


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



Victor H. Aguiar  Nail Kashaev[†]

This version: May 23, 2023 / 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 programming 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.

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 studies discrete choice models when the choice set that decision makers (DMs) face are unobserved by the researcher. We show how to nonparametrically identify and estimate the joint distribution of latent choice sets and choices when a cross-section of a small number of repeated choices is observed. 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 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.

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 advances in identification of these models.¹ In particular, we base our identification strategy

¹See, for instance, Hall and Zhou (2003), Hu (2008), Kasahara and Shimotsu (2009), Bonhomme et al. (2016), Kitamura and Laage (2018), and references therein.

on the insights from discrete nonclassical measurement error results in [Hu \(2008\)](#).² 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 can avoid the strict monotonicity condition because in our setting choice sets have a structural interpretation. We also do not need to know the number of possible choice sets, since we can identify and estimate it.

The structural interpretation of the latent choice sets also allows us to establish two new sufficient conditions for the linear independence condition, which is standard in the mixture literature. 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 the minimal amount of data, 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, sparsity allows for substantial dimensionality reduction when few choices are observed. Sparsity is justified with two primitive conditions: (i) the support of the random choice set is nested, and (ii) the support of the random choice set is such that there are excluded choices from its elements. We provide several theoretical and empirical examples (e.g., search and satisfy, brand loyalty, and the sleeping agent) where these two conditions are satisfied.

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*

²For applications in the context of auctions and discrete games see [Hu et al. \(2013\)](#), [Xiao \(2018\)](#), and [Luo \(2020\)](#).

problem (see, for instance, [Bertsimas et al., 2016](#)). As a result, the problem turns out to be a mixed-integer programming (MIP) problem 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. We exploit the high frequency of purchases 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. \(2021\)](#). They 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.³ We also differ in that 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.

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 using asymmetries in the substitution matrix. Crucially, they assume full support of prices, differentiability of the probability of choice with respect to prices, independence of preferences and random consideration sets, and a form of menu variation.⁴

³See also [Goeree \(2008\)](#) and [Barseghyan et al. \(2021b\)](#) for applications of consideration sets driven by item-dependent attention.

⁴In fact, since they allow prices to diverge to infinity they effectively allow for menu variation, because when the price of an alternative goes to infinity then the alternative becomes unavailable to the DMs.

Our setup does not impose any of these assumptions. This is advantageous because full support of prices could be a strong assumption on data availability since in many situations price variation is mainly spatial or temporal and hence it is discrete in nature (e.g., prices vary across markets). Moreover, menu variation is even rarer and does not usually exist except when the researcher collects this information explicitly (Hickman and Mortimer, 2016, Aguiar et al., 2023). In addition, independence of preferences and consideration sets falls in many situations such as when consumers follow a form of search-and-satisfy to form consideration sets, or when rational inattention plays a role (Caplin et al., 2019, Kashaev and Aguiar, 2022). Lu (2014) and Barseghyan et al. (2021a) use only cross-sectional variation and set-identify the parametric distribution of preferences only.⁵ 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 identification result. Section 4 presents the estimation procedure. Section 5, we present our empirical application. Finally, Section 6 concludes. All proofs can be found in Appendix A. Appendix B assesses the performance of our estimator in simulations. Appendix C contains additional estimation results.

2. Model

We consider an environment where choices are made from a random latent finite choice set \mathbf{D} .⁶ 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 $\mathbf{u} = \{\mathbf{u}_s\}_{s \in \mathcal{S}}$. 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

⁵Lu (2014) also provides a set of conditions that ensure that a system of moment inequalities he builds uniquely identifies the parameter of interest.

⁶We use boldface font (e.g. \mathbf{D}) to denote random objects and regular font (e.g. D) for deterministic ones.

not grow with sample size.⁷ 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 $(\mathbf{y}_s)_{s \in \mathcal{S}}$ 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 $\mathbf{x} = ((\tilde{\mathbf{x}}_{y,s}^\top)_{y \in \mathcal{Y}, s \in \mathcal{S}}, (\mathbf{r}_s)_{s \in \mathcal{S}})^\top$ 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 $\left\{ \left(\mathbf{y}_s^{(i)} \right)_{s \in \mathcal{S}}, \mathbf{x}^{(i)} \right\}_{i=1}^n$ drawn from a joint distribution of $(\mathbf{y}_s)_{s \in \mathcal{S}}$ and \mathbf{x} (a panel dataset).

Example 2. Suppose that there are n geographical markets (streets) and Y different fast-food restaurants as in [Currie et al. \(2010\)](#). 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} .

⁷ $|\mathcal{S}|$ denotes the cardinality of \mathcal{S} .

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 \mathcal{S} is (strictly) ordered by $<$. Without loss of generality, let $\mathcal{S} = \{1, 2, \dots, S\}$.

