TEXTO PARA DISCUSSÃO

No.707

Flexible Demand Estimation and Zero Market Shares

Lucas Lima

PUC RIO

DEPARTAMENTO DE ECONOMIA

Flexible Demand Estimation and Zero Market Shares*

Lucas Lima
PUC-Rio

March, 2024

click here for the latest version

Abstract

This paper develops a flexible discrete-choice demand framework for aggregate data sets that extends Berry, Levinsohn, and Pakes (1995) and the Pure Characteristics Demand Model of Berry and Pakes (2007). I provide a simple, computationally tractable, asymptotically normal estimator based on two contributions: a globally-convergent algorithm to recover utilities from observed demand and a Quasi-Bayes approach that minimizes simulation variance. The framework accommodates zero market shares, which are a challenge for alternative approaches. I show that zeros in demand generate an endogenously censored model, which leads to moment inequalities. As an application, I study moving costs US internal migration data. ***Keywords: Demand Estimation, Zero Market Shares, Moving Costs, Housing Policy, Endogenous Censoring, Moment Inequalities JEL Codes:***

*This article is based on my PhD thesis. An earlier version circulated under the title "Demand Estimation with Zeros: Moving Costs in the US." I wish to thank Myrto Kalouptsidi, Marcelo Moreira, Ariel Pakes, Nicola Rosaia, Eduardo Souza-Rodrigues, Elie Tamer and Allen Zhang for helpful comments and discussions. This paper greatly benefited from seminar participants at Queen Mary University of London, PUC-Rio, EPGE-FGV, Wisconsin-Madison, University of Toronto, and NYU Stern.

Introduction

Demand estimation is central to many economic questions, as it summarizes price elasticities and the impact of changes in choice sets. This knowledge is essential to predict the effect of subsidies, new goods, and other changes to the market structure. Furthermore, as the Industrial Organization literature has repeatedly emphasized, we need flexible demand specifications to allow the data to inform these predictions.[1] In this literature, the paper by Berry, Levinsohn, and Pakes (1995) (henceforth BLP) has introduced the most used method for flexible demand estimation with discrete-choice data sets.[2]

However, the assumptions in BLP can also be restrictive, and the literature has tried to relax them. Of particular concern is their assumption of idiosyncratic utility shocks with full support, which can have an undue influence on elasticities and welfare predictions. (Petrin, 2002; Bajari and Benkard, 2003; Ackerberg and Rysman, 2005; Ackerberg, Benkard, Berry, and Pakes, 2007) Utility shocks with full support imply strictly positive demand, and, whenever the data has zeros in the market shares, the BLP estimator fails.[3]

This limitation rules out interesting applications to demand models such as International Trade and Migration; for example, half of country pairs do not trade. (Helpman et al., 2008) It also limits the applicability of a standard demand model in more traditional markets, for example, electric vehicles, where Li (2019) observes models with no sales even at the state level, and online retailers, where Quan and Williams (2018) observe 13.5 million shoe transactions, but 85% of products have zero market shares at a state-month level.

This paper contributes to the literature by introducing an estimator for a demand model that extends BLP by relaxing their idiosyncratic utility shock assumption. In particular, my estimator applies to the Pure Characteristics Demand Model of Berry and Pakes (2007) (henceforth PC).[4] The PC model is theoreticaly appealing as it avoids counterintuitive implications of full-support utility shocks.[5] However, how to estimate it in large datasets has been an open problem in the literature, which has limited the number of its empirical applications.

My estimator uses a novel algorithm to compute the inverse demand. The algorithm is surprisingly simple compared to alternatives but works well even in applications with hundreds of options. I show it is globally convergent under very general assumptions.[6] Another novelty of

[1] Early examples of flexible demand specifications include Theil (1965); Christensen, Jorgenson, and Lau (1975); Deaton and Muellbauer (1980); McFadden (1974, 1981).

[2] Besides Industrial Organization, the BLP method has been successfully applied to many fields, like Public Finance, Health Economics, and International Trade. (Berry and Haile, 2016)

[3] The BLP estimator is not robust to close-to-zero market shares. Berry, Linton, and Pakes (2004b) provide an asymptotic approximation to the BLP estimator in a market with many options to understand the impact of small market shares on simulation and sampling errors.

[4] So-called because it assumes that idiosyncratic utility comes *purely* from the observed characteristics.

[5] See the discussion in Berry and Pakes (2007) for more details.

[6] the algorithm proceeds roughly as follows: from a starting guess, the utility for every product with excess demand (relative to the observed market share) is reduced *by the same* step size. This explores the separability of the utility model, which guarantees that demand moves in the right direction. Then, the step size decreases after the demand has crossed over the observed market share for every option. By waiting for every option to cross, the algorithm keeps the step size "well-calibrated" to the current approximation error.

my approach is to combine a Quasi-Bayes estimator with a simulation-based estimator to smooth over the objective function, minimizing the simulation error. The implementation uses a Sequential Monte Carlo Sampler, which efficiently computes the estimator. (Del Moral et al., 2006)

This paper's estimator and inference procedure rely on mapping a demand model with zero market shares to an endogenously censored model. We recover the product utility only when market shares are strictly positive; otherwise, the utility is censored, and we observe an upper bound to it. Under a conditional median assumption, this model leads to moment inequalities, as in Khan and Tamer (2009). I utilize the asymptotic approximation for many markets of Freyberger (2015) and the results for Quasi-Bayes estimators of Chernozhukov and Hong (2003).

To illustrate the performance of the proposed methodology, I apply it to US internal migration data using publicly available data sets from the Opportunity Insights[7] and the IRS.[8] Location choice is a natural application for my method because of the many options and the large number of zeros. Migration is defined at the commuting-zone level for families with children observed between 1996 and 2012. Thus, each household has 718 options to choose from; with that many options, naturally, there are many zeros: only about 3% of all the possible commuting-zone pairs have non-zero observations.[9]

In this empirical setup, it would be virtually impossible to apply the standard BLP approach. Nonetheless, my approach can handle the computational challenges and, even with many zeros, the data are highly informative and allow for a flexible specification of moving costs. An essential source of information, besides the families' choices, are (micro) moments on the average income for movers between large commuting zones, in the spirit of Petrin (2002) and Imbens and Lancaster (1994).

A common finding of the literature is very high, perhaps unreasonably high, moving costs. My procedure suggests high moving costs, consistent with the lack of mobility in the data, but not as high as previous estimates (Kennan and Walker, 2011; Bayer, McMillan, Murphy, and Timmins, 2016). The difference is due to the rich unobserved heterogeneity allowed by my approach. Without flexible unobserved heterogeneity, the variability in the data is explained by large idiosyncratic utility shocks. But with many options and large (independent) utility shocks, many people would prefer to move unless moving costs are very high.[10] In contrast, in my estimates, the significant heterogeneity in choices translates into highly variable moving costs.

Literature This paper is part of the vast literature on demand estimation for differentiated products following the seminal contributions of McFadden (1974, 1981) and Berry (1994); Berry, Levinsohn,

[7] https://opportunityinsights.org/data/

[8] https://www.irs.gov/statistics/soi-tax-stats-migration-data

[9] Due to confidentiality restrictions, observations are only reported if at least 25 children have moved between pairs of commuting zones.

[10] In the context of demand for health care insurance plans, Pakes, Porter, Shepard, and Calder-Wang (2021) get a similar reduction in switching costs when they allow for individual-by-option fixed effects. Their result highlights the importance of flexible unobserved heterogeneity to minimize the impact of distributional assumptions.

and Pakes (1995).[11]

My paper relates to a few different strands within this literature—first, it relates to the much smaller literature on demand models without the logit assumption. The closest to my approach is Berry and Pakes (2007). They develop a computational method to estimate a demand model with no logit utility shock, but otherwise similar to BLP. However, their method can be prohibitively costly to compute and assumes strictly positive demand. In comparison, my approach is simpler to implement, has a lower computational cost, and provides a unified method for PC and other similar demand models while allowing for zeros.[12]

Second, my paper contributes to the literature on computational methods for demand estimation. For example, Dubé, Fox, and Su (2012), Lee and Seo (2015), and Salanié and Wolak (2019) provide alternatives to the estimator developed in BLP to speed up or avoid the numerical inversion from demand to utilities. However, they all keep the logit assumption and extensions to other models are not trivial.

My algorithm to compute the inverse demand is related to the Market Share Adjusting algorithm of Bonnet, Galichon, Hsieh, O'Hara, and Shum (2022). However, their algorithm targets partially identified models with non-additive errors, while I focus on additive models. Additive models are less general but are essentially the only utility specification used in empirical research. Furthermore, by exploring the extra structure in them, my approach has better convergence properties.[13]

Third, my paper relates to the recent and active literature on solutions to BLP's inability to handle zeros. Thus far, the literature has dealt with zero demand in three ways: treating zeros as a statistical problem from a small number of consumers, removing zeros from the choice set, or aggregating different options until there are no zeros. Gandhi, Lu, and Shi (2020) assumes a (data-dependent) lower bound for choice probabilities and estimates the model using moment inequalities. Dubé, Hortaçsu, and Joo (2021) assumes a selection process for zeros that restricts the choice set before applying the BLP approach. Finally, Quan and Williams (2018) adds a parametric heterogeneity at the local level before aggregating to eliminate zero shares.[14] I deviate from this literature by allowing for bounded utilities; the bounds do not need to be known a priori. This approach can generate zero market shares for products dominated by other options besides zeros caused by a small sample of consumers or unavailable options.

My discussion on close-to-zero demand and its impact on the BLP model relates to the paper by Berry, Linton, and Pakes (2004b), which analyzes how small market shares affect the asymptotic variance of the BLP and PC estimators. My asymptotic results borrow from the many-market

[11] Recent surveys are Ackerberg, Benkard, Berry, and Pakes (2007); Nevo (2011); Berry and Haile (2016, 2021); Gandhi and Nevo (2021)

[12] Other approaches are the hedonic model by Bajari and Benkard (2005) and the "crowding-out" model by Ackerberg and Rysman (2005). Bajari and Benkard (2005) estimates the Pure Characteristic model by noticing that if two options have the same characteristics but different prices, then unobserved characteristics must exactly compensate the difference. Ackerberg and Rysman (2005) changes the utility of the BLP model by adding a term that depends on the number of options in the market.

[13] See Appendix A.2.

[14] A fourth approach is to replace the zero market shares with the posterior mean of a Bayesian model, as suggested in Li (2019) in the context of demand for electric vehicles.

approximation of Freyberger (2015), the results for endogenously censored models of Khan and Tamer (2009), and the results for Quasi-Bayes estimators of Chernozhukov and Hong (2003). The approach in Chernozhukov and Hong (2003) has been applied to demand estimation in a recent paper by Hong, Li, and Li (2021).

Finally, my empirical application relates to the extensive literature on the decision to move and moving costs. The discrete-choice models from Kennan and Walker (2011) and Bayer, McMillan, Murphy, and Timmins (2016) are closely related. However, in contrast to these papers, my method needs only aggregate data, allows for richer unobserved heterogeneity, and does not assume a logit utility shock. On the other hand, their models incorporate dynamic considerations, while I focus on a static model. The counterfactual housing policy analysis relies on the neighborhood impact on children's outcomes estimated in Chetty and Hendren (2018a,b).

Overview Section 1 introduces the demand model and Section 4 discusses why it is important to allow for zero market shares and how to handle them. Section 2 details the estimator's computational aspects, while Section 3 presents the asymptotic results. Sections 5 and 6 explore the migration data and empirical results. Section 7 concludes.

1 Demand Model

I start from a standard random utility discrete-choice demand model. Then, following BLP and PC, I consider heterogeneous consumers choosing an option among a set of differentiated products.

More precisely, consumer i in market m selects the option j from the choice set $\{0, \dots, J\}$ that maximizes the utility

$$u_{mji} = \xi_{mj} + \sum_{k=1}^{K} X_{mjk}\beta_{ki} + \varepsilon_{mji},$$

where ξ_{mj} denotes the product unobserved characteristic. X_{mjk} denotes the (observed) characteristic k, and β_{ki} represents i's "taste" for this characteristic. ε_{mji} represents the idiosyncratic "taste" for option j, which has a known distribution and is independent of everything else; thus, for example, in BLP, ε_{mji} has an extreme value type 1 distribution. In contrast, in the Pure Characteristics model of PC, ε_{mji} is identically zero. In general, I denote the distribution of ε_{mji} by G.

Heterogeneity in β_{ki} allows for consumers to differ in their tastes for characteristics. This flexibility is essential to generate realistic aggregate own- and cross-price elasticities. Because, with β_{ki} constant over i, elasticities depend only on the market shares, which leads to unreasonable substitution patterns, as it is well-known. (Berry et al., 1995) Therefore, I assume that β_{ki} follows a distribution $F(\boldsymbol{\theta})$ parametrized by a finite-dimensional $\boldsymbol{\theta}$.

