Identification and Estimation of Discrete Choice Models with Unobserved Choice Sets[*]

Victor H. Aguiar ⓡ Nail Kashaev[†]

This version: January 11, 2022 / First Version: July 9, 2019

Abstract We propose a framework for nonparametric identification and estimation of discrete choice models with unobserved choice sets. We recover the joint distribution of choice sets and preferences from a cross-section of repeated choices. We assume that either the latent choice sets are *sparse* or that the number of repeated choices is sufficiently large. Sparsity requires the number of possible choice sets to be relatively small. It is satisfied, for instance, when the choice sets are nested, or when they form a partition. Our estimation procedure is computationally fast and uses mixed-integer optimization to recover the sparse support of choice sets. Analyzing the ready-to-eat cereal industry using a household scanner dataset, we find that ignoring the unobservability of choice sets can lead to incorrect estimates of preferences due to significant latent heterogeneity in choice sets.

JEL classification numbers: C14, C5, D6

Keywords: random utility, discrete choice, random consideration sets, best subset regression

[*] The "ⓡ" symbol indicates that the authors' names are in certified random order, as described by Ray ⓡ Robson (2018). We would like to thank Roy Allen, Daniel Chaves, Mingshi Kang, Yuichi Kitamura, Mathieu Marcoux, Salvador Navarro, Joris Pinkse, David Rivers, Bruno Salcedo, Susanne Schennach, Tomasz Strzalecki, and David Wei for helpful comments and suggestions. We gratefully acknowledge financial support from the Western Social Science Faculty Grant (FRDF R5533A02) and Social Sciences and Humanities Research Council Insight Development Grant. Researchers own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

[†] Aguiar: Department of Economics, University of Western Ontario; vaguiar@uwo.ca. Kashaev: Department of Economics, University of Western Ontario; nkashaev@uwo.ca.

1. Introduction

This paper considers the nonparametric identification and estimation of discrete choice models when the choice set that decision makers (DMs) face are unobserved by the researcher.[1] We show how to nonparametrically identify and estimate the joint distribution of latent choice sets and choices when we observe a cross-section of a small number of repeated choices. We allow random preferences to be correlated with choice sets after conditioning on observed covariates and previous choices. We apply our methodology to analyze the ready-to-eat (RTE) cereal industry in the USA using a household scanner dataset (Nielsen Homescan). Our empirical findings suggest there is substantial latent choice set heterogeneity that can lead to very different estimates of parameters of interest if not taken into account.

The classical nonparametric treatment of discrete choice under random utility (McFadden and Richter, 1990)[2] uses exogenous choice set variation to identify the distribution of preferences nonparametrically. However, the researcher usually does not observe the choice sets from which DMs pick their most preferred alternative. As a response to this lack of observability, researchers usually impose parametric restrictions on the distribution of preferences and the distribution of choice sets, or assume that every DM faces the same choice set (Hickman and Mortimer, 2016). These assumptions are problematic as they may lead to inconsistent estimation of preferences.

We overcome the above issue by exploiting the repeated choice structure of our data. That is, we use variation in choices of the same DM in different instances. Assuming that the choice set of the same DM remains the same across choices, the observed sequence of choices from the same choice set reveals information about it. Intuitively, if we observe a DM who chooses repeatedly either alternative 1 or alternative 2, then we can conclude that, most likely, this DM considers only these two alternatives. Another DM who chooses alternatives

[1] See Manski (1977) for the early treatment of the problem.

[2] From a decision-theoretic tradition random utility was initially studied by Block and Marschak (1960) and Falmagne (1978).

1, 2, and 3, most likely (but not necessarily), has a bigger consideration set than the first DM. This variation in finitely many choices within an (unobserved) choice set and among many DMs allows us to pin down the choice sets. After the choice sets are identified, if every option within the choice set is picked with positive probability, then the variation in choices within a given choice set allows us to identify the distribution of choices conditional on choice sets.[3]

Formally, to establish our identification result, we show that the problem of discrete choice with unobserved choice sets can be framed as a finite mixture, thus permitting us to use recent advances in identification of these models.[4] In particular, we base our identification strategy on the insights from discrete nonclassical measurement error results in Hu (2008).[5] We require at least three observed choices from the same choice set that are conditionally independent conditional on the unobserved choice set, observed covariates, and observed history of previous choices. Using these three choices, we show that under the standard linear independence (rank) condition, the identification of both the distribution of choice sets and the distribution of preferences is possible. We differ from Hu (2008) in that we do not need to impose any strict monotonicity condition, and we do not need to know the number of possible choice sets, since we can identify and estimate it. Strict monotonicity restrictions are usually imposed to match anonymous functions to latent states. In our setting, the latent choice sets have a structural interpretation, and we do not need to rank them.

The structural interpretation of the latent choice sets also allows us to establish two new sufficient conditions for the linear independence condition. The first condition imposes no restrictions on the distribution of preferences or choice sets but requires observability of sufficiently large (but finite) number of choices. The second condition requires fewer data,

[3] Our work provides a methodological bridge between the decision-theoretic literature on stochastic choice that has been based fundamentally on choice set variation (e.g., Luce, 1959, Block and Marschak, 1960, Falmagne, 1978, McFadden and Richter, 1990, Gul and Pesendorfer, 2006, Manzini and Mariotti, 2014, Fudenberg et al., 2015, and Brady and Rehbeck, 2016) and the discrete choice literature that has exploited covariate variation to identify the parametric distribution of preferences (Train, 2009).

[4] See, for instance, Hall and Zhou (2003), Hu (2008), Kasahara and Shimotsu (2009), Bonhomme et al. (2016), Kitamura and Laage (2018), and references therein.

[5] For recent applications in the context of auctions and discrete games see Hu et al. (2013), Xiao (2018), and Luo (2020).

but imposes ***sparsity*** on the support of the latent choice sets. Given that the number of all possible choice sets grows exponentially with the number of alternatives, the sparsity condition allows for substantial dimensionality reduction.

We also provide a new consistent computationally efficient nonparametric estimator of the distribution of choice sets and choices conditional on choice sets. Our estimator is a two-step estimator. On the first step, we consistently estimate choice sets using our identification result. However, this estimator may not perform very well in finite samples. That is why, in the second step, we regularize it to achieve better finite sample properties. Using sparsity, we show that the problem of estimating a small number of sets can be cast to ***the best subset regression problem*** (see, for instance, Bertsimas et al., 2016). As a result, the problem turns out to be a mixed-integer program that can be solved very quickly with modern optimization routines.

We apply our estimator to the Nielsen Homescan dataset. We study the RTE cereal market. This market has been studied in Nevo (2001) using aggregate datasets. The RTE cereal market is known to be highly concentrated, with high differentiation, large advertisement-to-sale ratios, and product innovation. All of these factors suggest high variability of choice sets because of consumer loyalty, geographical variation in product availability, and targeted advertisement campaigns. We exploit the high frequency of purchases in the Nielsen Homescan with roughly weekly time variation to uncover substantial heterogeneity in choice set variation across markets. Furthermore, we find evidence in favor of our sparsity assumption and show that ignoring this latent choice set heterogeneity leads to higher in absolute value estimates of price elasticities in a simple model of demand in the spirit of Berry et al. (1995) and Nevo (2001).

The closest work to ours is Crawford et al. (2021). We differ from their work in several respects. Crawford et al. (2021) mainly consider settings where either choice sets do not change (stable choice sets) or become larger over time (growing choice sets). They do not put restrictions on how choice sets for different DMs are related. We work with settings where choice sets DMs face are stable, but sparse across DMs. Next, our approach is fully nonparametric while Crawford et al. (2021) work with the stylized multinomial logit model of choice and impose parametric restrictions on the choice set distribution.[6] We also differ in that we allow for the endogenous selection of DMs into choice sets. That is, we allow for correlation between preferences and choice sets even after conditioning on covariates. This is an important distinction in our empirical application, since we find that consumers with different choice sets exhibit different choice behavior. Dardanoni et al. (2020) recover jointly the distribution of preferences and the distribution of consideration sets under parametric restrictions on the distribution of consideration sets. We do not require such parametric restrictions.[7] Lu (2014) and Barseghyan et al. (2021a) use only cross-sectional variation and set-identify the parametric distribution of preferences only.[8] We point-identify and estimate the joint distribution of preferences and choice sets nonparametrically using a cross-section of repeated choices.

Section 2 presents the model. Section 3 contains our main identification result. Section 4 outlines the main idea of new our estimation procedure. The details of it can be found in Appendix B. In Section 5, we present our empirical application. Finally, Section 6 concludes. All proofs can be found in Appendix A. Appendix C assesses the performance of our estimator in simulations. Appendix D contains additional estimation results.

[6] See also Goeree (2008) and Barseghyan et al. (2021b) for applications of consideration sets driven by item-dependent attention.

[7] In an alternative strand of the literature, Abaluck and Adams (2021) exploit parametric restrictions on preferences and consideration to achieve identification without panel datasets and exclusion restrictions by assuming asymmetries in the substitution matrix. Allen and Rehbeck (2019) show how the important models of limited consideration of Manzini and Mariotti (2014) and Brady and Rehbeck (2016) can fail to satisfy this property.

[8] Lu (2014) also provides a set of conditions that ensure that a system of moment inequalities he builds uniquely identifies the parameter of interest.

2. Model

We consider an environment where choices are made from a random latent finite choice set $\mathbf{D}$.[9] Every choice instance, $\mathbf{y}_s$ with $s \in \mathcal{S}$, maximizes random preferences that are captured by the random strict preference orders represented by random (indirect) utility functions $\mathbf{u} = \{\mathbf{u}_s\}_{s \; S}$. The set $\mathcal{S}$ captures different choice instances such as experimental trials, shopping trips, time periods, agents, among others. We assume that $S = |\mathcal{S}|$ is finite and does not grow with sample size.[10] The utility functions are defined over some grand choice set that contains $\mathbf{D}$ with probability 1. Without loss of generality, we assume that the grand choice set is $\mathcal{Y} = \{1, 2, \dots, Y\}$, where Y is a finite constant.

Let $\mathbf{x} \in X \subseteq \mathbb{R}^{d_x}$ denote the vector of observed covariates. The set of covariates depends on a particular application and can include decision maker-specific characteristics (e.g., age and gender) and choice-problem-specific characteristics (e.g., zip code, location of the store, day of the year, month, or time of the day).

Assumption 1 (Observables)**.** *The researcher observes (can consistently estimate) the joint distribution of* $(\mathbf{y}_s)_{s \; S}$ *and* $\mathbf{x}$*.*

Here, we provide some examples of environments that fit our primitives.

Example 1. Suppose that Y brands of a product (e.g. cereal) are available in a given location (market). Let $\mathbf{x} = ((\tilde{\mathbf{x}}_{y,s}^\top)_{y \; Y, s \; S}^\top, (\mathbf{r}_s)_{s \; S})^\top$ be the vector of observed covariates, where $\tilde{\mathbf{x}}_{y,s}$ is the vector of characteristics of product y at time s (e.g., price and package size); $\mathbf{r}_s$ is the vector of characteristics of a decision maker (henceforth, DM) (e.g., age and income level) and market (e.g., market identifier). The DM draws a latent choice set $\mathbf{D} \subseteq \mathcal{Y}$ and purchases a product $\mathbf{y}_s$ from that set at every time period s. The analyst observes a sample of n independently and identically distributed (i.i.d.) across DMs observations $\left\{\left(\mathbf{y}_s^{(i)}\right)_{s \; S}, \mathbf{x}^{(i)}\right\}_{i=1}^{n}$ drawn from a joint distribution of $(\mathbf{y}_s)_{s \; S}$ and $\mathbf{x}$ (a panel dataset).

[9] We use boldface font (e.g. $\mathbf{D}$) to denote random objects and regular font (e.g. D) for deterministic ones.

[10] $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$.

Example 2. Suppose that there are n geographical markets (streets) and Y different fast-food restaurants as in Currie et al. (2010).[11] Assume that at every market there are at least S consumers that are choosing from $\mathbf{D}$, which represents the set of fast-food restaurants available on the street. That is, every $s \in \mathcal{S}$ represents a consumer. Let $\mathbf{x}$ be the vector of observed consumers and market characteristics (e.g., age, income level, zip-code, average market income). The analyst observes a sample of choices of at least S consumers from n independent markets $\left(\left(\mathbf{y}_s^{(i)}\right)_{s\in\mathcal{S}}, \mathbf{x}^{(i)}\right)_{i=1}^{n}$ drawn from a joint distribution of $(\mathbf{y}_s)_{s\in\mathcal{S}}$ and $\mathbf{x}$.

While we work with the environments where the choice sets do not change over s (the random choice set $\mathbf{D}$ is not indexed by s), we still allow agents to make different choices in different choice instances. We impose the following two restrictions on the dependence structure across s. First, we assume that S is (strictly) ordered by $<$. Without loss of generality, let $\mathcal{S} = \{1, 2, \ldots, S\}$.

Assumption 2 (Markovianity)**.** *For all $s \in \mathcal{S}$, and x, $y_1, \ldots, y_s$, and D in the support*

$$\mathbb{P}\left(\mathbf{y}_s = y_s \mid \mathbf{y}_{s-1} = y_{s-1}, \ldots, \mathbf{y}_1 = y_1, \mathbf{D} = D, \mathbf{x} = x\right) = \mathbb{P}\left(\mathbf{y}_s = y_s \mid \mathbf{y}_{s-1} = y_{s-1}, \mathbf{D} = D, \mathbf{x} = x\right).$$

Assumption 2 is a standard markovianity assumption in panel data settings. It requires future choices to be independent of the past choices as long as one conditions on the current choice. We assume the dependence on one lagged choice only to simplify the exposition. Our framework can be easily extended to cases when distribution over choices can depend on longer choice histories.[12] It can also be modified for structures where $\mathcal{S}$ denotes a network and $s - 1$ denotes a neighborhood.

Example 1. (continued) Suppose that a DM, who faces a choice set $\mathbf{D} = D$, obtains the

[11] Currie et al. (2010) analyzes the effect of fast-food restaurants availability on obesity of children and pregnant women.

[12] Mbakop (2017) uses markovianity of order statistics to identify the distribution of private valuations in auction settings.

following indirect utility from purchasing product y at time s:

$$\mathbf{u}_{y,D,s} = \alpha_{y,D,s}(\mathbf{x}, \mathbf{y}_{s-1}) + \varepsilon_{y,D,s},$$

where $\{\alpha_{y,D,s}\}$ are unknown functions that map $X \times \mathcal{Y}$ to $\mathbb{R}$; and $\{\varepsilon_{y,D,s}\}$ are taste shocks that are independent across s, but potentially correlated across y and D. Moreover, εs are allowed to be correlated with $\mathbf{x}$ and $\mathbf{y}_{s-1}$. Functions $\alpha_{y,D,s}$ may include unknown product-time-choice set fixed effects. In this example, Assumption 2 is satisfied. Also note that two DMs may get different utilities from the same product at the same moment of time if their choice sets are different.

Assumption 2 is trivially satisfied if one assumes that choices are conditionally independent conditional on covariates and choice sets (i.e., no need to condition in the past choices) as in the classical treatment of the Random Utility Model (RUM, McFadden, 1973) with observed choice sets. For instance, it is often assumed in the analysis of differentiated products demand systems using individual level data (e.g., Lu, 2014, Crawford et al., 2021).[13] Formally, the following assumption, which does not require $\mathcal{S}$ to be ordered, implies Assumption 2.