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

$$\mathbb{P}(\mathbf{y}_s = y_s \mid \{\mathbf{y}_{s'} = y_{s'}\}_{s'=1}^{s-1}, \mathbf{D} = D, \mathbf{x} = x) = \mathbb{P}(\mathbf{y}_s = y_s \mid \mathbf{y}_{s-1} = y_{s-1}, \mathbf{D} = D, \mathbf{x} = x).$$

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.⁸ 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 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, ε_s 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 as in the classical treatment of the Random Utility Model (RUM, [McFadden, 1973](#)) with observed choice sets. For instance, it is often assumed in

⁸[Mbakop \(2017\)](#) uses markovianity of order statistics to identify the distribution of private valuations in auction settings.

the analysis of differentiated products demand systems using individual level data (e.g., [Lu, 2014](#), [Crawford et al., 2021](#)).⁹ 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*

$$\mathbb{P}(\mathbf{y}_s = y_s \mid \{\mathbf{y}_{s'} = y_{s'}\}_{s' \in \mathcal{S} \setminus \{s\}}, \mathbf{D} = D, \mathbf{x} = x) = \mathbb{P}(\mathbf{y}_s = y_s \mid \mathbf{D} = D, \mathbf{x} = x).$$

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 : $\mathbf{u}_{y,D,s} = \tilde{\alpha}_{y,D,s}(\boldsymbol{\beta}_{y,D,s}, \mathbf{x}) + \varepsilon_{y,D,s}$, where $\{\tilde{\alpha}_{y,D,s}\}$ are unknown functions that map random coefficients, latent to the econometrician, $\boldsymbol{\beta}_{y,D,s}$ and covariates to \mathbb{R} ; and $\boldsymbol{\beta}_{y,D,s}$ and $\{\varepsilon_{y,D,s}\}$ are taste shocks that are independent across consumers s , but potentially are correlated across y and D and each other. Moreover, they 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*

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

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

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

⁹A 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.

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 $(\varepsilon_{D,s,y})_{y \in \mathcal{Y}, D \in \mathcal{D}_x}$, 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 $F^{\text{RUM}}(y \mid y', D, x) > 0$ 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 $(\varepsilon_{D,s,y})_{y \in \mathcal{Y}}$ is \mathbb{R}^Y for all D and s

(e.g. $\boldsymbol{\epsilon}_{D,s,y}$ 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 unobserved heterogeneity that is persistent across \mathcal{S} . 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 $F^{\text{RUM}}(y \mid y', D, x) = 0$. 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}

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} : $\mathcal{S}_1 = \{1, \dots, K\}$ and $\mathcal{S}_2 = \{K+2, \dots, 2K+1\}$. 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 $\mathbf{y}(\mathcal{S}_i) = (\mathbf{y}_s)_{s \in \mathcal{S}_i} \in \mathcal{Y}^K$. In other words, we partition the sequence $\{\mathbf{y}_s\}_{s \in \mathcal{S}}$ 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 .

Treating \mathbf{D} as a latent type in a finite mixture, we can apply the ideas in Hu (2008) as long as we can construct 3 independent measures of the latent type. In particular, the following lemma allows us to do this using the partition we defined in the previous paragraph.

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} .*

For all y^K, y, y', x , and D in the support, define $G(y^K | y, y', D, x, \mathcal{S}_i)$ as

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

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

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

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.¹⁰ Note that when K is small, in order to satisfy Assumption 5, it is necessary to restrict the size of $\mathcal{D}_{x,y,y'}$ (i.e. assume sparsity). 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 $m(\cdot | x)$ and $F^{\text{RUM}}(\cdot | y', D, x)$ are identified for all x, y' , and D in the support.*

¹⁰In the context of auctions, a similar assumption for $K = 1$ has been made in An (2017), Mbakop (2017), and Luo (2020).

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.

Intuitively, under Assumption 4, observing a decision-maker (DM) who over time selects either of two alternatives (e.g., y_1, y_2, y_1) suggests that she considers these two alternatives, whereas a DM selecting more alternatives (e.g., y_1, y_2, y_3) likely has a larger choice set. This variation in choices across instances helps pinpoint the choice sets, although we do not assume that S grows, meaning that a DM always picking y_1 may still consider all alternatives. Once choice sets are identified, the variation in choices within a given set (i.e. across DMs with the same choice set) enables the identification of the distribution of choices conditional on choice sets. Hence, our model embodies two competing forces: consideration and preferences. Without the repeated choice structure of the data, it is typically difficult to determine whether a good is selected less frequently due to being seldom considered or rarely chosen when considered. However, in our model, if many DMs select a good but not frequently across choice problems, we can infer that the good is often considered but typically dominated by another option. Conversely, if a good is chosen by few DMs but frequently when chosen, it indicates the good is rarely considered but often selected when it is.