Another essential feature is the unobserved characteristic ξ_{mj}. It aggregates unobserved (by the econometrician) product attributes or consistent taste variation within a market in a one-dimensional error term.[15] In practice, it explains why some options may have unexpected high or low demand

[15]The restriction to a one-dimensional unobserved characteristic is potentially restrictive but is standard in the literature

and may be correlated with some observed characteristics, such as the price.

More explicitly, given parameters θ and unobserved characteristics ξ,[16] we obtain a demand for each consumer. And, if we sum them up, we get an aggregate demand,

$$\sigma_j(\xi|\theta) = \int \mathbb{1}[u_{ji} \geq u_{j'i} \text{ for all } j'] \, \mathrm{d}F(\beta_i \mid \theta)\mathrm{d}G(\varepsilon), \tag{1}$$

that we can compare with the observed market shares, s.[17] The unobserved characteristics $\xi^*(\theta)$ that set aggregate demand equal to the observed market shares are the "error term." This is similar to how "$Y - X'\theta$" is the error term in a linear regression.[18]

An essential step in the estimation is to recover these unobserved characteristics, which must be done numerically.[19]

2 Computation: Estimation without the Logit Assumption

This section describes the two main computational aspects of my estimator. First, I present the algorithm to invert from observed demand to unobserved characteristics and discuss its convergence properties. Then, I outline the Quasi-Bayes estimation procedure. An interesting aspect is that it minimizes the simulation variance by averaging over simulation draws instead of keeping one draw fixed. The overall procedure has good behavior even in applications with hundreds of products and can be applied even when zero market shares are not an issue.

2.1 Algorithm for demand inversion

The algorithm is an iterative procedure to recover the unobserved characteristics from the observed market shares. It starts from an unobserved characteristic guess for each product. Then it increases the guess if demand is below the market share and decreases if above. At each iteration, the algorithm updates the unobserved characteristics by *the same step size*, which guarantees that demand for every product moves toward the market shares. Then, the step size decreases when *every* product demand has jumped over its market share *at least once* (since the last step update). This condition guarantees that we decrease the step size only when the unobserved characteristics are close to the target for every product.

The algorithm pseudocode follows.

(Athey and Imbens, 2007).

[16] For simplicity, I supress the market m from the notation until it is relevant.

[17] In general, this integral does not have a closed-form solution and needs to be approximated. I return to this point in Section 4.2.

[18] Under mild conditions, there is a unique $\xi^*(\theta)$ that solves the inversion problem (Berry, Gandhi, and Haile, 2013).

[19] Except for very special cases, there is no closed-form solution (Berry, 1994; Berry et al., 1995). See Section 2.1.

Algorithm: Demand Inversion Pseudocode

{1} Initialize $n \leftarrow 0$, a guess $\boldsymbol{\xi}^0$, a step size $\mathbf{step}_0$, and a scalar $\rho \in (0, 1)$.

{2} For *every* option, if $\sigma_j(\boldsymbol{\xi}^r) \leq s_j$, set

$$\xi_j^{r+1} \leftarrow \xi_j^r + \mathbf{step}_r,$$

else if $\sigma_j(\boldsymbol{\xi}^r) > s_j$, set

$$\xi_j^{r+1} \leftarrow \xi_j^r - \mathbf{step}_r.$$

{3} If demand for *every* option has changed side at least once, set

$$\mathbf{step}_{r+1} \leftarrow \rho \cdot \mathbf{step}_r.$$

Otherwise, keep the same step size.

{4} Set $r \leftarrow r + 1$ and go back to {2} until $\mathbf{step}_r$ is below a tolerance.

{5} Normalize $\boldsymbol{\xi}^r$ by subtracting the outside option unobserved characteristic

$$\xi_j^* \leftarrow \xi_j^r - \xi_0^r$$

Although simple, this algorithm has surprisingly good behavior for two reasons:
First, demand moves toward the market shares for every product at each step: Suppose we increase the unobserved characteristic for two options, j and j', by the same amount, $\Delta\xi$. Consumers that preferred j to j' before the increase will not change their minds and vice-versa because

$$u_j + \Delta\xi \geq u_{j'} + \Delta\xi \quad \iff \quad u_j \geq u_{j'}.$$

However, j and j' are now more desirable and may steal demand from other products, (weakly) increasing their demand. Naturally, the same argument extends to more than two products.
Second, the step size only decreases after the demand for *every* product has jumped over the observed market shares. After that happens, the demand is guaranteed to stay close enough to the observed market share, and it is the right moment to decrease the step size. This step size is calibrated to the current approximation error and is essential for convergence at a geometric rate. The argument is further elaborated in Lemma 3.

Remark. This algorithm is globally convergent at a geometric rate but not a contraction; the approximation error can increase at some iterations despite the demand always moving toward the market shares. The reason is that the demand may overshoot the target. However, the algorithm corrects itself in the next iteration and moves the demand in the opposite direction. Thus, if it

overshoots is by at most one step.

In practice, the algorithm performs quite well. Table 1 compares the average number of iterations until convergence of my method and the BLP contraction in a Monte Carlo experiment based on actual state-level migration data.

Both methods solve the mixed-logit specification exactly; however, the BLP contraction only works approximately for the Pure Characteristics specification. It relies on decreasing a multiplier, γ, to the logit's variance to approximate the no logit assumption of the Pure Characteristics model.[20] In contrast, my algorithm can work directly with this specification. Also, for the BLP contraction, the number of iterations until convergence rapidly increases as γ goes to zero, and even at crude approximations, it has poor performance.[21] My approach performs equally well with either specification.

		Algorithm			
		Proposed method	BLP contraction		
Specification		—	$\gamma = 1$	$\gamma = 0.1$	$\gamma = 0.01$
mixed-logit	# iter	118	72	—	—
	av. error (%)	0	0	—	—
pure char	# iter	150	72	2496	4737
	av. error (%)	0	21.4	3.2	0.5

Table 1: Comparison between algorithms. γ multiplies the logit shock for the BLP contraction approximation. Error is the average deviation from target $\boldsymbol{\xi}^*$ to algorithm output $\hat{\boldsymbol{\xi}}$ divided by the utility's standard error, $\frac{1}{J}\sum_j \frac{|\hat{\xi}_j - \xi_j^*|}{\sigma_U}$.

[20] See PC for details on the approximation and its limitations.

[21] It is well-known that the modulus of contraction of the BLP approaches 1 as the logit variance goes to zero (relative to the total variance of the utility). This explains the increase in the number of iterations as γ goes to zero. Nonetheless, this approximation was the least computationally expensive in the experiments of Berry and Pakes (2007), which is the reason for the comparison.

Convergence In this section, I show that the algorithm is globally convergent to the *unique* inverse to the demand in the following sense:

Assumption 1. There is a unique $\boldsymbol{\xi}^*$ such that:

1. for any $\epsilon > 0$ and option j we have[22]

$$\sigma_j(\boldsymbol{\xi}^* - \epsilon \mathbb{1}\{j\}) \leq s_j \leq \sigma_j(\boldsymbol{\xi}^* + \epsilon \mathbb{1}\{j\}), \tag{2}$$

2. and at least one of the inequalities is strict.

Remark. Under continuity, the first condition is equivalent to the existence of a unique solution to $\boldsymbol{\sigma}(\boldsymbol{\xi}^*) = \boldsymbol{s}$. However, for the simulated demand $\boldsymbol{\sigma}$ in (9), there are jumps whenever a consumer gets indifferent between two products, which this generalized definition of inverse allows for. The second condition relates to zeros in demand and defines the target as the $\boldsymbol{\xi}^*$ in (7). That is why it uses the same notation as Section 4.2.
Also, conditions that guarantee the uniqueness for inverse demand are known to be mild in this discrete-choice setting (Berry, Gandhi, and Haile, 2013).

With a well-specified target, the first step to convergence is to show that the step size decreases to zero.

Lemma 2. *The algorithm decreases the step size infinitely many times. Equivalently, $\sigma_j(\boldsymbol{\xi}^r) - s_j$ changes sign infinitely often for every product j.*

This result is natural because j's utility would increase without bound if $\sigma_j(\boldsymbol{\xi}^r)$ were always (weakly) below s_j. However, this is not enough for convergence as the step size could go to zero before $\boldsymbol{\xi}^r$ is close to $\boldsymbol{\xi}^*$. One way to avoid this possibility is to decrease the step size only after the excess demand, $\sigma_j(\boldsymbol{\xi}^r) - s_j$, has changed its sign for every product. The following lemma formalizes why that is enough and how $\boldsymbol{\xi}^r$ approximates $\boldsymbol{\xi}^*$.

Lemma 3. *After excess demand has changed sign at least once, for any product j it holds that*

$$\sigma_j\left(\boldsymbol{\xi}^r - 2\frac{\mathbf{step}_r}{\rho}\mathbb{1}\{j\}\right) \leq s_j \leq \sigma_j\left(\boldsymbol{\xi}^r + 2\frac{\mathbf{step}_r}{\rho}\mathbb{1}\{j\}\right) \tag{3}$$

Intuitively, suppose product j's utility was increased by $\mathbf{step}_{r-1}$ in the last iteration and its excess demand went from negative to positive. Now, decrease j's utility by $2 \cdot \mathbf{step}_{r-1}$ and compare with the other products:

$$u_j + \xi_j^r - 2 \cdot \mathbf{step}_{r-1} \geq u_{j'} + \xi_{j'}^r \qquad \Rightarrow \qquad u_j + \underbrace{\xi_j^r - \mathbf{step}_{r-1}}_{=\xi_j^{r-1}} \geq u_{j'} + \underbrace{\xi_{j'}^r + \mathbf{step}_{r-1}}_{\geq \xi_{j'}^{r-1}},$$

[22] I denote by $\mathbb{1}\{j\}$ the vector with one in the j-th entry and zeros in all the other entries.

where we used that, at each iteration, $\xi_{j'}^{r-1}$ can move by at most $\mathbf{step}_{r-1}$ in either direction. Therefore, the demand for j after we subtract $2 \cdot \mathbf{step}_{r-1}$ is lower than the demand at iteration $r-1$, which, in turn, was lower than the market share:

$$\sigma_j\left(\boldsymbol{\xi}^r - 2 \cdot \mathbf{step}_{r-1} \mathbb{1}\{j\}\right) \leq \sigma_j(\boldsymbol{\xi}^{r-1}) \leq s_j.$$

Also, no matter if the step size was decreased at iteration r or not, we have

$$-2\frac{\mathbf{step}_r}{\rho} \leq -2 \cdot \mathbf{step}_{r-1},$$

which completes the argument if j's excess demand changed sign at iteration r. On the other hand, if it changed sign before the iteration r, the argument still holds since the demand overshoots the observed market share by at most one step between step-size updates.

Now we can prove convergence, and, since the step size is going to zero by Lemma 2 and $\boldsymbol{\xi}^r$ satisfies (3) by Lemma 3, the algorithm converges to $\boldsymbol{\xi}^*$ in (2).

Proposition 4. *Under Assumption 1, $\boldsymbol{\xi}^r$ converges to $\boldsymbol{\xi}^*$ as r goes to infinity.*

Rate of convergence It is natural that the algorithm would converge at a geometric rate. After all, the step size is multiplied by a scalar $\rho < 1$ every time it is updated. However, it could be the case that each step size update takes more and more iterations to happen. With the following assumption, we can rule out this possibility and convergence is in fact geometric.

Assumption 5. There is M such that for all r large enough and any $\mathcal{J} \subset \{1, \ldots, J\}$ we have[23]

$$\sum_{j \in \mathcal{J}} \sigma_j(\boldsymbol{\xi}^r - M \cdot \mathbf{step}_r \cdot \mathbb{1}\{\mathcal{J}\}) \leq \sum_{j \in \mathcal{J}} s_j \leq \sum_{j \in \mathcal{J}} \sigma_j(\boldsymbol{\xi}^r + M \cdot \mathbf{step}_r \cdot \mathbb{1}\{\mathcal{J}\}).$$

Remark. The assumption implies that if we increase the utility for the products in $\mathcal{J}$ by an amount $M \cdot \mathbf{step}_r$, then these products will be in excess demand. From Lemma 3, this holds for a single product, $\mathcal{J} = \{j\}$, and $M = 2/\rho$. So, essentially, the assumption says that there is an M such that Lemma 3 extends to a set of products.
Furthermore, this assumption is as a quantitative version of the *connected substitutes* assumption of Berry, Gandhi, and Haile (2013), as it asks that for any subset of products if their utility is increased by M (relative to the step size), then they steal enough demand from the other options for their total demand to be higher than the observed market shares. That is, they substitute strongly enough with products outside the group $\mathcal{J}$.

