Estimating Preferences and Substitution Patterns from Second-Choice Data Alone*

Christopher Conlon† Julie Holland Mortimer‡ Paul Sarkis §

June 1, 2023

Abstract

We consider identification and estimation of a model of consumer choice where the main source of variation is in the set of products made available to consumers. Instead of relying on variation in the choice environment (prices, product characteristics) we require first-choice probabilities and a subset of (conditional) second-choice probabilities. We develop a semi-parametric low-rank approximation to the matrix of second-choice probabilities that is consistent with mixed logit models of demand but is defined in "product space" and does not require that product characteristics explain substitution patterns. In Monte Carlo experiments we show that our model can replicate a nested logit or random coefficients logit model. We apply our model to a single year of automobile data from Grieco et al. (2021) and show that we can fit substitution patterns with higher accuracy.

*[PRELIMINARY and INCOMPLETE] We thank Mark Stein and Mark Vend Company, and seminar participants at NYU Stern, University of Virginia, and the IIOC.

†New York University, Stern School of Business and NBER: cconlon@stern.nyu.edu

‡University of Virginia and NBER: juliemortimer@virginia.edu

§Boston College: sarkisp@bc.edu

1. Introduction

A large literature in empirical industrial organization (and economics more broadly) is concerned with estimating demand systems for differentiated products. These demand systems often have two key deliverables: the first is the own-price elasticity of demand, and the second is the pattern of substitution across products.

One key challenge in this literature is that the number of cross-product effects grows with the square of the number of products. A demand system with J products requires estimating J^2 cross-elasticities. This is true for common linear and log-linear demand systems, as well as the second-order flexible Almost Ideal Demand System of Deaton and Muellbauer (1980). This presents practical challenges both in terms of the required amount of variation in the data, and the need for additional instrumental variables.

The literature has primarily addressed this challenge by relying on parametric restrictions or treating products as bundles of characteristics, and formulating demand (and substitution patterns) in terms of these characteristics. For example, the logit model (McFadden, 1974) exhibits the independence of irrelevant alternatives (IIA) property, which restricts substitution to be proportional to observed market shares. The related nested logit (McFadden, 1978) exhibits proportional substitution within pre-determined groups and also across groups themselves. Finite (or continuous) mixtures of IIA logits can generate arbitrarily flexible substitution patterns under additional assumptions (McFadden and Train, 2000). In order to obtain that flexibility, the most common approach is to consider continuous mixtures of consumer preferences for observed characteristics and project substitution patterns onto product characteristics and their interaction with consumer demographics. This is the approach most commonly adopted in industrial organization (Berry et al., 1995, 1999; Nevo, 2000). While these models can be flexible, they have some drawbacks. The first is that their flexibility is generally limited by how well the underlying substitution patterns are explained by the observed product characteristics. The second is that even with modern tools, they can be computationally intensive and challenging to estimate, particularly as they become more flexible (Conlon and Gortmaker, 2020).

We propose a simple esimator, designed for an environment where a researcher observes aggre-

gate data on *market shares* (the probability that a consumer selects a particular product) and the matrix of *second choices* (the conditional probability that a consumer chooses a particular product if their first choice product is not available) from a *single market*. We provide a simple, easy to implement estimator based on a semiparametric mixture of logits to this first- and second-choice data. The key feature of our estimator is that we formulate the problem in *product space* rather than *characteristic space*, but avoid the problem of estimating J^2 elasticities by restricting the rank of the substitution matrix. Instead, we build on a result from Conlon and Mortimer (2021a) and write second-choice probabilities in terms of the first-choice probabilities for a finite number of types I. This reduces the number of parameters to $I \times J$, and restricts the rank of the matrix of substitutes to be no more than I.

Our approach is likely to be successful when underlying substitution patterns exhibit a *low rank structure* for which $I \ll J$. When this is the case, we can extend our estimator to the case where second-choice probabilities are only partially observed and redefine our problem as one of *matrix completion*. Our estimator also tends to produce choice probabilities that are *sparse*, so that not every individual type chooses each option with positive probability. This can be viewed as either a feature: it produces choice probabilities and substitution patterns that are more extreme than most mixed logits and is robust to not all individuals "considering" all products; or a bug: under full consideration it is no longer consistent with full support IID error terms.

Though our estimator recovers individual specific first-choice probabilities, we show it is straightforward to recover indirect utilities and tastes for product characteristics. This builds on a large literature of second-stage approaches in these kinds of models (Nevo, 2001; Bayer et al., 2007; Bayer and Timmins, 2007; Grieco et al., 2023). The most important parameters to recover in a second-stage are likely those governing (endogenous) prices. We show those parameters can be estimated using: (a) standard restrictions arising from instrumental variables; (b) auxiliary information on price-cost margins; or (c) auxiliary estimates of own-price elasticities. Together with the substitution patterns, these provide semiparametric estimates for most of the objects necessary to make counterfactual equilibrium predictions (mergers, taxes and subsidies, etc.). While academic researchers rarely have access to high quality data on price-cost margins, such information

is routinely provided to antitrust agencies as part of merger investigations.

What sets our estimator apart from the prior literature is that the data requirements are quite different. Rather than utilizing cross-market variation in the characteristics and assortment of products (see Berry and Haile (2014) on "aggregate data" for formal identification results), we rely on observing second-choice data for a single market (more in line with Berry and Haile (2022) on "micro data"). A valid concern is whether such second-choice data are readily available and what the quality of the data is likely to be. An obvious source of such data would be stated-preference second-choice data arising from consumer surveys. For example, the UK Competition and Markets Authority (CMA) frequently conducts surveys asking customers questions such as "if this supermarket were to close, where would you shop?" (Reynolds and Walters, 2008). Likewise, Berry et al. (2004a); Grieco et al. (2021) observe second-choice probabilities for a subset of automobile consumers and use these to inform mixed logit estimates of demand parameters. Another (inexpensive) alternative would be to design purpose-driven surveys such as Conlon and Gortmaker (2023) who survey consumers on second-choice soft-drink choices, or Magnolfi et al. (2022) on ready-to-eat breakfast cereal.[1] A long literature in marketing studies the design and implementation of *conjoints* for similar purposes Allenby et al. (2019). A (possibly more credible) alternative would be to rely on revealed preference approaches to estimate second choices. For example, Conlon and Mortimer (2013) show how to exploit exogenous variation in the timing of stock-out events, while Conlon et al. (2023) experimentally manipulate product assortment to estimate second-choice probabilities. An online alternative might be to experimentally manipulate the search results facing consumers in order to recover revealed-preference second-choice probabilities.

There are two interpretations of our approach: the first is that it provides a method to use data on second choices to estimate the parameters of a demand system without relying on product characteristics. The second is that it provides a way to rationalize surveys and other second-choice data in a way that is consistent with a discrete-choice demand framework. While academic researchers do not always need to make this connection, it is frequently relevant for antitrust

[1]The latter asks consumers to rank which products are more similar to one other rather than which product they prefer, though the design and implementation of the online survey could be easily adapted. The former survey was completed for less than $200 over the course of a week.

agencies. As an example, the CMA might be inclined to treat second-choice surveys as if they were the diversion ratios that enter the first-order conditions of merging parties and calculate UPP. Or, the DOJ or FCC may be inclined to use observational data on customer flows (sometimes called "win/loss data") in place of diversion ratios.[2] Likewise Farrell and Shapiro (2010) supposes that diversion ratios might be observed in the "normal course of business," implying that diversion ratios are data that firms might track internally, and could be requested by antitrust authorities as part of an investigation. Conlon and Mortimer (2021a) point out that diversion ratios measured from small price changes, quality changes, and second-choice data are related but not identical. Our framework provides a way to use second-choice data to recover the underlying primitives of the demand system, which can then be easily translated into the object(s) of interest, such as the appropriate UPP measure, or merger simulation.

In order to compare our semi-parametric method against other commonly used parametric approaches, we conduct two empirical exercises. First we estimate our semi-parametric model and mis-specified parametric models on data generated from a given model and compare the models' fit on out-of-sample second-choice probabilities. Using data from Conlon and Mortimer (2021b), we estimate a nested-logit model of demand with nests corresponding to product categories and an outside good in a separate nest, and a random-coefficients model of demand with independent normally distributed tastes β_i on the constant term and three observable product characteristics. We use these estimated models to generate two fake datasets using analytic formulas. We then compare our model's out-of-sample fit on predicted substitution patterns for rows of the second-choice matrix not used in estimation with three common parametric specifications: a simple logit, a random coefficients logit model with iid normal tastes over nests (RCN), and a random coefficients model with iid normal tastes over characteristics (RCC), estimated via Maximum Likelihood with and without additional information on second-choice probabilities.

When fitting the fake data generated by the nested-logit model, the parametric random-coefficients nested logit (RCN) model fits the data well. This model is almost correctly specified and differs only in whether the functional form follows a normal distribution or an extreme value distribution.

[2]See Qiu et al. (2021) for an in depth analysis of win/loss data.

The addition of second-choice moments improves the model's prediction error, resulting in nearly perfect fit. On the same data, the random-coefficients model based on characteristics (RCC) appears to be a poor fit while our semi-parametric model performs as well as, or better than, the RCN model estimated from observational data alone (and much better than the RCC model with or without additional moments). Similarly, when fitting fake data generated by the model based on characteristics, the RCC model fits well, the RCN model performs poorly, and our semi-parametric approach performs very well.

As a second exercise, we apply our estimator to a subset of data from Grieco et al. (2021), consisting of one year (2015) of aggregate market shares and the MaritzCX survey, which provides a matrix of second-choice probabilities.[3] The exercise demonstrates the estimator's effectiveness when dealing with a large number of products ($J = 318$) in a context for which substitution patterns may be influenced by endogenous characteristics, such as prices. To evaluate the estimator, we adopt the same cross-validation approach used in the previous exercise and compare our preferred specification to the parametric model estimated in the Grieco et al. (2021) study.

In this context, our semi-parametric estimator outperforms state-of-the-art parametric models that use significantly more data. With our preferred specification (selected by cross-validation), our model can rationalize both extremely high diversion ratios and more spread-out substitution patterns. For instance, our model can match the substitution between two cargo vans at around 67% and consistently keep the substitution between similar sedans below 10%. In contrast, logit-type parametric models are unlikely to achieve this, as shown by our comparison with the estimates from Grieco et al. (2021).

In related work (see Conlon et al. (2023)), we apply our estimator to experimental second-choice data for a subset of available products from Conlon and Mortimer (2021b).[4] That work demonstrates that even if one only observes second-choice substitution patterns from a subset of products, estimation of substitution patterns between remaining products can be accomplished as

[3]The MaritzCX survey includes recent car purchasers and is based on new vehicle registrations.

[4]The experiment involved multiple treatment arms for which one or two products were exogenously removed from vending machines in downtown Chicago office buildings. Thus, the resulting data consists of second-choice probabilities from only a few products relative to the whole set of products available, so that a significant portion of the matrix of second choices for all products is missing. This provides an illustration of the *low rank matrix completion* aspect of our approach.

a matrix completion exercise. This indicates that our estimator can be highly adaptable and easily estimated by antitrust authorities under time constraints and with limited data.

The remainder of the paper proceeds as follows: Section 2 reviews the results in Conlon and Mortimer (2021a) and describes the estimator. Section 3 provides the results of the Monte Carlo simulations. Section 4 provides the results of the estimation exercise on the survey second-choice data from the U.S. auto industry. Section 5 provides extensions and Section 6 concludes.

2. Model

2.1. Our Estimator

Throughout we use ***lower case bold*** *to denote the* $\dim(\mathcal{J})$ *vector of all choices (e.g.* $\mathbf{s_i}, \mathbf{p}$*) and* ***UPPER CASE BOLD*** *to denote the* $\dim(\mathcal{J}) \times \dim(\mathcal{J})$ *matrix of all choices (e.g.* $\mathbf{D}$*), and caligraphic font* $\mathcal{S}_j, \mathcal{D}_{j\to k}$ *to denote observed data. We will at times abuse notation and allow* $\mathcal{J}$ *to denote both a set and its cardinality* $\dim(\mathcal{J}) = \mathcal{J}$*.*

We begin with a researcher who has data on consumers who make a discrete choice among a set of products in $\mathcal{J}$ including an "outside" or "no purchase" option which we denote with $j = 0$. Furthermore, we assume that the researcher has access only to aggregate data on the first choices $\mathcal{S}_j$ and second choices $\mathcal{D}_{j\to k}$ for consumers facing a single choice environment (market). We denote the first and second-choice probabilities below:

$$\mathcal{S}_j = \mathbb{P}(\text{chooses } j \in \mathcal{J}) \tag{1}$$

$$\mathcal{D}_{j\to k} = \mathbb{P}(\text{chooses } k \in \mathcal{J} \setminus \{j\} \mid \text{chooses } j \in \mathcal{J}) \tag{2}$$

Later, we allow for a set of (j,k) tuples for which $\mathcal{D}_{j\to k}$ is observed, denoted P_Ω, and for which it is unobserved denoted $P_{\overline{\Omega}}$.