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.¹¹ 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.

¹¹Under Assumption 2', we need to observe at least 3 choices instead of 5.

3.2. Linear Independence and Choice Sets

To better understand Assumption 5 consider the following simple examples. Suppose that $Y = 3$ (i.e., $\mathcal{Y} = \{1, 2, 3\}$) 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 Assumption 5 is equivalent to checking whether the following matrix has full column rank:¹²

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

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 number of possible choice sets is bigger than the number of possible values y^K can take and Assumption 5 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)

$y^K \setminus D$	$\{3\}$	$\{1,3\}$	$\{2,3\}$	$\{1,2,3\}$
$(1, 1)^\top$	0	\times	0	\times
$(1, 2)^\top$	0	0	0	\times
$(2, 2)^\top$	0	0	\times	\times
$(3, 3)^\top$	1	\times	\times	\times

where \times denotes the nonzero elements of the matrix, and Assumption 5 is satisfied. As these examples demonstrate, for Assumption 5 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.

¹²The columns of this matrix correspond to different elements of $\mathcal{D}_{x,y'}$. The rows correspond to different values y^K can take.

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 Assumption 5. Proposition 1 is demanding in terms of observables – we need to observe DMs for at least $2Y + 3$ time periods. However, to the best of our knowledge, this is the only identification result available in the literature that does not impose *any* restrictions on $\mathcal{D}_{x,y'}$. Note that, using the Bayes rule, Assumption 5 can be rewritten in terms of the distribution over choice sets conditional on the history of choice. Thus, Proposition 1 implicitly states that if the history of choice is long enough, then it provides enough information to recover the choice sets.

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. These restrictions imposes additional structure on G that guarantees that the matrix associated with it has full rank.

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 $k = 2, \dots, d_{D,x,y'}$.*
- (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.

Next, we give several examples of environments when conditions in Proposition 2 hold.

Example 3 (Distance). Consider DMs that live in a location and who are choosing a restaurant for dinner. DMs have transportation technology with (heterogeneous) efficiency given by $\mathbf{e} \in \{e_k\}_{k=1}^{d_{D,x,y'}}$, $e_k < e_{k+1}$. The realization of \mathbf{D} captures the set of restaurants the DM can get to: $\mathbf{D} = \{y \in \mathcal{Y} : d(y) \leq \mathbf{e}\}$, where $d(y)$ is the distance to restaurant y . Condition (i) in Proposition 2 is satisfied. Each instance s can index choices of the same DM at different time periods whose mode of transportation does not change over time or choices of different DMs with the same mode of transportation.

Example 4 (Search Engine). Consider DMs that face the same list of recommendations by a search engine, for the same topic, but they differ in the number alternatives in the search results they check. For instance, some DMs only check the first 5 alternatives, others check the first 10, the rest check all search results. The realization of \mathbf{D} captures how far in the search results DM went. Condition (i) in Proposition 2 is satisfied due to the fact that DMs face the same search results and check them from top to bottom.

Example 5 (Brand/Product Loyalty). Suppose $\mathcal{D}_{x,y'} = \{\bar{D} \cup \{y_k\}\}_{k=1}^{d_{D,x,y'}}$, where $y' \in \bar{D}$ 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 $\bar{D} = \{1\}$.

Example 6 (Sleeping Agent). A DM is said to be a sleeping agent if she is in one of two possible states: (i) asleep, or (ii) awake. If the DM is asleep then she considers a default and fixed alternative $o \in \mathcal{Y}$, otherwise she considers \mathcal{Y} (see [Ho et al. \(2017\)](#), [Hortacsu et al. \(2017\)](#), [Abaluck and Adams \(2021\)](#)).¹³

Example 7 (Partitions). Consider a partition of \mathcal{P} of \mathcal{Y} . Note that a latent choice set with support on \mathcal{P} is consistent with excluded choices. Categorization as described in [Aguiar \(2017\)](#) and the nested stochastic choice function in [Kovach and Tserenjigmid \(2022\)](#) provide a behavioral foundation for this kind of latent choice set.

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

¹³[Abaluck and Adams \(2021\)](#) consider two models the alternative specific model in [Manzini and Mariotti \(2014\)](#) and a variant of the sleeping agent.

may draw their choice sets from completely different distributions. This is empirically relevant since it allows us to analyze consumers from very heterogeneous markets.

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 $\mathcal{D}_{x,y'}$ that depends on K . For example, when $K = 1$, it must be that for all x , $d_{D,x,y'} = |\mathcal{D}_{x,y'}| \leq Y$ (in comparison the total number of nonempty sets is $2^Y - 1$). In other words, the support of \mathbf{D} needs to be sparse if S is small relative to Y .¹⁴ As S grows, sparsity becomes less restrictive, until it can be completely dropped (Proposition 1). Nevertheless, sparsity can be satisfied in many empirical and theoretical settings as the following examples demonstrate.