Proposition 6. *If assumptions 1 and 5 hold, then there is C and $m < 1$ such that*

$$\|\boldsymbol{\xi}^r - \boldsymbol{\xi}^*\| \leq C \cdot (m)^r.$$

[23] By extension of the previous notation, I denote by $\mathbb{1}\{\mathcal{J}\}$ the vector with ones in the j-th entries such that $j \in \mathcal{J}$ and zeros in all the other entries.

With a general method to recover the unobserved characteristics ξ^*, let's turn to how to compute the estimator for θ_0.

2.2 Quasi-Bayes Estimator

The objective function (10) is discontinuous both because of simulations and the median assumption. A standard approach for discontinuous objective functions is the Quasi-Bayes (or Laplace) estimation, transforming the hard optimization problem into a simpler sampling problem. More precisely, I sample from the distribution π with density

$$\pi(\boldsymbol{\theta}, S) \propto \exp(-M\hat{Q}(\boldsymbol{\theta} \mid S))\phi(S), \tag{4}$$

where $\boldsymbol{\theta}$ are the parameters and S is the set of simulations, which follows a known distribution ϕ. Intuitively, the samples concentrate on the minimizer of $\hat{Q}(\boldsymbol{\theta} \mid S)$ because (4) is large whenever $\hat{Q}(\boldsymbol{\theta} \mid S)$ is close to the minimum.

Notice that, instead of keeping one set of simulations fixed, both parameters $\boldsymbol{\theta}$ and simulations S are sampled. However, the estimator is the average over the $\boldsymbol{\theta}$ samples only:

$$\hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta}\pi(\boldsymbol{\theta}, S)\, \mathrm{d}\boldsymbol{\theta}\mathrm{d}S. \tag{5}$$

This procedure reduces the impact of simulation error. To get an intuition, imagine there were sets of simulations $S_1, \ldots, S_D$ and estimators such that

$$\hat{\theta}_d = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \hat{Q}(\boldsymbol{\theta} \mid S_d).$$

Each estimator $\hat{\theta}_d$ is consistent, but some of its variance comes from the simulation error, while the average $\frac{1}{D}\sum_d \hat{\theta}_d$ is also consistent but with lower variance. The Quasi-Bayes approach approximately targets this average without computing the intermediate estimators $\hat{\theta}_d$.

For example, Table 2 compares my estimator and the BLP estimator based on a Monte Carlo experiment. The BLP estimator fixes one set of simulations, while the Quasi-Bayes estimator follows the procedure outlined before. In the mixed-logit specification, my approach takes longer to run than the BLP estimator with the same number of simulated consumers; however, the gain in precision more than compensates for that. In the Pure Characteristics specification, the Quasi-Bayes estimator outperforms the BLP approximation discussed in Footnote 20.

2.3 Sequential Monte Carlo Sampler Implementation

This section gives a heuristic discussion on the Sequential Monte Carlo Sampler and provides the algorithm.

		Estimator		
		Quasi-Bayes	BLP	
Specification	# Simulations	10^3	10^3	10^4
mixed-logit	time (seconds)	72	21	212
	av. error (%)	8	37	7
pure char	time (seconds)	122	712	5430
	av. error (%)	12	139	45

Table 2: Comparison between estimators. For the BLP Pure Characteristics approximantion, the variance multiplier is set to $\gamma = 0.1$. Error is the average absolute percentage deviation from true own-price elasticity across 1000 replications.

Sequential Monte Carlo Sampler Heuristic Sequential Monte Carlo (SMC) methods started as an approach to estimate state-space models, but has evolved to many other classes of problems. One such class is sampling from complicated distributions like the distribution in (4). The reason this distribution is hard to sample is two-fold. First, it can be multimodal, and other sampling methods can get stuck in low probability areas. The standard solution with SMC is to *temper* the distribution. Second, it is high-dimensional; the dimension of the sample space is the sum of the number of parameters and the number of simulated agents.

The standard SMC approach is to start with a large number of *particles*, $\{(\boldsymbol{\theta}_p, S_p)\}_{p=1}^P$, that provide a crude approximation to our target distribution. Then, the SMC proceeds in two alternating steps. First, it *selects* particles that have higher probability (according to the target distribution). Then, it *mutates* the surviving particles to explore the sample space.

By *tempering* the distribution to slowly change from a simple to sample distribution to our complicated target (4), we can overcome the issues with multimodality. More precisely, I start with a distribution $\pi_0(\boldsymbol{\theta}, S) \propto \mu(\boldsymbol{\theta})\phi(S)$, where μ is an easy to sample distribution. For example, μ can be a normal distribution centered at zero with large variance. Then, at each step, the target distribution is given by

$$\pi_t(\theta, S) \propto \pi_0(\theta, S)^{1-\alpha_t} \pi(\theta, S)^{\alpha_t}.$$

Where the sequence $0 = \alpha_0 < \cdots < \alpha_T = 1$ provides the tempering. These α can be select adaptively to ensure good performance.

Following Chopin and Papaspiliopoulos (2020) adaptive tuning parameters.

find $a \in [0, 1 - \alpha_{t-1}]$ such that

$$\frac{\left(\sum_{p=1}^P \exp\left(-a\hat{Q}(\boldsymbol{\theta}_p \mid S_p)\right) \pi_0(\boldsymbol{\theta}_p)^{-a}\right)^2}{\sum_{p=1}^P \exp\left(-2a\hat{Q}(\boldsymbol{\theta}_p \mid S_p)\right) \pi_0(\boldsymbol{\theta}_p)^{-2a}} = \text{ESS}_{\min}$$

and set $\alpha_t \leftarrow \alpha_{t-1} + a$.

equidistant α_t can have poor performance. We want the π_t to be "equidistant" in some sense. So that each step does not change much the distribution. This way to select α_t does exactly that in

terms of chi-square pseudo-distance.

Markov Kernel = k iterations of a Metropolis-Hasting kernel with Gaussian random walk proposal with covariance matrix $2.38d^{-\frac{1}{2}}\hat{\Sigma}_{t-1}$. d is the dimension of the parameter space. $\hat{\Sigma}_{t-1}$ is the empirical covariance matrix of the weighted particles from the previous iteration.

To approximate this average, we sample from $\pi(\theta, S)$ using a Sequential Monte Carlo Sampler.

algorithm The algorithm pseudocode is given below

Algorithm: SMC Sampler with Simulations Pseudocode

Setup:

{1} Define a sequence $0 = \alpha_0 < \cdots < \alpha_T = 1$ and set $t \leftarrow 0$.

{2} Sample P particles from distribution $\pi_0(\theta, S)$ and set weights $w_{p,t}$ to 1.

For $t = 1$ **to** T

{1} For each particle $p = (\theta, S)$, evolve the target distribution

$$\pi_t(\theta, S) \propto \pi_0(\theta, S)^{1-\alpha_t} \pi(\theta, S)^{\alpha_t}.$$

{2} Update the weights

$$w_{p,t} \propto w_{p,t-1} \frac{\pi_t(\theta, S)}{\pi_{t-1}(\theta, S)}$$

{3} If weights are too concentrated, sample P particles from the current set of particles with probability given by the weights.

Set weights to 1.

{4} If particles were not resampled, update only θ with a Metropolis-Hasting step.

{5} If particles were resampled, update both θ and S.

θ with the Metropolis-Hasting step and S from the simulation distribution ϕ.

Remark. The SMC approach only requires that the kernel used to update the particles have $\pi_t(\theta, S)$ as *a* fixed point but not necessarily the unique fixed point. In contrast, MCMC methods rely on ergodic kernels that have $\pi_t(\theta, S)$ as the unique fixed point. This distinction allows the algorithm to only update θ as long as S are sampled from the correct distribution ϕ.

3 Asymptotics

This section formalizes the estimator asymptotic behavior. The main result is that the estimator is asymptotically equivalent to the minimizer of the objective function without simulation error. Then, the asymptotically normality follows from standard arguments.

The asymptotic behavior of the estimator follows from two observations. First, that the estimator $\hat{\boldsymbol{\theta}}$ defined in (5) is asymptotically equivalent to the minimizer of

$$\hat{Q}(\boldsymbol{\theta}) := -\frac{1}{M} \ln \left(\int \exp(-M\hat{Q}(\boldsymbol{\theta} \mid S))\phi(S) \, \mathrm{d}S \right), \tag{6}$$

which is differentiable under weak conditions even if $\hat{Q}(\boldsymbol{\theta} \mid S)$ is discontinuous. Because of this smoothness, the result follows if the high-level conditions in Chernozhukov and Hong (2003, Lemma 2) are met. Figure 1 compares the smooth (thick black line) $\hat{Q}(\boldsymbol{\theta})$ with the discontinuous $\hat{Q}(\boldsymbol{\theta} \mid S)$ for many simulations draws, S.

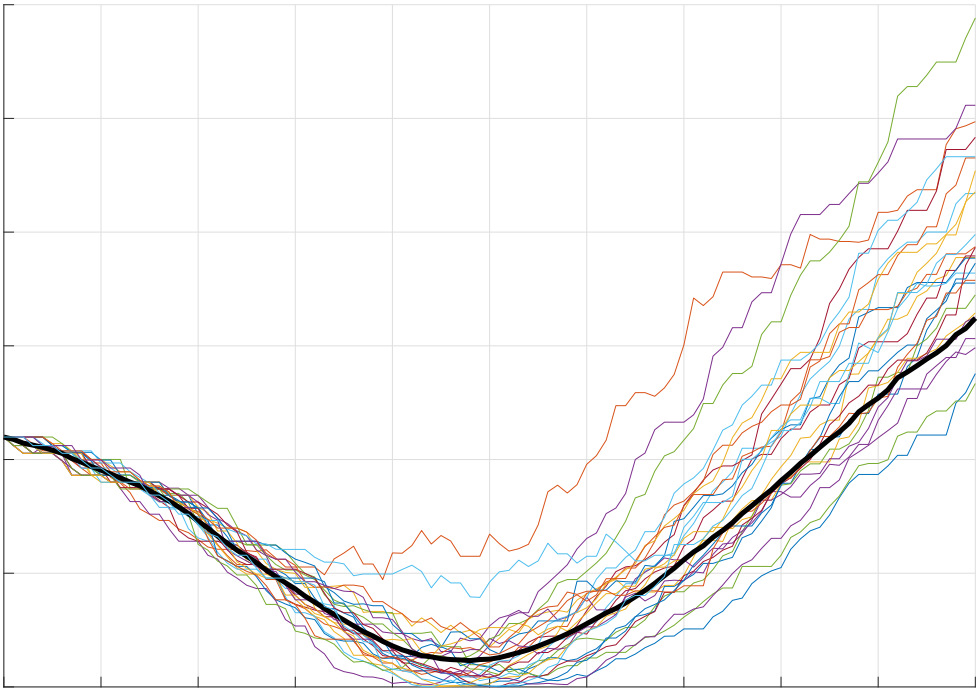

Figure 1: $\hat{Q}(\boldsymbol{\theta} \mid S)$ for different S and $\hat{Q}(\boldsymbol{\theta})$ (black)

Assumptions The asymptotic results rely on a series of standard assumptions. The assumptions closely follow the assumptions in Chernozhukov and Hong (2003).

Assumption 7 (Compactness)**.** The parameter space Θ is compact and θ_0 belongs to the interior of Θ.

Assumption 8 (Identification)**.** There is a function $Q(\boldsymbol{\theta})$, such that

1. Q is non-stochastic.
2. θ_0 is the unique minimizer of Q.
3. Q uniformly approximates $\hat{Q}$: For any $\epsilon > 0$

$$\limsup_M \mathbb{P}^* \left(\sup_{\theta \in \Theta} |\hat{Q}_M(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| > \epsilon \right) = 0$$

Assumption 9 (Quadratic expansion)**.** For some $\delta > 0$ and any θ with $|\theta - \theta_0| < \delta$,

1. Both $\hat{Q}_M(\theta)$ and $Q(\theta)$ are twice continuously differentiable.
2. There is $\Omega_M(\theta_0)$ bounded and uniformly strictly positive definite such that

$$M^{-\frac{1}{2}} \Omega_M(\theta_0)^{-\frac{1}{2}} \nabla_\theta \hat{Q}_M(\theta_0) \xrightarrow{\mathrm{d}} N(\mathbf{0}, \mathbf{I})$$

3. $J(\theta_0) = -\nabla_{\theta\theta'} Q(\theta_0)$ is strictly positive definite.
4. For any $\epsilon > 0$,

$$\limsup_M \mathbb{P}^* \left(\sup_{\theta:\, |\theta - \theta_0| < \delta} |\nabla_{\theta\theta'} \hat{Q}_M(\boldsymbol{\theta}) - \nabla_{\theta\theta'} Q(\boldsymbol{\theta})| > \epsilon \right) = 0$$

Remark. Usually asymptotic analysis of estimators behavior is decomposed in two steps (Newey and McFadden, 1994). First under compactness, identification, and a uniform law of large numbers the minimizer of the objective function is shown to converge to the minimizer of the limit of the objective function, θ_0. Then a quadratic expansion around θ_0 and a asymptotic normality assumption for $\nabla_\theta \hat{Q}_M(\theta_0)$ guarantees the asymptotic normality of the estimator.
In the context of Quasi-Bayes estimators, it is necessary to guarantee that the average

$$\hat{\theta} := \int \theta \pi(\theta, S) \, \mathrm{d}\theta \mathrm{d}S$$

behaves like the minimizer of $\hat{Q}_M(\theta)$. For this to be true, it is necessary to have a quadratic expansion around θ_0 with well-behaved approximation errors, so that the average and the mode (the minimizer of $\hat{Q}_M(\theta)$) are asymptotically equivalent (Chernozhukov and Hong, 2003). This almost implies asymptotic normality, so it is natural to treat both steps as one in the context of Quasi-Bayes estimators.