A common approach is to assume that these choice probabilities are generated by consumers i making discrete choices to maximize (random) indirect utilities: $u_{ij} = V_{ij} + \varepsilon_{ij}$. The most common specification assumes that ε_{ij} is IID and type I extreme value (Gumbel) distributed, so that the

probability that i chooses j (conditional on the vector of V_{ij}'s denoted $\mathbf{V_i}$) is given by:[5]

$$\mathbb{P}(u_{ij} > u_{ij'};\ \text{for all } j \neq j' \mid \mathbf{V_i}) = \frac{e^{V_{ij}}}{\sum_{j' \in \mathcal{J}} e^{V_{ij'}}} \equiv s_{ij}(\mathbf{V_i}). \tag{3}$$

The unconditional probabilities require integrating out over the distribution of V_{ij}, which is typically done by discretizing the distribution of heterogeneity with some probability weights corresponding to different vectors of $\mathbb{P}(\mathbf{V_i} = \mathbf{v_i}) = \pi_i$ so that $\pi_i \geq 0$ and $\sum_{i=1}^{I} \pi_i = 1$ (so that π constitutes a valid probability measure):

$$\mathbb{P}(u_{ij} > u_{ij'};\ \text{for all } j \neq j') = \int s_{ij}(\mathbf{V_i})\, f(\mathbf{V_i})\, \partial \mathbf{V_i} \approx \sum_{i=1}^{I} \pi_i\, s_{ij}(\mathbf{V_i}) \equiv s_j. \tag{4}$$

It is helpful to define some additional notation: let $\mathbf{s_i}$ be the $\dim(\mathcal{J})$ vector of type-specific (conditional) choice probabiltiies with entries s_{ij} defined in (3), let $\mathbf{s}$ be the vector of unconditional choice probabilities s_j defined in (4), and let $\mathbf{S}$ be a $\dim(\mathcal{J}) \times I$ matrix with column vectors $\mathbf{s_i}$. If π denotes an I vector with entries π_i then we can write $\mathbf{s} = \mathbf{S}\,\pi$.

In our previous work, Conlon and Mortimer (2021a), we show that for any mixed logit, the second-choice probabilities from (2) can be written in terms of the weights and the conditional and unconditional probabilities (π_i, s_{ij}, s_j) from (3) and (4):[6]

$$D_{j \to k} = \sum_{i=1}^{I} \pi_i \cdot \frac{s_{ik}}{1 - s_{ij}} \cdot \frac{s_{ij}}{s_j}. \tag{5}$$

It is convenient to interpret (5) as the (j,k)th entry in the second-choice matrix $\mathbf{D}$.

Our estimator simply matches the observed first and second choice data from (1) and (2) with the predicted versions from (4) and (5). We can accomplish this by minimizing the (potentially

[5] An alternative to the multinomial logit model is the multinomial probit model. This lets $\boldsymbol{\varepsilon}_i \sim N(0, \Sigma)$ where Σ is a $J \times J$ matrix. Like the log-linear model, this requires estimating on the order of $\frac{J \cdot (J-1)}{2}$ parameters, which makes it impractical for large J.

[6] For the case of "second choice data", Conlon and Mortimer (2021a) show it doesn't matter whether second choices are obtained by raising the price to the choke price, decreasing the quality such that no individuals purchase, or removing the product from the choice set. In all cases, we average over all individual diversion ratios $D_{jk,i} = \frac{s_{ik}}{1 - s_{ij}}$ and weight them in accordance with the fraction of j's sales they represent $w_i = \frac{s_{ij}}{s_j}$

weighted) least squares error (ℓ_2/Frobenius norm) so that:

$$\min_{(\mathbf{S},\pi)\geq 0} \|\mathcal{P}_\Omega(\mathcal{D}-\mathbf{D})\|_{\ell_2} + \lambda \|\mathcal{S}-\mathbf{S}\,\pi\|_{\ell_2} \text{ with } \pi\cdot\mathbf{1}_I = 1, \quad \mathbf{1}'_{\mathcal{J}}\mathbf{S} = \mathbf{1}_I. \tag{6}$$

The goal is to match the observed second choice probabilities in $\mathcal{D}$ and also the first-choice probabilities (market shares) $\mathcal{S}$ subject to some tuning parameter (Lagrange multiplier) λ. A typical challenge in computer science for problems like (6) is to avoid overfitting by restricting either the rank of $\mathbf{D}$ or its nuclear norm (sum of singular values). Below, we show that the rank of $\mathbf{D}$ is bounded by the number of types I, and we can restrict the rank of the matrix by directly limiting the number of types. More generally, we propose that the tuning parameters (λ, I) be chosen by cross-validation.

To see the rank restriction imposed by I we can simply re-write (5) in matrix form as:[7]

$$\begin{aligned}\mathbf{D} &= \left(\sum_{i=1}^{I} \pi_i \cdot \mathbf{s_i} \cdot \left[\frac{1}{(1-\mathbf{s_i})}\right]^T \cdot \text{diag}(\mathbf{s_i}/\mathbf{s})^{-1}\right)^T \\ &= \text{diag}(\mathbf{s})^{-1} \cdot \left(\sum_{i=1}^{I} \pi_i \cdot \left[\frac{\mathbf{s_i}}{(1-\mathbf{s_i})}\right] \cdot \mathbf{s_i}^T\right)\end{aligned} \tag{7}$$

This shows that we can write $\mathbf{D}$ as the sum of I rank-one matrices (outer product of vectors). The immediate result from (7) is that it shows us how to construct a low-rank $I \ll \mathcal{J}$ representation of a potentially large $\mathcal{J}\times\mathcal{J}$ matrix of substitution patterns. The plain IIA logit model corresponds to $I = 1$, in Appendix A.2 we show that the nested logit model limits the rank to be less than or equal to the number of nests.

We can also re-write the constraints in (6) as ℓ_1 constraints so that:

$$\min_{(\mathbf{S},\pi)\geq 0} \|\mathcal{P}_\Omega(\mathcal{D}-\mathbf{D})\|_{\ell_2} + \lambda \|\mathcal{S}-\mathbf{S}\,\pi\|_{\ell_2} \text{ with } \|\pi\|_{\ell_1} \leq 1, \quad \|\mathbf{s_i}\|_{\ell_1} \leq 1. \tag{8}$$

It should be clear that this represents a *non negative LASSO* problem (Wu et al., 2014). This

[7]Here $\text{diag}(\mathbf{s_i})$ is a diagonal matrix with entries s_{ij} and $\text{diag}(\mathbf{s})^{-1}$ is a diagonal matrix with entries $\frac{1}{s_j}$. The $\text{diag}(\mathbf{s})^{-1}$ term is serving to row-normalize the matrix so that $\sum_{j\neq k} D_{jk} = 1$ for each row. The diagonal entries $D_{jj,i} = \frac{s_{ij}}{1-s_{ij}}$, while not interpretable as a diversion ratios, are related to the willingness to pay (WTP) for product j (at least when s_{ij} isn't too large). See Conlon and Mortimer (2021a).

means we are likely to get *sparse solutions* to the program above so that $s_{ij} = 0$ for many (i, j). An economic interpretation might be that this is a product that type i really despises $V_{ij} \to -\infty$, or that type i is unaware of or does not consider j.[8] In a sense, we are agnostic about (and robust to) the particular reason for which $s_{ij} = 0$ (though it will largely be a result of the ℓ_1 penalty). Because second-choices depend on $\frac{s_{ik}}{1-s_{ij}}$, it is also worth noting that sparsity in $\mathbf{s_i}$ will tend to create sparsity in $\mathbf{D}$, particularly when the observed data $\mathcal{D}$ is sparse (or nearly sparse).[9] A common critique of logit and mixed logit demand systems is that all products are necessarily substitutable (at least a little) with one another.[10]

We should also note that the estimator in (6) or (8) is a *minimum distance* estimator, and the underlying asymptotic thought experiment would require either: (a) observing the complete matrix $\mathcal{D}$; or (b) taking the number of choices $\mathcal{J} \to \infty$ as in Berry et al. (2004b).

[Should we show that standard MD inference applies here?]

2.2. Second Stage

Our estimator recovers estimates of $\widehat{\mathbf{S}}$ or $\widehat{s}_{ij}$ and $\widehat{\pi}$, but these alone are not sufficient to calculate price-elasticities and consumer welfare. Following a long literature that separates the estimation of heterogeneous preferences from addressing the endogeneity of prices (Goolsbee and Petrin, 2004; Bayer et al., 2007; Bayer and Timmins, 2007; Grieco et al., 2023), we consider a second-stage in order to recover coefficients on prices.

In our first-step, we required the mixed logit only in the definition of second-choice probabilities in (5). In order to recover price sensitivities, we must lean harder on the assumption that we can write $u_{ij} = V_{ij} + \varepsilon_{ij}$ where ε_{ij} is an IID Type I extreme value error term. Following the logic in

[8] A large literature examines consideration sets in discrete choice models which may result from rational inattention (Matějka and McKay, 2015; Manzini and Mariotti, 2014), cognitive capacity, or random attention (Dardanoni et al., 2020; Masatlioglu et al., 2012). A significant challenge in this literature is to separate the impact of consumer preferences from heterogeneity in consideration sets. Typically, this requires either the presence of an exclusion restriction that shifts consideration independently of preference (Goeree, 2008), or exploiting differences in functional forms (Abaluck and Adams-Prassl, 2021). Other recent approaches can account for unobserved and limited consideration, but at the cost of potentially losing point identification (Barseghyan et al., 2021a,b).

[9] The expression in (5) implies $D_{jk,i} = 0$ IFF $s_{ij} = 0$ or $s_{ik} = 0$, which creates neither a theoretical nor practical problem.

[10] One way this has been addressed previously is the *pure characteristics model* of Berry and Pakes (2007).

Berry (1994) we can write:

$$V_{ij} - V_{i0} = \begin{cases} \ln \hat{s}_{ij} - \ln \hat{s}_{i0} & \text{if } \hat{s}_{ij} > 0, \\ \text{n.a.} & \text{if } \hat{s}_{ij} = 0. \end{cases} \tag{9}$$

This differs from the usual setup in two ways: (1) we recover i specific utility parameters V_{ij}; (2) in the case where we have sparse choice probabilities $s_{ij} = 0$ we need to be careful about interpretation. If the reason that $s_{ij} = 0$ is that the consumer i is unaware of j, then we can simply ignore the fact that $s_{ij} = 0$. If instead, $s_{ij} = 0$ because a consumer is highly price sensitive and never chooses a luxury car, then we have to model the selection rule more carefully.

Instrumental Variables

Most of the literature assumes that $V_{i0} = 0$ for each type i, though it isn't clear that is necessarily required here.[11] Our goal is to recover $\beta_i^p = \frac{\partial V_{ij}}{\partial p_j}$, and we present several approaches under different data environments. The most familiar approach would be to construct a second minimum distance estimator, where we stack up all (i, j):

$$\min_{\beta<0,\xi} \left\| z_j'(\ln \hat{s}_{ij} - \ln \hat{s}_{i0} - x_j\beta_i - p_j\beta_i^p + \xi_j) \right\| \tag{10}$$

This approach would require instruments for prices z_j as in Berry et al. (1995) or Berry and Haile (2014). It is unlikely is that our instruments will vary across i as well as across j. Though if we were willing to set $\xi_j = 0$, we could run I separate cross-product regressions (which is the primary source of variation here). Likewise if $\mathcal{J}$ is large, and we are not interested in the interpretation of non-price β_i coefficients then we can potentially estimate a partially linear model separately for each i:

$$\min_{\beta_i^p, f_i(\cdot)} \left\| z_j'(\ln \hat{s}_{ij} - \ln \hat{s}_{i0} - f_i(x_j) - p_j\beta_i^p) \right\|_{\ell_2}.$$

Observed Elasticities

[11] This would be sumsumed into the type-specific coefficient on the constant β_{i0} below.

An alternative approach would be to calibrate $\beta_i^p = \frac{\partial V_{ij}}{\partial p_j}$ using observed own-price elasticities. These might be estimated in any number of ways: from a (quasi)-experiment; from another study; or from a simpler (ie: plain logit) demand system on observational data.

For any mixed logit, the own-price elasticities follow a well-known format, which depends on objects we can estimate in the first-stage and β_i^p, and we can construct another minimum distance estimator:

$$\min_{\beta_i^p<0} \left\| \mathcal{E}_{jj} - \frac{p_j}{s_j} \sum_{i=1}^{I} \beta_i^p \cdot \widehat{\pi}_i \cdot \hat{s}_{ij} \cdot (1 - \hat{s}_{ij}) \right\|_{\ell_2}. \tag{11}$$

We need to observe at least as many elasticities as types I. Things simplify further if $\beta_i^p = \beta^p$, in which case a single elasticity (or the average elasticity) is sufficient to identify β^p. In our empirical example, we estimate β_i^p using the own-elasticities estimated in .

Observed Price-Cost Margins

Another possibility is that average price cost margins $(p_j - c_j)$ are often provided to antitrust authorities as part of merger review. Ideally these would be observed at the product level, but also possibly at the firm level. We need to construct Δ the $\mathcal{J} \times \mathcal{J}$ matrix of demand derivatives with entries $\frac{\partial q_j}{\partial p_k}$. We also need one additional piece of information, the ownership matrix $\mathcal{H}$ which typically has entries equal to 1 if products (j, k) have the same owner and zero otherwise. This allows us to write:

$$\mathbf{c} = \mathbf{p} - \left(\mathcal{H} \odot \left(\sum_{i=1}^{I} \pi_i \cdot \Delta_i(\beta_i^p) \right) \right)^{-1} \mathbf{s}, \tag{12}$$

$$\Delta_i(\beta_i^p) = \beta_i^p \left(-\mathbf{s_i}\, \mathbf{s_i}^T + \text{diag}[\mathbf{s_i}] \right).$$

This would enable us to match observed and predicted price-cost margins, or observed marginal costs to predicted marginal costs.

$$\min_{\beta_i^p<0} \left\| \mathcal{C}_{jj} - c_j(\beta_i^p, \mathcal{H}, \hat{S}, \widehat{\pi}) \right\|_{\ell_2}. \tag{13}$$

As with the elasticity we could match price-cost margins for a subset of products (or for the average

by firm, etc.).