Example 8 (Observable Bounds on Choice Sets). Suppose for every x , $\mathcal{D}_{x,y'} = \{D \subseteq \mathcal{Y} : L_{x,y'} \subseteq D \subseteq U_{x,y'}\}$, where $L_{x,y'}$ and $U_{x,y'}$ 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, $d_{D,x,y'} \leq |\mathcal{Y}|$ if and only if $2^{|\mathcal{U}_{x,y'} \setminus \mathcal{L}_{x,y'}|} \leq Y$. In the dataset used in Conlon and Mortimer (2013), $2^{|\mathcal{U}_{x,y'} \setminus \mathcal{L}_{x,y'}|} \leq 2^3 = 8$ (at most 3 stock-out events) while total number of products considered is 44.

Example 9 (Optimal Consideration with Sparsity). Consider the following modification of the optimal consideration introduced in Aguiar et al. (2023). A representative DM chooses her optimal attention strategy by solving

$$\begin{aligned} \max_{b \in \{0,1\}^{2^Y-1}, m \in \Delta^{2^Y-1}} \sum_{C \subseteq \mathcal{Y}, C \neq \emptyset} (\alpha(C)m(C) - K(m(C))) \\ \text{s.t. } m(C) \leq b_C, \text{ for all } C, \text{ and } \sum_C b_C \leq \tau, \end{aligned}$$

where $\Delta^K = \{x \in \mathbb{R}_+^K : \sum_k x_k = 1\}$; $\alpha(\cdot)$ measures the attractiveness of any set (e.g., the McFadden's surplus); $K(\cdot)$ is a convex and monotone cognitive cost function capturing difficulty

¹⁴In general, a sparsity condition is not needed 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.

of picking a consideration set; and an integer $\tau \leq Y$ capturing cognitive capacity of how many sets the DM can consider. The binary parameter b_C captures whether set C is active (i.e., if $b_C = 0$ implies that $m(C) = 0$). The constraint $\tau \leq Y$ induces sparsity, since at most Y sets can be active.¹⁵

Discussion of the Main Assumptions

The two main identifying assumptions we have maintained are (i) the random choice set is stable across choice instance; and (ii) either S is large enough or the support of the random choice set is sparse. The first assumption is new in the consideration sets literature because most of it is not focused on a repeated choices. When s is indexing short time periods and S is small, as in our application, we believe the first assumption is likely satisfied. Indeed, product loyalty and other forms of directed attention can be thought as a special case of system-1 thinking (Kahneman, 2003), hence, it is revised less often than utility maximization. Steiner et al. (2017) shows it can be optimal to acquire all the information immediately.¹⁶ These results, when combined with the results on limited consideration in Caplin et al. (2019), imply that choice sets are often formed at the very beginning of the time window. But if the time window is too long or S is large, the stability assumption may be violated. It is an open question how to relax it and allow for dynamics in the marginal distribution of the random choice set, specially if we want to maintain nonparametric point identification.

One may be concerned advertisement invalidates the choice set stability. We believe that this is not a central issue because of two reasons. First, elasticities of demand to advertisement may be very low.¹⁷ Second, advertisement campaigns are usually stable in a short time-window conditional on a market. Changing advertisement campaigns quickly in time is costly, and

¹⁵If one takes $\alpha(D) = \sum_{a \in D} \exp(u(a))/|D|$, where $\exp u(i) = i$, $i \in \{1, 2, 3, 4, 5\}$, as the normalized McFadden's surplus of a consumer that follows a logit additive random utility choice probability; $\tau = 5$; and $K(t) = t^2/2$, then the only active sets are $\{5\}$ and $\{4, 5\}$.

¹⁶Zhong (2022) formally describes under what conditions dynamic information acquisition is smoothed across time.

¹⁷Recent work on the Nielsen Homescan (Shapiro et al., 2021) have estimated that the elasticity of demand to TV-ads is very low in general and is statistically non-different from zero in two-thirds of the goods.

a short time window increases the likelihood DMs are exposed to the similar advertisement campaigns in time.¹⁸ From this discussion, it is also clear that the stability of sets can fail when the time window is too long and has low frequency. In those cases, it is unlikely that our assumption holds.

The sparsity assumption is natural in the consideration set literature as we witnessed from the examples illustrating it. However, it may fail in some setups. For example, in the random consideration set model of [Manzini and Mariotti \(2014\)](#) used in [Abaluck and Adams \(2021\)](#), the support of the random choice set is never sparse. Thus, if consideration at the individual level is not random but deterministic, then sparsity is a good assumption as long as we condition on enough observable covariates that limit the latent choice set heterogeneity. In that sense, our sparsity assumption may fail to hold if we do not condition on enough covariates. At the same time, if S is big enough, then sparsity is not needed and the framework of [Manzini and Mariotti \(2014\)](#) is covered by our results.