Proposition 10. *Under the regularity conditions previously stated, the average* $\hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta}\pi(\boldsymbol{\theta}, S)\, \mathrm{d}\boldsymbol{\theta}\mathrm{d}S$ *is asymptotically equivalent to the minimizer* $\hat{\boldsymbol{\theta}}_{\min}$ *of* (6)*. That is, their difference converges to zero faster than* $\frac{1}{\sqrt{M}}$*.*

$$\sqrt{M}\left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{\min}\right) = o_p(1).$$

Furthermore, $\hat{\theta}$ *is asymptotically normal.*

Proof of Proposition 10. The result follows from Lemma 2 and Theorem 1 from Chernozhukov and Hong (2003). □

Remark. The asymptotic normality in Assumption 9.2 follows under a standard Central Limit Theorem if $\Delta_\theta \hat{Q}_M(\theta_0)$ is asymptotically linear. Conditions for asymptotically linearity involve enough smoothness of the functional $\Delta_\theta \hat{Q}_M(\theta_0)$ with respect to the empirical distribution of the data and can be expressed in terms of its (Fréchet) derivative. E.g. see Ichimura and Newey (2022, Thm. 2) and their references.

Remark. The uniform convergence in Assumption 8.3 follows if an uniform law of large numbers hold for $\hat{Q}(\boldsymbol{\theta} \mid S)$ for almost every S. Conditions for ULLN are standard in the literature and can be expressed in terms of stochastic equicontinuity conditions, see Newey and McFadden (1994).

Under general conditions, $\hat{\boldsymbol{\theta}}_{\min}$ is asymptotically normal and equivalent to the estimator without simulation error. More precisely, Define $\hat{Q}^*(\boldsymbol{\theta})$ as the objective function (10) but computed without simulation error. If this estimator is well-defined and asymptotically normal, then it has the same asymptotic variance as $\hat{\boldsymbol{\theta}}$.

Corollary 11. *Suppose the number of simulations grows as fast as the sample size, then* $\hat{\boldsymbol{\theta}}_{\min}$ *is asymptotically equivalent to the minimizer of* $\hat{Q}^*(\boldsymbol{\theta})$*.*

Figure 2 compares $\hat{Q}(\boldsymbol{\theta})$ (in black) and $\hat{Q}^*(\boldsymbol{\theta})$ (lighter/red). In this example, it is clear that $\hat{Q}(\boldsymbol{\theta})$ is flatter than $\hat{Q}^*(\boldsymbol{\theta})$. In general, this property holds asymptotically: the limits of $\hat{Q}(\boldsymbol{\theta})$ and $\hat{Q}^*(\boldsymbol{\theta})$ are minimized at $\boldsymbol{\theta}_0$ but the limit of $\hat{Q}(\boldsymbol{\theta})$ has lower curvature.

Since $\hat{Q}^*(\boldsymbol{\theta})$ has no simulation error, the asymptotic behavior of its minimizer follows from standard Generalized Method of Moments results.

4 Why and How to Allow for Zero Market Shares

There are two reasons to prefer models that can generate zero market shares. First, whenever the demand for some products is too low, the conditional expectations used for estimation in BLP are unreliable. This problem affects any model which assumes strictly positive demand and does not go away even if we correct the observed market shares. Second, some markets have zero demand for some options, and allowing for zero market shares expands the possible applications. The following sections detail these two reasons and explain how the demand model can accommodate zero market shares.

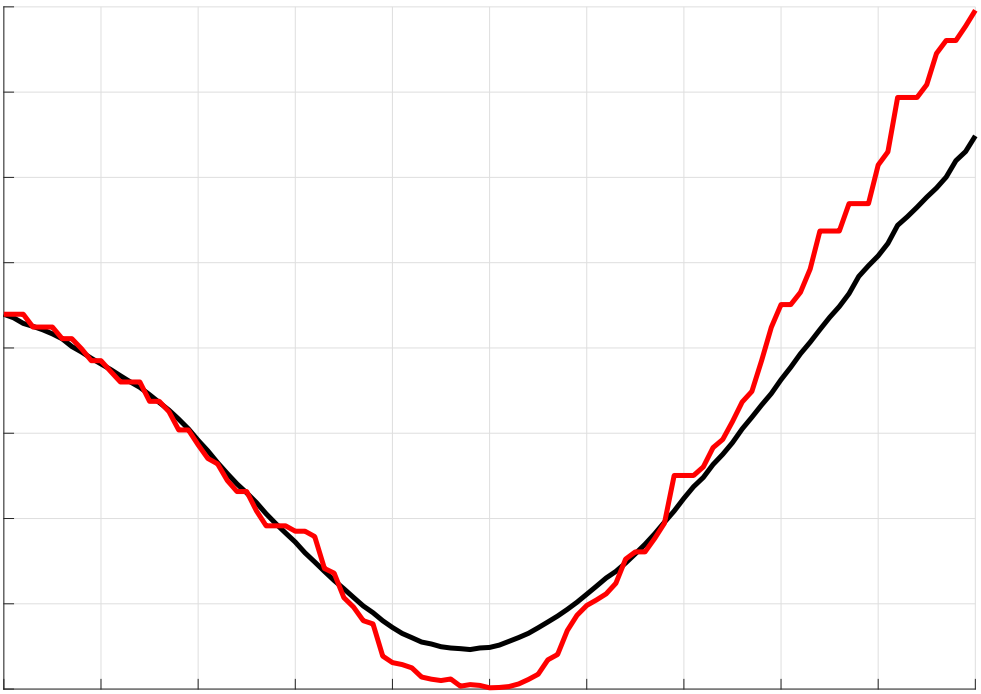

Figure 2: $\hat{Q}(\boldsymbol{\theta})$ (black) and $\hat{Q}^*(\boldsymbol{\theta})$ (lighter/red)

4.1 Close-to-zero market shares and strictly positive demand models

The BLP estimator relies on conditional expectations of the unobserved characteristics. However, these conditional expectations are very sensitive to sample uncertainty of market shares whenever market shares are close to zero. This sensitivity implies that parameters are estimated imprecisely, which, in turn, can affect counterfactual predictions. The issue is unavoidable whenever the model cannot generate zeros, which is equivalent to unbounded utilities.

Figure 3 illustrates the idea. the left panel shows all the possible market shares for three products, and the triangle boundary (in dark blue) represents the close-to-zero market shares for at least one product. In a simple logit model, the right panel shows the map from the market shares in the left to unobserved characteristics; more precisely, $\xi_j = \log(s_j) - \log(s_0)$. We see that, even though the dark blue area is small on the left panel, it maps to an unbounded region on the right. Moreover, as long as market shares are small enough, market share uncertainty is blown up to arbitrary uncertainty on the unobserved characteristics.

The same idea extends to a general model with strictly-positive market shares. The argument is that it is impossible to have a continuous function from the market shares (a compact set) to the unobserved characteristics (an unbounded set). In turn, this implies that there is no uniformly continuous function from the strictly-positive market shares to the unobserved characteristics.

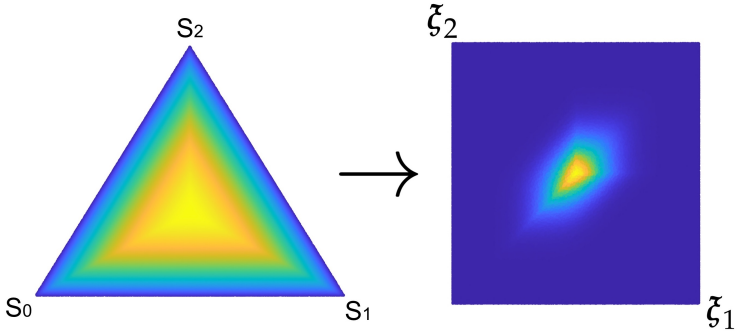

Figure 3: Map from Market Shares to Unobserved Characteristics.

Therefore, there are always two similar market shares that imply arbitrarily different unobserved characteristics.

To formalize the argument, given a ξ and parameters θ, define the choice probabilities π.

$$\pi_j = \text{Prob}\left(u_{ij} + \xi_j \geq \max_{k \neq j}\{u_{ik} + \xi_k\}\right).$$

These would represent the market shares if we observed infinitely many consumers. However, we only observe N consumers, and a multinomial distribution generates the observed market shares:

$$P(\boldsymbol{s} = \boldsymbol{a}) = \frac{N!}{(Na_0)! \cdots (Na_J)!} \pi_0^{Na_0} \cdots \pi_J^{Na_J}.$$

With these definitions, we can formalize the non-uniformity argument.

Proposition 12. *Suppose the demand model has unbounded utilities, which implies that for some products the choice probabilities are strictly positive. For any N, $\epsilon > 0$, and K there is a distribution for the unobserved characteristics, ξ, and a perturbation $\eta(\xi)$ such that the distributions induced on market shares by ξ and $\tilde{\xi} = \xi + \eta(\xi)$ are close,*

$$\|P - \tilde{P}\| := \sum_{\boldsymbol{a}} |P(\boldsymbol{s} = \boldsymbol{a}) - \widetilde{P}(\boldsymbol{s} = \boldsymbol{a})| < \epsilon,$$

but, for any instrument Z, the conditional expectations are well separated,

$$\left|\mathbb{E}[\xi \mid Z] - \mathbb{E}[\tilde{\xi} \mid Z]\right| = |\mathbb{E}[\eta(\xi) \mid Z]| > K.$$

In other words, the proposition says that it can be hard to know which one of the unobserved characteristics, ξ or $\tilde{\xi}$, generated the observed market share data. However, this is not a non-identification argument; if we fix the distributions of ξ and $\tilde{\xi}$ and let the number of consumers, N, go to infinity, we can identify the correct unobserved characteristics with probability 1. Instead, it is a non-uniformity result; no matter how many consumers, N, there are distributions of unobserved characteristics which are not separated with high probability, even if they have far apart conditional expectations.

Furthermore, as illustrated by the logit discussion, the result depends on arbitrarily close-to-zero choice probabilities and a model that restricts choice probabilities to be strictly positive (at least for some products). If we assume that choice probabilities are bounded away from zero, the result no longer holds. Therefore, the fundamental message from the proposition is that: If one believes that the correct model cannot generate zeros, then they need to restrict the parameter space to avoid close-to-zero probabilities. Naturally, how reasonable this restriction is will depend on the application.

Remark. The result does not depend on how market shares are inverted to unobserved characteristics. In particular, any correction to handle zero market shares, like adding a small number to the shares or ignoring the zero share products, will not solve the problem. In fact, notice that any inversion can only be close to either $\boldsymbol{\xi}$ or $\tilde{\boldsymbol{\xi}}$. Since either one could be the correct one, the inversion is guaranteed to have a large error.

4.2 Zero market shares and censored demand model

Another reason to consider models with zero demand is that some applications have zero market shares in the population. If utilities are bounded, we can explain the zeros as the option's cost is higher than its benefit for every consumer. However, this generates a new question: Which utility to assign to a product with zero market share? Since we allow for zeros in the population, any utility that is negative enough is consistent with what we observe. This section provides an answer by interpreting the question as a censoring problem.

If our model can generate zero demand, then there are many utilities consistent with zero market shares. Nonetheless, we can make progress by looking at the highest such utility. Furthermore, the demand model is mapped to an endogenously censored model. To simplify notation, the parameter $\boldsymbol{\theta}$ is left implicit until I discuss identifying assumptions.