2.3. Optional ℓ_2 Penalization

Our estimator is already using a non-negative LASSO ℓ_1 penalty term on s_{ij} and π as part of the "adding up" constraint that guarantees these are valid probability measures. There is a significant literature on joint ℓ_1 and ℓ_2 penalization (called *elastic net*) and see Heiss et al. (2022)

For all three cases in Section 2.2, we may want to add a penalty term to our minimum distance estimators of the form $\lambda_p \cdot \|\beta_i^p - \sum_i \pi_i \beta_i^p\|_{\ell_2}$ or $\lambda_p \cdot \|\beta_i^p\|_{\ell_2}$ to shrink the recovered β_i^p parameters towards the overall mean, or towards zero respectively. This would have the effect of penalizing "outliers" in β_i^p, which may be useful if we rely on cross-sectional variation with a limited number of observations, or to prevent extreme and imprecise β_i^p values for types where π_i is very small.

Likewise, even though we are already placing an ℓ_1 penalty term on s_{ij} such that $\sum_j s_{ij} = 1$, we may also want to consider an ℓ_2 penalty term of the form: $\|s_{ij} - \mathbf{s}\|_{\ell_2}$ or $\|s_{ij} - \mathcal{S}\|_{\ell_2}$. This would have the effect of shrinking the estimated s_{ij} back towards the plain logit "prior", and preventing the individual types from getting "too extreme". This might be useful as I becomes large and the variance of estimates increases, though it may be easier to estimate a model with fewer types. This may not be good idea if the true distribution of $\mathbf{s_i}$ includes extreme types.

Finally, we could consider an ℓ_2 penalty on the type-weights π_i so that $\sum_{i=1}^{I} \pi_i^2 \leq c_\pi$. This would effectively penalize the concentration/HHI of the types and push us towards estimating types with weights $\frac{1}{I}$ and away from types with very large or very small weights. This might be of practical use if models with large numbers of types "collapse", but it is otherwise not obviously useful.

2.4. Comparisons

Our semiparametric model is different from the fixed grid estimator of Fox et al. (2011); Heiss et al. (2022) who formulate the problem in characteristic space and construct a "prior" on mixture distribution for the coefficients $\beta_i^* \sim g(\beta)$. They draw a large number of points β_i (over 1000) from the prior, and then compute $s_{ij}^*(\beta_i)$ ahead of estimation on a fixed grid. They use a non-negative Lasso/L_1 penalty to impose sparsity such that a small number of types (20-50) receive positive

weight. They search over π_i in order approximate the true $f(\beta_i)$ via a finite mixture.

$$\min_{\pi} \sum_{j,t} \left(\mathcal{S}_{jt}(\mathbf{x_t}) - \sum_i \pi_i \cdot s^*_{ijt}(\beta^*_i, \mathbf{x_t}) \right)^2 \quad \text{subject to} \quad s^*_{ijt}(\beta^*_i, \mathbf{x_t}) = \frac{e^{\beta^*_i x_{jt}}}{1 + \sum_{j'} e^{\beta^*_i x_{j't}}}$$

$$0 \leq \pi_i \leq 1, \quad \sum_i \pi_i = 1 \tag{FKRB}$$

Both models estimate a constrained least squares problem for the aggregate shares, though theirs requires variation in $\mathbf{x_t}$ across markets, and ours requires second choices within a single-market. Another major difference is that we search over both the probabilities of types and the preferences of the types (s_{ij}, π_i) in *product space*, rather than considering a pre-specified grid of types in *characteristic space* and searching over π_i only.[12] The most important difference is that in their model generates a large number of types I for the fixed grid, while our approach is meant to keep I (and the rank of the second-choice matrix) as small as possible.

Another semiparametric model is Raval et al. (2017). Their semiparametric model for hospital demand groups consumers into $g \in G$ bins based on observable characteristics such as income, zip code, severity of diagnosis, and age. They have a tuning parameter on the minimum number of consumers per bin (which governs the number of bins). Within a group they assume all individuals have the same β_g but do not require anything about β_g and $\beta_{g'}$ other than that ξ_j is common across groups. They assume that preference follow a plain logit within each bin:

$$s_{g(i),j} = \frac{e^{\beta_g x_j + \xi_j}}{1 + \sum_{j'} e^{\beta_g x_{j'} + \xi_{j'}}}, \quad D_{kj,i} = \frac{s_{g(i),j}}{1 - s_{g(i),k}} \tag{RRT}$$

Raval et al. (2022) estimate diversion for hospitals using the diversion above and use observed second-choice probabilities (from natural disaster induced closures) to validate models of hospital demand, but do not use this variation to estimate the parameters of the model. The main difference between our models is that they are able to observe consumers and group them into demographic bins before estimation, whereas we consider aggregate data and must infer the mixing distributions.

In short, Fox et al. (2011) fix a grid of β^*_i (and hence s^*_{ij}) and estimate π_i using non-negative

[12] Arguably it would be straightforward to include product dummies in x_j and reformulate their approach in *product space* as well.

Lasso, while Raval et al. (2017) know the assignments of individuals to types $g(i)$, and type specific shares $s_{g(i),j}$, and estimate β_g for each group. We estimate both: (π_i, s_{ij}). However, our method requires observing at least some second choices, and our goal is to explain substitution with the smallest number of types possible.

Our approach in (6) is also related to a large class of problems in computer science known as "non-negative matrix factorization" which solve:

$$\min_{W \in \mathbb{R}_+^{(J+1)\times I}, H \in \mathbb{R}_+^{I\times(J+1)}} \left(\mathcal{D}_{jk} - (W\,H)_{jk}\right)^2. \tag{NMF}$$

This looks for a rank I approximation to $\mathcal{D}$ in terms of two compontent matrices X and W that each have non-negative elements. Absent the non-negativity constraints it is well known (Eckart–Young–Mirsky theorem) that the singular value decomposition gives the best rank I approximation to $\mathcal{D}$.

On one hand, our discrete choice setup places even more assumptions on (NMF) than simply non-negativity. Rather than factorize into two rank I matrices (W, H) we search for a single rank I matrix $\mathbf{S}$ and impose a particular relationship between the columns of $\mathbf{S}$ and the matrix $\mathbf{D}$. Our problem may be easier because $\mathcal{D}$ has a predictable structure and is amenable to a low-rank approximation, and discrete choice theory provides some structure on the component vectors $\mathbf{s_i}$, including that $\mathbf{s_j} = \sum_i \pi_i \cdot \mathbf{s_i}$.[13]

[Do we need a formal identification result?? – seems likely for $I \ll \mathcal{J}$ case.]

2.5. Computational Details

The minimum distance problem in (6) is non-convex, but easy to estimate relative to typical GMM or maximum likelihood estimators, and can generally be solved within a few seconds with a standard constrained L-BFGS optimization routine, or with an unconstrained Stochastic Gradient method (such as Adam (Kingma and Ba, 2017)). The objective is quadratic in parameters, and most of the

[13] One challenge of (NMF) is that it may not be identified (Ke et al., 2022) for the classic Latent Dirichlet Allocation (LDA) problem of Blei et al. (2003). It isn't clear whether or not the additional structure we impose is sufficient to restore identifcation.

constraints are quadratic or linear with the exception of the $\mathbf{D}$ matrix in (7) which are nonlinear and non-convex (though the derivatives are simple).

Typically, latent class logit models parameterize $V_{ij}(x_j) = \beta_i x_j + \xi_j$, and are estimated via full information maximum likelihood or via the EM algorithm as described in Greene and Hensher (2003). Estimation is often challenging for finite mixture models which estimate both π_i and β_i.[14] Our approach is both much easier to estimate, and is not necessarily restricted to interactions with observed covariates x_j. The main difference between our approach and the typical latent class logit is that we are estimating the model in *product space*. The summing up constraints are what enforce that the resulting model is consistent with mutually exclusive and exhaustive discrete choice.

In general, the problem scales with the number of products J and number of types I but not the number of "markets" or "observations". With I types there are $(J+1) \times I$ parameters.[15] If $I = 1$ then we are just fitting an IIA logit by least squares (with auxiliary moments for $\mathcal{D}$). If $I = 2$ then we have two latent classes like Berry and Jia (2010) use for business and leisure travelers. As we increase I, we increase the complexity of the model. In the limit we should be able to approximate any $\mathcal{D}$ as $I \to J$.[16] Our hope is that like many phenomenon, $\mathcal{D}$ may have a low-rank structure, or at least be amenable to low-rank approximation (Chen et al., 2021; Udell and Townsend, 2019).

3. Monte Carlo Simulations

In order to compare our semi-parametric method against other common parametric methods, we generate data from a given model and estimate both our semi-parametric model and mis-specified parametric models, then compare the models' fit on out-of-sample (on $\mathcal{P}_{\overline{\Omega}}$) second-choice probabilities $\left\|\mathcal{P}_{\overline{\Omega}}(\mathcal{D} - \mathbf{D}(\mathbf{S}, \pi))\right\|$ in both ℓ_2 (MSE) and ℓ_1 (MAD).

This exercise has two goals: verify whether our estimator can approximate common models, and identify how much data and model complexity is sufficient to achieve reasonable performance. In addition to our semi-parametric model, we estimate two mixed logit models with $V_{ij}(x_j) = \beta_i x_j + d_j$

[14] Maximum likelihood estimation of the latent class logit is notoriously difficult. The `gmnl R` package only offers FIML estimation for small problems and not EM estimation for large problems. Greene and Hensher (2003) propose an EM algorithm which appears to only be available in `NLOGIT/LIMDEP`.

[15] This ignores the parameters that are pinned down by the constraints in (6) such as s_j.

[16] We don't have a formal result here, though it seems like it would merely be a restatement of McFadden and Train (2000).

where $\beta_i \sim N(\mu, \Sigma)$ and Σ is diagonal. In the first, x_j is given by a set of dummies on product categories, which we label *random coefficients on nests* or (RCN). In the second x_j is given by the observed characteristics in the vending data (salt, sugar, and nut content), which we label *random coefficients on characteristics* (RCC).

3.1. Data Generating Process

We start with the data from Conlon and Mortimer (2021b) which included observational data on 66 vending machines in office buildings in downtown Chicago. For several weeks, the authors ran six experiments where the top-selling products in each category were removed.

We begin by estimating a nested logit model $V_{ijt} = \delta_j + \xi_t + \varepsilon_{ijt}(\rho)$ using the observational data for $\overline{J} = 45$ and $G = 6$ nests corresponding to each product category (Salty Snacks, Chocolate Candy, Non-Chocolate Candy, Cookies, Pastry, and Other) as well as an outside good in a separate nest. We calibrate $\rho = 0.25$ so that diversion to the outside good is approximately 30%.

We repeat this exercise on the same data where instead we calibrate a random coefficients model $V_{ijt} = \beta_i x_j + \delta_j + \xi_t$ with independent normally distributed tastes β_i on (constant, salt, sugar, and nut content).[17]

We then treat these parameter estimates as the "ground truth" θ_0, and use the estimated nested or random coefficients logit model to generate fake data. We construct two datasetes: one single market where all products are available and $T = 250$ markets of $M = 1{,}000$ consumers where only $J = 30$ products are available in any given market. For each dataset, we generate shares and second-choices following the analytic formulas given by the nested logit described in Appendix A.2.

1. Set the parameters of the model, product mean utilities and nonlinear parameters $\theta_0 = [\delta_j, \rho]$ or $\theta_0 = [\delta_j, \Sigma]$.

2. Generate the "true" shares $\mathcal{S}(\mathbf{x}, \theta_0)$ and second-choice probabilities $\mathcal{D}(\mathbf{x}, \theta_0)$ assuming full availability $\overline{J} = 45$ using the nested logit formulas from Appendix A.2.

3. For each market $t = 1, ..., T$, draw $J = 30$ products (plus an outside good) to be available in

[17] We calibrate the diagonal elements to be somewhat more heterogeneous than those estimated in Conlon and Mortimer (2021b) $\sigma_0 = 3.52, \sigma_{salt} = 1.00, \sigma_{sugar} = 1.8, \sigma_{nuts} = 0.38$

that market $\mathbf{x_t}$, making sure each nest g contains at least one product.

(a) Compute shares and multiply by market size $M \cdot s_j(\mathbf{x_t})$ to obtain quantities $q_j(\mathbf{x_t})$.

(b) Generate the $\mathcal{J} \times \mathcal{J}$ market specific second-choice matrix $\mathcal{D}(\mathbf{x_t})$.

3.2. Comparing semi-parametric and parametric models

To compare our model with common parametric specifications, we estimate two random coefficients models, RCC and RCN on the $T = 250$ markets from our fake data. We estimate the parametric models via Maximum Likelihood. In alternative specifications, we augment these data with additional moments matching selected rows from $\mathcal{D}_{j,\cdot}(\mathbf{x})$ to the second-choice probabilities predicted by the model for the full sample of $J = 45$ products $D_{j,\cdot}(\mathbf{x}, \theta)$.

Our semiparametric estimator from (6) uses less data than the parametric models. We don't use any of the observed variation in choice sets across markets t, nor any product characteristics. Instead, we use only the aggregate shares when all products are available $\mathcal{S}(\mathbf{x}, \theta_0)$ and a subset of $L \in \{6, 8, 10\}$ rows from the matrix of second-choice probabilities $\mathcal{D}_{j,\cdot}(\mathbf{x}, \theta_0)$.[18]

Because the models are parametrized differently, rather than compare $\widehat{\theta}$, we instead compare models based on out-of-sample predicted substitution patterns $\left\| P_{\overline{\Omega}} \left(\mathcal{D}(\mathbf{x}, \theta_0) - \mathbf{D}(\mathbf{x}, \widehat{\theta}) \right) \right\|$. We compare models based only on row not used in estimation $P_{\overline{\Omega}}$. We report the out-of-sample MSE and mean absolute deviation (MAD) in Figure 1. What is immediately obvious is that the RCN model fits the data quite well in terms of MAD and RMSE. Because the data generating process is a nested logit, this model is nearly correctly specified and differs only as to whether the functional form follows a normal distribution or a nested logit distribution. The additional second-choice moments improve the prediction error of the RCN model (in terms of RMSE) so that it fits nearly perfectly. The RCC model appears badly misspecified; the additional moments improve the RMSE signficantly but the MAD very little.