Assumption 2 . *For all $s \in \mathcal{S}$, and x, $y_1, \dots, y_s$, and D in the support*

$$\mathbb{P}\left(\mathbf{y}_s = y_s \mid \mathbf{y}_{s-1} = y_{s-1}, \dots, \mathbf{y}_1 = y_1, \mathbf{D} = D, \mathbf{x} = x\right) = \mathbb{P}\left(\mathbf{y}_s = y_s \mid \mathbf{D} = D, \mathbf{x} = x\right).$$

Example 2. (continued) Suppose that consumer s, who faces a set of restaurants $\mathbf{D} = D$, obtain the following indirect utility from going to restaurant y:

$$\mathbf{u}_{y,D,s} = \tilde{\alpha}_{y,D,s}(\mathbf{x}) + \varepsilon_{y,D,s},$$

[13] A condition that requires independence of observed data across time periods is also standard in the analysis of differentiated products demand systems using market level data (e.g., Berry et al., 1995, Nevo, 2000, 2001). For instance, in this literature, the independent markets are often defined using a time interval (e.g., week, quarter, or year) and location (e.g, town or zip-code). As a result, it is often assumed that the market shares of a product in the same location, but different time periods conditional on observables are independent draws from the same distribution.

where $\{\tilde{\alpha}_{y,D,s}\}$ are unknown functions that map X to $\mathbb{R}$; and $\{\varepsilon_{y,D,s}\}$ are taste shocks that are independent across consumers s, but potentially are correlated across y and D. Moreover, εs are allowed to be correlated with $\mathbf{x}$. In this example, Assumption 2′ is satisfied.

The next assumption imposes stability on the conditional distribution of choices.

Assumption 3 (Distribution Stability). *For all $s, k \in \mathcal{S}$, and x, y, y', and D in the support*

$$\mathbb{P}\left(\mathbf{y}_s = y \mid \mathbf{y}_{s-1} = y', \mathbf{D} = D, \mathbf{x} = x\right) = \mathbb{P}\left(\mathbf{y}_k = y \mid \mathbf{y}_{k-1} = y', \mathbf{D} = D, \mathbf{x} = x\right).$$

This assumption is a form of stationarity of choice. Without Assumption 3 we can only identify the conditional distribution of choices conditional on the previous decision, choice sets, and covariates. In particular, for all y, y', x, and D in the support define

$$F_s^{\text{RUM}}(y \mid y', D, x) = \mathbb{P}\left(\mathbf{y}_s = y \mid \mathbf{y}_{s-1} = y', \mathbf{D} = D, \mathbf{x} = x\right).$$

Given Assumptions 2 and 3, F_s^{RUM} does not depend on s and we can drop the s subscript. Assumption 3 can be dropped if Assumption 2 is replaced by Assumption 2′.

Example 1. (continued) If for every $x \in X$ the conditional on $\mathbf{x} = x$ distribution of $(\varepsilon_{D,s,y})_{y \in Y, D \in \mathcal{D}_x}$, where $\mathcal{D}_x$ is the conditional support of $\mathbf{D}$ conditional on $\mathbf{x} = x$, does not depend on s, then Assumption 3 is satisfied.

Assumptions 2 and 3 imply that, after the choice set is realized, the choices of DMs are consistent with the classic RUM. In other words, after conditioning on the choice set, the choices at $s - 1$, and covariates, we can rewrite the conditional distribution of observed choices at any $s \in \mathcal{S}$ as a finite mixture model:

$$\Pr(\mathbf{y}_s = y \mid \mathbf{y}_{s-1} = y', \mathbf{x} = x) = \sum_{D \in \mathcal{D}_{x,y'}} m(D \mid x, y') F^{\text{RUM}}(y \mid y', D, x)$$

for all x, F, y', and y, where $m(D \mid x, y')$ is the conditional probability of $\mathbf{D} = D$ conditional on $\mathbf{x} = x$ and $\mathbf{y}_{s-1} = y$. Using the data on choices and covariates, the researcher is interested in recovering the conditional distribution of choice sets captured by m and the random utility maximization aspects of the model captured by F^{RUM}.

Next, we impose the following regularity condition on F^{RUM}.

Assumption 4 (Full Support)**.** *For every x, y , and D in the support $F^{\mathsf{RUM}}(y \mid y , D, x) > 0$ for every $y \in D$.*

Assumption 4 is a standard assumption in discrete choice literature: every alternative in every choice set is chosen with positive probability. McFadden (1973) pointed out that in finite samples, Assumption 4 is not testable, since zero market shares are not distinguishable from arbitrarily small but positive market shares. Additionally, if an alternative is never observed in the data, then it may be that this alternative either does not belong to any choice set or is always dominated by another alternative. Assumption 4 excludes such cases.

Example 1. (continued) If the support of random vector $(\varepsilon_{D,s,y})_{y \quad Y}$ is $\mathbb{R}^{Y}$ for all D and s (e.g. $\varepsilon_{D,s,y}$ are independent identically Type I extreme-value distributed), then Assumption 4 is satisfied.

We conclude this section by noting that Assumptions 2 and 2 imply that, apart from $\mathbf{D}$, there are no other sources of persistent across $\mathcal{S}$ unobserved heterogeneity. Similar to Crawford et al. (2021), we can easily extend our analysis to cases when this persistent unobserved heterogeneity has a discrete distribution (e.g., in panel settings we can allow for random coefficients with discrete support). In this case, however, we would need larger S. In this paper, we focus on choice sets as the main source of persistent latent heterogeneity because they have economic meaning and impose additional restrictions on the model (i.e., they are not just abstract latent types) that affect identification and estimation. Namely, under Assumption 4, $y \notin D$ implies that $F^{\mathsf{RUM}}(y \mid y , D, x) = 0$. We show how we use this special structure of choice sets in Sections 3 and 4.

3. Identification

3.1. Identification of m and F^{RUM}

Without observing previous choices, given the grand choice set $\mathcal{Y}$, the biggest possible support of $\mathbf{D}$ is $2^{Y}\setminus\{\emptyset\}$.[14] Unfortunately, the information contained in $Y=|\mathcal{Y}|$ outcomes over S choice instances may not be enough to identify the distribution supported on $2^{Y}-1=\left|2^{Y}\setminus\{\emptyset\}\right|$ points. Depending on the number of choice instances S, to pin down the distribution of choice sets nonparametrically, we may need to assume that some subsets of $2^{Y}\setminus\{\emptyset\}$ are never considered. That is, we may need to impose *sparsity* assumptions.

Let K denote the biggest integer that is less than or equal to $(S-3)/2$. If $K\geq 1$, then we can construct two nonoverlapping subsets of $\mathcal{S}$: $\mathcal{S}_1=\{1,\ldots,K\}$ and $\mathcal{S}_2=\{K+2,\ldots,2K+1\}$. Assume for a moment that $K=(S-3)/2$. Then $s=K+1$ separates $\mathcal{S}_1$ and $\mathcal{S}_2$; and $s=2K+2$ separates $\mathcal{S}_2$ from the last observation $\mathbf{y}_S$. For any $\mathcal{S}_i$, $i=1,2$, define $\mathbf{y}(\mathcal{S}_i)=(\mathbf{y}_s)_{s\ \mathcal{S}_i}\in\mathcal{Y}^{K}$. In other words, we partition the sequence $\{\mathbf{y}_s\}_{s\ S}$ into 2 random vectors and 3 random variables: $\mathbf{y}(\mathcal{S}_1),\mathbf{y}_{K+1},\mathbf{y}(\mathcal{S}_2),\mathbf{y}_{2K+2}$, and $\mathbf{y}_S$.

This partition allows us to establish the following lemma.

Lemma 1. *Under Assumption 2, $\mathbf{y}(\mathcal{S}_1),\mathbf{y}(\mathcal{S}_2)$, and $\mathbf{y}_{2K+3}$ are conditionally independent conditional on $\mathbf{y}_{K+1}$, $\mathbf{y}_{2K+2}$, $\mathbf{x}$, and $\mathbf{D}$.*

Lemma 1 allows us to construct 3 conditionally independent sets of choices from unobserved D.

Let G be the conditional probability mass function of $\mathbf{y}(\mathcal{S}_i)$ conditional on $\mathbf{y}_{K+1}$,$\mathbf{y}_{2K+2}$, $\mathbf{x}$, and $\mathbf{D}$. That is,

$$G\left(y^{K}\mid y,y\ ,D,x,\mathcal{S}_i\right)=\mathbb{P}\left(\mathbf{y}(\mathcal{S}_i)=y^{K}\mid\mathbf{y}_{K+1}=y,\mathbf{y}_{2K+2}=y\ ,\mathbf{D}=D,\mathbf{x}=x\right)$$

[14] $2^{\mathcal{Y}}$ denotes the set of all subsets of $\mathcal{Y}$.

for all y^K , y, y , x, and D in the support.

Assumption 5 (Linear Independence)**.** *For every x, y, y in the support, and $i \in \{1, 2\}$*

$$\{G\,(\cdot \mid y, y\ , D, x, \mathcal{S}_i)\}_{D\ \ D_{x,y,y'}}$$

is a collection of linearly independent functions.

Note that $K \geq 1$ if and only if $S \geq 5$. Hence, $S \geq 5$ is a *necessary* condition for Assumption 5 to be well-defined. Moreover, if one assumes that Assumption 2 is satisfied, then one can set K to be the biggest integer that is less than or equal to $(S - 1)/2$, $\mathcal{S}_1 = \{1, \dots, K\}$, and $\mathcal{S}_2 = \{K + 1, \dots, 2K\}$. In this case $S \geq 3$ becomes a necessary condition.

The rank conditions similar to Assumption 5 are standard in the missclassification literature and the literature on finite mixtures (see, for instance, Hu, 2008, Allman et al., 2009, Kasahara and Shimotsu, 2009, An et al., 2010, Bonhomme et al., 2014, Kasahara and Shimotsu, 2014). It essentially means that the variation in choice sets induces sufficient variation in the implied distributions over choices.[15] We discuss Assumption 5 in greater detail and provide sufficient conditions for it in the next section.

We are ready to state our main result.

Theorem 1. *Suppose Assumptions 1-5 hold. Then $m(\cdot \mid x)$ and $F^{\mathrm{RUM}}(\cdot \mid y\ , D, x)$ are identified for all x, y , and D in the support.*

Theorem 1 recovers nonparametrically the joint distribution of choice sets and choices. To the best of our knowledge, no other work on this topic achieves this. We emphasize that (i) we do not impose any structure on the statistical dependence between preferences and choice sets; (ii) we do not need to know the number of possible choice sets.

[15] In the context of auctions, a similar assumption for $K = 1$ has been made in An (2017), Mbakop (2017), and Luo (2020).

Intuitively, if we observe a DM who over time chooses either of two alternatives (e.g., $\{y_1, y_2, y_1\}$), most likely, under Assumption 4, she considers these two alternatives. Another DM who chooses more alternatives (e.g., $\{y_1, y_2, y_3\}$), most likely has a bigger choice set than the previous DM. This variation in choices across choice instances allows us to pin down the choice sets. We, however, do not assume that S grows, so even if a DM in the data always picks y_1 she may still consider all available alternatives. After the choice sets are identified, the variation in choices within a given choice set (i.e. across DMs that have the same choice set) allows us to identify the distribution of choices conditional on choice sets.

In other words, we have two competing forces in the model: consideration and preferences. Without the repeated choice structure of the data, in general, it is hard to say whether a good is picked less often because it is rarely considered or because it is rarely picked when considered. However, in our setting, if the good is picked by many DMs, but these DMs do not pick it often across choice problems, then we can conclude that this good is frequently considered, but on average is dominated by something else. In contrast, if a good is picked by a small number of DMs, but these DMs choose it often across choice problems, then the good is rarely considered, but if considered, then picked frequently.

In one of the steps of the proof of Theorem 1, we use the eigendecomposition argument of Hu (2008) and Hu et al. (2013). In Hu (2008), one needs to observe at least 3 choices. Since in our setting we allow for dependence between choices across decision problems, we need to observe the choices at least five times.[16] However, we do not need to impose any monotonicity restrictions on F^{RUM}. Another difference from Hu (2008) and Hu et al. (2013) is that we do not need to know the number of possible choice sets. It can be identified from the data.[17]

[16] Under Assumption $2'$, we need to observe at least 3 choices instead of 5.

[17] Also, in contrast to Hu et al. (2013), in general, we do not have a natural normalization for eigenvectors and cannot directly use them in our identification argument. Thus, we need to use a different argument to prove our result. See Appendix A for further details.

3.2. Linear Independence and Choice Sets

To better understand Assumption 5 consider the following simple examples. Suppose that $Y = \{1, 2, 3\}$ and $K = 1$ (i.e., S equals to 5 or 6). Then $\mathcal{S}_1 = \{1\}$ and for a fixed x (we drop it from the notation) and $y = 3$, the support of $\mathbf{D}$ conditional on x and y, $\mathcal{D}_{x,y'}$, is a subset of $\{\{3\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$ (it is known that $y = 3$ was chosen before). If we assume that $\mathcal{D}_{x,y'} = \{\{3\}, \{1, 3\}, \{2, 3\}\}$, then checking the linear independence condition is equivalent to checking whether the following matrix has full column rank:[18]

$y^K \setminus D$	{3}	{1,3}	{2,3}
1	0	$F^{\mathrm{RUM}}(1 \mid y, \{1, 3\})$	0
2	0	0	$F^{\mathrm{RUM}}(1 \mid y, \{2, 3\})$
3	1	$F^{\mathrm{RUM}}(3 \mid y, \{1, 3\})$	$F^{\mathrm{RUM}}(3 \mid y, \{2, 3\})$

This matrix has full column rank as long as Assumption 4 is satisfied. Using similar argument, we can conclude that if $|\mathcal{D}_x| \leq 2$, then Assumption 5 is generically satisfied in this example. However, if $\mathcal{D}_{x,y'} = \{\{3\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$, then the above matrix has less rows than columns (the number of possible choice sets is bigger than the number of possible values y^K can take) and the linear independence condition fails to hold. If, next, we increase K to $K = 2$ (i.e. S equals to 7 or 8), then for $\mathcal{S}_1 = \{1, 2\}$ the matrix becomes of the form (we only display 4 rows out of 9)

$y^K \setminus D$	{3}	{1,3}	{2,3}	{1,2,3}
$(1, 1)^{\mathsf{T}}$	0	×	0	×
$(1, 2)^{\mathsf{T}}$	0	0	0	×
$(2, 2)^{\mathsf{T}}$	0	0	×	×
$(3, 3)^{\mathsf{T}}$	1	×	×	×

[18] The columns of this matrix correspond to different elements of $\mathcal{D}_{x,y'}$. The rows correspond to different values y^K can take.

where × denotes the nonzero elements of the matrix, and Assumption 5 is satisfied. As these examples demonstrate, for the linear independence condition to hold, we need to have either enough choices from the same choice set or sparse $\mathcal{D}_{x,y'}$ (i.e. relatively small support of $\mathbf{D}$). The former can be formalized as the following sufficient condition for Assumption 5.

Proposition 1. *If Assumption 4 holds and $K \geq Y$, then Assumption 5 is satisfied.*

Proposition 1 provides a new, purely data-driven sufficient condition for the linear independence assumption. Proposition 1 is demanding in terms of observables – we need to observe DMs for at least $2Y + 3$ time periods. However, it does not impose *any* restrictions on $\mathcal{D}_{x,y'}$. For instance, $\mathcal{D}_{x,y'}$ is allowed to include all subsets of $\mathcal{Y}$ (including singleton sets) that contain y . To the best of our knowledge, this is the only identification result available in the literature that does not impose any restrictions on the random choice set. We note that having $d_{D,x,y'} = |\mathcal{D}_{x,y'}|$ points in the support of $\mathbf{y}(\mathcal{S}_i)$ is a necessary condition for Assumption 5 to hold. Thus, if $Y^K \geq 2^Y - 1$, then this necessary condition is satisfied. Hence, even if $K < Y$, Assumption 5 still may hold.

