \mathcal{L}_{\pi\pi}^\blacksquare(s - \pi^*)$, see L8. 15

R

ρ_D rate governing $\Delta\hat{\Omega}^\blacklozenge(\theta, \pi) - \hat{\Omega}^\blacklozenge(\theta^*, s)(\theta, \pi)$, $\rho_D(\theta, \pi) = \eta \max\{\eta\rho_{\text{id}}(\theta), \rho^\blacksquare(\pi)\}$. 13
$\rho_{\textbf{id}}$ 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

$s_{ijm}(\nu; \theta, \delta)$ choice probability before integrating out random coefficients. 5
s_{jm} observed market share. 6
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
∇ 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, $\text{plim}_{M\to\infty}\left(\mathcal{L}_{\theta\pi}\mathcal{L}_{\pi\pi}^{-1}\mathbb{D}_\pi^\nabla \mathcal{P}B(B^\nabla B)^{-1}\right)$. 15
Ξ_θ at the truth, $\text{plim}_{M\to\infty}\left[\mathbb{D}_\theta^\nabla \mathcal{P}B(B^\nabla B)^{-1}\right]$. 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