For $I \geq 4$ consumers, the semi-parametric model performs as well or better than the RCN model estimated from observational data alone (and much better than the RCC model with or without the additional moments) in terms of both MAD and RMSE. The number of "observed"

[18] When including $L = \{5, 6, 8, 10\}$ rows, we use the same ordering of products and try to select the top product from each category.

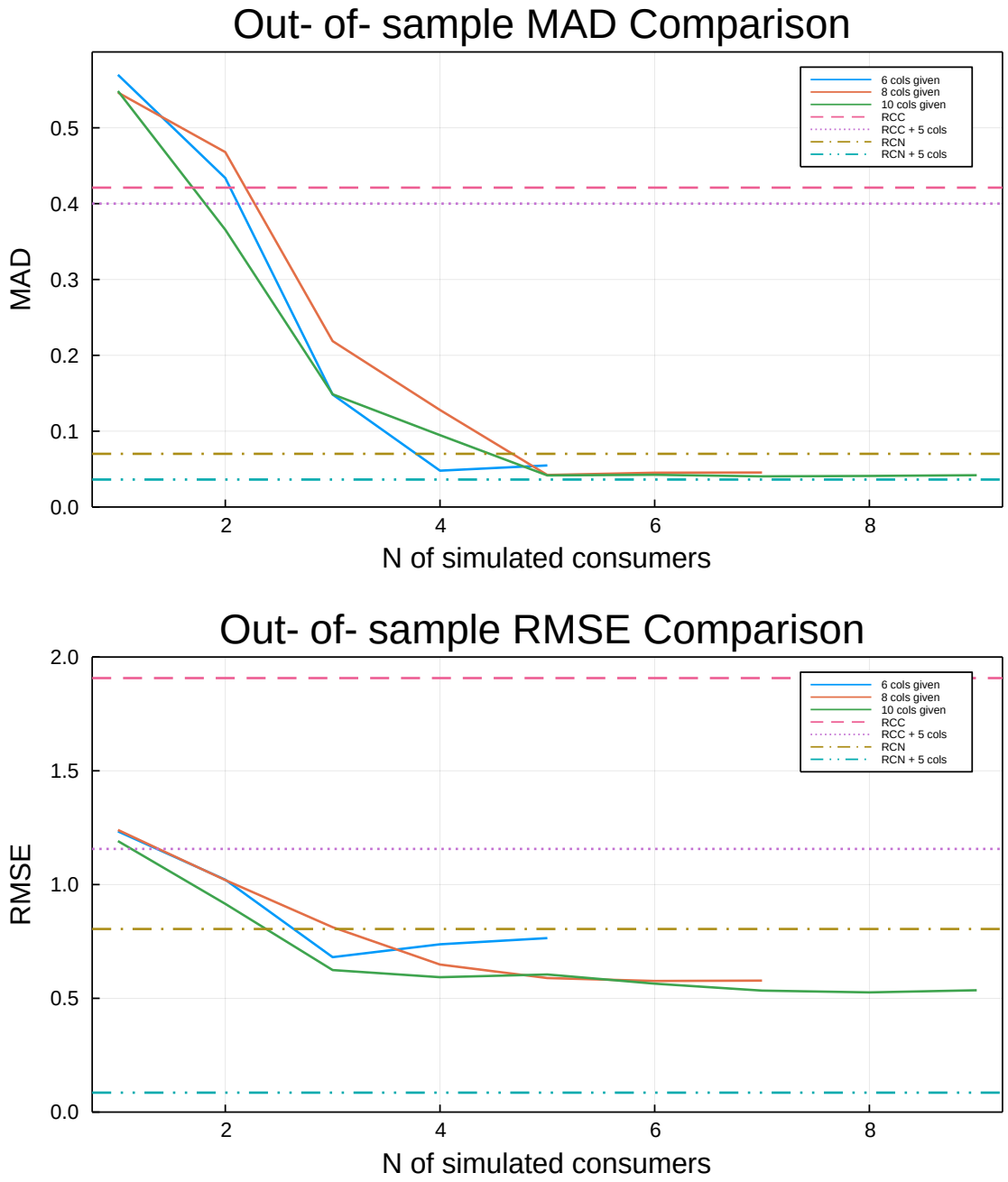
rows $L \in \{6, 8, 10\}$ does not have a signficant impact on the prediction error of the model, suggesting that $L = 6$ rows is "enough" to explain the substitution patterns in the remaining $J = 45 - L$ rows of the matrix.[19]

We get a similar result in Figure 2 when we use the RCC model as the data generating process. In this case the semi-parametric model substantially outperforms the RCN model. The addition of the second-choice moments does little to improve the predictive power of the RCN model.

We calculate the "true" rank of the $\mathcal{D}(\mathbf{x}, \theta_0)$ from the nested logit DGP to be $G = 6$ (the number of nests) and the approximate rank (nuclear norm) to be 2.33. For the RCC DGP the approximate rank 4.5. This helps explain why our semi-parametric approximation of rank $I = 4$ or $I = 5$ performs so well—the underlying matrix has a low-rank structure which makes it amenable to our approximation.

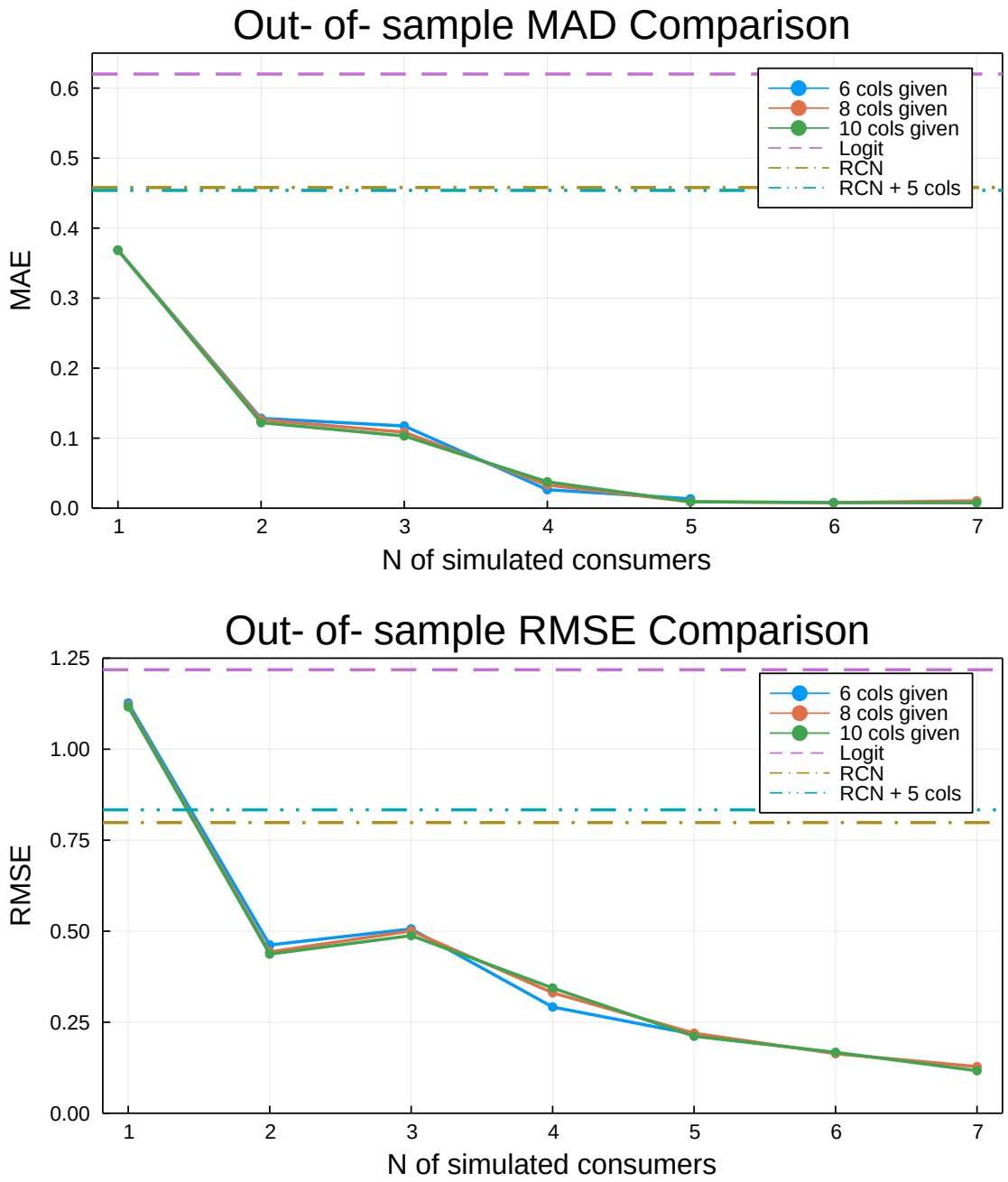
[19] Figure 1 does not report (or estimate) cases where the rank of the matrix I exceeds the number of observed rows L.

Figure 1: Out-of-sample fit for nested logit: Parametric and Semi-parametric Models



Notes: The data generating process is a nested logit model with six categories (rank $G = 6$, nuclear norm 2.33). Parametric models (Logit, RCC, and RCN) are fit on full sample $T = 250$ with or without additional moments based on (five) observed rows of second choice matrix $\mathcal{D}$.
Semi-parametric model (CMS) fit only on shares $\mathcal{S}(\mathbf{x}, \theta_0)$ and L rows of second-choice matrix $\mathcal{D}_{l,\cdot}(\mathbf{x}, \theta_0)$.
Out of sample fit is now on remaining $J - L$ rows where $L \in \{6, 8, 10\}$.

Figure 2: Out-of-sample fit for RCC: Parametric and Semi-parametric Models



Notes: The DGP is a random coefficients logit ($\sigma_0 = 3.52, \sigma_{salt} = 1.00, \sigma_{sugar} = 1.8, \sigma_{nuts} = 0.38$) with rank X, nuclear norm 4.5.

Parametric models (Logit, RCN) are fit on full sample $T = 250$ with or without additional moments based on (five) observed rows of second choice matrix $\mathcal{D}$.

Semi-parametric model (CMS) fit only on shares $\mathcal{S}(\mathbf{x}, \theta_0)$ and L rows of second-choice matrix $\mathcal{D}_{l,\cdot}(\mathbf{x}, \theta_0)$.

Out of sample fit is on remaining $J - L$ rows where $L \in \{6, 8, 10\}$.

4. U.S. Auto Industry Application

In this section, we estimate our model using data provided by Grieco et al. (2021) on the US automobile industry in 2015. These data include both market share (first choice) data $\mathcal{S}$, and second-choice data second-choice data $\mathcal{D}$ from the 2015 MaritzCX survey. MaritzCX is an automobile industry research and marketing firm that surveys recent car purchasers based on new vehicle registrations. The survey includes a question about the cars that the 53,328 respondents considered but did not purchase. We consider the first listed car as the purchaser's second choice to construct $\mathcal{D}$. For market shares, $\mathcal{S}$, we use the 2015 sales for all U.S. vehicles matched with models observed in the survey data and a market size of 20,765,000 as per GMY's methodology.

Our goal is to compare our predictions of $\mathbf{D}(\theta)$ to those in GMY. The GMY model incorporates elements of Berry et al. (1995), Petrin (2002), and Berry et al. (2004a), and represents the state-of-the-art in terms of BLP-style demand estimation, and uses data from many years (not only 2015). It includes demand, supply, and micro moments based on demographic information and the second-choice data presented above. Consumer utility is determined by a linear index of car characteristics (e.g. price, footprint, segment).[20] It also accounts for year-to-year variation in the utility of the outside good and average unobserved quality of new cars. Consumer heterogeneity is generated by interacting household characteristics and unobserved preferences with car attributes, allowing for different substitution patterns by demographics. GMY compute both first-choice probabilities (market shares) $\mathbf{s_t}(\theta)$ and second-choice shares conditional on the first choice $\mathbf{D_t}(\theta)$ using this consumer-choice model. The supply model assumes simultaneous multi-product pricing given a constant marginal cost function using a Bertrand-Nash assumption.

To properly compare our model with GMY, we must first understand how the latter model constructs micro-moments related to second-choice data. GMY measures correlations in car characteristics between observed first and second choices from survey data and matches them using the same correlations implied by their model (see Conlon and Gortmaker (2023)) for a description of

[20] A complete list of characteristics entering the utility function includes: price, footprint, horsepower, mpg, height, curbweight, number of trims, years since last redesign and dummies for van, SUV, truck, luxury, sport, electric, european brand, US brand, new release. Some of these characteristics are interacted with demographic variables such as income, age, rural status and family size.

these types of moments). This differs from our estimator, as GMY matches summary statistics instead of the matrix of second-choice probabilities. Our model may have an unfair advantage in that we are assessing performance on its ability to fit second-choice probabilities, while the GMY model is estimating parameters on a much larger set of data while trying to fit additional moments. Still, this serves as a useful benchmark, as we should think of GMY as the state-of-the-art and the best one can hope to do with a BLP-type model.