Suppose our model can generate zeros and we have a solution to the demand inversion, say $\boldsymbol{\xi}$. Let ℓ be a product with zero market share. If we decrease ξ_ℓ and make ℓ less desirable, then, of course, demand for this product stays at zero. Also, since no one demanded ℓ to begin with, making it less desirable does not affect the demand for any other product. Thus, we can construct infinitely many solutions to the demand inversion whenever there are zero market shares. However, there is a natural largest solution: for each product ℓ with zero market share, define its largest unobserved characteristic:

$$\xi_\ell^* := \max \xi_\ell \tag{7}$$

such that $\boldsymbol{\xi}$ solves the demand inversion, $\boldsymbol{\sigma}(\boldsymbol{\xi}) = \boldsymbol{s}$.

For a product j with positive market share, ξ_j is the same for all solutions to $\boldsymbol{\sigma}(\boldsymbol{\xi}) = \boldsymbol{s}$. Therefore, we can define $\xi_j^* := \xi_j$ for any such solution, which implies the following characterization.

Proposition 13. *The set of solutions to the demand inversion $\sigma(\xi) = s$ is given by*

$$\left\{ \xi \quad \text{such that} \quad \begin{array}{ll} \xi_j \leq \xi_j^* & \text{if} \quad s_j = 0 \\ \xi_j = \xi_j^* & \text{if} \quad s_j > 0 \end{array} \right\}.$$

In particular, ξ^* is a solution.

Remark. In this formulation, we observe the "error term," ξ_j, whenever demand is strictly positive. Otherwise, it is censored, and we only observe an upper bound to it.

To further relate these solutions to a standard censored model, it is helpful to define the threshold at which demand becomes zero: For each j, define $\bar{\xi}_j := \max \xi_j$ such that ξ solves

$$\sigma_j(\xi) = 0 \qquad \text{and} \qquad \sigma_{j'}(\xi) = s_{j'}, \quad j' \neq j.$$

Corollary 14. *Given the true unobserved characteristics ξ, what can be observed is*

$$\xi_j^* = \max\{\xi_j, \bar{\xi}_j\}.$$

Furthermore, $\bar{\xi}_j \geq \xi_j$ if and only if $s_j = 0$.

Demand is zero at this threshold, but any larger unobserved characteristic would steal demand from other products. Naturally, this threshold is random and depends on the characteristics of all the products, observed and unobserved. Because of that, in the parlance of censored models, the censoring is endogenous (Khan and Tamer, 2009).

To complete the model, we need to impose an identifying assumption.

Conditional median assumption The demand $\sigma(\xi)$ implicitly depends on parameters θ, and so does the inverse $\xi = \sigma^{-1}(s)$. In this section, I assume there are M markets and make explicit that the unobserved characteristics are a function of parameters, $\xi_m(\theta)$.

As discussed in Section 1, the unobserved characteristic may be correlated with prices or other observed characteristics. If there were no censoring, the endogeneity would be addressed with instruments Z and a conditional *mean* assumption at the true parameter θ_0:

$$\mathbb{E}[\xi_{m,j}(\theta_0) \mid Z_{m,j}] = 0.$$

The identifying assumption would be that this equation only holds at θ_0.

In censored models, conditional mean assumption are well-known to have little or no identifying power: many parameters satisfy the conditional mean restriction. Intuitively, changes in the distribution of ξ below the threshold $\bar{\xi}$ affect the conditional mean but are not observable. The standard alternative is to assume a conditional *median* assumption.[24] Therefore, I assume that the

[24] See the handbook chapter by Powell (1994) for a discussion about identifying assumptions. Medians work with

(conditional) median of $\xi_{m,j}(\boldsymbol{\theta})$ is identically zero if an only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$:

$$\mathbb{P}\left(\xi_{m,j}(\theta_0) \leq 0 \mid Z_{m,j}\right) = 0.5. \tag{8}$$

Following Khan and Tamer (2009), this assumption implies two conditional moment inequalities:

1. $\mathbb{P}\left(\xi^*_{m,j}(\theta_0) \leq 0 \mid Z_{m,j}\right) \leq 0.5$,
2. $\mathbb{P}\left(\xi^*_{m,j}(\theta_0) \leq 0 \text{ or } s_{m,j} = 0 \mid Z_{m,j}\right) \geq 0.5$

where $\xi^* = \max\{\xi, \bar{\xi}\}$ as in Corollary 14. The first inequality holds because[25]

$$\xi^* = \max\{\xi, \bar{\xi}\} \leq 0 \quad \Rightarrow \quad \xi \leq 0.$$

And the second one holds because if $s > 0$ there is no censoring, thus $\xi = \xi^*$; therefore,

$$\xi \leq 0 \quad \Rightarrow \quad \xi^* \leq 0 \text{ or there is censoring} \quad \Leftrightarrow \quad \xi^* \leq 0 \text{ or } s = 0.$$

For estimation, it is necessary to replace the conditional moment, a function (of $Z_{m,j}$), with a finite number of conditional moments inequalities. Thus, let $\{\psi_q(Z_{m,j})\}$ be a finite set of weakly positive functions, then we have the unconditional moment inequalities

$$\mathbb{E}\left[\left(\mathbb{1}[\xi^*_{m,j}(\theta) \leq 0] - 0.5\right)\psi_q(Z_{m,j})\right] \leq 0 \quad \text{and}$$
$$\mathbb{E}\left[\left(\mathbb{1}[\xi^*_{m,j}(\theta) \leq 0 \text{ or } s_{m,j} = 0] - 0.5\right)\psi_q(Z_{m,j})\right] \geq 0.$$

It is essential that ψ_q are weakly positive to preserve the inequalities.

Finally, estimation is based on the sample analogs of these inequalities,[26]

$$\frac{1}{M}\sum_{m,j}\left(\mathbb{1}[\xi^*_{m,j}(\theta) \leq 0] - 0.5\right)\psi_q(Z_{m,j}) \leq 0 \quad \text{and}$$
$$\frac{1}{M}\sum_{m,j}\left(\mathbb{1}[\xi^*_{m,j}(\theta) \leq 0 \text{ or } s_{m,j} = 0] - 0.5\right)\psi_q(Z_{m,j}) \geq 0.$$

Accounting for simulation error Until now, it was implicitly assumed that the demand could be computed exactly given $\boldsymbol{\xi}$ and parameters $\boldsymbol{\theta}$. In reality, the random coefficients imply that the demand in (1) is an integral without a closed-form solution, which needs to be approximated with simulation draws (Pakes, 1986; Berry et al., 1995). This, in turn, makes the demand inversion, $\boldsymbol{\xi}^*_{m,j}(\boldsymbol{\theta} \mid S)$, also a function of the simulation draws, S.

censored models because they are less sensitive than means to extreme observations; however, they are more sensitive to observations near zero. In particular, the asymptotic variance of median-based estimator is usually inversely proportional to the density at zero of the error term.

[25] See Khan and Tamer (2009) for details.

[26] In practice, it is usual to normalize each moment inequality by dividing it by its standard deviation. For simplicity, I keep it implicit in this section.

More explictly, the simulations $S = \{\varepsilon_i, \nu_i\}_{i=1}^{N_s}$ are drawn from a known distribution ϕ such that

$$\varepsilon_i \sim G \qquad \text{and} \qquad \beta_i = f(\nu_i \mid \theta) \sim F(\theta).$$

That is, we can construct the randomness in (1) from the simulation draws S. For example, if β_i is distributed normally with mean θ_1 and standard deviation θ_2, we could take ν_i to be draws from a standard normal, and $f(\nu_i \mid \theta) = \theta_1 + \theta_2 \nu_i$.

Then, the demand in (1) is approximated by

$$\sigma_j(\xi \mid S) = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbb{1}[u_{ji} \geq u_{j'i} \text{ for all } j'], \tag{9}$$

which is used to compute the demand inversion, $\xi^*_{m,j}(\theta \mid S)$.[27]

Objective function Finally, the objective function in the parameters θ conditional on a particular set of simulations S is given by

$$\begin{aligned}\widehat{Q}(\theta \mid S) := \sum_q \max \left\{ \frac{1}{M} \sum_{m,j} \left(\mathbb{1}[\xi^*_{m,j}(\theta \mid S) \leq 0] - 0.5 \right) \psi_q(Z_{m,j}), 0 \right\} \\ - \min \left\{ \frac{1}{M} \sum_{m,j} \left(\mathbb{1}[\xi^*_{m,j}(\theta \mid S) \leq 0 \text{ or } s_{m,j} = 0] - 0.5 \right) \psi_q(Z_{m,j}), 0 \right\}\end{aligned} \tag{10}$$

This function is discontinuous for two reasons: The unobserved characteristics, $\xi^*(\theta \mid S)$, may be a discontinuous function of θ, and $\hat{Q}$ is a discontinuous function of ξ. In both cases, the reason for the lack of continuity is the indicator function $\mathbb{1}$. This complicates the optimization procedure and motivates the Quasi-Bayes approach, which replaces the complex optimization problem with a more straightforward sampling problem.

5 US Internal Migration Data

My empirical application uses the Opportunity Insights and the IRS Migration data. These data are publicly available and provide observations at the commuting-zone (CZ) and state levels on several economic and demographic statistics, including labor market conditions, housing costs, and income distribution. The Opportunity Insights data contain tax information and migration decisions from 1996 to 2012 aggregated by CZ on families with children born between 1980 and 1986. The data also provide estimates of the *causal* effect on children's outcomes of growing in different CZs.[28] In particular, I focus on the children's income percentile at age 26, which is the

[27] Since $\sigma(\xi \mid S)$ is discontinuous in ξ, the definition of inversion needs to be generalized. See the discussion after Assumption 1 in Section 2.1.

[28] Chetty and Hendren (2018b) restricts attention to the 718 CZs with population larger than 25,000 out of the total 741 CZs, and I do the same. This restriction excludes 0.36% of the population. See their paper for more detail about the data

preferred outcome measure of Chetty and Hendren (2018b). The IRS Migration data provide the average household income for movers between counties, which I aggregate by CZs.[29] These data provide extra moments for the estimation.

Zeros in Migration Migration data at fine granularity generally have many origin-destination pairs with no observations. For example, in my setting, families can decide to stay in their CZ or move to any 717 other ones, and of the $718 \times 718 = 515,524$ origin-destination pairs, only about 3% (16,253) have positive flow.[30] Figure 4 depicts the sparsity of the transition matrix, where the blue points represent positive flow between an origin-destination pair; The origin and destination are ordered according to their longitude. There is an apparent geographic concentration of migration, but there are still significant long moves, especially between the east and west coasts.

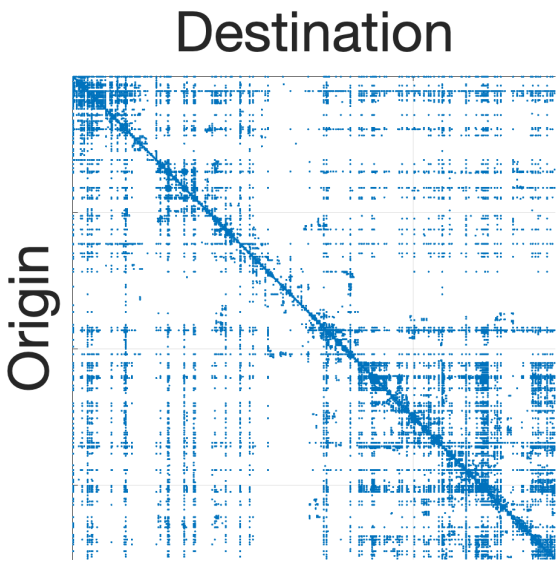

Figure 4: Commuting-zone pairs with any movers. Commuting zones ordered by their longitude; west-east correspond to left-right (Destination) and up-down (Origin).

Zero flows are also evidence of the lack of mobility in the data: Out of 26 million possible movers, only a little more than 5% moved across CZs in the period considered. Alos, Figure 5 shows the fraction of movers by CZ; in most CZs, less than 8% of the families move.

5.1 Empirical model

This section maps the model from Section 1 to a migration decision problem.

processing.

[29] This sample includes all the households, not only the families with children. To minimize the discrepancy, I adjust the income based on the 2010 American Community Survey (ACS) ratio between married-couples and household incomes by state.

[30] Because of confidentiality reasons, pairs with fewer than 25 movers are reported as 0.