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. We conclude this section by noting that the markovianity assumption can be extended to longer choice histories. In this case, Propositions 1 and 2 are still valid. One will only need to redefine K . However, conditioning on longer histories would require larger S . As a result, one would need to think carefully whether choice sets are still stable across choices.

¹⁸The elasticities estimates of [Shapiro et al. \(2021\)](#) take into account dynamics by considering the stock of advertisement. They estimate long-run elasticities that are low and in many cases close to zero.

4. Estimation

Here we provide a computationally efficient and consistent estimator of m and F^{RUM} . To simplify the exposition and because of its relevance in our empirical application, we assume that $K = 1$. That is, the rank condition implies sparsity. We discuss how to adjust the procedure at the end of this section. Let P be the conditional probability mass function of choices conditional on covariates. That is,

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

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

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

Assumption 6 requires the analyst to have access to a consistent estimator of P . For instance, if a sample of i.i.d. observations is available and X is finite, then one can use the standard empirical conditional probability mass function as \hat{P} . 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](#)).¹⁹ The rate of convergence α_n depends on the asymptotic behavior of \hat{P} . For instance, the above estimator \hat{P} with discrete \mathbf{x} 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 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

$$d_D = \text{rank} \left(\left[\sum_{k=1}^Y P(i, j, k) \right]_{i, j \in \mathcal{Y}} \right),$$

¹⁹For a recent application of a sieve estimator with continuous covariates see, for instance, [Kashaev \(2020\)](#).

where $\text{rank}(A)$ is the rank of matrix A . Thus, under Assumption 6, we can estimate the upper bound for d_D , \hat{d}_D , by replacing P by \hat{P} in the above formula. Instead of using \hat{d}_D , 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. In our estimation routine, we take \hat{P} and \hat{d}_D as given.

Ideal Estimator

Let $\Delta^{a \times b}$ denote the space of matrices of size $a \times b$ with nonnegative entries and columns summing up to 1. Also, denote the distance between the estimated \hat{P} and the implied by $F \in \Delta^{Y \times d_D}$ and $M \in \Delta^{d_D}$ distribution as

$$\text{dist}(\hat{P}|F, M) = \sqrt{\sum_{y_1, y_3, y_5} \left\{ \hat{P}(y_1, y_3, y_5) - \left[\sum_{j=1}^{d_D} F_{y_1, j}^1 \cdot F_{y_3, j}^3 \cdot F_{y_5, j}^5 \cdot M_j \right] \right\}^2}.$$

Fix a collection of \hat{d}_D subsets of \mathcal{Y} , $\mathfrak{D} = \{D_j\}_{j=1}^{\hat{d}_D}$. Then we can compute

$$\begin{aligned} T(\mathfrak{D}) &= \min_{F^s \in \Delta^{Y \times \hat{d}_D}, M \in \Delta^{\hat{d}_D}} \text{dist}(\hat{P}|F, M) \\ \text{s.t. } &F_{y, j}^s \leq 1 \text{ (} y \in D_j \text{) for all } y, j, s, \end{aligned}$$

where the set of constraints forces the probability of choosing an alternative that is not considered (i.e., not in the choice set) to be zero.

The collection \mathfrak{D}^* that minimizes $T(\cdot)$ (and contains y and y') would deliver a consistent estimator of the support of the choice sets, and, thus, would give us a consistent estimator of F^{RUM} and m .²⁰ Unfortunately, when the number of possible choice sets has to be small, this procedure becomes computationally prohibitive for even relatively small Y . For instance, if, as

²⁰Since 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.

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 $\{1, 2, \dots, 5\}$. Even if $T(\mathfrak{D})$ is computed within 0.01 sec, finding \mathfrak{D}^* 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.²¹ However, if K is such that \hat{d}_D is close to $2^Y - 1$, and one can check all possible \mathfrak{D} , then one should use this estimator. That is, the estimator we propose in the next section should be used in cases when the optimization of $T(\mathfrak{D})$ over all possible \mathfrak{D} is not feasible.

Step-1 Estimator

The Step-1 estimator, is similar to the previous estimator, but it does not force the constraints captured by \mathfrak{D} . In particular, define \bar{F} and \bar{M} as²²

$$\bar{F}, \bar{M} = \arg \min_{F^s \in \Delta^{Y \times \hat{d}_D}, M \in \Delta^{\hat{d}_D}} \text{dist}(\hat{P}|F, M).$$

\bar{F} and \bar{M} are consistent estimators of F^{RUM} and m .²³ However, because of the sampling uncertainty and numerical optimization errors, the elements of \bar{F} that correspond to $F^{\text{RUM}}(y | D) = 0$ (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} that are smaller than a prespecified $\varepsilon > 0$.²⁴ Formally, we need the following strengthening of Assumption 4.