4.1. Cross-validation exercise

We utilize a cross-validation method to determine the rank of our matrix completion estimator. The 318 products (car models) are randomly split into 20 folds, and we estimate our model 20 times, leaving out one fold each time. To assess the out-of-sample fit on the omitted second-choices $\left\|P_{\overline{\Omega}}(\mathcal{D} - \mathbf{D}(\theta))\right\|$, we compare the Mean Absolute Deviation (MAD) ℓ_1 and Root Mean Squared Error (RMSE) ℓ_2 errors. In Figure 3, we present the performance of our model for various rank choices and compare it to the second-choice probabilities estimated with the state-of-the-art parametric model used in GMY.

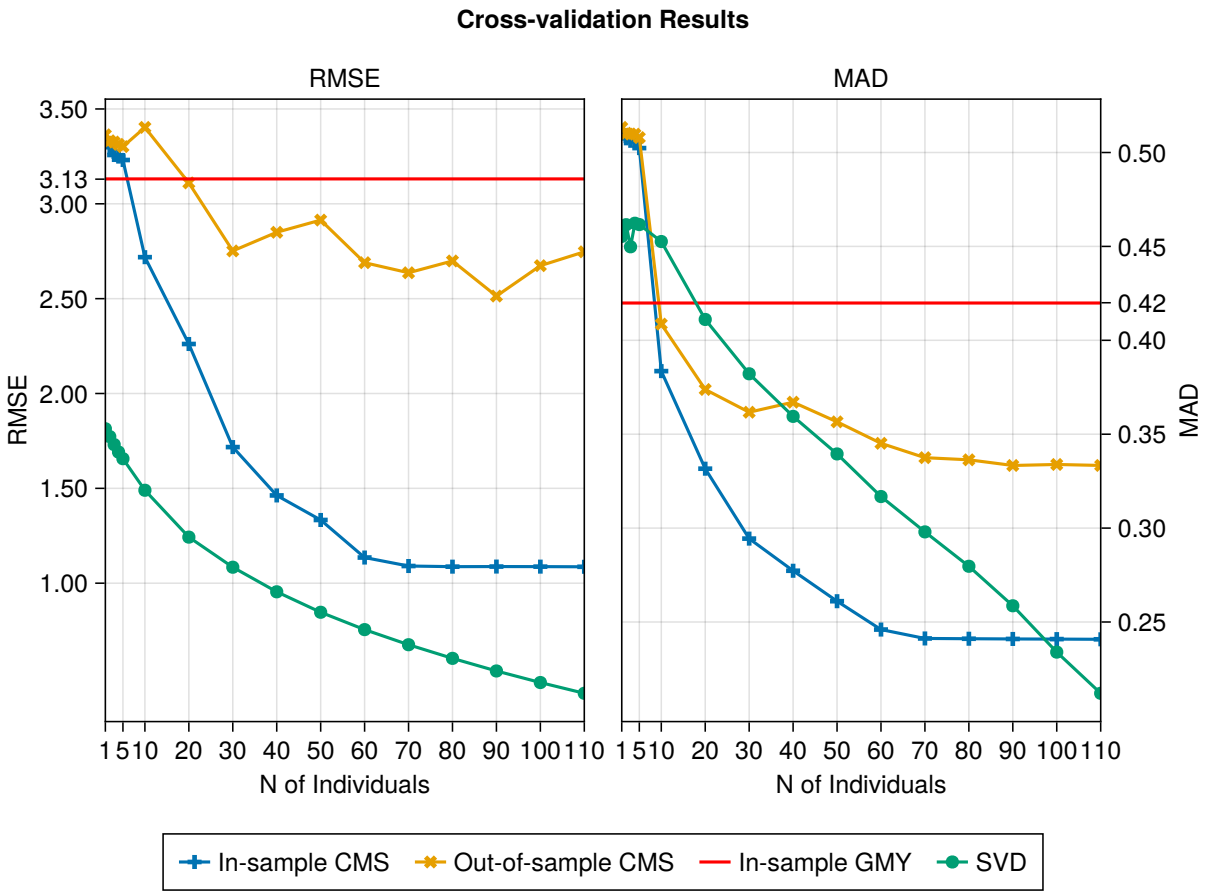


Figure 3: Out-of-sample fit on $\mathcal{D}$ for different rank I

Cross-validation exercise selects the model with lowest out-of-sample prediction error: $I = 90$

Our semiparametric model shows superior in-sample performance compared to its parametric counterpart, even with a relatively low rank. Out-of-sample, our model outperforms the parametric model starting from a rank of $I = 20$ for RMSE fit, and $I = 10$ for MAD fit. It is important to remember here that the parametric model used in GMY employs a vast amount of data while our estimator uses only second choices and aggregate sales for a single year. According to our cross-validation exercise, our preferred specification is a rank of $I = 90$. In the following subsections, we will present results based on estimating this specification, as well as $I = 1$ (logit) and $I = 30$ as it could be a local minimum if the researcher were to stop the search at a lower rank than we did.

Figure 4 shows a comparison between our preferred specification and the data, with an additional matrix (far right) displaying the error that remains from prediction. In the following plots, we adopt

the visual convention that the product j removed corresponds to a column, and each cell represents second-choice probabilities $D_{j\to k}$. The blocks that show up on the diagonal of the matrix are clusters of high substitution among products. They are organized in segments with respect to their size, starting from convertibles (top left) to cargo vans (bottom right). This organization of the data is purely visual and does not impair the performance of our estimator.

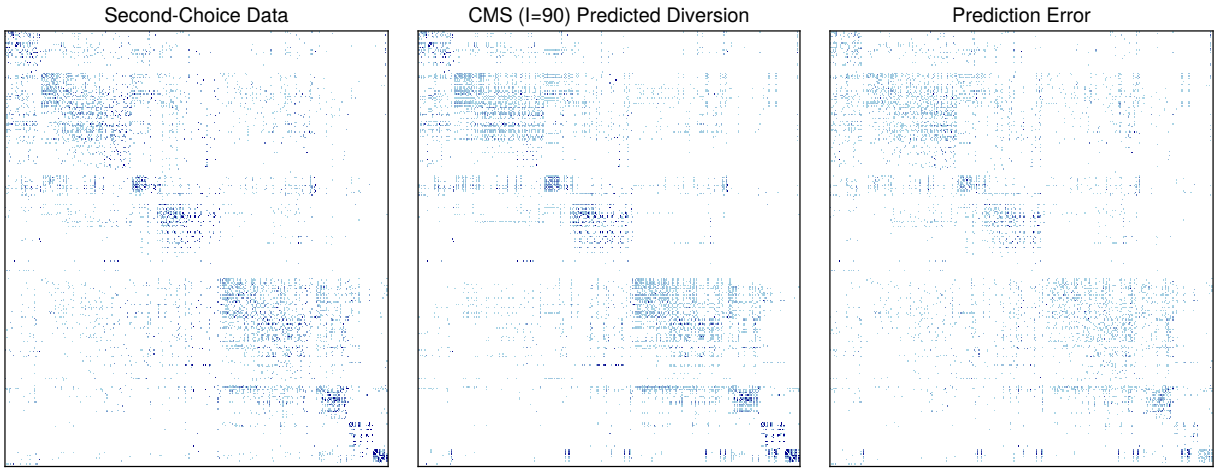


Figure 4: Side-by-side comparison between data $\mathcal{D}$, predicted second-choice probabilities $\mathbf{D}$, and absolute error $\|\mathcal{D} - \mathbf{D}\|$

As is generally the case with image processing (another field that uses matrix completion methods), prediction errors tend to occur along "edges" of the image: in our case, products for which substitution is high relative to surrounding products.

4.2. Comparison of implied second-choice probabilities

Figure 5 presents a comparison between the second-choice survey data and the estimated second-choice probabilities from three models: GMY, our preferred specification, and a logit model. Visually, our semiparametric model with a rank of $I = 90$ is the closest to the data used for estimation. GMY does a good job of identifying "groups" of vehicles with strong substitution (visible as blocks on the diagonal–mostly because it includes normally distributed random coefficients on product categories such as: sedans, SUV's, Pickup Trucks, CUV's, etc.). One challenge for GMY (and most parametric logit models) is that it produces logit-like substitution within each category where the most popular SUV is the best substitute for other SUV's (visually this appears as a gradient within each block). At the same time, because of the logit error, it also predicts too much substitution

to the most popular overall products such as the Camry, Accord, and F-150 even from different groups. This phenomenon is also observed in other contexts such as the vending experimental data (presented in the next section), although in this case, GMY does a significantly better job, in part because the model has quite a few parameters, and categories do a good job at explaining substitution. The main challenge of these parametric models is that while they over-predict substitution to the field, they under-predict substitution to the closest substitutes.

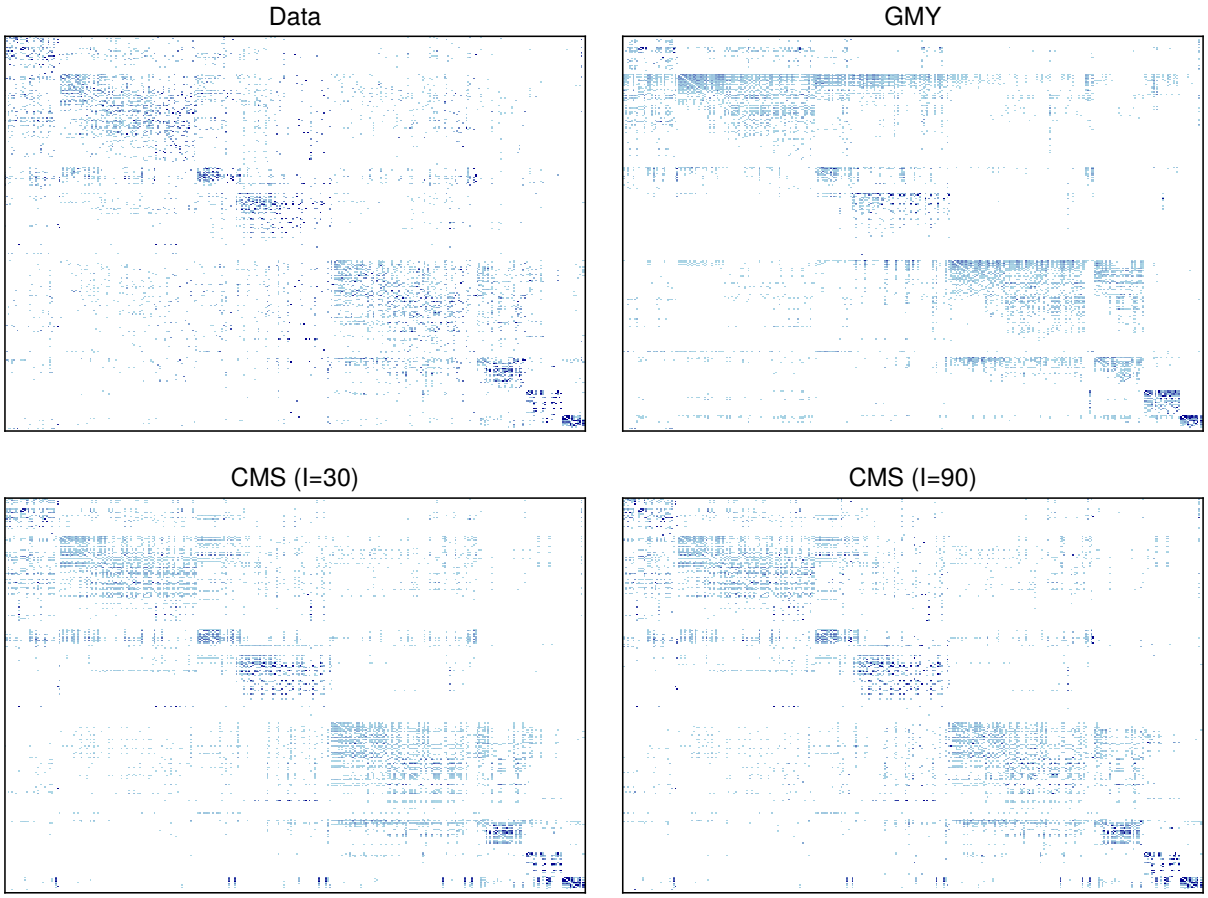


Figure 5: Estimated Second-Choice Probability Matrices

Visually, our preferred specification matches the sparsity patterns of the data quite closely, while lower rank models and GMY tend to show "fat tails" in that second-choices off the diagonal are higher than in the data.

Our model is capable of rationalizing high second-choice probabilities within groups of highly substitutable products and zero diversion with products outside the groups. To illustrate this fact, we compare estimated second-choice probabilities from multiple estimators for three vehicles: the

Honda Accord, the Ford F-Series (the best selling vehicle in the U.S.), and the Mercedes-Benz Sprinter Van. The first vehicle displays relatively spread out second-choice probabilities to other models in its category, while the second displays substitution concentrated between only a few vehicles. Finally, the third vehicle has one substitute with two thirds of recorded substitution. Tables 1, 2 and 3 show that our preferred specification is able to rationalize these three different substitution patterns with a low error rate compared to the parametric model in GMY or a simple logit model.