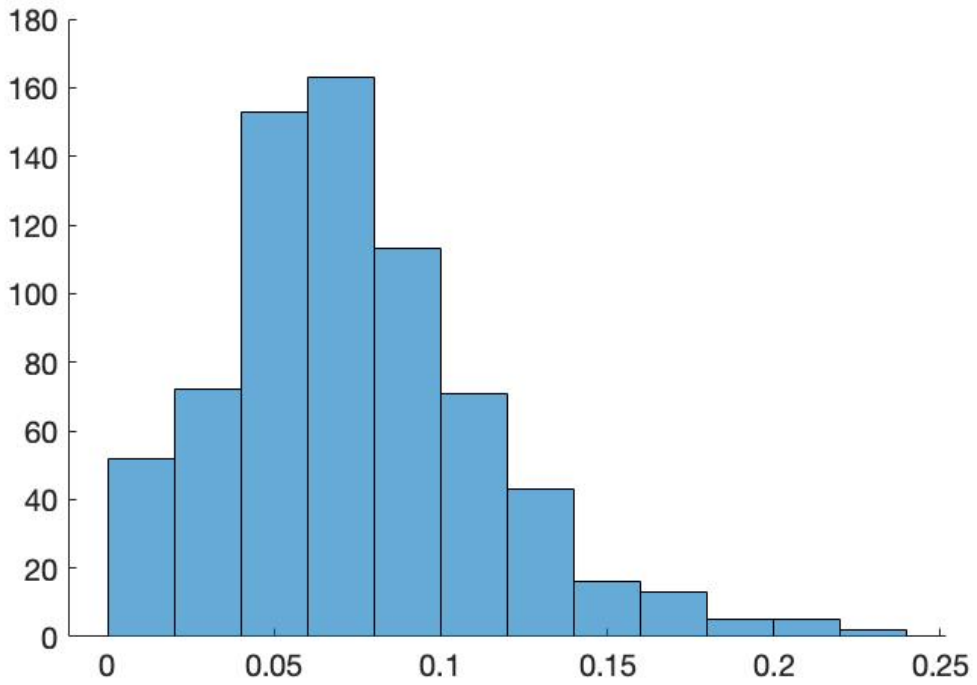

Figure 5: Histogram with fraction of movers by CZ.

For a family i from location o, the outside option is to stay in o and receive utility u_{ioo}, which I normalize to zero. On the other hand, if they decide to move to $d \neq o$, utility is given by

$$u_{ido} = \underbrace{-\mu_i}_{\{1\}} \underbrace{- \frac{\alpha}{\text{inc}_i} \log\left(\frac{\text{rent}_{id}}{\text{rent}_{io}}\right)}_{\{2\}} + \underbrace{\boldsymbol{X}_{do}\boldsymbol{\beta}_i}_{\{3\}} + \xi_{do} + \varepsilon_{ido}.$$

where

{1} μ_i represents potential psychological and monetary moving costs of leaving the origin CZ, including the fees from selling a house, and work and community disruptions. The distribution of μ_i is random and allowed to change with the family income, inc_i. More precisely, it has a different mean and variance for each income bin, and the randomness is represented by the ν_i^m variable:

$$\mu_i = \begin{cases} m_1 + \sigma_1 \nu_i^m & \text{if } \text{inc}_i \in \text{inc\_bin}_1 \\ \vdots & \vdots \\ m_7 + \sigma_7 \nu_i^m & \text{if } \text{inc}_i \in \text{inc\_bin}_7 \end{cases}.$$

{2} Income inc_i is drawn from the income distribution at the origin,[31] and the housing costs for

[31] I tried incorporating a change in income after migration by matching the education distribution and income by

each income is assumed to be proportional to the average 2-bedroom rent at each commuting zone;

$$\text{rent}_{id} = \text{rent}_d \times f(\text{inc}_i).$$

This specification allows the income fraction spent on housing costs to vary with income but simplifies the term $\{2\}$ because

$$\frac{\text{rent}_{id}}{\text{rent}_{io}} = \frac{\text{rent}_d}{\text{rent}_o}$$

{3} The variables in $\boldsymbol{X}_{do}$ represent the many controls available, including employment rate, fraction of white households, and distance between d and o. Most variables enter as the change between the value at the destination and origin. However, the change in some variables is split into the positive and negative parts.[32] This formulation allows for some sorting in taste across commuting zones. For example, if families from white commuting zones prefer to stay in mostly white areas and vice-versa for non-white families, then families on average will dislike both changes up and down in the fraction of white families.

The random coefficients $\boldsymbol{\beta}_i$ have the format

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \Lambda \boldsymbol{\nu}_i$$

where $\boldsymbol{\nu}_i$ represents the random draw in $\boldsymbol{\beta}_i$, and Λ is the covariance matrix, which is restricted to be diagonal. Also, some entries have zero variance to represent controls without random coefficients.

The sources of randomness ν_i^m, $\boldsymbol{\nu}_i$, and $\boldsymbol{\varepsilon}_i$ are independent and follow a triangular distribution from -1 to 1.[33] This specification guarantees that utility is bounded.

Moments The first set of moments used for estimation is based on the conditional median assumption in (8). Following Bayer, Ferreira, and McMillan (2007), the key assumption is that the housing costs, rent_d and rent_o, may be endogenous, but the other variables, $\boldsymbol{X}$ are valid instruments. These instruments enter the objective function $\hat{Q}(\boldsymbol{\theta} \mid S)$ through a normal kernel:

$$\psi_k(Z_{do}) = \frac{1}{716} \sum_{\ell \neq d,o} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{|X_{k,\ell o} - X_{k,do}|^2}{2\sigma_k^2} \right),$$

education at different states. The effect is mostly that families with high income get richer after moving and families with low income get poorer. Since these changes can be captured by the moving costs, the impact on the estimates and counterfactual experiments was small and I decided to use only the income at origin.

[32] That is, the maximum and minimum between the change and zero.

[33] That is, each random variable is independent and has the pdf

$$f(v) = (v+1)\mathbb{1}\{-1 \leq v \leq 0\} + (1-v)\mathbb{1}\{0 < v \leq 1\}$$

where k is each of the exogenous characteristics, and σ_k is the standard deviation of these characteristics over all pairs of CZs. These functions are positive, which preserves the inequalities in the objective function. Intuitively, if k represents fraction of college graduates in the CZ, then the value of $\psi_k(Z_{do})$ is higher if there are many CZs with a similar fraction of graduates as CZ d has, and vice-versa. That is, $\psi_k(Z_{do})$ is higher, when CZ d has many substitutes along the dimension k.[34]

This first set of moments is used to compute objective function (10) in Section 4.2. However, we can get more precise estimates if we add other sources of information. The one I consider are (micro) moments, which are a increasingly popular way to add information to BLP-type estimators following Petrin (2002). The extra moments are the correlation between families choices and their incomes. More precisely, I consider two set of moments: First, the average income of movers to each state from each CZ. Second, the share of movers from each state for each income bin. These moments are particularly important to pin down the distribution of moving costs by income.

If we denote by $\boldsymbol{m}(\boldsymbol{\theta} \mid S)$ the difference between these observed averages and the ones implied by the model, the objective function becomes

$$\hat{Q}(\boldsymbol{\theta} \mid S) + (\boldsymbol{m}(\boldsymbol{\theta} \mid S))^{\mathrm{T}} \widehat{W}(\boldsymbol{\theta} \mid S) (\boldsymbol{m}(\boldsymbol{\theta} \mid S)),$$

where $\hat{Q}(\boldsymbol{\theta} \mid S)$ is defined in (10) and $\widehat{W}(\boldsymbol{\theta} \mid S)$ is taken as a diagonal matrix with the inverse of the variance for each moment in $\boldsymbol{m}$.

6 Moving Costs and Counterfactual Housing Policy

This section discusses the moving costs estimates and their implications for voucher-based housing policies. Moving costs are large and highly variable, which suggests large benefits to targeted housing policies. I also utilize the counterfactuals to discuss how my estimates have different implications from the higher moving costs estimates found by the literature.

Because the moving costs μ_i are measured in utils, we first need to convert them into a more meaningful unit. One such unit is by how much housing cost would have to be discounted to induce a move to a particular destination. And, to isolate the moving cost, it is natural to take the destination to be somewhere identical to the origin. Thus, we can measure the moving costs as the percentage discount $\delta\%$ such that

$$\mu_i = \frac{\alpha}{\text{inc}_i} \log(\delta\%) \quad \Rightarrow \quad \delta\% = 1 - \exp\left(\frac{\mu_i \cdot \text{inc}_i}{\alpha}\right).$$

Figure 6 presents how moving costs measured in housing discount vary with income. The middle blue line represents the discount for a household with median moving cost at each income bracket. While the red lines above and below represent the maximum and minimum discounts,

[34] See Gandhi and Houde (2019) for a more formal discussion on using the difference between characteristics as instruments.

respectively. The dashed lines around are bootstrap 95% confidence intervals clustered by origin.[35] It is salient that higher discounts are needed to move richer families, which is intuitive for two compounding reasons: Higher income is tipically related to a lower price sensitivity, and richer families spend a smaller share of their income in housing. However, there is substantial variation between the maximum and minimum moving costs especially for richer families. That is, some people are easily convinced to move independent of income.

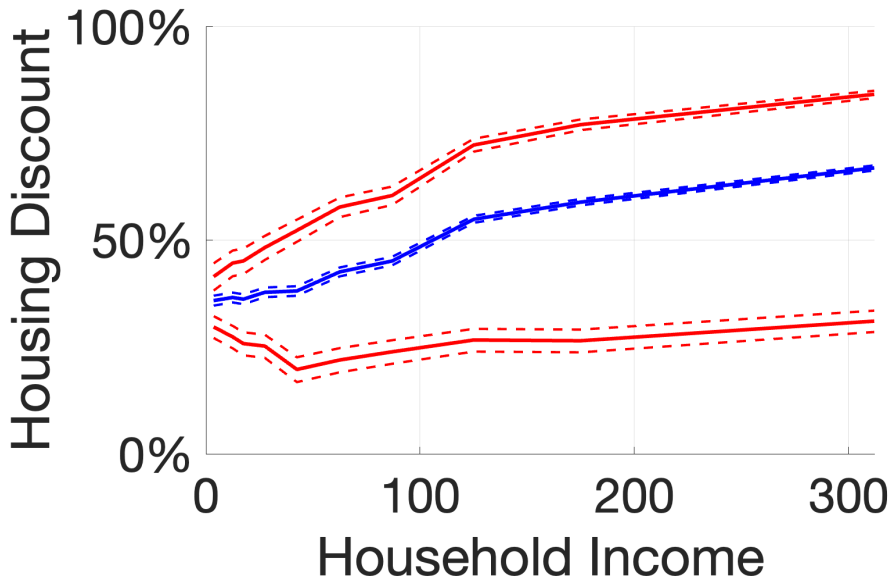

Figure 6: Moving costs measured in housing costs discount. Blue line represents the median moving costs at each income, while red lines represent the maximum and minimum moving costs. Dashed lines represent pointwise 95% confidence intervals.

[35] It is not clear what is the correct assumption about the data sampling process here. That is, what is the superpopulation that generated the observed CZ. However, in this context, clustering by origin is the same as clustering by "market" in a standard demand application, and, even if hard to interpret, the confidence intervals do give a sense of how sensitive are the estimates. See the discussion by Manski and Pepper (2018).

Counterfactual Housing Policy With the estimates of moving costs, we can consider counterfactual voucher-based housing policies. I utilize an approximation to the current voucher policy. Vouchers are available to families below the US poverty line and pay at most 30% of their income in housing; whatever is left is paid by the government.

I implement three counterfactuals, the first allows any family below the poverty line to use the voucher anywhere with no restrictions.[36] The second counterfactual restrict vouchers to be used only to move to high-benefit commuting zones. The set of high-benefit commuting zones is decided by selecting the commuting zones that most increase the average benefit to children until the budget expenditure is the same as the first counterfactual. Finally, the third counterfactual allows for vouchers to target both origin and destination. Since moving costs and preferences are highly heterogenous, this targeted approach can be useful. Similarly to the second counterfactual, the expenditure is the same as the first counterfactual.

Figure 7 presents the impact of each counterfactual on the average outcome for movers. Following Chetty and Hendren (2018b), I use the impact in the children's income percentile at age 26 as the relevant outcome. Each counterfactual is only partially idenfied because we only observe an upper bound to utility for commuting zones with no flow. To compute the Figure 7, I use a greedy search to find the highest and lowest possible value for the counterfactual. We see that the first counterfactual induces more families to move relative to the baseline, but has only a small impact in the outcome of the children. That is, families do use the voucher, but prefer to not move to a high-opportunity neighborhood. The more targeted counterfactuals solve this problem by restricting which commuting zones are eligible to use the vouchers in. We see that the targeted voucher are better at inducing families to move to high-opportunity commuting zones. However, vouchers that only target the destination do not generate as much migration as the unrestricted voucher. A voucher matching high-opportunity destinations to origins with low moving costs increases both the average outcome for movers and the number of movers.