Now we investigate conditions that do not impose any restrictions on S, but restrict the support of choice sets by assuming a form of sparsity.

Proposition 2. *Suppose Assumption 4 and one of the following conditions hold:*

(i) *(Nestedness) For every x and y in the support, $\mathcal{D}_{x,y'}$ is a collection of nested sets. That is, $\mathcal{D}_{x,y'} = \{D_k\}_{k=1}^{d_{D,x,y'}}$ such that $D_{k-1} \subseteq D_k$ for $k = 2, \ldots, d_{D,x,y'}$.*

(ii) *(Excluded Choices) For every x and y in the support, $\mathcal{D}_{x,y'} = \{D_k\}_{k=1}^{d_{D,x,y'}}$ is such that for every k there exists $y_k \in \mathcal{Y}$ such that $y_k \in D_k$, but $y_k \notin D_{k'}$ for all $k \neq k$.*

Then Assumption 5 is satisfied.

Conditions in Proposition 2 are sufficient for Assumption 5 to hold because they impose enough structure on the matrix constructed from G to guarantee that it is of full rank. For instance, when $K = 1$, nestedness implies that the matrix $[G(y \mid y, D, x, \mathcal{S}_1)]_{y\ \ Y, D\ \ D_{x,y'}} =$

$[F^{\mathrm{RUM}}(y \mid y, D, x)]_{y \quad Y,D \quad D_{x,y'}}$ is triangular for any s.[19] In Appendix E, we give two examples of environments when conditions in Proposition 2 may hold.

Note that Propositions 1 and 2 do not assume that the identity or the number of support points $d_{D,x,y'}$ is known or is the same for all x and y. Distinct in terms of covariates DMs may draw their choice sets from completely different distributions. This is empirically relevant since it allows us to analyze consumers from very heterogeneous markets.

Propositions 1 and 2 display two complementary and non-overlapping identifying features of the model. Proposition 1 states that if you observe enough choices from the same choice set, then no restrictions on the support of the choice sets are needed. In contrast, Proposition 2 is not demanding in terms of data availability (i.e., S can be as small as 5), but requires more structure on the support of random consideration sets. We believe this trade-off is inevitable if one wants to achieve nonparametric identification of *both* the distribution of unobserved choice sets and the distribution of choices conditional on choice set.

Importantly, Assumption 5 can be satisfied in settings beyond the ones considered in Propositions 1 and 2. In particular, one of the implications of Assumption 5 is a restriction on the cardinality of $\mathcal{D}_{x,y'}$. Indeed, for $K = 1$, it must be that for all x, the number of points in the support of the random choice set, $d_{D,x,y'} = |\mathcal{D}_{x,y'}|$, has to be smaller than or equal to the total number of alternatives, i.e., $d_{D,x,y'} \leq Y$. In other words, the support of $\mathbf{D}$ needs to be sparse. In general, there is no need to impose a sparsity condition for identification of finite mixtures if the dependent variable is continuously distributed and the latent heterogeneity is discrete (e.g., Hu et al., 2013). In our setting, the dependent variable has finite support, thus, we have to reduce the dimensionality of the problem by bounding the cardinality of the support of the latent choice sets. This sparsity restriction can be satisfied in many empirical settings. Suppose for every x

$$\mathcal{D}_{x,y'} = \{D \subseteq \mathcal{Y} : \mathrm{L}_{x,y'} \subseteq D \subseteq \mathrm{U}_{x,y'}\},$$

[19] We use $[a_{ij}]_{i \in I, j \in J}$ to denote a matrix of the size $|I| \times |J|$ with entries of the form a_{ij}.

where $\mathrm{L}_{x,y'}$ and $\mathrm{U}_{x,y'}$ are some observed sets. Such a restriction on the choice sets appears in Conlon and Mortimer (2013), Lu (2014), and Gentry (2016). In this case,

$$d_{D,x,y'} \leq |\mathcal{Y}| \iff 2^{\left|\mathrm{U}_{x,y'} \setminus \mathrm{L}_{x,y'}\right|} \leq Y.$$

In the dataset used in Conlon and Mortimer (2013), $2^{\left|\mathrm{U}_{x,y'} \setminus \mathrm{L}_{x,y'}\right|} \leq 2^3 = 8$ (at most 3 stock-out events) while total number of products considered is 44.[20]

4. Outline of the Estimation Procedure

In this section, we informally describe the main idea behind our new estimator. The full details about the estimator and its performance in simulations can be found in Appendices B and C.

For simplicity, consider the case with $S = 5$ and drop dependence on covariates. Recall that under Assumption 5, Lemma 1 implies that $\mathbf{y}_1$, $\mathbf{y}_3$, and $\mathbf{y}_5$ are conditionally independence conditional on $\mathbf{y}_2$, $\mathbf{y}_4$, and $\mathbf{D}$. Hence, the conditional distribution of $\mathbf{y}_1$, $\mathbf{y}_3$, and $\mathbf{y}_5$ conditional on $\mathbf{y}_2 = y$ and $\mathbf{y}_3 = y$ can be written as

$$P(y_1, y_3, y_5) = \sum_{D} P(y_1|D)P(y_3|D)P(y_5|D)m(D),$$

where we dropped the dependence on y and y to simplify the notation. This is essentially a finite mixture. However, it has an important restriction: latent D is not an abstract type, but a set. Hence, under Assumption 4, $P(y_k|D) = 0$ if and only if $y_k \notin D$. Moreover, if S is not big enough (see Proposition 1), then we need the support for $\mathbf{D}$ to be sparse.

We show that both these features of our environment ($\mathbf{D}$ is a choice set and its support is

[20] Conlon and Mortimer (2013) study demand estimation using data on product availability from vending machines.

sparse) can be accommodated in an estimation procedure. Essentially, our procedure consists of two steps. First, we estimate the finite mixture model without imposing sparsity, but restricting D to be a set. Second, to impose sparsity, we pick a small number of sets from the first step that minimize a sum of squared residuals. The second step can be formulated as a *best subset regression problem*. Our estimator is computationally fast and performs well in simulations (see Appendix C for further details).

5. Illustrative Empirical Application: Brand Choice Set Variation and Price Elasticity in the Ready-to-Eat Cereal Market.

In this section, we illustrate the applicability of our framework by studying the effects of brand choice set variation at the market level on consumption of the RTE cereal market using the Nielsen Homescan Panel (Homescan). The RTE cereal industry has been previously analyzed under the assumption that DMs consider all available brands (e.g., Nevo, 2000, 2001). We want to analyze to what extent this assumption is valid and what implications it has on parameters of interest such as price elasticity. (However, our results are not fully comparable to these previous results since we have a different dataset with richer variation needed in this setup.) The RTE cereal market is known to be highly concentrated, with high differentiation, large advertisement-to-sale ratios, and product innovation (Nevo, 2001). All of these factors suggests high variability in choice sets because of consumer loyalty, geographical variation in product availability, and targeted advertisement campaigns. We confirm this insight in our quantitative analysis and uncover substantial choice set variation. Moreover, we show that ignoring this latent choice set heterogeneity leads to higher in absolute value estimates of price elasticities in a simple model of demand.

To simplify the analysis and due to data limitations, in our application we strengthen

Assumption 2 by assuming Assumption 2 (i.e. conditional on the choice sets and covariates choices today do not depend on choices made yesterday). We believe this assumption is reasonable in our illustrative empirical application.[21] Importantly, it allows us to use shorter panels ($S = 3$ instead of $S = 5$), which greatly reduces our data requirements.

First, we describe the dataset and empirical validity of our assumptions. Next, we estimate the unobserved distribution of the choice sets and the distribution of choices given the choice sets by applying our new estimator outlined in the previous section. See Appendices B and C for further details. Finally, in the spirit of Berry et al. (1995), Nevo (2000, 2001), we estimate a simple parametric model of the RTE cereal demand.

Data Construction

We consider $Y = 5$ brands of RTE cereal: Store brand (CTL), General Mills (GM), Kellogg (K), Quaker (Q), and other brands of RTE cereal (O). We record only purchases of households that buy 1 brand per trip.[22] We focus on households that are frequent buyers. We define frequent buyers as households that buy at least one RTE cereal in $S = 3$ consecutive trips.[23] The majority of households in our sample makes 1 trip per week. Thus, the predominant time frequency of our dataset is weekly. We focus on trips and households present in the Homescan in 2016-2018.[24] We include only the 3 earliest consecutive trips per household. Each household appears only once in the cross-section.[25] We consider a balanced panel by dropping any household that does not have 3 consecutive trips in a given year. We end up

[21] We consider a short time window, which we believe allows to disregard habit formation. Also, we consider frequent buyers that usually buy a few units of cereal each shopping trip and then repeat their shopping trip weekly.

[22] The households that buy more than 1 brand in a given trip are dropped from the sample to avoid dealing with bundling.

[23] A trip is an instance of a household member going to a store and purchasing at least one item that is recorded in the Homescan.

[24] We eliminate from our sample trips happening in December and January, because of their strong seasonality effects on RTE cereal consumption.

[25] To ensure this we consider the first 3 trips per year per household. Then we create a unified panel with the information of years $2016 - 2018$, and we balance the panel keeping only the first 3 trips. Hence, if any household appears in all three years, we keep only its 2016 observations.

having $S = 3$ consecutive choices of $n = 47,509$ households.

There are only 2 product characteristics available in the Homescan and Nielsen Retail Scanner: price of a unit (USD) and size of it (ounces).[26] The dataset also contains information on zip-codes for every household/purchase in the sample. We use the Nielsen Retail Scanner and the Homescan to construct the dataset on prices and sizes by pooling all the information on prices per UPC code (barcode) of all RTE cereals by week and location (3 digit zip-code).[27] Then we compute the mean price of every brand at every location.[28] Brand-location size variable is built similarly. As a result, given the information on the location of every household, we match every purchase with the price and size.

Despite having a relatively large sample, there are too many 3-digit level zip-codes to treat them as markets. Thus, to increase the number of observations per market, similar to Nevo (2001), we use prices and geographic coordinates (i.e. longitude and latitude) of every location to define markets. In particular, we define a market by employing K-means clustering with the Euclidean norm using centroids based on prices and geographic location. In other words, we group together households that live close to each other and face similar prices. We initialize the K-means and fix the number of markets using the 3-digit zip-code. All locations with less than 2000 households are collapsed to a single dummy location.[29] In total we obtain 34 markets.[30] The map depicting the geographical locations of the markets can be found in Appendix D. Finally, we aggregate prices on the market-brand level.

Since the dataset contains information on the household's income, the age of the household's head, and the size of the household, we also compute the average (on the market level) income, the age of the head of the household, and the household size. We use these

[26] We also know barcodes of every purchase. Unfortunately, it is hard to match these barcodes with actual products to obtain additional product characteristics since these barcodes change over time and some products are not produced anymore.

[27] To obtain prices in the Homescan we use the paid price (including discounts) divided by the number of units, and we drop from our sample those that pay a zero price after discount.

[28] We average across weeks to diminish measurement error in prices and because there are some missing prices per brand.

[29] This quantity was chosen on the basis of simulations, to ensure a sufficiently high number of observations per market.

[30] We obtain qualitatively the same results when we increased the number of markets to 72.

demographics in our analysis of own-price elasticity. Summary statistics for demographic variables are provided in Table 1

Table 1 – Summary Statistics of Demographic Variables

Variable	Mean	Median	Std	Min	Max
Average Age (years)	54.33	54.27	1.55	49,87	57.12
Average Income (USD)	23,333	22,503	4,506.25	14,543.4	32,580
Average HH Size	2.7	2.7	0.12	2.49	3.18

Notes: These summary statistics are computed for 34 markets. For instance, the minimum market average age is the smallest among 34 markets market-average age, not the age of youngest head of the household in the sample.

Since we mainly focus on frequent buyers (i.e., weekly purchases) of RTE cereals, we believe that the assumptions that choice sets are stable across time and choices are conditionally independent across time are satisfied. It is less likely that the choice set changes within a short time horizon[31] and there are unobserved shocks to preferences over cereals that are correlated across time. That is, we believe that after controlling for observed characteristics and choice sets, all variation in choices is driven by idiosyncratic taste shocks.[32] Moreover, similar assumptions (or their stronger versions) are usually made in the literature on estimation of demand systems using individual and market level data. For instance, Nevo (2001) assumes independence of measurements of market shares across markets, where markets are defined as a pair of location and time window, and known fixed choice sets.

[31] For instance, product loyalty can be thought as a special case of system-1 thinking (Kahneman, 2003), hence, it is revised less often than utility maximization. Also, any product innovation within brands may take more time than three weeks which is the modal time window in our dataset.

[32] In our application we use an additive random utility framework where the mean utility is assumed to be stable in the time-window, but taste shocks are idiosyncratic. Arguably, in a short time-window it is less likely that the DM adapts her mean utility due to structural environmental changes. In addition, Saito and Lu (2020) show that the continuation problem can be ignored if preferences are separable and additive in time, which is a standard assumption in applied work.

Nonparametric Estimation of Consideration Sets and True Market Shares

We estimate m and F^{RUM} for every market. Let $\hat{m}(D|j)$ denote the estimated probability that set D is considered in market j. Since after averaging there is no within market variation, conditioning on the market also conditions on the product and consumers characteristics on top of unobserved market fixed effects.

Using the estimated $\hat{m}$, first, we find that *all* markets have less than 5 sets that are considered by more than 10 percent of households in the market. That is,

$$\sum_{j} \mathbb{1}\left(\sum_{D} \mathbb{1}\left(\hat{m}(D|j) > 0.1\right) \geq 5\right) = 0.$$

Even if we lower the threshold to 5 percent, more than 20 percent of markets have less than 5 choice sets. That is, among 5 estimated sets at least 1 set is considered by less than 10 percent of population in every market, and a sizable fraction of markets has at least one set that is faced by less than 5 percent of consumers. These findings lend support to our sparsity assumption.

Next, we compute the estimated proportion of individuals in the sample who considered sets of a given cardinality l as $\sum_{j,D} \mathbb{1}\left(|D| = l\right)\hat{m}(D|j)w_j$, where $w_j = N_j/N$ is a fraction of the whole sample (of size N) that is coming from market j (of size N_j). As Figure 1 demonstrates, sets of all sizes are considered. The vast majority (about 70 percent) of the sample considered sets of cardinality 4 and 5. Given that most likely all 5 brands are usually present, these individuals can be thought of as full consideration individuals that we usually work with in discrete choice settings. However, about 16 percent of DMs only considered one brand. These are super loyal consumers that always purchase the same brand no matter what.

Next, we consider the composition of sets of cardinality 1 and 4 (there is only one set of size 5). The results are presented in Figure 2. Interestingly, Store brand (CTL) is never considered alone. The rest of the brands are almost equally likely considered by those who

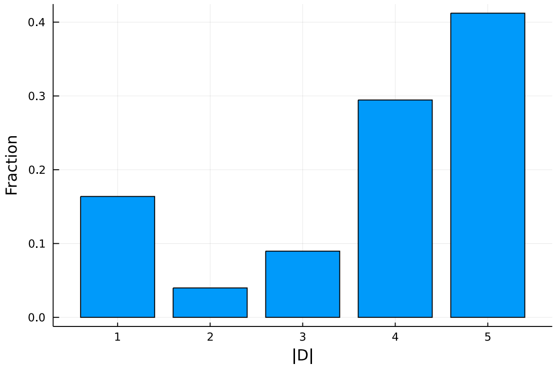

Figure 1 – Proportion of Individuals Considering Sets of Given Cardinality: $|D|$ denotes the size of the choice set.

only look at one brand (Quaker has the smallest share of about 19 percent). Among those who considered sets of size 4, almost half considered everything but Quaker. The rest of sets of cardinality 4 have similar shares.