²¹Our 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.

²²Instead, 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 worse in our simulations.

²³If covariates are discrete, then instead of minimizing the Euclidean distance, one can also minimize the Kullback-Leibler divergence and obtain maximum-likelihood estimates.

²⁴In our application and simulations, we set $\varepsilon = 0.01$.

Assumption 7. For every $x \in X$, $D \in \mathcal{D}_x$, $y, y' \in D$, and some known $\varepsilon > 0$

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

Note that Assumption 7 does not require existence of such lower bound since it always exists by Assumption 4. We only need to know this bound.²⁵ Given \bar{F} and \bar{M} , let the Step-1 estimator of m and F^{RUM} , be $M^{s1} = \bar{M}$ and

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

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 Appendix B). 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 samples.²⁶ 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

No sparsity. Given the collection of all nonempty subsets of \mathcal{Y} , $\{D_j\}_{j=1}^{2^Y-1}$, let F^* and M^* be the solutions to

$$\begin{aligned} & \min_{F^s \in \Delta^Y \times 2^{Y-1}, M \in \Delta^{2^Y-1}} \text{dist}(\hat{\mathbb{P}} \mid F, M) \\ & \text{s.t. } F_{y,j}^s \leq \mathbb{1}(y \in D_j) \text{ for all } y, j, s, \end{aligned}$$

²⁵In applications, one can always conduct sensitivity analysis and make ε smaller until the results do not change. One can also use 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.

²⁶If some of the estimated sets appear more than ones (i.e., two columns of \tilde{F}^{s1} has the same zero components), then we can just drop one of them.

that is closest to F^{s1} and M^{s1} . This optimization procedure is similar to the one in the previous section. However, it does not impose the sparsity condition – all possible nonempty 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. If the Step-1 estimator is ignored at this step, then the resulting estimator may not be consistent. This is why, obtaining the Step-1 estimator is necessary for our procedure. The sparsity condition is essential for our identification result, thus, one has to enforce it in estimation.

Mixed-Integer Programming, MIP. Note that in contrast to the Step-1 estimator, F^* is forced to assign zeros at proper positions. But, in finite samples, since no restrictions on the number of choice sets is imposed, it may assign positive mass to more than \hat{d}_D sets. To solve this issue, we propose to solve the following MIP problem:

$$\begin{aligned} \hat{B}^{s2}, \tilde{M} = & \arg \min_{B \in \{0,1\}^{2^Y-1}, M \in \Delta^{2^Y-1}} \text{dist}(\hat{P}|F^*, M) \\ \text{s.t. } & M_j \leq B_j, \text{ for all } j, \text{ and } \sum_j^{2^Y-1} B_j \leq \hat{d}_D. \end{aligned}$$

Note that $B \in \{0,1\}^{2^Y-1}$ and the objective function is similar to the least-squares objective since F^* is fixed. Informally, one can think of the above estimation problem as a regression problem with $2^Y - 1$ “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). 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 \mathfrak{D}^{s^2} be the collection of sets estimated by \hat{B}^{s^2} . Now, since we consistently estimated the choice sets on the previous step, F^{s^2} and M^{s^2} that minimize $T(\mathfrak{D}^{s^2})$ would be consistent estimators of F^{RUM} and m .²⁷

4.1. Discussions and Computational Aspects

In cases, when the ideal estimator can not be computed, one should use the proposed two-step estimator. Note that after the choice sets are found, in principle, any parametric or nonparametric estimation of F^{RUM} can be conducted. Our procedure is extremely fast and can be easily applied to choice sets of moderate size (e.g., in our empirical application— $Y = 5$ —our procedure completes the estimation in less than one minute on a single core computer). Our main advantage is coming from employing MIP to our problem. Modern MIP 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 possible choice sets in our estimation procedure by restricting M (e.g., ruling out singleton choice sets, and the lower and upper bound restrictions $L_{x,y'}$ and $U_{x,y'}$ discussed in Section 3.2).

In addition we remark that, given the discrete nature of the estimator of the latent choice set and the use of MIP, deriving confidence sets for the true choice sets, and, thus, for m , and F^{RUM} is nontrivial. 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 . We leave the problem of constructing confidence sets for model parameters when the choice sets are also estimated for future work.

²⁷Similar to the Step-1 estimator, instead of minimizing the Euclidean distance, we can conduct maximum likelihood estimation here when covariates are discrete.

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

Here 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 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 assume Assumption 2'. We believe this assumption is reasonable in our illustrative empirical application.²⁸ Importantly, it allows us to use shorter panels ($S = 3$ instead of $S = 5$), which greatly reduces our data requirements.