Model	Raw	Logit	CMS I=30	CMS I=90	GMY
Subaru Legacy	10.27	1.01	8.05	8.21	1.3
Toyota Camry	9.1	0.84	6.7	6.85	9.48
Acura Tlx	6.07	0.71	1.83	2.07	0.46
Honda Civic	5.97	0.91	2.9	2.75	3.89
Mazda Mazda6	5.68	0.52	4.77	4.87	1.32
Volkswagen Passat	4.01	0.74	4.34	4.41	1.22
Nissan Altima	3.52	0.6	3.87	3.89	7.22
Hyundai Sonata	3.52	0.68	5.61	5.68	5.09
Volkswagen Jetta	3.33	0.97	4.4	4.23	1.48
Mazda Mazda3	2.15	1.08	2.08	1.61	1.49
Toyota Corolla	1.96	0.71	2.46	2.32	4.66

Table 1: Top Substitutes: Honda Accord

Model	Raw	Logit	CMS I=30	CMS I=90	GMY
Ram Pickup	24.59	1.36	23.38	23.37	19.4
Gmc Sierra	20.29	1.28	21.0	21.02	17.27
Chevrolet Silverado	15.62	1.21	16.73	16.75	33.62
Toyota Tundra	12.98	0.76	12.69	12.69	2.29
Toyota Tacoma	6.31	1.13	3.6	3.62	2.83
Chevrolet Colorado	4.64	1.08	3.37	3.38	2.87
Gmc Canyon	2.3	0.62	1.71	1.73	1.02
Nissan Frontier	1.63	0.67	0.83	0.84	0.61
Jeep Wrangler	1.59	0.62	0.96	0.81	0.06
Nissan Titan	0.7	0.07	0.79	0.81	0.18
Ford Explorer	0.63	0.4	0.05	0.03	0.71

Table 2: Top Substitutes: Ford F-Series

Model	Raw	Logit	CMS I=30	CMS I=90	GMY
Ford Transit Wagon	66.67	0.19	47.63	66.19	0.04
Ram Promaster	16.67	0.02	0.0	16.19	1.76
Ford Transit Connect	8.33	0.18	7.33	7.85	0.01
Nissan Nv	8.33	0.17	29.96	7.58	6.52
Mini Cooper	0.0	0.45	0.0	0.0	0.0
Volkswagen Beetle Ii Cabrio	0.0	0.25	0.0	0.0	0.0
Audi A5	0.0	0.28	0.0	0.0	0.02
Mazda Mx-5 Miata	0.0	0.19	0.0	0.0	0.0
Audi S5	0.0	0.15	0.0	0.0	0.0
Porsche Boxster	0.0	0.07	0.0	0.0	0.0
Volkswagen Eos	0.0	0.07	0.0	0.0	0.0

Table 3: Top Substitutes: Mercedes-Benz Sprinter Van

Figure 6 presents a comparison of in-sample performance for various ranks using different fit measures: RMSE, MAE, the fraction of correctly predicted top 10 substitutes, and the fraction of correctly predicted pairwise comparisons. We selected the latter two measures because they represent important features of the second-choice matrix $\mathbf{D}$, although they are not directly targeted by either class of models in estimation. The % Correct Top 10 fit measures how many top 10 substitutes (unranked) were correctly predicted (0/1) on average across products. The pairwise comparisons ask for two substitutes k and k' whether $\mathbb{I}[D_{jk} > D_{jk'}]$ is predicted correctly (0/1) and averages across all (j, k, k'). As an example, consider second-choice probabilities for the Honda Accord in Table 1: the Subaru Legacy dominates the Toyota Camry in the data, which is correctly predicted by both CMS $I = 30$ and $I = 90$, but not by GMY nor the Logit model. The % Correct Pairwise Fit is the average proportion of such correctly predicted pairwise comparisons across all products. One challenge is for these comparisons is that because there are $\mathcal{J} = 318$ products, for many choices (k, k') both may represent rarely chosen substitutes $\mathcal{D}_{j\to k} \approx 0$.

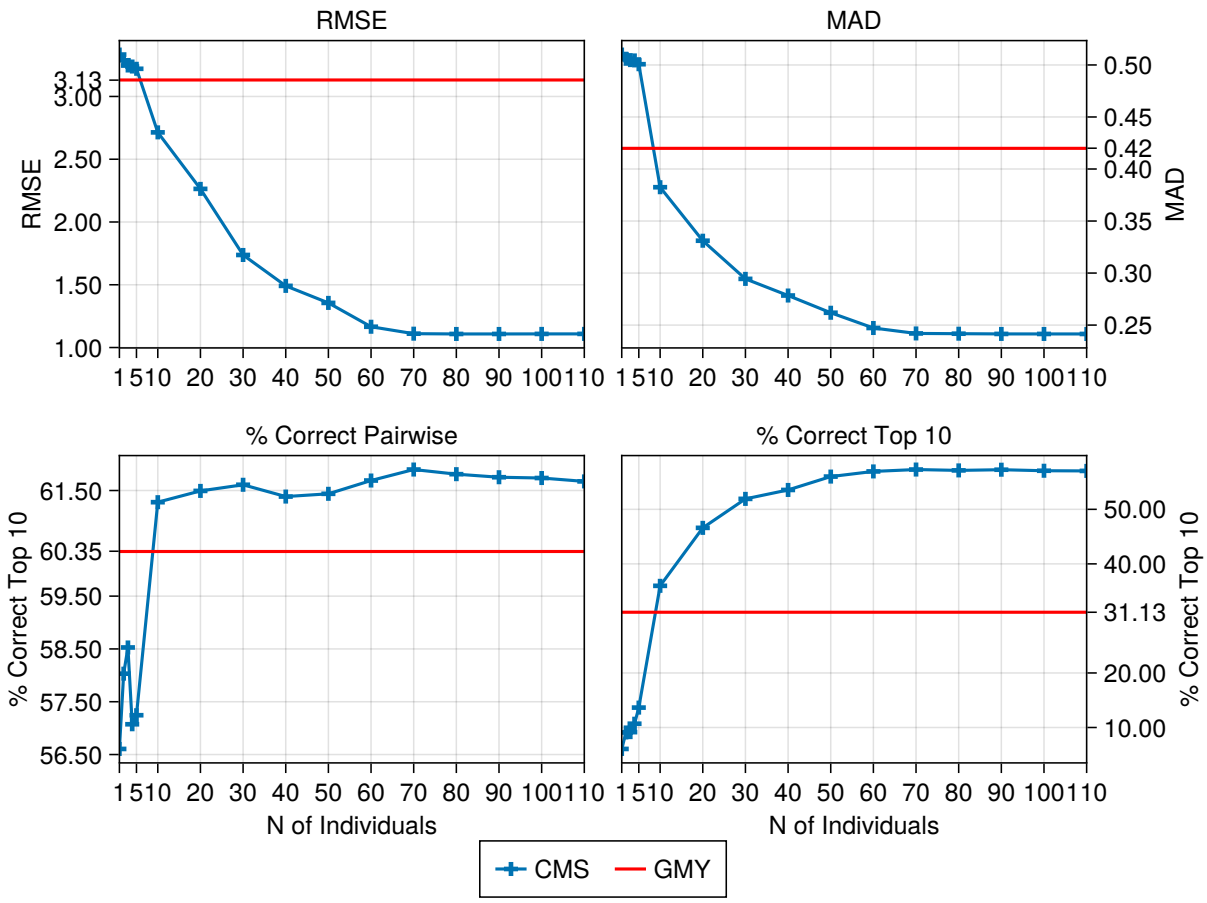


Figure 6: In-sample fit comparison across ranks

Our semiparametric model outperforms its parametric counterpart in all measures starting from relatively low ranks (between 5 and 10). Note: the two bottom fit measures are explained in the previous paragraph.

Our model is able to produce these extreme patterns for a few reasons. One is that it tends to produce highly sparse $\mathbf{s_i}$ vectors with no substitution to a large number of products at the individual level. The second is that individual diversion ratios are $D_{j,k,i} = \frac{s_{ik}}{1-s_{ij}}$, so we can get extremely high rates of substitution when for some individual i, there is a high value of both s_{ij} and s_{ik} (such as in the Mercedes-Benz Sprinter Van example Table 3). This is particularly true if $s_{ij} = 0$ for most of the other types, such that only a small number of types have non-trivial choice probabilties for the Sprinter Van. However, our model still places some significant restrictions on substitution patterns. In Table 1, we see that the Honda Accord is highly substitutable both the Toyota Camry ($D_{j\to k}$) and the Subaru Legacy ($D_{j\to k'}$). For small values of I, this will imply and without strong

substitution between $D_{k\to j}$ (ie: from Camry and Legacy to Accord) or $D_{k\to k'}$ (between Legacy and Camry). If this pattern doesn't arise in the data, we can rationalize it by having one type with high s_{ij} for Camry and Accord (but not Legacy) and another with high s_{ij} for Legacy and Accord (but not Camry). As the rank I gets larger, it becomes easier to accomodate this kind of behavior, however this may also be the source of "overfitting" the observed substitution patterns, one advantage of the low-rank structure might be that the model is able to "learn" these kinds of patterns.

5. Extensions

5.1. Multiple Product Removals

In some settings, we might have data on substitution patterns from multiple product removals. For example, in the vending machine experiment described in this paper, two treatment arms removed two products simultaneously: Snickers and M&M's in one arm and Doritos and Cheetos in the other. In this case we know don't know the identity of the first choice j, only that it was from some subset $\mathcal{J}_{rem} \subset \mathcal{J}$. We show in Appendix A.1 that the diversion ratio to remaining product k from removing multiple products $j \in \mathcal{J}_{\text{rem}}$ with total choice probability: $s_{i,\mathcal{J}_{\text{rem}}} \equiv \sum_{j\in\mathcal{J}_{\text{rem}}} s_{ij}$ has a structure similar to (5), and is given by:

$$\begin{aligned}\mathcal{D}_{\mathcal{J}_{rem}\to k} &= \mathbb{P}(\text{chooses } k \in \mathcal{J} \setminus \{\mathcal{J}_{rem}\} \mid \text{chooses } \mathcal{J}_{rem} \subset \mathcal{J}) \\ D_{\mathcal{J}_{\text{rem}}\to k} &= \sum_{i=1}^{I} \pi_i \cdot \frac{s_{ik}}{1 - s_{i,\mathcal{J}_{\text{rem}}}} \cdot \frac{s_{i,\mathcal{J}_{\text{rem}}}}{s_{\mathcal{J}_{\text{rem}}}} \end{aligned} \tag{14}$$

A similar scenario arises if the consumer is instead restricted in the set of their second-choice options. For example, in the MaritzCX survey, consumers are required to list a second-choice option (and are unable to list "no purchase" as their second-choice). In this case, we are interested in the following probability, which we can compute under the model using an augmented "individual

diversion ratio" $D_{jk,i} = \frac{s_{ik}}{1-s_{i,\mathcal{J}rem}}$, while leaving the other terms unchanged:

$$\begin{aligned}\mathcal{D}_{j\to k\setminus\{0\}} &= \mathbb{P}(\text{chooses } k \in \mathcal{J} \setminus \mathcal{J}_{rem} \mid \text{chooses } j \in \mathcal{J}) \\ D_{j\to k\setminus\{0\}} &= \sum_{i=1}^{I} \pi_i \cdot \frac{s_{ik}}{1 - s_{i,\mathcal{J}_{\text{rem}}}} \cdot \frac{s_{ij}}{s_j}. \end{aligned} \tag{15}$$

In the case of the MaritzCX suvey used in Section 4, the restricted set $\mathcal{J}_{rem} = \{0, j\}$, where the outside option is not available as a second-choice.

Conceptually neither of (14) and (15) is different from (5), and these represent minor modifications for the $\mathbf{D}$ matrix in (6). In fact, one could even combine the two modifications and consolidate or restrict both the first- and second-choices. One tradeoff is that the more we consolidate first and second-choices, the less information we have to inform our estimates. As an example, we might know how many sedan buyers have a second-choice product that is a SUV, which would be less informative than model level second-choice data.

5.2. Estimating Market Size

In most applications the number or share of individuals choosing the outside option q_0 or $\mathcal{S}_0$ is unobserved, but assumed to be known by the researcher. In practice this is often calibrated to census or Current Population Survey (CPS) data on the number of individuals or households.[21] In Backus et al. (2021), the market size is estimated by trying to capture foot traffic to stores by projecting total category sales (breakfast cereal) on sales of other product categories (milk and eggs) and taking the fitted values.

However, since we are using data from only a single market, we can treat the number of consumers who are choosing the outside option as a free parameter and modify (6):

$$\min_{(\mathbf{S},\pi,q_0,\mathcal{S})\geq 0} \left\|\mathcal{P}_\Omega(\mathcal{D} - \mathbf{D})\right\|_{\ell_2} + \lambda \left\|\mathcal{S} - \mathbf{S}\,\pi\right\|_{\ell_2} \text{ with } \left\|\pi\right\|_{\ell_1} \leq 1, \quad \left\|\mathbf{s_i}\right\|_{\ell_1} \leq 1. \tag{16}$$

$$\mathcal{S}_j = \frac{q_j}{q_0 + \sum_{k\in\mathcal{J}\setminus\{0\}} q_k}.$$

If we are overidentified in the original problem (6), we may be able to estimate a single additional pa-

[21] See Berry et al. (1995); Nevo (2001) as examples.

rameter q_0 without much difficulty. Technically, we add $\mathcal{J}+1$ pararameters for $\mathcal{S}$ and $\mathcal{J}$ constriants in (16). The constraints for $\mathcal{S}_j$ are not linear but are very simple since $\frac{\partial \mathcal{S}_j}{\partial q_0} = -\frac{q_j}{(q_0+\sum_{k\in\mathcal{J}\setminus\{0\}} q_k)^2}$.

5.3. Adding Parametric Restrictions

There may be cases where we are interested not only in substitution among existing products, but also what might happen if we were to introduce a new product, or if prices or characteristics were to change. In this case we might want to construct a prediction for $\mathcal{D}(\mathbf{x}')$ or $\mathcal{S}(\mathbf{x}')$ at some $\mathbf{x}' \neq \mathbf{x}$.