Next, we compute the fraction of DMs who paid attention to a set that contains a given brand b as $\square_{j,D} \mathbb{1}(b \in D)\hat{m}(D|j)w_j$. Similar to Figure 2, Figure 3 indicates that Quaker is considered less often (about 58 percent of DMs) than other brands (about 80 percent of DMs).

Finally, there are just 2 sets that attract more than 5 percent of DMs: the set that contains all 5 brands (about 40 percent of DMs) and the set that contains all brands but Quaker (about 15 percent).

Overall, we can conclude that although most DMs seem to consider almost all brands, there is a sizeable fraction of those who only consider one brand. Moreover, Quaker is considered less often than other brands and Store brand is always considered with other alternatives.

Next, we consider the estimates of F^{RUM} per market and consideration set. To simplify the exposition, we focus on the 5 biggest (in terms of numbers of observations)markets ($j \in \{1, 2, 13, 16, 18\}$). Tables 2 and 3 display estimated market shares of all brands assuming that every DM considered all brands (i.e. model with full consideration) and estimated market shares obtained via our procedure, respectively. The largest difference between the

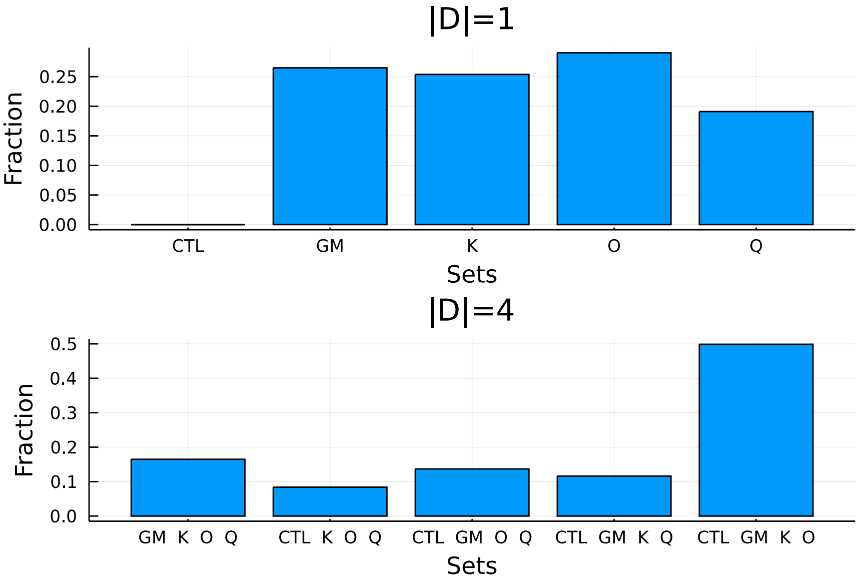

Figure 2 – distribution of Sets of Given Cardinality: $|D|$ denotes the size of the choice set. CTL=Store brand, GM=General Mills, K=Kellogg, Q=Quaker , O=other brands.

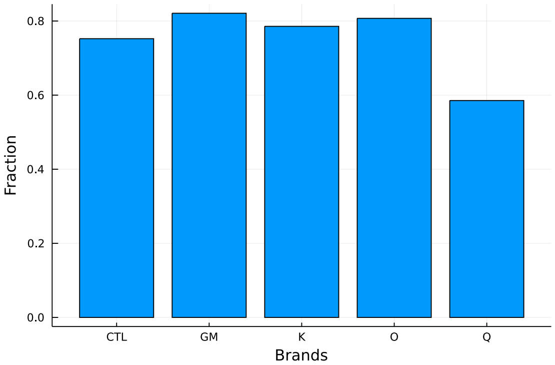

Figure 3 – Proportion of Individuals Considering a Brand: CTL=Store brand, GM=General Mills, K=Kellogg, Q=Quaker , O=other brands.

two shares is for Kellogg in market 16. It is about 33 percentage points or 200 percent. Interestingly, the results of our estimations imply that in market 18 there were no DMs who considered all 5 brands. If we assume that rankings of shares correspond to mean utilities of brands (i.e. additive random utility), then Table 2 implies that, on average, General Mills is the most preferred brand in 4 out of 5 biggest markets if one assumes full consideration. If one allows for unobserved choice sets, then General Mills is the best alternative only in market 16. Kellogg is the best brand in markets 1 and 2.

Table 2 – Full Consideration Market Shares Assuming Observed Choice Sets

Brand/Market	1	2	13	16	18
CTL	0.15	0.152	0.207	0.161	0.123
GM	0.316	0.293	0.269	0.311	0.292
K	0.283	0.286	0.293	0.279	0.26
O	0.183	0.219	0.169	0.186	0.249
Q	0.069	0.05	0.062	0.063	0.076

Notes: Results are rounded to 3 digits.

Table 3 – Full Consideration Market Shares Assuming Unobserved Choice Sets

Brand/Market	1	2	13	16	18
CTL	0.1	0.099	0.123	0.117	n/a
GM	0.334	0.319	0.219	0.372	n/a
K	0.346	0.32	0.097	0.226	n/a
O	0.132	0.202	0.497	0.196	n/a
Q	0.088	0.061	0.064	0.089	n/a

Notes: Results are rounded to 3 digits.

Moreover, different in terms of choice sets DMs display different preferences over brands. For instance, in market 1, those who consider all 5 brands prefer Kellogg over all other brands. At the same time, those who do not consider Quaker predominantly buy other brands of cereals (see Table 4). This emphasizes the importance of allowing for correlation between preferences and choice set when estimating the model. Overall, our estimates suggest that allowing for unobserved choice sets affects not only point estimates of market shares, but

Table 4 – Market Shares for Two Choice Sets in Market 1

Brand/Set	{CTL,GM,L,Q,O}	{CTL,GM,L,O}
CTL	0.1	0.081
GM	0.334	0.161
K	0.346	0.11
O	0.132	0.648
Q	0.088	0.0

Notes: Results are rounded to 3 digits.

also rankings of brands. Moreover, DMs that consider different choice sets may have different preferences.

Parametric Estimation of Price Elasticity with Hidden Choice Set Variation

Note that the estimated F^{RUM} for every market is a collection of market shares of every alternative in a given market for different groups of consumers. For instance, $\mathbf{sh}_{y,D,j,s} = \hat{F}^{\mathsf{RUM}}_s(y|D,j)$ is the estimated market share of brand y among those consumers that face the choice set D at time s in market j and decided to purchase something.[33] So, we can proceed as if the estimated market shares are the true market shares and parametrize F^{RUM}.[34] In our application, we take the standard logit specification of F^{RUM}. In particular, following Nevo (2001), we assume that the random utility that consumer i gets from brand y in choice set D at time s in market j is

$$\alpha_{y,D} + \beta_D \mathbf{p}_{y,j} + \mathbf{r}_j^{\mathsf{T}} \gamma_{y,D} + \xi_{y,D} + \Delta\xi_{y,D,j,s} + \varepsilon_{i,y,D,s},$$

where $\mathbf{p}_{y,j}$ is the average market price of brand y in market j; $\mathbf{r}_j$ is the vector of market demographics that consists of the average household income, the average age of the household head, and the average household size in the market. Unobserved by the analyst, the

[33] Our DMs are RTE cereal frequent buyers. Hence, we do not allow for the option of not buying anything.

[34] The market shares computed directly from the data are not the true market shares but rather a mixture of the market shares from different choice sets.

market/choice set/time specific quality of brand y, which is potentially correlated with $\mathbf{p}_{y,j}$, is captured by $\xi_{y,D} + \Delta\xi_{y,D,j,s}$. The first term is the mean quality of a product y in a choice set D. The second term is the mean-zero choice set/time/brand specific deviation from that quality; $\varepsilon_{y,D,s}$ is the additive random shock that is independent from all other variables. These shocks are i.i.d. with a Type I extreme-value distribution. These assumptions reduce the model to the well-known (multinomial) Logit model.

Note that in this setup, Assumption 2 is satisfied if $\Delta\xi_{y,D,j,s}$s are conditionally independent across time conditionally on choice sets, the observables, and the market identity. Again, given the high-frequency of our dataset, we think the conditional independence assumption is reasonable in our setting as roughly 3 weeks of purchases would not be enough time to affect habit formation, which is the main source of inertia.

Here, for simplicity, we assume the Logit model and that $\alpha_{y,D}$, $\gamma_{y,D}$, and β_D are fixed parameters. The model can be extended to the Generalized Extreme Value model (McFadden et al., 1978), which includes the Nested Logit model, and to the case when coefficients are random (e.g, Nevo, 2001). Moreover, since we are estimating shares conditional on the market, after the shares are estimated, one can add as many market level covariates to the model as she wants.

The parameter of interest is β_D that can vary with choice sets. This parameter captures the price elasticity of demand for all brands. Note that since it is indexed by the choice set D we allow for correlation between preferences and choice sets. We use variation across time and markets to estimate β_D.[35]

Our parametric specification implies that for any $y, \bar{y} \in D$

$$\Delta\xi_{y,\bar{y},D,j,s} = \Delta\xi_{y,D,j,s} - \Delta\xi_{\bar{y},D,j,s} = \log\left(\frac{\mathbf{sh}_{y,D,j,s}}{\mathbf{sh}_{\bar{y},D,j,s}}\right) - \alpha_{y,\bar{y},D} - \beta_D(\mathbf{p}_{y,j} - \mathbf{p}_{\bar{y},j}) - \mathbf{r}_j^{\mathsf{T}}\gamma_{y,\bar{y},D},$$

where $\alpha_{y,\bar{y},D} = \alpha_{y,D} - \alpha_{\bar{y},D} + \xi_{y,D} - \xi_{\bar{y},D}$ and $\gamma_{y,\bar{y},D} = \gamma_{y,D} - \gamma_{\bar{y},D}$. Thus, if $\Delta\xi_{y,\bar{y},D,j,s}$ was not correlated with prices, then we could have used the ordinary least squares estimator to

[35] One can allow β to vary across time. In this case, one would have to use only the market variation.

consistently estimate β_D. However, because of the price endogeneity, we use instruments and the two-step efficient Generalized Method of Moments (GMM) estimator. In particular, following Berry et al. (1995), Nevo (2000), and Nevo (2001), we construct two instruments: average product characteristics (i.e., size) of competing brands and average across neighboring markets price of the brand.[36]

First, we estimate the model assuming that there is no variation in choice sets. That is, every DM faces all 5 brands (Direct estimator). Next we estimate β_D using our procedure for two most popular choice sets: {CTL, GM, K, O, Q} and {CTL, GM, K, O}. In both cases, we use other brands of cereal (O) as the base option $\bar{y}$. The results of estimation are presented in Table 5.[37]

Table 5 – Estimates of β_D

	Direct	{CTL, GM, K, O, Q}	{CTL, GM, K, O}
β_D	-16.11	-5.84	-56.46

Notes: Results are rounded to 2 digits.

There is substantial heterogeneity between 2 types of DMs. The ones that consider all 5 brands have the price coefficient approximately 3 times smaller in magnitude than the direct estimate. The DMs who consider all brands but Quaker have substantially larger in magnitude price coefficient. Given these estimates of the price coefficient, we can compute the implied own-price elasticities under the assumption that the distribution over the choice sets m does not depend on prices.[38] Formally, the own-price elasticity of brand y in market j at time t is defined as[39]

$$\text{Elas}_{y,j,s} = \frac{p_{y,j}}{sh_{y,j,s}} \frac{\partial sh_{y,j,s}}{\partial p_{y,j}} = \frac{p_{y,j}}{sh_{y,j,s}} \underset{D}{\square} \frac{\partial sh_{y,D,j,s}}{\partial p_{y,j}} m(D|j),$$

[36] The details of construction of instruments can be found in our replication files.

[37] In Appendix D, we construct the standard errors for these estimates under the assumption that there is no estimation error in the estimated market shares and all uncertainty is coming from variation in observed covariates.

[38] See Goeree (2008) for similar exclusion restrictions.

[39] Without exclusion restrictions, one would also need to take into account the derivatives of m with respect to prices.

where $sh_{y,j,s}$ is the observed share of brand y in market j at time s. If there is no choice set variation, then, under our parametrization, $\text{Elas}_{y,j,s} = \beta_{\text{Naive}} p_{y,j}(1 - sh_{y,j,s})$. With choice set variation and under the assumption that prices do not affect choice set probabilities,

$$\text{Elas}_{y,j,s} = \sum_{D} \frac{sh_{y,D,j,s}}{sh_{y,j,s}} \text{Elas}_{y,D,j,s} m(D|j),$$

where $\text{Elas}_{y,D,j,s} = \beta_D p_{y,j}(1 - sh_{y,D,j,s})$.

Since elasticities may not be constant across markets we report own-price elasticities for the largest in terms of observations market (Market 1) in Table 6. In the first column, we use estimates of the price coefficient assuming that there is no choice set variation (Direct). The second column is computed using our estimates of the price coefficients for different choice sets. The third and the last column report elasticities for those who consider all 5 brands or do not consider Quaker, respectively (i.e., $\text{Elas}_{y,D,j,t}$).

Table 6 – Estimates of Own-Price Elasticities in Market 1

	Direct	Choice Set Variation	{CTL, GM, K, O, Q}	{CTL, GM, K, O}
CTL	-1.98	-0.98	-0.76	-7.51
GM	-2.34	-1.39	-0.83	-10.06
K	-2.03	-1.08	-0.67	-8.81
O	-2.25	-2.35	-0.87	-3.39
Q	-2.4	-0.7	-0.85	0

Notes: The first column is computed assuming that all consumers face all 5 brands. The second column is computed assuming choice set variation. The third column is computed for those consumers who consider all 5 brands. The last column is computed for those consumers who do not consider Quaker. Results are rounded to 2 digits.

The estimates of the own-price elasticities that assumes no choice set variation are similar to ones in Nevo (2001).[40] However, the demand of those considering all brands is substantially less elastic than those who do not consider Quaker. That is, we find substantial unobserved heterogeneity in how consumers react to price changes. As a result of this heterogeneity, the implied own-price elasticity that takes into account the choice set variation is smaller for

[40] In Appendix D, we report the median across markets own-price elasticities. The results are qualitatively the same.

almost all brands. For Quaker, given that it is not considered by a large group of consumers (about 83 percent), the difference is more than three fold.

Estimating own-price elasticities without considering hidden categorization/menu variation could lead to higher in absolute values estimates. Here, frequent buyers purchase a cereal from a particular category of brands. For instance, some consumers could always avoid Quaker cereal or only consider General Mills. Others, however, may consider everything. In general, these frequent buyers of RTE cereal may have strong opinions about what they like to consider and what they avoid. Also, they are exposed to advertisement, promotions, and other factors that affect the category of brands they consider. Our approach remains completely flexible with respect to the particular story that leads to the formation of a category of brands but imposes a sparsity restriction. Namely, conditional on a market, there can be at most 5 choice sets in each market.

6. Conclusion

In this paper, we show that observing three or more choices from the same latent choice set is sufficient to nonparametrically identify and consistently estimate the joint distribution of choice sets and choices in discrete-choice models when choice sets are not observable. Our main result requires a linear independence condition on the conditional distribution of choices. This condition is satisfied when either there are enough observed choices from the same choice set or the support of choice sets is sparse. The application of our computationally efficient estimator to a scanner dataset indicates that there is a substantial unobserved choice set heterogeneity and correlation between preferences and choice sets that can contaminate estimates of the standard parameters of interest (e.g., own-price elasticities).