²⁸We 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.

Data

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 in $S = 3$ consecutive trips in 2016-2018 (most of them are weekly purchases). We end up having a balanced panel of $S = 3$ consecutive choices of $n = 47,509$ households. (See Appendix C for further details on data construction)

Using Nielsen Retail Scanner and Homescan Panel data, we construct brand prices and sizes, and gather household demographics like income, head age, size, and zip-code. We then match each purchase with corresponding brand price and size. Due to a large sample but many 3-digit zip-codes, we follow Nevo (2001) in grouping nearby households with similar prices, resulting in 34 markets. Further details on market construction are in Appendix C. Lastly, we aggregate prices at the market-brand level.

We already mentioned that since we focus on weekly purchases, 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 and there are unobserved shocks to preferences over cereals that are correlated across time. Also, any product innovation within brands may take more time than three weeks which is the modal time window in our dataset. That is, we believe that after controlling for observed characteristics and choice sets, all variation in choices is driven by idiosyncratic taste shocks.²⁹ 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.

²⁹Here 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.

Nonparametric Estimation of Consideration Sets and Market Shares

We estimate m and F^{RUM} conditioning on every market and covariate value. Since we aggregate the covariates on the market level, variation in covariates is only driven by variation across markets. Let $\hat{m}(D|j, x)$ denote the estimated probability that set D is considered in market j given covariate value x . Using the estimated \hat{m} , first, we find that 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 support our sparsity assumption.

Next we find that only 41 percent of DMs consider all brands. This shows that the majority of DMs violate the classical assumption of full consideration. About 30 percent of households considered all but one brands, and there is a sizeable fraction of those who only consider one brand (about 16 percent). This evidence supports the existence of important latent set heterogeneity.

We consider the estimates of F^{RUM} per market and consideration set. To simplify the exposition, we focus on the 4 biggest (in terms of numbers of observations) markets ($j \in \{1, 2, 13, 16\}$). Table 1 displays estimated market shares assuming that every DM considered all brands and estimated market shares obtained via our procedure, respectively. This table suggests that ignoring latent set heterogeneity may lead to substantially biased estimates of market shares (e.g., about 33 percentage points for other brands in market 13) or even reverse ranking of alternatives in terms of market shares (e.g., on average, GM leads in 4 out of 5 markets under full consideration and is the largest only in market 16 when we take into account latent choice sets).

Moreover, DMs that are different in terms of choice sets, display different preferences over brands. For instance, in market 1, those who consider all 5 brands prefer K over all other brands. At the same time, those who do not consider Q predominantly buy other brands of cereals (see Table 12 in the appendix). This emphasizes the importance of allowing for correlation between preferences and choice set when estimating the model.

Table 1 – Full Consideration Market Shares

Assuming Observed Choice Sets					Assuming Unobserved Choice Sets				
Brand/Market	1	2	13	16	Brand/Market	1	2	13	16
CTL	0.15	0.152	0.207	0.161	CTL	0.1	0.099	0.123	0.117
GM	0.316	0.293	0.269	0.311	GM	0.334	0.319	0.219	0.372
K	0.283	0.286	0.293	0.279	K	0.346	0.32	0.097	0.226
O	0.183	0.219	0.169	0.186	O	0.132	0.202	0.497	0.196
Q	0.069	0.05	0.062	0.063	Q	0.088	0.061	0.064	0.089

Parametric Estimation of Price Elasticity with Hidden Choice Set Variation

Note that $\mathbf{sh}_{y,D,j,s} = \hat{F}_s^{\text{RUM}}(y|D,j)$ 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.³⁰ So, we can proceed as if the estimated market shares are the true market shares and parametrize F^{RUM} .³¹ In our application, we take the standard logit specification of F^{RUM} .³² 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 $\alpha_{y,D} + \beta_D \mathbf{p}_{y,j} + \mathbf{r}_j^T \gamma_{y,D} + \xi_{y,D} + \Delta \xi_{y,D,j,s} + \epsilon_{i,y,D,s}$, 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 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; $\epsilon_{i,y,D,s}$ 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.

³⁰Our DMs are RTE cereal frequent buyers. Hence, we do not allow for the option of not buying anything.

³¹The 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.

³²The 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](#)).

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

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

where $\alpha_{y,\bar{y},D}^* = \alpha_{y,D} - \alpha_{\bar{y},D} + \xi_{y,D} - \xi_{\bar{y},D}$ and $\gamma_{y,\bar{y},D}^* = \gamma_{y,D} - \gamma_{\bar{y},D}$. 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.³³

The parameter of interest is own-price elasticities under the assumption that the distribution over the choice sets m does not depend on prices:³⁴

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

where $\text{Elas}_{y,D,j,s} = \beta_D p_{y,j} (1 - sh_{y,D,j,s})$, and $sh_{y,j,s}$ is the observed share of brand y in market j at time s . If there is no choice set variation, then, under our parametrization, $\text{Elas}_{y,j,s} = \beta_{\text{Direct}} p_{y,j} (1 - sh_{y,j,s})$, where β_{Direct} is the price coefficient under the assumption that there is no choice set variation. Note that since in our model β_D is indexed by D , we allow for correlation between preferences and choice sets. We use variation across time and markets to estimate β_D and then obtain own price elasticities.