In this case having a parametric structure on characteristics like $V_{ij}(x_j) = \beta_i x_j + d_j$ would be helpful. We can simply impose an additional set of constraints on (6), and search over β_i and d_j as well as π_i and s_{ij}:

$$s_{ij}(\mathbf{x}) = \frac{e^{\beta_i x_j + d_j}}{1 + \sum_k e^{\beta_i x_k + d_k}}.$$

This creates some challenges: the bound $\text{rank}(D) \leq I$ may no longer be informative; for any fixed I the fit of the model is likely to be worse. However, we've effectively reduced the number of free parameters from $(J+1) \times I$ to $J + (K+1) \times I$ where $\dim(\beta) = K$.

The advantage of these additional restrictions is that it enables us to test whether or not the characteristics x_j span the space of substitution patterns $\mathcal{D}$. An obvious choice would be a non-nested model comparison like Rivers and Vuong (2002) which compares the fit of the model with and without parametric restrictions and adjusted for the degrees of freedom.

5.4. Other Variation: Prices, Quality, and Characteristics

We could derive a similar result where instead of second-choice diversion, we consider diversion with respect to an infinitesimal change in some characteristic (such as price) where $\frac{\partial V_{ij}}{\partial z_j} = \beta_i^z$:[22]

$$\frac{\partial s_j}{\partial z_k}(\mathbf{x}) = \sum_{i=1}^{I} \beta_i^z \cdot \pi_i \cdot s_{ik}(\mathbf{x}) \cdot (1[j=k] - s_{ij}(\mathbf{x})) \tag{17}$$

$$\mathbf{D}_z(\mathbf{x}) = \text{diag}(\mathbf{s_z})^{-1} \cdot \sum_{i=1}^{I} \pi_i \cdot \beta_i^z \cdot \mathbf{s_i}(\mathbf{x}) \cdot \mathbf{s_i}^T(\mathbf{x}). \tag{18}$$

There are three ways to view (18). The first is that in order to predict elements of $\mathbf{D}_z(\mathbf{x})$, we now need an estimate of β_i^z for each type i. This means that counterfactuals which depend on $\mathbf{D}_z(\mathbf{x})$ or $\frac{\partial s_j}{\partial z_j}(\mathbf{x})$ are not identified from our simple estimator in (6) alone. The second is that if we observe part or all of $\mathcal{D}_z(\mathbf{x})$, we can use this variation to estimate β_i^z. The third is that if all individuals agree on $\beta_i^z = \beta^z$, as in the case of the quality index ξ_j, then the expression simplifies substantially.

5.5. Multiple Markets

Our estimator in (6) is conditioned on a single set of observables $\mathbf{x}$ so that we observe aggregate shares $\mathcal{S}(\mathbf{x})$ and some elements of $\mathcal{D}(\mathbf{x})$ to recover $\mathbf{s_i}(\mathbf{x})$ and π_i. Implicitly, this means everything is conditional on $\mathbf{x}$. If we observed data from multiple markets $t = 1, \ldots, T$ with different $\mathbf{x_t}$, we could either estimate separately market by market, or parameterize $V_{ij}(\mathbf{x_t})$ and use the parametric structure to pool parameters across markets.[23]

5.6. Identification and Inference

There are different ways to think about asymptotic inference in (6). One approach would be to take the observed elements of $P_\Omega \to \infty$ and implicitly $J \to \infty$ and treat (6) as a GMM estimator. This would be similar to the approach taken in Berry et al. (2004b).

A more practical approach might be to think about (6) as a minimum distance estimator where

[22] Here $\text{diag}(\mathbf{s_z})^{-1}$ is a diagonal matrix where entries are given by $\left(\frac{\partial s_j}{\partial z_j}\right)^{-1}$ for a particular characteristic z. This term row-normalizes the matrix so that $\sum_{j \neq k} D_{kj} = 1$.

[23] The former is used in applications like hospital demand. The latter is effectively the identification approach in much of the rest of the IO literature Berry et al. (1995); Nevo (2001), etc.

the observed second-choice probabilities are themselves estimated from some sample. This would be the case if we were using a survey of n individuals to estimate the second-choice probabilities. In that case we would need: $\left\|\mathcal{D}^n_{jk} - \mathcal{D}^0_{jk}\right\| \xrightarrow{p} 0$ (the sample estimates converge in probability to the true population second-choice probabilities.)

The standard conditions for minimum distance estimators are straightforward to verify for our constrained least squares problem. Compactness of $\theta \in \Theta$ would be guaranteed by the constraints in (6), all parameters are constrained to the unit interval. We provide the derivatives $\frac{\partial D_{jk}}{\partial \theta}$ in Appendix A.3. For the derivatives to be bounded we need that $\mathbf{s_i}$ is non-degenerate and has at least two nonzero elements for each i. The objective in (6) is quadratic and all other constraints are linear or quadratic in parameters, which should guarantee that $Q(\theta)$ is twice continuously differentiable with bounded derivatives.

Identification of the model in (6) is (likely) straightforward so long the rank of the observed $\mathcal{D}$ matrix is sufficiently large: $\text{rank}(\mathcal{D}) \gg I$. This guarantees that the objective function is never equal to zero. We must also have more observed elements of $\mathcal{D}$ than unknowns $(J+1) \cdot I$. Because $D_{jk}(\theta)$ is a non-convex function of the parameters, at best we can only hope to establish local identification at some values of $(\mathbf{S}, \pi)$. We need a rank condition on the Jacobian (with respect to parameters $\theta = [\mathbf{s_i}, \pi_i]$) for there to be a unique solution in the least-squares sense. A necessary though not sufficient condition is that the vectors $\mathbf{s_i}$ be linearly independent and $\pi_i > 0$ strictly for all i. We provide the derivatives $\frac{\partial D_{jk}}{\partial \theta}$ in Appendix A.3.

The easiest way to construct confidence intervals for $D_{jk}(\widehat{\theta})$ (or other outputs) is to bootstrap the underlying survey data used to construct $\mathcal{D}^n$ and re-estimate the model in (6). Even on relatively large problems, estimation only takes a few seconds in our application.

[Is this kind of informal sketch useful at all?]

6. Conclusion

We develop a semi-parametric estimator of the full matrix of diversion ratios based on matrix completion methods commonly used in computer science. Our approach uses data on aggregate market shares and (potentially partially observed) second-choice diversion ratios, requires no information

on product characteristics, and is computationally easy to estimate. We demonstrate the approach in Monte Carlo simulations and compare it to commonly-used but potentially misspecified parametric models, and we apply the method to the US automobile market, for which we have diversion on all inside goods in a setting with more than 300 products.

References

Abaluck, J. and A. Adams-Prassl (2021): "What do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses," *The Quarterly Journal of Economics*, 136, 1611–1663.

Allenby, G. M., N. Hardt, and P. E. Rossi (2019): "Chapter 3 - Economic foundations of conjoint analysis," in *Handbook of the Economics of Marketing*, ed. by J.-P. Dubé and P. E. Rossi, North-Holland, vol. 1 of *Handbook of the Economics of Marketing, Volume 1*, 151–192.

Backus, M., C. Conlon, and M. Sinkinson (2021): "Common Ownership and Competition in the Ready-to-Eat Cereal Industry," Working Paper 28350, National Bureau of Economic Research.

Barseghyan, L., M. Coughlin, F. Molinari, and J. C. Teitelbaum (2021a): "Heterogeneous Choice Sets and Preferences," *Econometrica*, 89, 2015–2048.

Barseghyan, L., F. Molinari, and M. Thirkettle (2021b): "Discrete Choice under Risk with Limited Consideration," *American Economic Review*, 111, 1972–2006.

Bayer, P., F. Ferreira, and R. McMillan (2007): "A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of Political Economy*, 115, 588–638.

Bayer, P. and C. Timmins (2007): "Estimating Equilibrium Models Of Sorting Across Locations," *The Economic Journal*, 117, 353–374.

Berry, S. (1994): "Estimating discrete-choice models of product differentiation," *RAND Journal of Economics*, 25, 242–261.

Berry, S. and P. Jia (2010): "Tracing the Woes: An Empirical Analysis of the Airline Industry," *American Economic Journal: Microeconomics*, 2, 1–43.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

——— (1999): "Voluntary Export Restraints on Automobiles: Evaluating a Trade Policy," *American Economic Review*, 89, 400–430.

——— (2004a): "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 112, 68–105.

Berry, S., O. B. Linton, and A. Pakes (2004b): "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems," *The Review of Economic Studies*, 71, 613–654.

Berry, S. and A. Pakes (2007): "The Pure Characteristics Demand Model," *International Economic Review*, 48, 1193–1225.

Berry, S. T. and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

——— (2022): "Nonparametric Identification of Differentiated Products Demand Using Micro Data," Tech. Rep. arXiv:2204.06637, arXiv.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003): "Latent dirichlet allocation," *The Journal of Machine Learning Research*, 3, 993–1022.

Cardell, N. S. (1997): "Variance Component Structures for the Extreme Value and Logistic Distributions," *Econometric Theory*, 13, 185–213.

Chen, Y., Y. Chi, J. Fan, and C. Ma (2021): "Spectral Methods for Data Science: A Statistical Perspective," *Foundations and Trends® in Machine Learning*, 14, 566–806, citation:SpectralMethodsBook2021.

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

——— (2023): "Incorporating Micro Data into Differentiated Products Demand Estimation with PyBLP," .

Conlon, C. and J. H. Mortimer (2021a): "Empirical properties of diversion ratios," *The RAND Journal of Economics*, 52, 693–726.

Conlon, C., J. H. Mortimer, P. Sarkis, and R. Gonzalez Valdenegro (2023): "Effects of Product Availability: Experimental Evidence," .

Conlon, C. T. and J. H. Mortimer (2013): "Demand Estimation under Incomplete Product Availability," *American Economic Journal: Microeconomics*, 5, 1–30.

——— (2021b): "Efficiency and Foreclosure Effects of Vertical Rebates: Empirical Evidence," *Journal of Political Economy*, 129, 3357–3404, publisher: The University of Chicago Press.

Dardanoni, V., P. Manzini, M. Mariotti, and C. J. Tyson (2020): "Inferring Cognitive Heterogeneity From Aggregate Choices," *Econometrica*, 88, 1269–1296.

Deaton, A. and J. Muellbauer (1980): "An Almost Ideal Demand System," *The American Economic Review*, 70, 312–326, publisher: American Economic Association.

Farrell, J. and C. Shapiro (2010): "Antitrust Evaluation of Horizontal Mergers: An Economic Alternative to Market Definition," *The B.E. Journal of Theoretical Economics*, 10, publisher: De Gruyter.

Fox, J. T., K. i. i. Kim, S. P. Ryan, and P. Bajari (2011): "A simple estimator for the distribution of random coefficients," *Quantitative Economics*, 2.

Goeree, M. S. (2008): "Limited Information and Advertising in the U.S. Personal Computer Industry," *Econometrica*, 76, 1017–1074.

Goolsbee, A. and A. Petrin (2004): "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica*, 72, 351–381.

Greene, W. H. and D. A. Hensher (2003): "A latent class model for discrete choice analysis: contrasts with mixed logit," *Transportation Research Part B: Methodological*, 37, 681–698.

Grieco, P. L. E., C. Murry, J. Pinkse, and S. Sagl (2023): "Conformant and Efficient Estimation of Discrete Choice Demand Models," .

Grieco, P. L. E., C. Murry, and A. Yurukoglu (2021): "The Evolution of Market Power in the US Auto Industry," .

Heiss, F., S. Hetzenecker, and M. Osterhaus (2022): "Nonparametric estimation of the random coefficients model: An elastic net approach," *Journal of Econometrics*, 229, 299–321.

Ke, S., J. L. Montiel Olea, and J. Nesbit (2022): "Robust Machine Learning Algorithms for Text Analysis," .

Kingma, D. P. and J. Ba (2017): "Adam: A Method for Stochastic Optimization," .

Magnolfi, L., J. McClure, and A. T. Sorensen (2022): "Triplet Embeddings for Demand Estimation," .

Manzini, P. and M. Mariotti (2014): "Stochastic Choice and Consideration Sets," *Econometrica*, 82, 1153–1176.

Masatlioglu, Y., D. Nakajima, and E. Y. Ozbay (2012): "Revealed Attention," *American Economic Review*, 102, 2183–2205.

Matějka, F. and A. McKay (2015): "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model," *American Economic Review*, 105, 272–298.

McFadden, D. (1974): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka, New York: Academic Press, 105 – 142.

——— (1978): "Modelling the Choice of Residential Location," in *Spatial Interaction Theory and Planning Models*, ed. by A. Karlqvist, L. Lundsqvist, F. Snickars, and J. Weibull, North-Holland.

McFadden, D. and K. Train (2000): "Mixed MNL models for discrete response," *Journal of Applied Econometrics*, 15, 447–470.

Nevo, A. (2000): "Mergers with Differentiated Products: The Case of the Ready-to-Eat Cereal Industry," *The RAND Journal of Economics*, 31, 395–421.

——— (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69, 307–342.

Petrin, A. (2002): "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 110, 705–729, publisher: The University of Chicago Press.

Qiu, J., M. Sawada, and G. Sheu (2021): "Win/Loss Data and Consumer Switching Costs: Measuring Diversion Ratios and the Impact of Mergers," SSRN Scholarly Paper 3957662, Social Science Research Network, Rochester, NY.

Raval, D., T. Rosenbaum, and S. A. Tenn (2017): "A Semiparametric Discrete Choice Model: An Application to Hospital Mergers," *Economic Inquiry*, 55, 1919–1944.

Raval, D., T. Rosenbaum, and N. E. Wilson (2022): "Using Disaster Induced Closures to Evaluate Discrete Choice Models of Hospital Demand," 88.

Reynolds, G. and C. Walters (2008): "The Use of Customer Surveys for Market Definition and the Competitive Assessment of Horizontal Mergers," *Journal of Competition Law and Economics*, 4, 411–431, publisher: Oxford University Press.