References

Abaluck, Jason and Abi Adams (2021) "What do consumers consider before they choose? Identification from asymmetric demand responses," *The Quarterly Journal of Economics*, Accepted.

Allen, Roy and John Rehbeck (2019) "Revealed Stochastic Choice with Attributes," *Available at SSRN 2818041*.

Allman, Elizabeth S, Catherine Matias, John A Rhodes et al. (2009) "Identifiability of parameters in latent structure models with many observed variables," *The Annals of Statistics*, 37 (6A), 3099–3132.

An, Yonghong (2017) "Identification of first-price auctions with non-equilibrium beliefs: A measurement error approach," *Journal of econometrics*, 200 (2), 326–343.

An, Yonghong, Yingyao Hu, and Matthew Shum (2010) "Estimating first-price auctions with an unknown number of bidders: A misclassification approach," *Journal of Econometrics*, 157 (2), 328–341.

Barseghyan, Levon, Maura Coughlin, Francesca Molinari, and Joshua C Teitelbaum (2021a) "Heterogeneous choice sets and preferences," *Econometrica*, forthcoming.

Barseghyan, Levon, Francesca Molinari, and Matthew Thirkettle (2021b) "Discrete choice under risk with limited consideration," *American Economic Review*, forthcoming.

Berry, Steven, James Levinsohn, and Ariel Pakes (1995) "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, 841–890.

Bertsimas, Dimitris, Angela King, Rahul Mazumder et al. (2016) "Best subset selection via a modern optimization lens," *Annals of statistics*, 44 (2), 813–852.

Block, Henry David and Jacob Marschak (1960) “Random orderings and stochastic theories of responses,” *Contributions to probability and statistics*, 2, 97–132.

Bonhomme, Stéphane, Koen Jochmans, and Jean-Marc Robin (2014) “Nonparametric estimation of finite mixtures,”Technical report, cemmap working paper.

——— (2016) “Non-parametric estimation of finite mixtures from repeated measurements,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (1), 211–229.

Brady, Richard L. and John Rehbeck (2016) “Menu-Dependent Stochastic Feasibility,” *Econometrica*, 84 (3), 1203–1223.

Chen, Xiaohong (2007) “Large sample sieve estimation of semi-nonparametric models,” *Handbook of econometrics*, 6, 5549–5632.

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

Cragg, John G and Stephen G Donald (1997) “Inferring the rank of a matrix,” *Journal of econometrics*, 76 (1-2), 223–250.

Crawford, Gregory S, Rachel Griffith, and Alessandro Iaria (2021) “A survey of preference estimation with unobserved choice set heterogeneity,” *Journal of Econometrics*, 222 (1), 4–43.

Currie, Janet, Stefano DellaVigna, Enrico Moretti, and Vikram Pathania (2010) “The effect of fast food restaurants on obesity and weight gain,” *American Economic Journal: Economic Policy*, 2 (3), 32–63.

Dardanoni, Valentino, Paola Manzini, Marco Mariotti, and Christopher J Tyson (2020) “Inferring cognitive heterogeneity from aggregate choices,” *Econometrica*, 88 (3), 1269–1296.

Falmagne, J. C. (1978) “A representation theorem for finite random scale systems,” *Journal of Mathematical Psychology*, 18 (1), 52–72, 10.1016/0022-2496(78)90048-2.

Fudenberg, Drew, Ryota Iijima, and Tomasz Strzalecki (2015) “Stochastic choice and revealed perturbed utility,” *Econometrica*, 83 (6), 2371–2409.

Gentry, Matthew L (2016) “Displays, sales, and in-store search in retail markets,” *Working paper*.

Goeree, Michelle Sovinsky (2008) “Limited information and advertising in the US personal computer industry,” *Econometrica*, 76 (5), 1017–1074.

Gul, Faruk and Wolfgang Pesendorfer (2006) “Random expected utility,” *Econometrica*, 74 (1), 121–146.

Hall, Peter and Xiao-Hua Zhou (2003) “Nonparametric estimation of component distributions in a multivariate mixture,” *The annals of statistics*, 31 (1), 201–224.

Hickman, William and Julie Holland Mortimer (2016) “Demand estimation with availability variation.,” *Handbook on the Economics of Retailing and Distribution*, 306.

Hu, Yingyao (2008) “Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution,” *Journal of Econometrics*, 144 (1), 27–61.

Hu, Yingyao, David McAdams, and Matthew Shum (2013) “Identification of first-price auctions with non-separable unobserved heterogeneity,” *Journal of Econometrics*, 174 (2), 186–193.

Kahneman, Daniel (2003) “A perspective on judgment and choice: mapping bounded rationality.,” *American psychologist*, 58 (9), 697.

Kasahara, Hiroyuki and Katsumi Shimotsu (2009) “Nonparametric identification of finite mixture models of dynamic discrete choices,” *Econometrica*, 77 (1), 135–175.

——— (2014) “Non-parametric identification and estimation of the number of components in multivariate mixtures,” *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 97–111.

Kashaev, Nail (2020) “Identification and estimation of discrete outcome models with latent special covariates,” *working paper*.

Kitamura, Yuichi and Louise Laage (2018) “Nonparametric analysis of finite mixtures,” *arXiv preprint arXiv:1811.02727*.

Li, Qi and Jeffrey Scott Racine (2007) *Nonparametric econometrics: theory and practice*: Princeton University Press.

Lu, Zhentong (2014) “A Moment Inequality Approach to Estimating Multinomial Choice Models with Unobserved Consideration Sets,” *Working paper*.

Luce, R Duncan (1959) *Individual choice behavior: A theoretical analysis*: Wiley.

Luo, Yao (2020) “Unobserved heterogeneity in auctions under restricted stochastic dominance,” *Journal of Econometrics*, 216 (2), 354–374.

Manski, Charles F (1977) “The structure of random utility models,” *Theory and decision*, 8 (3), 229.

Manzini, Paola and Marco Mariotti (2014) “Stochastic Choice and Consideration Sets,” *Econometrica*, 82 (3), 1153–1176, 10.3982/ECTA10575.

Mbakop, Eric (2017) “Identification of auctions with incomplete bid data in the presence of unobserved heterogeneity,”Technical report, Working paper, Northwestern University.

McFadden, Daniel et al. (1978) “Modelling the choice of residential location.”

McFadden, Daniel (1973) “Conditional Logit Analysis of Qualitative Choice Behavior,” *Frontiers in Econometrics*, 105–142.

McFadden, Daniel and Marcel K Richter (1990) “Stochastic rationality and revealed stochastic preference,” *Preferences, Uncertainty, and Optimality, Essays in Honor of Leo Hurwicz, Westview Press: Boulder, CO*, 161–186.

Nevo, Aviv (2000) “A practitioner’s guide to estimation of random-coefficients logit models of demand,” *Journal of economics & management strategy*, 9 (4), 513–548.

——— (2001) “Measuring market power in the ready-to-eat cereal industry,” *Econometrica*, 69 (2), 307–342.

Ray, Debraj ⓡ Arthur Robson (2018) “Certified random: A new order for coauthorship,” *American Economic Review*, 108 (2), 489–520.

Saito, Kota and Jay Lu (2020) “Repeated Choice: A Theory of Stochastic Intertemporal Preferences.”

Train, Kenneth E (2009) *Discrete choice methods with simulation*: Cambridge university press.

Xiao, Ruli (2018) “Identification and estimation of incomplete information games with multiple equilibria,” *Journal of Econometrics*, 203 (2), 328–343.

Appendices for "Identification and Estimation of Discrete Choice Models with Unobserved Choice Sets."

A. Proofs

A.1. Proof of Lemma 1

Assume that $K = 1$. Also, to simplify the exposition, we drop $\mathbf{D}$ and $\mathbf{x}$ and write $\mathbb{P}(y_s)$ instead of $\mathbb{P}(\mathbf{y}_s = y_s)$. First, note that after applying several times Bayes' theorem and Assumption 2, we get that

$$\begin{aligned}\mathbb{P}(y_1 \mid y_4, y_3, y_2) &= \frac{\mathbb{P}(y_4, y_3, y_2, y_1)}{\mathbb{P}(y_4, y_3, y_2)} = \frac{\mathbb{P}(y_4 \mid y_3, y_2, y_1)\,\mathbb{P}(y_3 \mid y_2, y_1)\,\mathbb{P}(y_2, y_1)}{\mathbb{P}(y_4 \mid y_3, y_2)\,\mathbb{P}(y_3 \mid y_2)\,\mathbb{P}(y_2)} \\ &= \frac{\mathbb{P}(y_4 \mid y_3)\,\mathbb{P}(y_3 \mid y_2)\,\mathbb{P}(y_2, y_1)}{\mathbb{P}(y_4 \mid y_3)\,\mathbb{P}(y_3 \mid y_2)\,\mathbb{P}(y_2)} = \frac{\mathbb{P}(y_2, y_1)}{\mathbb{P}(y_2)} = \mathbb{P}(y_1 \mid y_2).\end{aligned}$$

Similarly, $\mathbb{P}(y_1 \mid y_4, y_2) = \mathbb{P}(y_1 \mid y_2)$. Hence,

$$\begin{aligned}\mathbb{P}(y_3 \mid y_4, y_2, y_1) &= \frac{\mathbb{P}(y_3, y_1 \mid y_4, y_2)}{\mathbb{P}(y_1 \mid y_4, y_2)} = \frac{\mathbb{P}(y_1 \mid y_4, y_3, y_2)\,\mathbb{P}(y_3 \mid y_4, y_2)}{\mathbb{P}(y_1 \mid y_2)} \\ &= \frac{\mathbb{P}(y_1 \mid y_2)\,\mathbb{P}(y_3 \mid y_4, y_2)}{\mathbb{P}(y_1 \mid y_2)} = \mathbb{P}(y_3 \mid y_4, y_2).\end{aligned}$$

As a result,

$$\begin{aligned}\mathbb{P}(y_5, y_3, y_1 \mid y_4, y_2) &= \mathbb{P}(y_5 \mid y_4, y_3, y_2, y_1)\,\mathbb{P}(y_3 \mid y_4, y_2, y_1)\,\mathbb{P}(y_1 \mid y_4, y_2) \\ &= \mathbb{P}(y_5 \mid y_4, y_2)\,\mathbb{P}(y_3 \mid y_4, y_2)\,\mathbb{P}(y_1 \mid y_4, y_2).\end{aligned}$$

Thus, $\mathbf{y}_5$, $\mathbf{y}_3$, and $\mathbf{y}_1$ are conditionally independent conditional on $\mathbf{y}_4$ and $\mathbf{y}_2$. For $K > 1$ one just need to relabel y_5 as y_{2K+3}, y_4 as y_{2K+2}, y_3 as $y(\mathcal{S}_2)$, y_2 as y_{K+1}, and y_1 as $y(\mathcal{S}_1)$, and apply the above arguments.

A.2. Proof of Theorem 1

Fix some x, y, y , and $s = S$. Take two disjoint subsets of size K $\mathcal{S}_1$ and $\mathcal{S}_2$ that do not contain s. Let $g : \mathcal{Y}^K \to \bar{\mathcal{Y}} = \{1, 2, \ldots, Y^K\}$ be any one-to-one mapping. Define two random variable on $\bar{\mathcal{Y}}$: $\mathbf{z}_1 = g(\mathbf{y}(\mathcal{S}_1))$ and $\mathbf{z}_2 = g(\mathbf{y}(\mathcal{S}_2))$. To simplify the exposition, we drop x, y, y , and $\mathcal{S}_i$ from the notation. All the probabilities below are defined conditional on $\mathbf{x} = x$, $\mathbf{y}_{K+1} = y$, and $\mathbf{y}_{2K+2} = y$. Define the following matrices

$$
\begin{aligned}
L_{1,2} &= [\mathbb{P}(\mathbf{z}_1 = i, \mathbf{z}_2 = j)]_{i,j \ \ \bar{Y}}\,,\\
L_{1|D} &= [\mathbb{P}(\mathbf{z}_1 = i \mid \mathbf{D} = D_k)]_{i \ \ \bar{Y}, k=1,\ldots,d_D}\,,\\
L_{2|D} &= [\mathbb{P}(\mathbf{z}_2 = i \mid \mathbf{D} = D_k)]_{i \ \ \bar{Y}, k=1,\ldots,d_D}\,,\\
A_D &= \operatorname{diag}\left((\mathbb{P}(\mathbf{D} = D_k))_{k=1,\ldots,d_D}\right) = \operatorname{diag}\left((m(D_k))_{k=1,\ldots,d_D}\right),
\end{aligned}
$$

where diag(z) is a diagonal matrix with vector z on the diagonal.

Step 1. In this step, we show how to identify the number of choice sets that are considered with positive probability. By the law of total probability, Lemma 1 (recall that we are also conditioning on $\mathbf{x} = x$, $\mathbf{y}_{K+1} = y$, and $\mathbf{y}_{2K+2} = y$) implies that

$$
\begin{aligned}
\mathbb{P}(\mathbf{z}_1 = i, \mathbf{z}_2 = j) &= \underset{k}{\square}\ \mathbb{P}(\mathbf{z}_1 = i, \mathbf{z}_2 = j \mid \mathbf{D} = D_k)\,\mathbb{P}(\mathbf{D} = D_k)\\
&= \underset{k}{\square}\ \mathbb{P}(\mathbf{z}_1 = i \mid \mathbf{D} = D_k)\,\mathbb{P}(\mathbf{z}_2 = j \mid \mathbf{D} = D_k)\,\mathbb{P}(\mathbf{D} = D_k)\,.
\end{aligned}
$$

Or in matrix notation

$$
L_{1,2} = L_{1|D} A_D L_{2|D}^{\mathsf{T}}.
$$

Under Assumption 5 the maximal number of the points in the support of $\mathbf{D}$ is equal to the number of the possible outcomes. That is, $d_D \leq \square\bar{\mathcal{Y}}\square$.

Next, note that Assumption 5 implies that $L_{1|D}$ and $L_{2|D}$ have full column rank (d_D). Hence, using the properties of the rank operator we can conclude that

$$\text{rank}\left(L_{1,2}\right)=\text{rank}\left(L_{1|D}A_DL_{2|D}^{\intercal}\right)=\text{rank}\left(A_DL_{2|D}^{\intercal}\right)=\text{rank}\left(A_D\right)=d_D.$$

That is, the rank of $L_{1,2}$ is equal to $d_D=|\mathcal{D}_x|$. Hence, since $L_{1,2}$ is observed (can be consistently estimated), we can identify (consistently estimate) the number of choice sets that DMs are using.

Step 2. Knowing d_D and the fact that $L_{1|D}$ and $L_{2|D}$ have full column rank, we take a collection of alternatives in $\bar{\mathcal{Y}}$, $\{\tilde{z}_k\}_{k=1}^{d_D}$, such that the following observable modification of $L_{1,2}$ is nonsingular (have full rank):

$$\tilde{L}_{1,2}=\left[\mathbb{P}\left(\mathbf{z}_1=\tilde{z}_i,\mathbf{z}_2=\tilde{z}_j\right)\right]_{i,j\ \ \{1,\ldots,d_D\}}.$$

Such collection $\{\tilde{z}_k\}_{k=1}^{d_D}$ always exists since one can always find d_D linearly independent rows of $L_{1|D}$. Indeed, similar to Step 1

$$\tilde{L}_{1,2}=\tilde{L}_{1|D}A_D\tilde{L}_{2|D}^{\intercal},$$

where

$$\tilde{L}_{1|D}=\left[\mathbb{P}\left(\mathbf{z}_1=\tilde{z}_i|\mathbf{D}=D_k\right)\right]_{i,k\ \ \{1,\ldots,d_D\}},$$
$$\tilde{L}_{2|D}=\left[\mathbb{P}\left(\mathbf{z}_2=\tilde{z}_i|\mathbf{D}=D_k\right)\right]_{i,k\ \ \{1,\ldots,d_D\}}.$$

Since $\tilde{L}_{1|D}$ and $\tilde{L}_{2|D}$ are nonsingular, it implies that $\tilde{L}_{1,2}$ is nonsingular as well (A_D has rank d_D).