To compare with previous results in the literature, I take the moving costs estimates of Kennan and Walker (2011). They find moving costs higher than $300,000 for the average mover. I inflate the moving costs for every income bin to match this number for the average income. Then, I recompute the unobserved characteristics that would generate the observed migration flows at this higher moving costs. In practice, these changes imply that migration is less explained by observable characteristics. In turn, families will be less price sensitive. We can see that in the counterfactuals in Figure 8. The voucher introduction induces fewer families to move because of the higher moving costs. And the targeted vouchers are much less effective at inducing families to change their destination. In this alternative model, the reason to move is explained mostly by unobserved factors, which implies that changes in the relative housing costs have a smaller impact in their decisions.

[36] In practice, vouchers can be transfered anywhere in the United States. However, there can be bureucratic challenges and prejudice against voucher-holders that I do not incorporate in the model. Also, even though the current policy in effect in the US is similar to this first counterfactual, it is not as comprehensive because of budget constraints; Only about 25% of eligible families get a housing voucher.

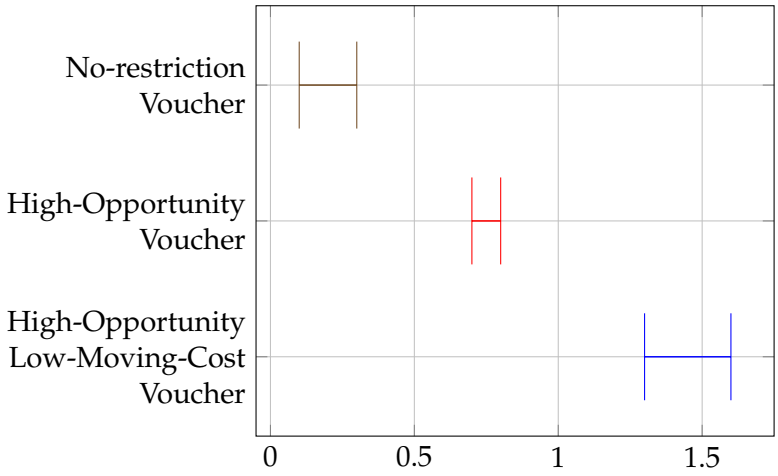

Figure 7: Average change in income percentile at age 26 for movers relative to baseline for different counterfactuals.

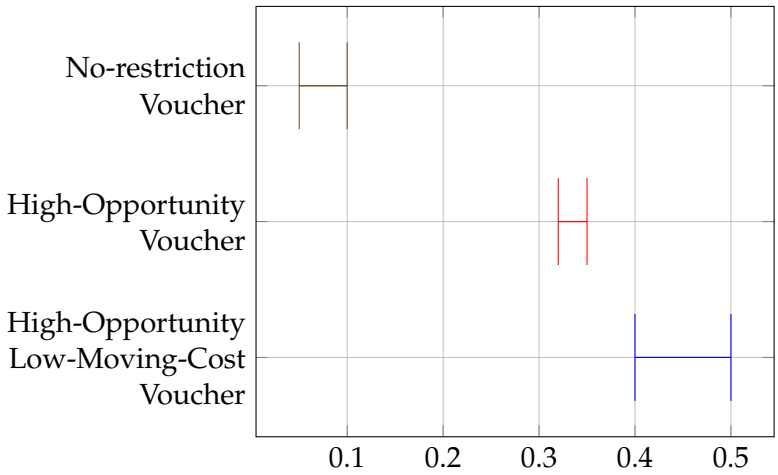

Figure 8: Counterfactuals with higher moving costs: Average change in income percentile at age 26 for movers relative to baseline for different counterfactuals.

7 Conclusion

Flexible demand estimation is essential to many policy questions. By generalizing the approach of Berry, Levinsohn, and Pakes (1995), I provide a simple estimator to a general discrete-choice demand model, including the Pure Characteristics demand model of Berry and Pakes (2007). The general model allows for zero market shares, which are common in many relevant applications like migration and international trade. Moreover, zero market shares lead naturally to an endogenously censored model.

As an example, this paper applies the estimator to an aggregate US internal migration data set. Because of the large number of zero migration flows, alternative methods would not handle these data. By allowing for a flexible moving cost specification, I find high and highly variable moving costs, implying a significant return to targeted voucher-based housing policies. However, my moving cost estimates are substantially lower than some previous estimates, and the difference has a large impact on the counterfactual evaluation.

The analysis of location choice is an important application of my method, but not the only one. Another potential application is trade data sets, which have large trading costs and many zeros (Helpman et al., 2008). Furthermore, the method's applications go beyond zero market shares. For example, allowing a flexible demand specification can be used to analyze product development incentives, which are sensitive to the logit assumption if products are similar in characteristics. Thus, estimating models without the logit assumption may be essential to analyze this question correctly.

The migration application has some simplifications that would be useful to address in future research. In particular, there are some institutional aspects about vouchers that I did not incorporate in my setting. Some of the salient ones are the bureaucratic work to transfer vouchers to different locations and the documented prejudice against voucher holders. An avenue to incorporate these considerations is to add microdata on voucher recipients in the spirit of Berry, Levinsohn, and Pakes (2004a). Another limitation is that I only consider monetary incentives, and there is evidence that more informational costs are a sizeable aspect of moving costs (Bergman et al., 2019; Fujiwara et al., 2020). Furthermore, my analysis considers only partial equilibria. It would be helpful to extend the analysis and incorporate how migration affects local housing and labor market conditions, such as Diamond (2016).

Another interesting future development is to analyze the impact of zeros in dynamic discrete-choice models. In a dynamic setting, the number of observations per state can be small, and zeros are even more common. Moreover, choice probabilities are usually close to zero for some options, such as in a simple entry and exit setting: firms outside the market rarely enter, and firms inside rarely leave. This extension is not straightforward because the map between (conditional) choice probabilities and utilities needs to be is trivial to compute. However, conditional on such a map for a model that allows for zero choice probabilities, the analysis of zeros as a censoring problem would still be relevant.

References

Daniel Ackerberg, C Lanier Benkard, Steven Berry, and Ariel Pakes. Econometric tools for analyzing market outcomes. *Handbook of econometrics*, 6:4171–4276, 2007. 2, 4

Daniel A. Ackerberg and Marc Rysman. Unobserved product differentiation in discrete-choice models: Estimating price elasticities and welfare effects. *The RAND Journal of Economics*, 36(4): 771–788, 2005. 2, 4

Susan Athey and Guido W Imbens. Discrete choice models with multiple unobserved choice characteristics. *International Economic Review*, 48(4):1159–1192, 2007. 6

Patrick Bajari and C Lanier Benkard. Discrete choice models as structural models of demand: Some economic implications of common approaches. *Working Paper*, 2003. 2

Patrick Bajari and Lanier Benkard. Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. *Journal of political economy*, 113(6): 1239–1276, 2005. 4

Patrick Bayer, Fernando Ferreira, and Robert McMillan. A unified framework for measuring preferences for schools and neighborhoods. *Journal of political economy*, 115(4):588–638, 2007. 25

Patrick Bayer, Robert McMillan, Alvin Murphy, and Christopher Timmins. A dynamic model of demand for houses and neighborhoods. *Econometrica*, 84(3):893–942, 2016. 3, 5

Peter Bergman, Raj Chetty, Stefanie DeLuca, Nathaniel Hendren, Lawrence F Katz, and Christopher Palmer. Creating moves to opportunity: Experimental evidence on barriers to neighborhood choice. *NBER Working Paper Series*, 2019. 30

Steven Berry and Philip Haile. Identification in differentiated products markets. *Annual review of Economics*, 8:27–52, 2016. 2, 4

Steven Berry and Ariel Pakes. The pure characteristics demand model. *International Economic Review*, 48(4):1193–1225, 2007. 1, 2, 4, 5, 8, 30

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, pages 841–890, 1995. 1, 2, 3, 4, 5, 6, 8, 11, 12, 16, 17, 21, 26, 30

Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of political Economy*, 112(1): 68–105, 2004a. 30

Steven Berry, Oliver B. Linton, and Ariel Pakes. Limit theorems for estimating the parameters of differentiated product demand systems. *The Review of Economic Studies*, 71(3):613–654, 2004b. 2, 4

Steven Berry, Amit Gandhi, and Philip Haile. Connected substitutes and invertibility of demand. *Econometrica*, 81(5):2087–2111, 2013. 6, 9, 10

Steven T Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262, 1994. 3, 6

Steven T Berry and Philip A Haile. Foundations of demand estimation. *NBER Working Paper Series*, September 2021. 4

Odran Bonnet, Alfred Galichon, Yu-Wei Hsieh, Keith O'Hara, and Matt Shum. Yogurts choose consumers? estimation of random-utility models via two-sided matching. *The Review of Economic*

Studies, 89(6):3085–3114, 2022. 4, 37, 38

Victor Chernozhukov and Han Hong. An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, 2003. 3, 5, 14, 15, 16

Raj Chetty and Nathaniel Hendren. The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *The Quarterly Journal of Economics*, 133(3):1107–1162, 2018a. 5

Raj Chetty and Nathaniel Hendren. The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, 133(3):1163–1228, 2018b. 5, 22, 23, 28

Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020. 12

Laurits R Christensen, Dale W Jorgenson, and Lawrence J Lau. Transcendental logarithmic utility functions. *The American Economic Review*, 65(3):367–383, 1975. 2

Angus Deaton and John Muellbauer. An almost ideal demand system. *The American economic review*, 70(3):312–326, 1980. 2

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006. 3

Rebecca Diamond. The determinants and welfare implications of us workers' diverging location choices by skill: 1980-2000. *American Economic Review*, 106(3):479–524, 2016. 30

Jean-Pierre Dubé, Jeremy T Fox, and Che-Lin Su. Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. *Econometrica*, 80 (5):2231–2267, 2012. 4

Jean-Pierre Dubé, Ali Hortaçsu, and Joonhwi Joo. Random-coefficients logit demand estimation with zero-valued market shares. *Marketing Science*, 2021. 4

Joachim Freyberger. Asymptotic theory for differentiated products demand models with many markets. *Journal of Econometrics*, 185(1):162–181, 2015. 3, 5

Thomas Fujiwara, Eduardo Morales, and Charly Porcher. A revealed-preference approach to measuring information frictions in migration decisions. *Working Paper*, 2020. 30

Amit Gandhi and Jean-François Houde. Measuring substitution patterns in differentiated-products industries. *NBER Working Paper Series*, 2019. 26

Amit Gandhi and Aviv Nevo. Empirical models of demand and supply in differentiated products industries. *NBER Working Paper Series*, September 2021. 4

Amit Gandhi, Zhentong Lu, and Xiaoxia Shi. Estimating demand for differentiated products with zeroes in market share data. *Working paper, SSRN*, 2020. 4

Elhanan Helpman, Marc Melitz, and Yona Rubinstein. Estimating trade flows: Trading partners and trading volumes. *The quarterly journal of economics*, 123(2):441–487, 2008. 2, 30

Han Hong, Huiyu Li, and Jessie Li. Blp estimation using laplace transformation and overlapping simulation draws. *Journal of Econometrics*, 222(1):56–72, 2021. 5

Hidehiko Ichimura and Whitney K Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022. 16

Guido W Imbens and Tony Lancaster. Combining micro and macro data in microeconometric

models. *The Review of Economic Studies*, 61(4):655–680, 1994. 3

John Kennan and James R Walker. The effect of expected income on individual migration decisions. *Econometrica*, 79(1):211–251, 2011. 3, 5, 28

Shakeeb Khan and Elie Tamer. Inference on endogenously censored regression models using conditional moment inequalities. *Journal of Econometrics*, 152(2):104–119, 2009. 3, 5, 20, 21

Jinhyuk Lee and Kyoungwon Seo. A computationally fast estimator for random coefficients logit demand models using aggregate data. *The RAND Journal of Economics*, 46(1):86–102, 2015. 4

Jing Li. Compatibility and investment in the us electric vehicle market. *Working Paper*, 2019. 2, 4

Charles F Manski and John V Pepper. How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, 100(2): 232–244, 2018. 27

Daniel McFadden. The measurement of urban travel demand. *Journal of public economics*, 3(4): 303–328, 1974. 2, 3

Daniel McFadden. Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications*, 1981. 2, 3

Aviv Nevo. Empirical models of consumer behavior. *Annual Review of Economics*, 3(1):51–75, 2011. 4

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994. 15, 16

Ariel Pakes. Patents as options: Some estimates of the value of holding european patent stocks. *Econometrica*, 54(4):755, Jul 01 1986. 21

Ariel Pakes, Jack R Porter, Mark Shepard, and Sophie Calder-Wang. Unobserved heterogeneity, state dependence, and health plan choices. *NBER Working Paper Series*, 2021. 3

Amil Petrin. Quantifying the benefits of new products: The case of the minivan. *Journal of political Economy*, 110(4):705–729, 2002. 2, 3, 26

James L Powell. Estimation of semiparametric models. *Handbook of econometrics*, 4:2443–2521, 1994. 20

Thomas W Quan and Kevin R Williams. Product variety, across-market demand heterogeneity, and the value of online retail. *The RAND Journal of Economics*, 49(4):877–913, 2018. 2, 4

Bernard Salanié and Frank A Wolak. Fast, "robust", and approximately correct: estimating mixed demand systems. Technical report, National Bureau of Economic Research, 2019. 4

H. Theil. The information approach to demand analysis. *Econometrica*, 33(1):67–87, 1965. 2

A Appendix

A.1 Additional results

Definition. For a given sequence of the algorithm $(\boldsymbol{\xi}_n)$, define

$$\mathcal{I}^+ := \{i \in \mathcal{I} \mid \forall\, M\, \exists\, m \text{ s.t. } \boldsymbol{\xi}_{i,n} > M\, \forall\, n > m\}$$
$$\mathcal{I}^- := \{i \in \mathcal{I} \mid \forall\, M\, \exists\, m \text{ s.t. } \boldsymbol{\xi}_{i,n} < -M\, \forall\, n > m\}$$

and

$$n^{\geq} := \sup\{n \in \mathbb{N} \mid \sigma_0(\boldsymbol{\xi}_n) \geq s_0\}$$
$$n^{<} := \sup\{n \in \mathbb{N} \mid \sigma_0(\boldsymbol{\xi}_n) < s_0\}$$

Notice that $n^{\geq} + n^{<} = \infty$.