We report own-price elasticities for the largest in terms of observations market (Market 1) in [Table 2](#). 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 Q, respectively (i.e., $\text{Elas}_{y,D,j,t}$).

The “Direct” estimates of the own-price elasticities are similar to ones in [Nevo \(2001\)](#).³⁵

³³The details of construction of instruments can be found in our replication files.

³⁴See [Goeree \(2008\)](#) for similar exclusion restrictions.

³⁵In [Appendix C](#), we report the median across markets own-price elasticities with similar results.

Table 2 – 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 Q. Results are rounded to 2 digits.

However, the demand of those considering all brands is substantially less elastic than those who do not consider Q. 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 almost all brands. For Q, 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 category 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 Q cereal or only consider GM. Others, however, may consider everything. In general, these frequent buyers may have strong opinions about what they like to consider and what they avoid. 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.

6. Conclusion

We showed that observing 3 or more choices from the same latent choice set can be sufficient to nonparametrically identify the joint distribution of choice sets and choices in discrete-choice

models when choice sets are not observable. We require 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 substantial unobserved choice set heterogeneity and correlation between preferences and choice sets that can contaminate estimates of the own-price elasticities of demand.

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.
- Aguiar, Victor H (2017) “Random categorization and bounded rationality,” *Economics Letters*, 159, 46–52.
- Aguiar, Victor H., Maria Jose Boccardi, Nail Kashaev, and Jeongbin Kim (2023) “Random utility and limited consideration,” *Quantitative Economics*, 14 (1), 71–116.
- 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.
- 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.
- Caplin, Andrew, Mark Dean, and John Leahy (2019) “Rational inattention, optimal consideration sets, and stochastic choice,” *The Review of Economic Studies*, 86 (3), 1061–1094.
- 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.
- Gentry, Matthew L (2016) “Displays, sales, and in-store search in retail markets.”
- Goeree, Michelle Sovinsky (2008) “Limited information and advertising in the US personal computer industry,” *Econometrica*, 76 (5), 1017–1074.
- 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.
- Ho, Kate, Joseph Hogan, and Fiona Scott Morton (2017) “The Impact of Consumer Inattention on Insurer Pricing in the Medicare Part D Program,” *The RAND Journal of Economics*, 48 (4), 877–905.
- Hortacsu, Ali, Seyed Ali Madanizadeh, and Steven L Puller (2017) “Power to choose? An analysis of consumer inertia in the residential electricity market,” *American Economic Journal: Economic Policy*, 9 (4), 192–226.
- 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*.
- Kashaev, Nail and Victor H Aguiar (2022) “A random attention and utility model,” *Journal of Economic Theory*, 204, 105487.
- Kitamura, Yuichi and Louise Laage (2018) “Nonparametric analysis of finite mixtures,” *arXiv preprint arXiv:1811.02727*.
- Kovach, Matthew and Gerelt Tserenjigmid (2022) “Behavioral Foundations of Nested Stochastic Choice and Nested Logit,” *Journal of Political Economy*, 130 (9), 2411–2461, [10.1086/720399](https://doi.org/10.1086/720399).
- 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*.
- Luo, Yao (2020) “Unobserved heterogeneity in auctions under restricted stochastic dominance,” *Journal of Econometrics*, 216 (2), 354–374.
- Manzini, Paola and Marco Mariotti (2014) “Stochastic Choice and Consideration Sets,” *Econometrica*, 82 (3), 1153–1176, [10.3982/ECTA10575](https://doi.org/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.
- 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.
- Shapiro, Bradley T., Günter J. Hitsch, and Anna E. Tuchman (2021) “TV Advertising Effectiveness and Profitability: Generalizable Results From 288 Brands,” *Econometrica*, 89 (4), 1855–1879, <https://doi.org/10.3982/ECTA17674>.
- Steiner, Jakub, Colin Stewart, and Filip Matějka (2017) “Rational Inattention Dynamics: Inertia and Delay in Decision-Making,” *Econometrica*, 85 (2), 521–553.
- Xiao, Ruli (2018) “Identification and estimation of incomplete information games with multiple equilibria,” *Journal of Econometrics*, 203 (2), 328–343.
- Zhong, Weijie (2022) “Optimal Dynamic Information Acquisition,” *Econometrica*, 90 (4), 1537–1582.