Step 3. This step is based on Hu (2008) and Hu et al. (2013). Fix some $y \in \mathcal{Y}$ and define

$$\tilde{L}_{1,D} = \left[\mathbb{P}\left(\mathbf{z}_1 = \tilde{z}_i, \mathbf{D} = D_k\right)\right]_{i,k \in \{1,\dots,d_D\}},$$

$$\tilde{L}_{2,1,y} = \left[\mathbb{P}\left(\mathbf{z}_2 = \tilde{z}_i, \mathbf{z}_1 = \tilde{z}_j, \mathbf{y}_t = y\right)\right]_{i,j \in \{1,\dots,d_D\}},$$

$$A_{y|D} = \operatorname{diag}\left(\left(\mathbb{P}\left(\mathbf{y}_s = y | \mathbf{D} = D_k\right)\right)_{k \in \{1,\dots,d_D\}}\right) = \operatorname{diag}\left(\left(F^{RUM}(y|D_k)\right)_{k \in \{1,\dots,d_D\}}\right).$$

By the law of total probability, Lemma 1 implies that

$$\begin{aligned}\mathbb{P}\left(\mathbf{z}_1 = \tilde{z}_i, \mathbf{z}_2 = \tilde{z}_j\right) &= \sum_k \mathbb{P}\left(\mathbf{z}_1 = \tilde{z}_i, \mathbf{z}_2 = \tilde{z}_j | \mathbf{D} = D_k\right)\mathbb{P}\left(\mathbf{D} = D_k\right) \\ &= \sum_k \mathbb{P}\left(\mathbf{z}_2 = \tilde{z}_j | \mathbf{D} = D_k\right)\mathbb{P}\left(\mathbf{z}_1 = \tilde{z}_i, \mathbf{D} = D_k\right).\end{aligned}$$

Hence, in matrix notation we get

$$\tilde{L}_{1,2}^{\mathsf{T}} = \tilde{L}_{2|D}\tilde{L}_{1,D}^{\mathsf{T}}.$$

Since, by construction in Step 2, $\tilde{L}_{2|D}$ is nonsingular, we have that

$$\tilde{L}_{1,D}^{\mathsf{T}} = \tilde{L}_{2|D}^{-1}\tilde{L}_{1,2}^{\mathsf{T}}. \tag{1}$$

Similarly to the previous calculations, Lemma 1 implies that

$$\tilde{L}_{2,1,y} = \tilde{L}_{2|D}A_{y|D}\tilde{L}_{1,D}^{\mathsf{T}}.$$

Combining the latter with equation (1) we get the following eigenvector-eigenvalue decomposition of $R_y = \tilde{L}_{2,1,y}\left(\tilde{L}_{1,2}^{\mathsf{T}}\right)^{-1}$

$$R_y = \tilde{L}_{2|D}A_{y|D}\tilde{L}_{2|D}^{-1}. \tag{2}$$

Step 4. Note that in the decomposition (2) the change in y does not affect eigenvectors of R_y, but affects its eigenvalues. For R_y let $\{(\eta_k, \lambda_{y,k})\}_{k=1}^{d_D}$ denote the set of its eigenvectors and eigenvalues. To pin down eigenvectors uniquely note that it suffices to pick those that belong to a simplex (each one of them should sum up to 1). In contrast to the existing results (e.g. Hu et al., 2013), we do not use these eigenvectors to identify $L_{2|D}$ since $\tilde{L}_{2|D}$ is only a submatrix of $L_{2|D}$.

Take $y = 1$ and fix the set of eigenvectors of R_1, $\{\eta_k\}_{k=1}^{d_D}$. Stack them in any order to a matrix Λ_1. Then we can compute

$$A_{y|D} = \Lambda_1^{-1} R_y \Lambda_1$$

for every y. Since the order of eigenvalues is fixed, the diagonal entries of $A_{y|D}$ correspond to the same sets. Note that $y \in D$ if and only if $F^{\mathrm{RUM}}(y|D) > 0$. Thus, we can identify the identity of choice sets and $F^{\mathrm{RUM}}(y|D)$ for every y and D. Hence, we can identify the conditional distributions of $\mathbf{z}_1$ and $\mathbf{z}_2$ conditional on $\mathbf{D}$. Thus, we identify $L_{2|D}$.

Step 5. Finally, let $m = (m(D_k))_{k \ \{1,\ldots,d_D\}}$, then

$$m = L_{1,D}^{\mathsf{T}} \cdot \iota,$$

where ι is the vector of ones. Hence,

$$L_{1,2}^{\mathsf{T}} \iota = L_{2|D} L_{1,D}^{\mathsf{T}} \cdot \iota = L_{2|D} m.$$

Since $L_{1,2}$ is observed (can be consistently estimated), and $L_{2|D}$ is constructively identified and has full column rank, we also identify the distribution of choice sets. The fact that the choice of x, y, and y was arbitrary completes the proof.

A.3. Proof of Proposition 1

Fix some $x \in X$, $y, y \in \mathcal{Y}$, and $\mathcal{S}_i$. To simplify the exposition, we drop x, y, y, and $\mathcal{S}_i$ from the notation. Note that if the linear independence condition is satisfied when $\mathcal{D} = \{A \cup \{y, y\} : A \subseteq \mathcal{Y}\}$, then it is automatically satisfied for any smaller $\mathcal{D}$. Hence, without loss of generality, we assume that $\mathcal{D} = \{A \cup \{y, y\} : A \subseteq \mathcal{Y}\}$.

First, order all elements in $\mathcal{D}$ according to their cardinality (with arbitrary order among sets with the same cardinality). For any two sets D_k and D_m such that $k < m$, D_m is never a subset of D_k.

Given the above order over elements in $\mathcal{D}$, consider the following sequence of 2^{Y-2} vectors in $\mathcal{Y}^K$. For any D_k take y_k^K such that every element in D_k is some component of y_k^K. Since $K \geq Y$ and $|D_k| \leq Y$ such y^K always exists. Moreover, for any $D_k \neq D_l$ it has to be true that $y_k^K \neq y_l^K$.

Consider the matrix $\mathcal{G}$ of size $2^{Y-2} \times 2^{Y-2}$ such that (j, k)-element of it is

$$\mathcal{G}_{j,k} = G(y_j^K \mid D_k).$$

Note that this matrix is upper-triangular. Indeed, take any j, k such that $j > k$. Since $j > k$, then D_j is never a subset of D_k. Hence, there exists a component of y_j^K that is not an element of D_k. This means that the probability of observing a sequence y_j^K given set D_k is zero. That is, $\mathcal{G}_{j,k} = 0$ if $j > k$. Assumption 4 implies that the diagonal elements, $\mathcal{G}_{j,j}$, are nonzero since any element of D_j can be observed with positive probability. Since $\mathcal{G}$ is upper-triangular with nonzero diagonal elements, it is of full column rank (the determinant of $\mathcal{G}$ equals to the product of the diagonal elements).

Adding more rows to $\mathcal{G}$ does not change its column rank. Hence, the linear independence condition is satisfied. The fact that the choice of x, y, y, and $\mathcal{S}_i$ was arbitrary completes the proof.

A.4. Proof of Proposition 2

(i). Fix some $x \in X$, $y, y \in \mathcal{Y}$, and $\mathcal{S}_i$. To simplify the exposition, we drop x, y, y , and $\mathcal{S}_i$ from the notation. Note that if the linear independence condition is satisfied when $\mathcal{D} = \{\{y, y\}, \{y, y, y_1\}, \{y, y, y_1, y_2\}, \{y, y, y_1, y_2, y_3\}, \dots, \mathcal{Y}\}$, then it is automatically satisfied for any smaller $\mathcal{D}$. So, without loss of generality, we assume that $|\mathcal{D}| = Y - 1$.

Nestedness implies that $D_k \subseteq D_{k+1}$ for all k. Let $\{y_j^K\}_{j=1}^{Y}$ be a sequence in $\mathcal{Y}^K$ such that $y_j^K = (y_j, y_j, \dots, y_j)^\top$. Recall that $D_k = \{y, y, y_1, y_2, \dots, y_k\}$. Consider the matrix $\mathcal{G}$ of size $Y - 1 \times Y - 1$ such that (j, k)-element of it is

$$\mathcal{G}_{j,k} = G(y_j^K \mid D_k).$$

Note that this matrix is upper-triangular. Indeed, take any j, k such that $j > k$. Hence, none of the components of y_j^K are elements of D_k. This means that the probability of observing a sequence y_j^K given set D_k is zero. That is, $\mathcal{G}_{j,k} = 0$ if $j > k$. Assumption 4 implies that the diagonal elements, $\mathcal{G}_{j,j}$, are nonzero since any element of D_j can be observed with positive probability. Since $\mathcal{G}$ is upper-triangular with nonzero diagonal elements, it is of full column rank (the determinant of $\mathcal{G}$ equals to the product of the diagonal elements). Adding more rows to $\mathcal{G}$ does not change its column rank. Hence, the linear independence condition is satisfied. The fact that the choice of x, y, y , and $\mathcal{S}_i$ was arbitrary completes the proof.

(ii). Fix some $x \in X$, $y, y \in \mathcal{Y}$, and $\mathcal{S}_i$. To simplify the exposition, we drop x, y, y , and $\mathcal{S}_i$ from the notation. Let $\{y_k^K\}_{k=1}^{|D|}$ be a sequence in $\mathcal{Y}^K$ such that $y_k^K = (y_k, y_k, \dots, y_k)^\top$, where $y_k \in \mathcal{Y}$ are from condition (ii) of the proposition. Consider the matrix $\mathcal{G}$ of size $|\mathcal{D}| \times |\mathcal{D}|$ such that (j, k)-element of it is

$$\mathcal{G}_{j,k} = G(y_j^K \mid D_k).$$

Note that this matrix is diagonal. Indeed, take any j, k such that $j \neq k$. Hence, none of the components of y_j^K are elements of D_k. This means that the probability of observing a sequence y_j^K given set D_k is zero. That is, $\mathcal{G}_{j,k} = 0$ if $j \neq k$. Assumption 4 implies that the diagonal elements, $\mathcal{G}_{j,j}$, are nonzero since any element of D_j can be observed with positive probability. Since $\mathcal{G}$ is diagonal with nonzero diagonal elements, it is of full column rank (the determinant of $\mathcal{G}$ equals to the product of the diagonal elements). Adding more rows to $\mathcal{G}$ does not change its column rank. Hence, the linear independence condition is satisfied. The fact that the choice of x, y, y , and $\mathcal{S}_i$ was arbitrary completes the proof.

B. Estimation

In this appendix, we provide a computationally efficient and consistent estimator of m and F^{RUM}. First, we introduce some notation and objects that are necessary for estimation. Next, we discuss two straightforward consistent estimators and discuss their drawbacks – one is computationally infeasible in moderate size problems, the other one (we call it the Step-1 estimator) is fast but does not perform well in finite samples. We then propose a new estimator that overcomes these two issues. This estimator regularizes the fast Step-1 estimator to achieve a better performance in finite samples.

To simplify the exposition, we assume that $K = 1$. Let P be the conditional probability mass function of choices conditional on covariates. That is,

$$\mathrm{P}(y_1, y_3, y_5 \mid y, y\ , x) = \mathbb{P}\left(\mathbf{y}_1 = y_1, \mathbf{y}_3 = y_3, \mathbf{y}_5 = y_5 \mid \mathbf{y}_2 = y, \mathbf{y}_4 = y\ , \mathbf{x} = x\right)$$

for every $y_1, y_2, y_3, y, y\ \in \mathcal{Y}$ and $x \in X$. We assume that the analyst has access to a consistent estimator of P, $\hat{\mathrm{P}}$. For instance, if the analyst observes a sample of size n of i.i.d. observations coming from the joint distribution of $(\mathbf{y}_s)_{s\ S}$ and $\mathbf{x}$, $\left(\left(\mathbf{y}_s^{(i)}\right)_{s\ S}, \mathbf{x}^{(i)}\right)_{i=1}^{n}$, and X is a finite set, then for any y_1, y_2, y_3, y, y' and x one can use

$$\hat{\mathrm{P}}(y_1, y_3, y_5 \mid y, y', x) = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\mathbf{y}_1^{(i)} = y_1, \mathbf{y}_3^{(i)} = y_3, \mathbf{y}_5^{(i)} = y_5, \mathbf{y}_2^{(i)} = y, \mathbf{y}_4^{(i)} = y', \mathbf{x}^{(i)} = x\right\}}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\mathbf{y}_2^{(i)} = y, \mathbf{y}_4^{(i)} = y', \mathbf{x}^{(i)} = x\right\}}.$$

For continuously distributed x, one can use any nonparametric estimator of a conditional expectation based on sieves or kernels (see Chen, 2007 and Li and Racine, 2007).[41] In our estimation routine, we take $\hat{P}$ as given.

Assumption 6. *There exists an estimator of* P, $\hat{\mathrm{P}}$, *and a diverging sequence of positive natural numbers* α_n *such that* $\alpha_n\left(\hat{\mathrm{P}}(y_1, y_3, y_5 \mid y, y', x) - \mathrm{P}(y_1, y_3, y_5 \mid y, y', x)\right)$ *is stochastically bounded in probability for any* y_1, y_2, y_3, y, y' *and* x *in the support. That is,*

$$\hat{\mathrm{P}}(y_1, y_3, y_5 \mid y, y', x) - \mathrm{P}(y_1, y_3, y_5 \mid y, y', x) = O_P(\alpha_n^{-1})$$

for all $y_1, y_2, y_3, y, y' \in \mathcal{Y}$ *and* $x \in \mathcal{X}$.

The rate of convergence α_n depends on the asymptotic behavior of $\hat{\mathrm{P}}$. For instance, the estimator $\hat{\mathrm{P}}$ with discrete $\mathbf{x}$ above is $\sqrt{n}$-consistent estimator (i.e. $\alpha_n = \sqrt{n}$).

We fix some x, y, and y' and conduct the analysis below "conditional on $\mathbf{x} = x$, $\mathbf{y}_2 = y$, and $\mathbf{y}_4 = y'$." To simplify the exposition we drop x, y, and y' from the notation (in our empirical application we estimate the model for all values of covariates). Note that the fact that we condition on y, y' implies that only sets that contain y, y' will be considered with positive probability.

In the proof of Theorem 1, we show that the cardinality of the support of $\mathbf{D}$ (conditional on x, y, and y') is equal to

$$d_D = \operatorname{rank}\left(\left(\sum_{k=1}^{Y} \mathrm{P}(i, j, k)\right)_{i,j \in Y}\right),$$

[41] For a recent application of a sieve estimator with continuous covariates see, for instance, Kashaev (2020).

where rank(A) is the rank of matrix A. Thus, given that $\hat{\mathrm{P}}$ converges in probability to P, we can estimate the upper bound for the cardinality of $\mathcal{D}$ as

$$\hat{d}_D = \text{rank}\left(\left(\sum_{k=1}^{Y} \hat{\mathrm{P}}(i,j,k)\right)_{i,j \in Y}\right).$$

Instead of using this upper bound, we can infer the rank by applying any standard procedure (e.g., Cragg and Donald, 1997). However, since in the second step of our method, we pick no more than $\hat{d}_D$ sets that explain the data, asymptotically our procedure is still consistent.