Assumption 15. For a given sequence $(\boldsymbol{\xi}_n)$, assume that if $\mathcal{I}^+ \neq \emptyset$, then

$$\lim \sum_{\mathcal{I}^+} \sigma_i(\boldsymbol{\xi}_n) = 1$$

and for all $i \in \mathcal{I}^-$

$$\lim\ \sigma_i(\boldsymbol{\xi}_n) = 0$$

With this, we have a few lemmas.

Lemma 16. *Let $(\boldsymbol{\xi}_n)$ be a sequence generate by the algorithm.*

$$\mathcal{I}^+ = \emptyset \qquad \textit{and} \qquad \mathcal{I}^- = \emptyset$$

Proof. We can go by contradiction. Suppose $\mathcal{I}^+ \neq \emptyset$, then, by assumption 15

$$\sigma_0(\boldsymbol{\xi}_n) < s_0$$

But this implies that $n^{\geq} < \infty$, which implies that $\boldsymbol{\xi}_n \leq \boldsymbol{\xi}_{n^{\geq}}$. In particular, for $i \in \mathcal{I}^+$ we have

$$\boldsymbol{\xi}_{i,n} \leq \boldsymbol{\xi}_{i,n^{\geq}} < \infty$$

Suppose $\mathcal{I}^- \neq \emptyset$, then for $i \in \mathcal{I}^-$ we have, by assumption 15

$$\limsup\ \sigma_i(\boldsymbol{\xi}_n) < s_i$$

Which implies that for large enough m we have that for all $n > m$ that

$$\boldsymbol{\xi}_{i,n} \geq \boldsymbol{\xi}_{i,m} > -\infty$$

□

The distinction between the outside good and the inside good is what creates this artificial division between $\boldsymbol{\xi}_{i,n}$ going to infinity or to minus infinity. We could do the same thing with equivalence classes of $\boldsymbol{\xi}$ and reduce the two arguments to one.

Corollary 17. *Let $(\boldsymbol{\xi}_n)$ be a sequence generate by the algorithm.*

$$n^{\geq} = \infty \qquad \textit{and} \qquad n^{<} = \infty$$

Proof. Suppose $n^{\geq} = \infty$ but $n^{<} < \infty$, then we have $\mathcal{G}^{+} \neq \emptyset$. Similarly, if $n^{\geq} < \infty$ but $n^{<} = \infty$, then $\mathcal{G}^{-} \neq \emptyset$. □

Now we go to the first step to understand how the algorithm behaves.

Definition. For a given sequence of the algorithm $(\boldsymbol{\xi}_n)$, define

$$\mathcal{G}^{<} := \{i \in \mathcal{G} \mid \exists\, m \text{ s.t. } \sigma_i(\boldsymbol{\xi}_n) < s_i \;\forall\, n > m\}$$

$$\mathcal{G}^{>} := \{i \in \mathcal{G} \mid \exists\, m \text{ s.t. } \sigma_i(\boldsymbol{\xi}_n) > s_i \;\forall\, n > m\}$$

Proof of Lemma 2. Suppose the step size is updated only finitely many times. This is equivalent to

$$\mathcal{G}^{<} \cup \mathcal{G}^{>} \neq \emptyset$$

We know from corollary 17 that $n^{\geq} = \infty$ and $n^{<} = \infty$, and this implies that

$$\mathcal{G}^{<} \subset \mathcal{G}^{+} \qquad \text{and} \qquad \mathcal{G}^{>} \subset \mathcal{G}^{-}$$

But from assumption 15 we have

$$\mathcal{G}^{<} \cup \mathcal{G}^{>} \subset \mathcal{G}^{+} \cup \mathcal{G}^{-} = \emptyset$$

□

Proof of Lemma 3. Without loss of generality, we can assume that every step increases the unobserved characteristic for options that have shares below their targets. Suppose an option j changed side from below to above its target at step m before step n and did not went below again between m and n. Since its unobserved characteristic was not updated after m, we have

$$\xi_{n,j} = \xi_{m,j} + \mathbf{step}_m.$$

At the same time, because it did not change side between m and n, we know the step sized could have been updated at most once, that is $\mathbf{step}_n \geq \mathbf{update} \cdot \mathbf{step}_m$.

Therefore,

$$\xi_{n,j} - \frac{\mathbf{step}_n}{\mathbf{update}} \leq \xi_{m,j}.$$

And, because we only increase the unobserved characteristics, $\xi_{n,k} \geq \xi_{m,k}$ for all k. It follows that

$$\sigma_j\left(\boldsymbol{\xi}_n - \frac{\mathbf{step}_n}{\mathbf{update}}\mathbb{1}\{j\}\right) \leq \sigma_j(\boldsymbol{\xi}_m) \leq s_j \leq \sigma_j(\boldsymbol{\xi}_n) \leq \sigma_j\left(\boldsymbol{\xi}_n + \frac{\mathbf{step}_n}{\mathbf{update}}\mathbb{1}\{j\}\right).$$

The same is true if j changed side from above to below. To see this, change the perspective and assume that every step decreases the unobserved characteristics. Then the same argument implies that

$$\xi_{n,j} + \frac{\mathbf{step}_n}{\mathbf{update}} \geq \xi_{m,j}.$$

And the result follows. □

Proof of Proposition 4. By Lemma 2, the step size converges to zero. Therefore, by lemma 3, for any $\epsilon > 0$

$$\limsup \sigma_j(\boldsymbol{\xi}_n - \epsilon\mathbb{1}\{j\}) \leq s_j \leq \liminf \sigma_j(\boldsymbol{\xi}_n + \epsilon\mathbb{1}\{j\})$$

Take a convergent subsequence of $\boldsymbol{\xi}_n$, call its limit $\tilde{\boldsymbol{\xi}}$. It follows that

$$\sigma_j(\tilde{\boldsymbol{\xi}} - 2\epsilon\mathbb{1}\{j\}) \leq s_j \leq \sigma_j(\tilde{\boldsymbol{\xi}} + 2\epsilon\mathbb{1}\{j\}).$$

Then, by assumption 1, we have $\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^*$. Since the sequence $\boldsymbol{\xi}_n$ is bounded and every convergent subsequence converges to $\boldsymbol{\xi}^*$, it follows that $\boldsymbol{\xi}_n$ converges to $\boldsymbol{\xi}^*$. □

Proof of Proposition 6. The idea is to show by induction in the number of products, J, that we can take $m = (\text{update})^{\frac{1}{M^J}}$, for the same M as in assumption 5. To do that, we first show that under assumption 5 it takes at most M^J steps to update the step size.

It is easy to see that if $J = 1$, then it would take at most M iterations for the step size to be update. For an n such that the step has been updated at least once, define $\mathcal{G}^{<}_{n,n+M^{J-1}}$ as the set of options that are below they share between n and $n + M^{J-1}$.

$\mathcal{G}^{<}_{n,n+M^{J-1}}$ can be at most a singleton. To see this, notice that any two options within $\mathcal{G}^{<}_{n,n+M^{J-1}}$ can be treated as one, since they always update together. Therefore, the sum of their shares changed sides by the induction step. Now repeat the process M times. If

$$\mathcal{G}^{<}_{n,n+M^J} = \bigcap_{k=1}^{M} \mathcal{G}^{<}_{n+(k-1)M^{J-1},n+kM^{J-1}}$$

is empty, we are done. If it is not, then the only option in all these sets was update at least M times more than the other options, and, by assumption 5, it implies that it changed side.

Now suppose it take at most $S = M^J$ steps to update the step size. Therefore for any n we have

$$\|\xi_n - \xi\| \leq \sum_{k=1}^{\infty} \|\xi_{n+(k-1)S} - \xi_{n+kS}\| \leq \text{step}_n \times S \sum_{k=1}^{\infty} (\text{update})^{k-1} = \frac{\text{step}_n \times S}{1 - \text{update}}.$$

But since for some constant C' we have $\text{step}_n \leq C'(\text{update})^{\frac{n}{S}}$ it follows that

$$\|\xi_n - \xi\| \leq \frac{C'S}{1 - \text{update}} (\text{update})^{\frac{n}{S}} = Cm^n,$$

for $C = \frac{C'S}{1-\text{update}}$ and $m = (\text{update})^{\frac{1}{S}}$. □

A.2 Example: Bonnet et al. (2022) Market Share Adjusting algorithm

Consider a market with 3 goods, $j \in \{0, 1, 2\}$. Utility for the outside good 0 is normalized to zero. For consumer i, utility for goods 1 and 2 is given by

$$u_{i1} = \delta_1 + \sigma\nu_{i1} \qquad \text{and} \qquad u_{i2} = \delta_2 + \nu_{i2}$$

where $(\nu_{i1}, \nu_{i2}) \overset{iid}{\sim} N(0, I_2)$. I set $\delta_2 = -1$ and, given σ, set δ_1 so that the outside good has a share of 0.5. That is, δ_1 solves the equation

$$\mathbb{P}\left(0 \geq \delta_1 + \sigma\nu_{i1}, 0 \geq -1 + \nu_{i2}\right) = \Phi\left(-\frac{\delta_1}{\sigma}\right)\Phi(1) = 0.5$$

where Φ is the standard normal cdf.

σ	10^{-0}	10^{-2}	10^{-4}
Proposed algorithm	37	37	59
Bonnet et al. MSA	97	993	45125

Table 3: Iterations until convergence

If we take to the limit and set $\sigma = 0$, the MSA algorithm fails to converge. My proposed method still converges to $\delta = (0, 0, -1)$, as expected.

A.3 Example: DGP 1

Consider a market with $J \in \{50, 500\}$ goods. Utility for the outside good 0 is normalized to zero. The good j characteristics $x_j = (x_{j1}, x_{j2}, x_{j3})$ follow a multivariate normal distribution with

$$\mu = \begin{pmatrix} 1.5 \\ 1.5 \\ 1.5 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & -0.7 & 0.3 \\ \cdot & 1 & 0.3 \\ \cdot & \cdot & 1 \end{pmatrix}$$

Consumer i has taste shocks $\nu_i = (\nu_{i1}, \nu_{i2}, \nu_{i3})$ such that

$$\begin{pmatrix} \nu_{i1} \\ \nu_{i2} \\ \nu_{i3} \end{pmatrix} \overset{d}{\sim} \begin{pmatrix} 0.5 \\ 0.5 \\ 0.2 \end{pmatrix} + \begin{pmatrix} \text{Unif}_1 \\ \text{Unif}_2 \\ \text{Unif}_3 \end{pmatrix}$$

where Unif_k are iid uniform distributions on $[0, 1]$.

Utility is given by

$$U_{ij} = \sum_{k=1}^{3} x_{jk}\nu_{ik} + \delta_j$$

And δ is such that the outside good has about 50% of the market and the other goods share the remaining market approximately uniformly.[37]

J		50	500
Proposed algorithm	mean time (s)	0.05	1.95
	median time (s)	0.04	1.46
	max time (s)	0.15	5.74
Bonnet et al. MSA	mean time (s)	0.74	8.72
	median time (s)	0.07	1.72
	max time (s)	29.12	312.23

Table 4: Time until convergence (seconds), 100 MC replications, 10000 simulated consumers.

...

[37] More precisely, δ is set so that

$$S_0 \propto \frac{J}{2}, \quad S_1 \propto \text{Unif}_1^*, \ \ldots, \quad S_J \propto \text{Unif}_J^*$$

where Unif_j^* are iid uniform distributions on $[0, 1]$.