Discrete Choice and Welfare Analysis with Unobserved Choice Sets*

Victor H. Aguiar　　Nail Kashaev†

vaguiar@uwo.ca　　nkashaev@uwo.ca

This version: July 14, 2019

Abstract We propose a framework for doing sharp nonparametric welfare analysis in discrete choice models with unobserved variation in choice sets. We recover jointly the distribution of choice sets and the distribution of preferences. To achieve this we use panel data on choices and assume *nestedness* of the latent choice sets. Nestedness means that choice sets of different decision makers are ordered by inclusion. It is satisfied, for instance, when the choice set variation is the result of either a search process or unobserved feasibility. Using variation of the uncovered choice sets we show how to do ordinal (nonparametric) welfare comparisons. When one is willing to make additional assumptions about preferences, we show how to nonparametrically identify the ranking over average utilities in the standard multinomial choice setting.

JEL classification numbers: C14, C5, D6

Keywords: welfare analysis, random utility, discrete choice, random consideration sets.

*We would like to thank Roy Allen, Salvador Navarro, David Rivers, and Bruno Salcedo 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.

†Department of Economics, University of Western Ontario.

1. Introduction

This paper proposes a fully nonparametric framework for doing welfare analysis in discrete choice models when the choice sets that decision makers (DMs) face are unobserved by the researcher. Our main contribution is the nonparametric identification of the joint distribution of choice sets and preferences when we have access to a panel dataset on choices. This distribution allows us to do sharp counterfactual analysis of welfare.

Our work provides a methodological bridge between the decision theoretic literature on stochastic choice that has been based fundamentally on choice set variation[1] and the discrete choice literature that has exploited covariate variation to identify the distribution of preferences under parametric assumptions (Train (2009)). We recover the unobserved choice set variation using a panel dataset and studying the joint distribution of choices across time. This data is richer than the usual (spot) market shares used in the parametric discrete choice literature but crucially does not presume that any choice set variation is observed.

The classical nonparametric treatment of discrete choice under random utility (McFadden and Richter (1990))[2] uses the choice set variation and the observed substitution patters that arise from it to identify the distribution of preferences. The key advantage of using choice set variation is that it imposes essentially no restrictions on the distribution of preferences beyond stability of this distribution across choice sets. However, the econometrician 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,

[1]See 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).

[2]From a decision theoretic tradition random utility was initially studied by Block and Marschak (1960) and Falmagne (1978). They provided necessary and sufficient conditions for a probabilistic choice rule to be consistent with random utility in an environment with full choice set variation.

and 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 propose a new approach to jointly identify the distribution of unobserved choice sets and the distribution of preferences without any parametric assumptions. In particular, we identify the probability that any given choice set is considered and the probability that an alternative is picked given a choice set. If the recovered choice set variation is full (i.e. any choice set is faced with positive probability by some DM), then, in terms of identification power, our setup is as informative as the ideal data from stochastic choice models where researchers have full observable choice set variation.

We provide two applications of our main identification result. They exploit choice set variation. First, we focus on doing nonparametric ordinal welfare analysis. We provide a methodology to compute the fraction of DMs that are better-off (worse-off) when *moving* from one choice set to another one. This analysis imposes essentially no restriction on preferences and takes the form of a linear program. It also can be used to rank choice sets in a Pareto sense making the analysis robust to critiques of cardinal approaches of welfare evaluation. Second, we provide a way to identify average welfare under additional restrictions on preferences (i.e., additive random utility with independent identically distributed (i.i.d.) or exchangeable shocks). This second application allows cardinal comparisons that may be more informative than the ordinal results in our first application while keeping the nonparametric nature of the exercise. In both cases, we show how uncovering the latent choice set variation using our methodology provides a way to do robust welfare analysis.

Though we focus on exploiting full variation of choice sets without making any (semi)parametric assumptions about preferences, our analysis does not preclude the use of variation in other observables coupled with some (semi)parametric restrictions on preferences. Moreover, the otherwise problematic, unobserved choice set variation may bring additional information. This information can deliver the identification of random utilities under assumptions that are substantially weaker than the assumptions that are usually made in models without choice set variation (see Section 4.2).

The key assumptions that we make are: (i) the distribution of choice sets and the distribution of preferences are independent, (ii) the distribution of choice sets vary with observable covariates, and for a fixed value of these covariates, the collection of unobservable choice sets satisfies *nestedness*.