B.1. Straightforward Estimator

Given that our identification result implies that there is a unique distribution over choice sets m and the distribution over choice conditional on choice sets F^{RUM}, we can directly minimize the Euclidean distance to the observed distribution of choices over all possible combinations of $\hat{d}_D$ subsets of $\mathcal{Y}$. The collection of subsets that minimizes this distance is a consistent estimator of the support of the choice sets.[42] Formally, let $\mathbb{R}_+^{a\times b}$ denote the space of matrices of size $a \times b$ with nonnegative real entries. Fix a collection of $\hat{d}_D$ different subsets of $\mathcal{Y}$ encoded by some matrix $\tilde{A} \in \mathbb{R}_+^{Y\times \hat{d}_D}$, where $\tilde{A}_{y,j}$ equals to 1 if alternative y belongs to a choice set D_j, and equals to 0 otherwise. Then we can compute

$$
\begin{aligned}
T(\tilde{A}) = \min_{F^s \in \mathbb{R}_+^{Y\times \hat{d}_D},\, M \in \mathbb{R}_+^{\hat{d}_D}} \; & \sum_{y_1,y_3,y_5} \left(\hat{\mathrm{P}}(y_1,y_3,y_5) - \sum_{j=1}^{\hat{d}_D} F^1_{y_1,j} \cdot F^3_{y_3,j} \cdot F^5_{y_5,j} \cdot M_j\right)^2 \\
\text{s.t. } & \sum_{y \in Y} F^s_{y,j} = 1,\; j = 1,2,\ldots,\hat{d}_D,\; s = 1,2,3, \\
& \sum_{j=1}^{\hat{d}_D} M_j = 1, \\
& F^s_{y,j} \le \tilde{A}_{y,j},\; y \in \mathcal{Y},\; j = 1,2,\ldots,\hat{d}_D,\; s = 1,3,5.
\end{aligned}
$$

[42] Since there are finitely many collections of subsets (i.e., the parameter space is discrete), this estimator of the support of choice sets will converge arbitrary fast.

The first and the second set of constrains require F^s and M to be proper probability distributions.[43] The last set of constraints restricts the probability of choosing an alternative that is not considered (i.e., not in the choice set) to be zero.

The collection $\tilde{A}$ that minimizes $T(\cdot)$ and contains y and y would deliver a consistent estimator of the choice sets, and, thus, give us a consistent estimator of F^{RUM}. Unfortunately, this procedure becomes computationally prohibitive for even relatively small Y. For instance, if, as in our empirical application, one assumes that choices are independent conditional on choice sets and covariates and that $Y = 5$, then, without any restrictions on choices sets, there are $31!/(31-5)! > 2\times 10^7$ possible combinations of 5 different sets out 31 nonempty subsets of $\{1,2,\ldots,5\}$. Even if every $T(A)$ is computed within 0.01 sec, finding $\tilde{A}$ on a single core computer would take more than two days. In stark contrast, the procedure that we propose in the next section takes less than one minute on a single core computer.[44]

The next estimator, which we call the Step-1 estimator, is similar to the previous estimator, but it does not force the constraints captured by matrix $\tilde{A}$. In particular, let the Step-1 estimator of F^{RUM} and m be

$$\bar{F}_{s1},\bar{M}_{s1} = \arg\min_{F^s \in \mathbb{R}_+^{Y\times \hat{d}_D},\, M \in \mathbb{R}_+^{\hat{d}_D}} \sum_{y_1,y_3,y_5}\left(\hat{\mathrm{P}}(y_1,y_3,y_5) - \sum_{j=1}^{\hat{d}_D} F^1_{y_1,j}\cdot F^3_{y_3,j}\cdot F^5_{y_5,j}\cdot M_j\right)^2$$

$$\text{s.t.} \quad \sum_{y\in Y} F^s_{y,j} = 1,\ j = 1,2,\ldots,\hat{d}_D,\ s = 1,3,5,$$

$$\sum_{j=1}^{\hat{d}_D} M_j = 1,$$

where $\bar{F}_{s1} = (\bar{F}^1_{s1},\bar{F}^2_{s2},\bar{F}^3_{s3})$.[45]

[43] We cannot force F^1 to be equal to F^3 since F^1 and F^3 represent $F^{\mathrm{RUM}}(\cdot|y,D,x)$ and $F^{\mathrm{RUM}}(\cdot|y',D,x)$ respectively.

[44] Our simulations indicate that the estimation time of our procedure grows exponentially with Y. However, it is substantially faster than the procedure that checks all sets. For instance, our method takes about 6 hours to estimate a model with $Y = 10$. The alternative would require solving $1023!/(1023-10)! > 10^{18}$ optimization problems.

[45] Instead the described estimator, one can also use the estimator based on diagonalization argument as in Hu et al. (2013). Unfortunately, it suffers from the same issues in finite samples and performs a little worse in our simulations.

Given our identification result, the estimated $\bar{F}^{5}_{s1,y,j}$ and $\bar{M}_{s1,j}$ are consistent estimators of $F^{\mathrm{RUM}}(y \mid D_j)$ and $m(D_j)$ since they are minimizing the Euclidean distance between the observed distribution and the distribution implied by the parameters.[46] However, because of the sampling uncertainty and numerical optimization errors the elements of matrix $\bar{F}^{t}_{s1}$ that correspond to $F^{\mathrm{RUM}}(y \mid D) = 0$ (i.e. $y \notin D$) may not be exactly equal to zero. That is why, to recover the identity of choice sets, we trim the elements of $\bar{F}^{t}_{s1,y,j}$ that are smaller than a prespecified $\varepsilon > 0$. In our application and simulations, we set $\varepsilon = 0.01$. Formally, we need the following strengthening of Assumption 4.

Assumption 7. *For every $x \in X$, $D \in \mathcal{D}_x$, and $y, y \in D$*

$$F^{\mathrm{RUM}}(y \mid y, D, x) \geq \varepsilon$$

for some known $\varepsilon > 0$.

Thus, for every y and j we define the Step-1 estimator as

$$F^{s}_{s1,y,j} = \frac{\bar{F}^{s}_{s1,y,j}\mathbb{1}\left(\bar{F}^{s}_{s1,y,j} \geq \varepsilon\right)}{\sum_{y'=1}^{Y} \bar{F}^{s}_{s1,y',j}\mathbb{1}\left(\bar{F}^{s1,s}_{y',j} \geq \varepsilon\right)}.$$

Note that, instead of a fixed threshold, one can use a threshold ε_n that converges to 0 sufficiently slowly (e.g. $\varepsilon_n = \log(\log(n))/\sqrt{n}$ if $\alpha_n = \sqrt{n}$). In this case, Assumption 7 is not needed.[47]

The Step-1 estimator does not require checking all possible collections of subsets of the grand choice set, however, it may perform poorly in finite samples (see Section C). For instance, one problem of the Step-1 estimator is that it trims the unconstrained estimator of F^{RUM} to get the identity of consideration sets. This trimming, while delivering correct identities of the choices sets asymptotically, may be sensitive to the choice of ε in finite

[46] If covariates are discrete, then instead of minimizing the Euclidean distance, one can also minimize the Kullback-Leibler divergence and obtain maximum-likelihood estimates.

[47] If some of the estimated sets appear more than one time (i.e., two columns of $\tilde{F}^{s}$ has the same zero components), then we can just drop one of them.

samples. Thus, in finite samples, we propose to regularize the Step-1 estimator by using it as the starting point in the procedure described in the next section.

Step-2 Estimator

Unconstrained Estimation. Let $A \in \mathbb{R}^{Y\times(2^Y-2)}$ be the matrix of zeros and ones that encodes all subsets of $\mathcal{Y}$ that contain y and y' (if $y = y'$, then $A \in \mathbb{R}^{Y\times(2^Y-1)}$). That is, $A_{y,j} = \mathbb{1}(y \in D_j)$, $D_j \in 2^Y \setminus \emptyset$ and $y, y' \in D_j$. Let

$$\bar{F}_{s2}, \bar{M}_{s2} = \underset{F^s \in \mathbb{R}_+^{Y\times(2^Y-1)},\, M \in \mathbb{R}_+^{2^Y-1}}{\arg\min} \sum_{y_1,y_3,y_5} \left(\hat{\mathrm{P}}(y_1,y_3,y_5) - \sum_{j=1}^{\hat{d}_D} F^1_{y_1,j}\cdot F^3_{y_3,j}\cdot F^5_{y_5,j}\cdot M_j \right)^2$$

$$\text{s.t.} \quad \sum_{y \in Y} F^s_{y,j} = 1,\ j = 1,2,\ldots,2^Y-1,\ s = 1,3,5,$$

$$\sum_{j=1}^{2^Y-1} M_j = 1,$$

$$F^s_{y,j} \leq A_{y,j},\ y \in \mathcal{Y},\ j = 1,2,\ldots,2^{Y-2},\ s = 1,3,5.$$

This optimization procedure is similar to the one in the previous section. However, it does not impose the sparsity condition – all possible subsets of $\mathcal{Y}$ are allowed. As a result, this optimization problem, in general, may have several global minima since no assumptions on the number of choice sets are imposed. However, since the Step-1 estimator is consistent, there is a unique global minima to which the Step-1 estimator converges in probability. Hence, if we search for the optimum in the neighborhood of the Step-1 estimator, then the minimizer is still a consistent estimator of m and F^{RUM}.

Mixed Integer Optimization. Note that in contrast to the Step-1 estimator, $\bar{F}_{s2}$ is forced to assign zeros at proper positions because of the constraints associated with matrix A. But, in finite samples, since no restrictions on the number of choice sets is imposed, it may assign positive mass to more that $\hat{d}_D$ sets. To solve this issue, we propose to solve the following mixed-integer problem:

$$\begin{aligned}
\hat{B}_{s2}, \tilde{M}_{s2} = & \operatorname*{arg\,min}_{B \in \{0,1\}^{2^{Y}-2},\ M \in \mathbb{R}_{+}^{2^{Y}-2}} \sum_{y_1,y_3,y_5} \left(\hat{\mathrm{P}}(y_1, y_3, y_5) - \sum_{j=1}^{2^{Y}-2} \bar{F}^{1}_{s2,y_1,j} \cdot \bar{F}^{3}_{s2,y_3,j} \cdot \bar{F}^{5}_{s2,y_5,j} \cdot M_j \right)^2 \\
\text{s.t.} \quad & \sum_{j}^{2^{Y}-2} M_j = 1, \\
& M_j \leq B_j,\ j = 1, 2, \ldots, 2^{Y}-2, \\
& \sum_{j}^{2^{Y}-2} B_j \leq \hat{d}_D.
\end{aligned}$$

Note that $B \in \{0, 1\}^{2^{Y}-2}$ is a vector of zeros and ones and the objective function is similar to the least-squares objective since $\bar{F}_{s2}$ is fixed. Informally, one can think of $\bar{F}_{s2}$ as being a collection of regressors and $\sum_{j}^{2^{Y}-1} B_j \leq \hat{d}_D$ being a sparsity constraint: at most $\hat{d}_D$ regressors have to be active. As we discussed before, the model selection procedures, in general, are not consistent. But since we use a consistent estimator as a starting point in optimization the resulting $\hat{B}_{s2}$ correctly recovers the choice sets with probability approaching 1. Also since the last constraint is an inequality constraint, we may end up having less than $\hat{d}_D$ active choice sets.

Final Step. Finally, let $\hat{A} \in \mathbb{R}_{+}^{Y \times \hat{d}_D}$ be the matrix of zeros and ones that encodes the choice sets estimated by $\hat{B}_{s2}$. Now, since we consistently estimated the choice sets on the previous step, we can finally estimate F^{RUM} and m:[48]

$$\begin{aligned}
F_{s2}, M_{s2} = & \operatorname*{arg\,min}_{F^{s} \in \mathbb{R}_{+}^{Y \times \hat{d}_D},\ M \in \mathbb{R}_{+}^{\hat{d}_D}} \sum_{y_1,y_3,y_5} \left(\hat{\mathrm{P}}(y_1, y_3, y_5) - \sum_{j=1}^{\hat{d}_D} F^{1}_{y_1,j} F^{3}_{y_3,j} F^{5}_{y_5,j} M_j \right)^2 \\
\text{s.t.} \quad & \sum_{y \in Y} F^{s}_{y,j} = 1,\ j = 1, 2, \ldots, \hat{d}_D,\ s = 1, 3, 5, \\
& \sum_{j=1}^{\hat{d}_D} M_j = 1,
\end{aligned}$$

[48] Similar to the Step-1 estimator, instead of minimizing the Euclidean distance, we can conduct maximum likelihood estimation here when covariates are discrete.

$$F^s_{y,j} \leq \hat{A}_{y,j},\ y \in \mathcal{Y},\ j = 1, 2, \ldots, \hat{d}_D,\ s = 1, 3, 5.$$

B.2. Discussions and Computational Aspects

The key part of the proposed two-step estimator is recovering the choice sets (i.e., the support of $\mathbf{D}$). After the choice sets (or $\hat{B}_{s2}$) are found, in principle, any parametric or nonparametric estimation of F^{RUM} can be conducted (e.g., one can maximize the sample likelihood).

Our procedure is extremely fast and can be easily applied to choice sets of moderate size. For instance, in our empirical application with $Y = 5$, our procedure completes the estimation in less than one minute on a single core computer. The main advantage is coming from employing mixed-integer programming to our problem. In particular, one can think of the search for $\hat{B}_{s2}$ as a regression problem with $2^Y - 2$ "regressors" and at most $\hat{d}_D$ nonzero coefficients. In the statistical literature, this problem is known as *the best subset problem* (see Bertsimas et al., 2016 and references therein for extensive discussion). Modern mixed-integer optimization algorithms can solve the best subset problem with thousands of observations and hundreds of active regressors within minutes.

It is easy to impose any restrictions on the set of possible choice sets in our estimation procedure. For instance, if one wants to rule out singleton choice sets, one just needs to set the columns in matrix A that correspond to singleton sets to zero columns and remove all repeated columns. Similarly, the lower and upper bound restrictions $\mathrm{L}_{x,y'}$ and $\mathrm{U}_{x,y'}$ discussed in Section 3.2 are easy to impose. If, for example, one wants to assume that, say, alternative y is always considered, it is sufficient to set every element of the y-th row of A to 1 and again remove all repeated columns.

We conclude this section by noting that, given the discrete nature of the estimator of the latent choice set and the use of mixed-integer optimization, deriving confidence sets for the true choice sets, and, thus, for m, and F^{RUM} is nontrivial.[49] However, if we assume that the choices sets are known, then the problem of estimation of m and F^{RUM} is standard, and under the standard regularity conditions, one can conduct inference either by using normal approximations or bootstrap depending on the way one estimates P.[50] We leave the problem of constructing confidence sets for model parameters when the choice sets are also estimated for future work.

C. Finite-sample Performance of the Estimator

This section aims to analyze the finite sample performance of the estimator we propose in Appendix B.