Rivers, D. and Q. Vuong (2002): "Model selection tests for nonlinear dynamic models," *The Econometrics Journal*, 5, 1–39.

Train, K. E. (2009): *Discrete Choice Methods with Simulation*, Cambridge: Cambridge University Press, 2 ed.

Udell, M. and A. Townsend (2019): "Why Are Big Data Matrices Approximately Low Rank?" *SIAM Journal on Mathematics of Data Science*, 1, 144–160.

Wu, L., Y. Yang, and H. Liu (2014): "Nonnegative-lasso and application in index tracking," *Computational Statistics & Data Analysis*, 70, 116–126.

Appendices

A. Theoretical Results

A.1. Multiple Product Removals

From Conlon and Mortimer (2021a), we know that, in a random coefficients logit demand framework, diversion from good j to good k when j is no longer available is:

$$D_{jk} = \frac{s_k(\mathcal{J}, x) - s_k(\mathcal{J}\backslash j, x)}{s_j(\mathcal{J}, x)} = -\frac{1}{s_j(\mathcal{J}, x)} \int \frac{s_{ij}(\mathcal{J}, x) s_{ik}(\mathcal{J}, x)}{(1 - s_{ij}(\mathcal{J}, x))}$$

For clarity, we use the subscript $\backslash j$ as an equivalent for writing $(\mathcal{J}\backslash j)$ and we remove the x. We can rewrite the previous expression to get the share of k when j is not available:

$$s_{k\backslash j} = s_k + \int \frac{s_{ij} s_{ik}}{(1 - s_{ij})} \tag{A1}$$

In addition, we know that for each individual type within the mixed logit model, diversion is written as:

$$D_{ijk} = \frac{s_{ik} - s_{ik\backslash j}}{s_{ij}} = -\frac{s_{ik}}{(1 - s_{ij})}$$

which we can rewrite to find individual i's share of k when j is not available:

$$s_{ik\backslash j} = s_{ik} + \frac{s_{ik} s_{ij}}{(1 - s_{ij})} = s_{ik} \cdot \left(1 + \frac{s_{ij}}{(1 - s_{ij})}\right) = \frac{s_{ik}}{(1 - s_{ij})} \tag{A2}$$

For the multiple product removals case, we want to prove that the individual share of good k when all products $j \in \mathcal{J}_{\text{rem}}$ are removed is:

$$s_{ik\backslash \mathcal{J}_{\text{rem}}} = \frac{s_{ik}}{1 - \sum_{j \in \mathcal{J}_{\text{rem}}} s_{ij}}$$

We can proceed by induction. Equation (2) is the our first case, when removing only one product. Assume that after removing p products we have

$$s_{ik\backslash 1,\ldots,p} = \frac{s_{ik}}{1 - \sum_{j=1}^{p} s_{ij}},$$

then when we remove $p + 1$ products, we have

$$\begin{aligned} s_{ik\backslash 1,\ldots,p+1} = \frac{s_{ik\backslash 1,\ldots,p}}{1 - s_{ip+1\backslash 1,\ldots,p}} = \frac{s_{ik}}{1 - \sum_{j=1}^{p} s_{ij}} \cdot \frac{1}{1 - \frac{s_{ip+1}}{1 - \sum_{j=1}^{p} s_{ij}}} &= \frac{s_{ik}}{1 - \sum_{j=1}^{p} s_{ij}} \cdot \frac{1 - \sum_{j=1}^{p} s_{ij}}{1 - \sum_{j=1}^{p+1} s_{ij}} \\ &= \frac{s_{ik}}{1 - \sum_{j=1}^{p+1} s_{ij}} \end{aligned}$$

Therfeore, for a particular instance of IIA logit i, we have:

$$s_{ik\backslash\mathcal{J}_{\text{rem}}} = \frac{s_{ik}}{1 - \sum_{j\in\mathcal{J}_{\text{rem}}} s_{ij}}$$

For the mixed logit, we want to prove that:

$$s_{k\backslash\mathcal{J}_{\text{rem}}} = s_k + \int s_{ik} \cdot \frac{\sum_{j\in\mathcal{J}_{\text{rem}}} s_{ij}}{1 - \sum_{j\in\mathcal{J}_{\text{rem}}} s_{ij}}$$

We already know that this is the case for a single removed product, as expressed in (1). And, if after removing p goods we have:

$$s_{0\backslash 1,\ldots,p} = s_k + \int s_{ik} \cdot \frac{\sum_{j=1}^{p} s_{ij}}{1 - \sum_{j=1}^{p} s_{ij}}$$

then for $p+1$, we have

$$\begin{aligned}
s_{k\backslash 1,\ldots,p+1} &= s_{k\backslash 1,\ldots,p} + \int \frac{s_{ik\backslash 1,\ldots,p} \cdot s_{ip+1\backslash 1,\ldots,p}}{1 - s_{ip+1\backslash 1,\ldots,p}} \\
&= s_k + \int s_{ik} \cdot \frac{\sum_{j=1}^{p} s_{ij}}{1 - \sum_{j=1}^{p} s_{ij}} + \int \frac{\frac{s_{ik}}{1-\sum_{j=1}^{p} s_{ij}} \cdot \frac{s_{ip+1}}{1-\sum_{j=1}^{p} s_{ij}}}{1 - \frac{s_{ip+1}}{1-\sum_{j=1}^{p} s_{ij}}} \\
&= s_k + \int s_{ik} \cdot \frac{\sum_{j=1}^{p} s_{ij}}{1 - \sum_{j=1}^{p} s_{ij}} + \int s_{ik} \cdot \frac{s_{ip+1}}{(1 - \sum_{j=1}^{p} s_{ij})(1 - \sum_{k=1}^{p+1} s_{ik})} \\
&= s_k + \int s_{ik} \cdot \left[\frac{\sum_{j=1}^{p} s_{ij}}{1 - \sum_{j=1}^{p} s_{ij}} + \frac{s_{ip+1}}{(1 - \sum_{j=1}^{p} s_{ij})(1 - \sum_{k=1}^{p+1} s_{ik})} \right] \\
&= s_k + \int s_{ik} \cdot \left[\frac{(1 - \sum_{k=1}^{p+1} s_{ik}) \cdot (\sum_{j=1}^{p} s_{ij}) + s_{ip+1}}{(1 - \sum_{j=1}^{p} s_{ij})(1 - \sum_{k=1}^{p+1} s_{ik})} \right] \\
&= s_k + \int s_{ik} \cdot \left[\frac{(1 - \sum_{k=1}^{p} s_{ik}) \cdot (\sum_{j=1}^{p} s_{ij}) + (1 - \sum_{k=1}^{p} s_{ik}) \cdot s_{ip+1}}{(1 - \sum_{j=1}^{p} s_{ij})(1 - \sum_{k=1}^{p+1} s_{ik})} \right] \\
&= s_k + \int s_{ik} \cdot \left[\frac{\sum_{k=1}^{p+1} s_{ik}}{(1 - \sum_{k=1}^{p+1} s_{ik})} \right]
\end{aligned}$$

Therefore,

$$s_{k\backslash\mathcal{J}_{\text{rem}}} = s_k + \int s_{ik} \cdot \frac{\sum_{j\in\mathcal{J}_{\text{rem}}} s_{ij}}{1 - \sum_{j\in\mathcal{J}_{\text{rem}}} s_{ij}}$$

and we can write diversion from multiple products removed to k as:

$$D_{\mathcal{J}_{\text{rem}},k} = \frac{s_{k\backslash\mathcal{J}_{\text{rem}}} - s_k}{\sum_{j\in\mathcal{J}_{\text{rem}}} s_j} = \frac{1}{\sum_{j\in\mathcal{J}_{\text{rem}}} s_j} \int \frac{s_{ik} \sum_{j\in\mathcal{J}_{\text{rem}}} s_{ij}}{1 - \sum_{j\in\mathcal{J}_{\text{rem}}} s_{ij}}$$

A.2. Nested Logit Details

We use the Cardell (1997); Berry (1994) parameterization of the nested logit model. (This is not the same as the Train (2009); McFadden (1978) version).

A consumer i purchasing product j in a market where it is available obtains utility given by:

$$u_{ij} = \delta_j + \zeta_{ig}(\rho) + (1-\rho)\varepsilon_{ij}$$

where δ_j and ρ are parameters, ε_{ij} is i.i.d. Type-I Extreme-Value and ζ_{ig} is the idiosyncratic nest preference, such that $\zeta_{ig}(\rho) + (1-\rho)\varepsilon_{ij}$ is also extreme value. Thus, if we denote $\mathcal{J}_g$ as the set of products in nest g, the logit inclusive value IV_g for nest g is given by:

$$IV_g = \sum_{k \in \mathcal{J}_g} \exp\left(\frac{\delta_k}{1-\rho}\right)$$

The choice probabilities can be written as the product of the logit for choice j conditional on category choice g and the logit probability of choosing category g:

$$s_{j|g} = \frac{\exp\left(\frac{\delta_j}{1-\rho}\right)}{IV_g}, \quad s_g = \frac{IV_g^{(1-\rho)}}{\sum_{g'} IV_{g'}^{(1-\rho)}} \quad s_j = s_{j|g} \cdot s_g.$$

Following the results in the appendix of Conlon and Mortimer (2021a), and defining $Z(\rho, s_g) = [\rho + (1-\rho)s_g] \in (0,1]$, we get two formulas for diversion from product j to product k:

$$\text{(Same Nest g:)} D_{jk} = \frac{s_{k|g}}{Z^{-1}(\rho, s_g) - s_{j|g}}$$
$$\text{(Different Nests:)} D_{jk} = \frac{s_k \cdot (1-\rho)}{1 - Z(\rho, s_g) \cdot s_{j|g}}.$$

It is helpful to define the $J \times 1$ vector $\mathbf{s_{|g}}$ as having entries $s_{j|g}$ if j is a member of nest g and 0 if it is not. This allows us to write the transposed diversion matrix in terms of:

$$\mathbf{D}^T_{cross} = \sum_{g=1}^{G} (1-\rho)\, \mathbf{s} \cdot \left[\frac{Z^{-1}(\rho, s_g)}{Z^{-1}(\rho, s_g) - \mathbf{s_{|g}}}\right]^T$$
$$\mathbf{D}^T_{same} = \sum_{g=1}^{G} \mathbf{s_{|g}} \cdot \left[\frac{1}{Z^{-1}(\rho, s_g) - \mathbf{s_{|g}}}\right]^T.$$

We can combine both so that:

$$\mathbf{D}^T = \sum_{g=1}^{G} \left[(1-\rho)\, Z^{-1}(\rho, s_g)\, \mathbf{s} + \mathbf{s_{|g}}\right] \cdot \left[\frac{1}{Z^{-1}(\rho, s_g) - \mathbf{s_{|g}}}\right]^T$$

Because the term inside the summation can be written as the product of two vectors this diversion matrix will have at most rank G, the number of nests.

A.3. Jacobian

Our objective in (6) is $Q(\theta) = \|\mathcal{P}_\Omega(\mathcal{D} - \mathbf{D})\|_{\ell_2} = \sum_{(j,k)\in \text{OBS}} (\mathcal{D}_{jk} - D_{jk})^2$. We can treat $\mathbf{S}\pi = \mathbf{s}$ as a constraint that is linear in parameters. Likewise $\|(\mathcal{S} - \mathbf{s})\|_{\ell_2}$ is straightforward so we focus on the second-choices. We set $\theta = [\mathbf{s_i}, \pi_i, \mathbf{s}]$. We can write the Jacobian as :

$$\frac{\partial\, Q(\theta)}{\partial\, \theta} = \sum_{(j,k)\in \text{OBS}} 2 \cdot (\mathcal{D}_{jk} - D_{jk}) \frac{\partial\, D_{jk}}{\partial\, \theta}$$

We can look at this element-by-element for $j \neq k \neq l$:

$$\begin{aligned}
\frac{\partial\, D_{jk}}{\partial\, s_{il}} &= 0, \quad \frac{\partial\, D_{jk}}{\partial\, s_k} = 0 \\
\frac{\partial\, D_{jk}}{\partial\, s_{ik}} &= \sum_{i=1}^{I} \frac{\pi_i}{s_j} \cdot \frac{s_{ij}}{1 - s_{ij}} = \sum_{i=1}^{I} \pi_i \cdot D_{jk,i} \cdot \frac{1}{s_{ik}} \\
\frac{\partial\, D_{jk}}{\partial\, s_{ij}} &= \sum_{i=1}^{I} \frac{\pi_i}{s_j} \cdot \frac{s_{ik}}{(1 - s_{ij})^2} = \sum_{i=1}^{I} \pi_i \cdot D_{jk,i} \cdot \frac{1}{s_{ij}(1 - s_{ij})} \\
\frac{\partial\, D_{jk}}{\partial\, s_j} &= \sum_{i=1}^{I} -\frac{\pi_i}{s_j^2} \cdot \frac{s_{ik}\, s_{ij}}{1 - s_{ij}} = -\frac{1}{s_j} \sum_{i=1}^{I} \pi_i \cdot D_{jk,i} \\
\frac{\partial\, D_{jk}}{\partial\, \pi_i} &= \frac{s_{ij}}{s_j} \cdot \frac{s_{ik}}{1 - s_{ij}} = D_{jk,i}
\end{aligned}$$

Each of these derivatives are bounded because $(\pi_i, s_{ij}, s_j) \in [0, 1)$. Since we expect $s_j > 0$ from the data, the only risk is that $s_{ij} = 1$ (sparsity $s_{ij} = 0$ does not cause any issues). In this case it is sufficient that $D_{jk} > 0$ for all j and at least one k so that $\mathbf{s_i}$ has at least two nonzero elements.