Independence between preferences and the choice sets allows us to disentangle the preference maximization part of the choice process from the randomness that is coming from random feasibility. The independence assumption also allows us to compare individuals that faced different choice sets without making any parametric assumptions. Starting with McFadden and Richter (1990), the independence assumption is central to most work in discrete choice and is implicitly assumed even when choice sets are observed (Kitamura and Stoye (2018)).[3]

Nestedness requires the unobserved choice sets faced by DMs that have the same observable covariates to be ordered by set inclusion. That is, for any two choice sets, D_1 and D_2, either D_1 is contained in D_2 or vice verse. This assumption holds naturally in many situations of interest.

First, the unobserved choice set variation can arise due to limited consideration. In this setup, our nestedness restriction can be interpreted as search behavior with random thresholds. In contrast, previous works on consideration sets were driven by item-dependent attention.[4] Consider, for instance, DMs that scroll down the catalog of an online store. Some DMs may see more items in the catalog because of unobserved patients level. Since all DMs see the same catalog but only differ in the time they stop scrolling, their unobserved choice sets naturally satisfy the nestedness condition.

Second, (random) unobserved feasibility can also generate choice set variation that

[3] Kitamura and Stoye (2018) further discuss this point and proposes a way to deal with departures from independence, when choice sets are observable.

[4] See Goeree (2008), Barseghyan et al. (2019), and Dardanoni et al. (2019).

satisfies the nestedness condition. Consider DMs maximizing their preferences over a discrete (discretized) choice set given budget constraints. Every budget constraint is characterized by prices and an unobserved disposable income (Polisson and Quah (2013)). Different realizations of disposable income will imply different sets of feasible alternatives. In many situations prices are observed but disposable income is not. Nonetheless, the unobserved budgets satisfy the nestedness restriction, thus making them amenable to our treatment.

Another example is restaurant choice. One can think of the set of available restaurants being determined by the distance DMs are able to travel. In other words, the distance a DM can travel to a restaurant of her choice is driven by the mode of transportation she has access to. So the choice set of those who, for instance, have a car contains the choice set of those who can only walk. More generally, different DMs have different costs of transportation per unit of distance.

We base our identification strategy on the insights from discrete nonclassical measurement error results in Hu (2008). We require at least three observed choices over a unit of measurement (time or location) that are conditionally independent conditional on the unobserved choice set. Using these three choices we show that the nestedness assumption on latent choice sets and the substitution patterns arising from utility maximization behavior provide primitive conditions that ensure the identification of both the distribution over choice sets and the distribution of preferences. We differ from Hu (2008) in that we do not need to impose any strict monotonicity (ranking) condition and we do not need to know the number of possible choice sets. Strict monotonicity restrictions are usually imposed to match anonymous functions to latent states. In our settings, the latent choice sets have a structural interpretation and we do not need to rank them.

Our identification strategy is constructive, thus leading to a natural and easy to implement estimation procedure that we outline. The estimation procedure is essentially based on computations of ranks and eigenvalues of different matrices. We also do not require knowledge of the number of possible choice sets. Moreover, we show how to identify and estimate this number in practice.

Although nestedness is a flexible condition that covers many cases of interest, in order to expand the scope of our analysis, we provide a generalization of our main result for nonnested cases. In the general case, we only require a simple linear independence restriction imposed on the conditional choice distribution. This restriction is satisfied, for instance, when choice sets form a partition, but it is also satisfied under alternative restrictions on the choice sets. We also provide a counterexample that fails to satisfy this linear independence condition.

The closest work to ours is Crawford et al. (2019). They also study discrete choice with unobservable but heterogeneous choice sets. Crucially, we differ from their work in that our approach is fully nonparametric. In particular, we do not impose any restriction on the distribution of preferences while they work with the stylized multinomial logit model of choice. They also impose parametric restrictions on the choice set distribution. Their main contribution is the parametric identification of a particular model of random utility. In contrast, we focus on doing welfare analysis nonparametrically. Another related work is Dardanoni et al. (2019). They 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 and only rely on the nestedness condition.[5]

[5] In an alternative strand of the literature, Abaluck and Adams (2017) exploit parametric restrictions on preferences and consideration to achieve identification without panel datasets and without exclusion restrictions. In practice, they need the substitution matrix associated with their model to exhibit asymmetries. This may be restrictive, in fact, Allen and Rehbeck (2019) show how the important models of limited consideration of Manzini and Mariotti (2014) and Brady and Rehbeck (2016) can exhibit a symmetric substitution matrix. In our nonparametric framework these issues are irrelevant.