First, we define the data generating processes (DGPs) used in simulations. In all experiments we assume that there are no covariates, Assumption 2 is satisfied, and that $S = 3$ (this setup is equivalent to the one with choices depending on the previous choice and $S = 5$). There are $Y = 5$ alternatives and $d_D = 5$ choice sets. Every DGP is characterized by two matrices: $Pyd \in \mathbb{R}^{5\times 5}$ and $Pd \in \mathbb{R}^5$. Pyd and Pd is such that $Pyd_{y,j} = \mathbb{P}\left(\mathbf{y}_s = y \mid D_j\right)$ and $Pd_j = \mathbb{P}\left(\mathbf{D} = D_j\right)$. In other words, every column of Pyd corresponds to a choice set. For instance, the fourth column of Pyd in DGP1 indicates that the fourth choice set has two elements, $\{1, 4\}$, and that conditional on $\mathbf{D} = \{1, 4\}$ alternative 1 is picked with probability 0.4. Since the fifth element of Pd is 0.15, the probability of considering $\{1, 4\}$ is 0.15.

[49] The existing methods conduct inference on parameters of interest under parametric assumptions about the distribution of choice sets. If these parametric assumptions are not valid, then the resulting standard error are incorrect.

[50] In general, asymptotic properties of estimated m and F^{RUM} depend on whether one has continuous covariates and uses kernels or sieves.

DGP1:

$$Pyd = \begin{bmatrix} 1 & 0.6 & 0.5 & 0.4 & 0.2 \\ 0 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0 & 0.8 \end{bmatrix}, \quad Pd = \begin{bmatrix} 0.2 \\ 0.15 \\ 0.3 \\ 0.15 \\ 0.2 \end{bmatrix}.$$

DGP2:

$$Pyd = \begin{bmatrix} 1 & 0.6 & 0.5 & 0.25 & 0.1 \\ 0 & 0.4 & 0.2 & 0.35 & 0.25 \\ 0 & 0 & 0.3 & 0.25 & 0.15 \\ 0 & 0 & 0 & 0.15 & 0.3 \\ 0 & 0 & 0 & 0 & 0.2 \end{bmatrix}, \quad Pd = \begin{bmatrix} 0.2 \\ 0.15 \\ 0.3 \\ 0.15 \\ 0.2 \end{bmatrix}.$$

The cardinality of the choice sets in DGP1 does no vary much. It assigns positive probability to the choice sets of cardinality less than 3 only. For example, alternative 1 enters all choice sets.[51] All other options enter only one choice set. DGP2 is very heterogeneous in terms of the size of the choice sets. The smallest one contains only alternative 1. The biggest one contains all alternatives. Note that these DGPs satisfy conditions of Proposition 2 (Excluded Choices and Nestedness).

We evaluate the ability of the Step-1 and Step-2 estimators to correctly recover sets. The results are presented in Table 7. As expected, the performance of both estimators improves with the sample size. The Step-2 estimator outperforms the Step-1 estimator in all experiments. The gains from the Step-2 estimator are especially striking for DGP2. For example, for a sample size of 2000 the Step-1 estimator correctly recovers all 5 sets only in 4 percent of cases. However, the Step-2 estimator recovers all 5 sets in 29 percent of cases.

To see how noisy the estimates of choice sets can be we also compute the average number of correctly recovered sets. As Table 8 shows, our estimator on average finds correctly more

[51] We impose this restriction in the estimation step.

Table 7 – Percent of Correctly Estimated Sets

Sample Size		2000	5000	10000	50000
DGP1	Step-1	63.0	72.6	80.2	95.8
	Step-2	64.1	86.8	93.9	99.9
DGP2	Step-1	4	9.9	24.8	77.6
	Step-2	29	45.7	63.4	93.5

Notes: Number of replications=1000, $\varepsilon = 0.01$. Results are rounded to 1 digit.

than 4 out of 5 sets. The Step-1 estimator recovers at least 3 out of 5 sets correctly on average. The performance of both estimators improves with the sample size.

Table 8 – Average Number of Correctly Estimated Sets

Sample Size		2000	5000	10000	50000
DGP1	Step-1	4.63	4.73	4.8	4.96
	Step-2	4.60	4.86	4.94	5
DGP2	Step-1	3.33	3.56	3.92	4.77
	Step-2	4.14	4.39	4.61	4.93

Notes: Number of replications=1000, $\varepsilon = 0.01$, number of sets=5. Results are rounded to 2 digits.

Tables 9-16 present the bias and the mean-squared-error of estimation of m and F^{RUM}. Every experiment was conducted 1000 times. Since some elements of matrix Pyd are zeros and elements of every column sum up to one, for estimates of F^{RUM} we only report the nonzero, linearly independent elements. For instance, for DGP1 we only report estimates of $(1,2)$, $(1,3)$, $(1,4)$, and $(1,5)$ elements of Pyd. The estimators perform well even in the samples of a moderate size. As expected, both the bias and the root-mean-squared-error decrease with the sample size.

Table 9 – Bias in Estimating m. DGP1 ($\times 10^{-5}$)

Sample Size\Set	D_1	D_2	D_3	D_4	D_5
2000	-112.6	35.1	43.8	35.7	-2
5000	-30.4	18.7	8.7	1.8	1.2
10000	-2.2	-8.1	8.3	-1.3	3.3
50000	4.8	-12.5	-0.004	6.2	1.6

Notes: Number of replications=1000. Results are rounded to 6 digits.

Table 10 – Root Mean Squared Error in Estimating m. DGP1 ($\times 10^{-3}$)

Sample Size\Set	D_1	D_2	D_3	D_4	D_5
2000	13.1	10.5	12.5	8.5	9.2
5000	8	6.6	7.7	5.4	5.9
10000	5.8	4.8	5.2	3.7	4.2
50000	2.7	2.1	2.4	1.7	1.9

Notes: Number of replications=1000. Results are rounded to 4 digits.

D. Results from the Application Omitted From the Main Text

Figure 4 depicts the geographical location of our markets and the fractions of the population in these markets that consider the two most frequent sets. We also add those who only consider General Mills alone for a better glimpse of the heterogeneity in choice sets.

Given that the estimators of β_D in Section 5 are GMM estimators, if we treat the estimated shares $sh_{Py,D,j,s}$ as the true shares (i.e. no estimation error), then we can easily construct the 2-step efficient GMM standard errors. Table 17 displays the estimates of β_D together with their standard errors. Following Nevo (2001), here we report the median across markets own-price elasticities in Table 18. In the first column we use estimates of the price coefficient assuming that there is no choice set variation (Naive). The second and third columns were constructed by using the price coefficients for those consumers who considered all 5 brands or $\{CTL, GM, K, O\}$, respectively.

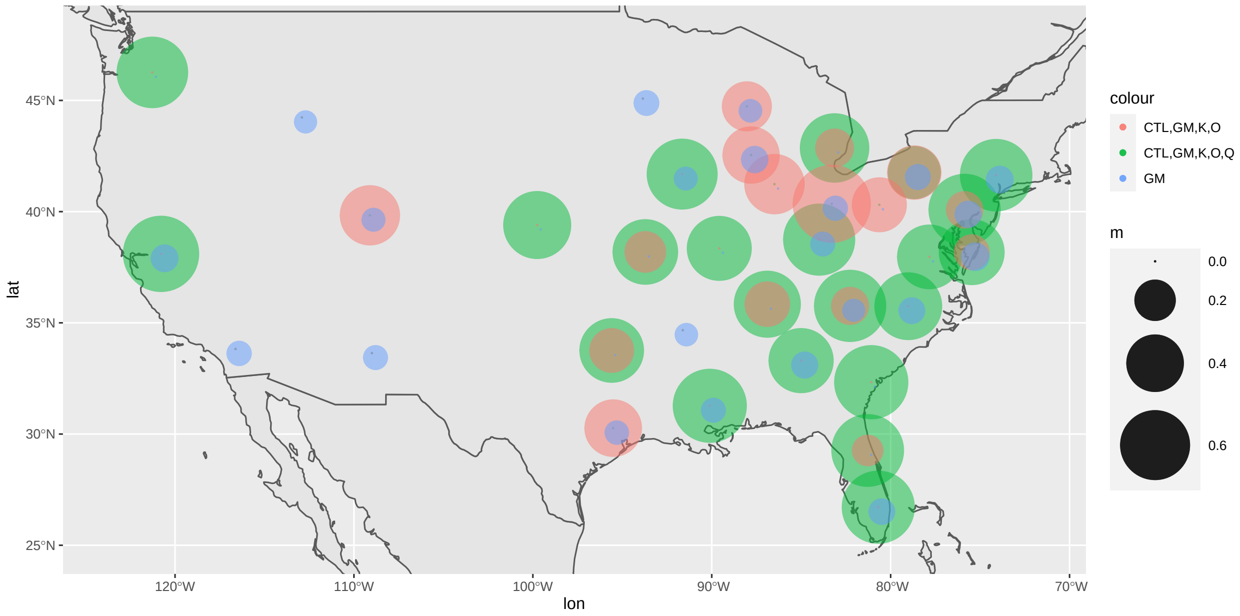

Figure 4 – Location of Markets and Consideration Probabilities of Some Choice Sets: CTL=Store brand, GM=General Mills, K=Kellogg, Q=Quaker , O=other brands. The size of every circle corresponds to the estimates of consideration probability $\hat{m}$. The color of every circle corresponds to a different choice set: Red={CTL,GM,K,O}, Green={CTL,GM,K,O,Q}, Blue={GM}.

Table 11 – Bias in Estimating m. DGP2 ($\times 10^{-3}$)

Sample Size\Set	D_1	D_2	D_3	D_4	D_5
2000	-16.7	-9.7	-54.2	37.7	42.9
5000	-11.8	-7.4	-65.8	53.4	31.6
10000	-6.7	-2.3	-50.1	40.7	18.4
50000	-0.6	0.4	-6.5	4.3	2.4

Notes: Number of replications=1000. Results are rounded to 4 digits.

Table 12 – Root Mean Squared Error in Estimating m. DGP2 ($\times 10^{-2}$)

Sample Size\Set	D_1	D_2	D_3	D_4	D_5
2000	5.3	4.7	10.4	10.7	5.1
5000	4.0	3.3	10.6	10.1	4.0
10000	2.6	1.8	8.3	7.5	2.6
50000	0.3	0.5	1.5	1.2	0.6

Notes: Number of replications=1000. Results are rounded to 3 digits.

E. Examples for Proposition 2

Example 3 (Distance)**.** Consider a collection of DMs that live close to each other and who pick a restaurant. DMs live in the same location but are different in terms of the unobserved type of transportation they have access to. For instance, some DMs can only walk to restaurants, others can bike or take a cab. The realization of the random choice set **D** captures the set of restaurants the DM can get to. Condition (i) in Proposition 2 is satisfied due to geographical nestedness: a person with a car can get to any place attainable by a bicyclist; a person with a bicycle can get to any place attainable by a pedestrian. In this example, s can index choices of the same DM at different time periods whose mode of transportation does not change over time. Alternatively, s can index choices of different DMs with the same mode of transportation.

Example 4 (Brand/Product Loyalty)**.** Suppose $\mathcal{D}_{x,y'} = \{\bar{D} \cup \{y_k\}\}_{k=1}^{d_{D,x,y'}}$, where $y \in \bar{D}$,

Table 13 – Bias in Estimating F^{RUM}. DGP1 ($\times 10^{-4}$)

Sample Size\(i, j)	$1, 2$	$1, 3$	$1, 4$	$1, 5$
2000	0.9	2.7	-0.4	-1.8
5000	3.7	-4.4	8.5	1.3
10000	0.9	-1.6	2.9	-0.6
50000	-2.3	-2.1	-0.0	-0.3

Notes: Number of replications=1000. Results are rounded to 5 digits. Only estimated elements of Pyd are displayed. Only linearly independent estimated elements of $\hat{F}^{\mathsf{RUM}}$ are displayed.

Table 14 – Root Mean Squared Error in Estimating F^{RUM}. DGP1 ($\times 10^{-2}$)

Sample Size\(i, j)	$1, 2$	$1, 3$	$1, 4$	$1, 5$
2000	3.3	2.3	3.0	2.3
5000	2.1	1.4	2.0	1.5
10000	1.5	1.0	1.4	1.0
50000	0.7	0.4	0.6	0.4

Notes: Number of replications=1000. Results are rounded to 3 digits. Only linearly independent estimated elements of $\hat{F}^{\mathsf{RUM}}$ are displayed.

$y_k \neq y_{k'}$ for $k \neq k$, and $y_k \notin \bar{D}$ for all k. Then condition (ii) in Proposition 2 is satisfied. Alternatives y_k represent a particular brand or product that DM is loyal to. $\bar{D}$ can be thought of as the set of products considered by every DM (e.g., store brand). Brand or product loyalty is characterized by the fact that if DM considers y_k, she ignores all other options not present in $\bar{D}$. For instance, let $Y = \{1, 2, 3, 4\}$ be a set of different brands of cereals, where 1 represents a store brand. Then the choice sets $\{\{1, 2\}, \{1, 3\}, \{1, 4\}\}$ represents three types of consumers that look at brands 2, 3, and 4. However, everyone pays attention to the default store brand option $\bar{D} = \{1\}$.

Table 15 – Bias in Estimating F^{RUM}. DGP2 ($\times 10^{-2}$)

Sample Size\\(i, j)	$1, 2$	$1, 3$	$2, 3$	$1, 4$	$2, 4$	$3, 4$	$1, 5$	$1, 5$	$3, 5$	$4, 5$
2000	0.9	3.3	-2.3	7.7	-0.7	-0.1	1.4	1.6	1.3	-0.7
5000	1.9	3.7	-2.9	8.1	-0.6	-0.4	0.9	1.2	1.0	-0.2
10000	1.2	3.3	-2.7	5.4	-0.4	0.0	0.4	0.7	0.7	-0.0
50000	0.2	0.3	-0.4	0.6	0.2	0.2	0.1	0.1	0.0	0.1

Notes: Number of replications=1000. Results are rounded to 3 digits. Only estimated elements of Pyd are displayed. Only linearly independent estimated elements of $\hat{F}^{\mathsf{RUM}}$ are displayed.

Table 16 – Root Mean Squared Error in Estimating F^{RUM}. DGP2 ($\times 10^{-2}$)

Sample Size\\(i, j)	$1, 2$	$1, 3$	$2, 3$	$1, 4$	$2, 4$	$3, 4$	$1, 5$	$1, 5$	$3, 5$	$4, 5$
2000	15.9	15.2	10.3	18.6	17.2	14.2	3.3	4.2	3.6	3.5
5000	11.0	12.2	8.5	14.4	12.3	10.2	2.2	2.9	2.5	2.1
10000	5.7	9.3	6.2	9.6	7.1	6.3	1.5	2.1	1.9	1.5
50000	1.3	1.3	1.2	2.6	2.4	2.0	0.7	0.8	0.7	0.7

Notes: Number of replications=1000. Results are rounded to 3 digits. Only estimated elements of Pyd are displayed. Only linearly independent estimated elements of $\hat{F}^{\mathsf{RUM}}$ are displayed.

Table 17 – Estimates of β

	Naive	{CTL, GM, K, O, Q}	{CTL, GM, K, O}
$\hat{\beta}$	-16.11	-5.84	-56.46
std. error	2.30	7.1	20.04

Notes:Standard errors are computed assuming that there is no estimation error in shares. Results are rounded to 2 digits.

Table 18 – Estimates of Median Across Markets Own-Price Elasticities

	Naive	{CTL, GM, K, O, Q}	{CTL, GM, K, O}
CTL	-1.83	-0.72	-7.32
GM	-2.41	-0.86	-9.51
K	-2.04	-0.7	-8.15
O	-2.07	-0.77	-7.83
Q	-2.38	-0.86	0

Notes: The first column is computed assuming that consumers face all 5 brands. The second column is computed assuming choice set variation for those consumers who consider all 5 brands. The last column is computed for those consumers who do not consider Quaker. Results are rounded to 2 digits.