Outline

Section 2 formally presents the model. Section 3 contains our main identification result for the distribution of choice sets and preferences. Section 4 presents two applications of our main result. We provide a methodology to do nonparametric ordinal welfare analysis with minimal assumptions, and also study the case of average welfare analysis within our setup. Section 5 presents a generalization of our main identification result covering nonnested choice sets. Finally, Section 6 concludes. All proofs and the omitted details about estimation can be found in Appendix A.

2. Model

2.1. Choice Sets and Preferences

We consider an environment where at time $t = 0$ a decision maker (DM) is faced (at random) with a finite choice set, $\mathbf{D}$, and then at every $t \in \mathcal{T} = \{1, 2, \ldots, T\}$ chooses the alternative in $\mathbf{D}$ that maximizes her random preferences.[6] The preferences at every moment of time are captured by the random strict preference orders represented by random utility functions $\mathbf{u} = \{\mathbf{u}_t\}_{t\in\mathcal{T}}$ that 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 $Y = \{1, 2, \ldots, d_y\}$, where d_y is a finite constant.

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

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

index, in different applications it may have different meanings. For instance, t may index different DMs facing the same choice set. We only require the realization of the unobserved choice set $\mathbf{D}$ to be the same across t (see Example 3 for more details).

Assumption 1 (Observables)**.** *The researcher observes (can consistently estimate) the joint distribution of* $\{\mathbf{y}_t\}_{t\in\mathcal{T}}$ *and* $\mathbf{x}$*.*

In order to disentangle random preferences (i.e., preference heterogeneity) from variation in choice sets (i.e., latent choice sets) we impose the following independence condition between preference heterogeneity and the distribution of random choice sets.

Assumption 2 (Independence)**.** $\mathbf{D}$ *and* $\mathbf{u}$ *are independent conditional on* $\mathbf{x}$*. That is, there is a conditional probability measure* $m(\cdot|x)$ *such that*

$$\mathbb{P}\left(\mathbf{D} = D|\mathbf{x},\mathbf{u}\right) = m(D|\mathbf{x}) \quad \text{a.s.} \tag{1}$$

for all $D \in \mathcal{D}_x$*, where* $\mathcal{D}_x$ *is the conditional support of* $\mathbf{D}$ *conditional on* $\mathbf{x} = x$*.*

Independence is a natural restriction in this environment as we want to decompose the observed distribution of choices into its choice set variation (captured by m) and preference heterogeneity (captured by $\mathbf{u}$) components without making any parametric assumptions.

Also, we are interested in modelling decision making that is the result of a two-stage process where the DM's choice set is realized first and only then she chooses rationally from her choice set. There are two unobserved sources of heterogeneity: feasibility and preferences. Assumption 2 emphasizes that, after conditioning on covariates $\mathbf{x}$, preferences of DMs do not affect the choice sets they face. This allows us to separate randomness that comes from a random utility from a random choice set uncertainty and to avoid any parametric assumptions about the distribution of preferences and choice sets. We want to remark that the independence assumption is central to most work in discrete choice, starting with McFadden and Richter (1990).[7]

While we work with the environments where the choice sets do not change over time (the random choice set $\mathbf{D}$ is not indexed by t), we still want to allow agents to make different choices in different time periods. Following the classical treatment of the Random Utility Model (RUM) with observed choice sets we make the following assumption.

Assumption 3 (Conditionally independent preferences)**.** *Conditional on the realization of covariates, preferences are i.i.d. across time. That is,*

$$\{\mathbf{u}_t\}_{t=\mathcal{T}}|(\mathbf{x}=x) \sim \text{i.i.d.},$$

for all $x \in X$.

Assumptions 2 and 3 together 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 and covariates, choices of the DM are also i.i.d.. We want to emphasize that Assumptions 2 and 3 do not restrict the dependence structure between $\mathbf{u}$ and $\mathbf{x}$, or between $\mathbf{D}$ and $\mathbf{x}$ (See examples below). Moreover, the choices $\{\mathbf{y}_t\}_{t\in\mathcal{T}}$ are correlated over time through the latent choice set $\mathbf{D}$.

Since, after conditioning on the realization of the choice set and covariates, the choices are i.i.d., we can define

$$y^{\text{RUM}}(D,u) = \arg\max_{y'\in D} u(y')$$

for every choice set D and utility function u, and

$$F^{\text{RUM}}(y|D,x) = \mathbb{P}\left(y = y^{\text{RUM}}(\mathbf{D},\mathbf{u})|\mathbf{x}=x,\mathbf{D}=D\right)$$

[7]See Kitamura and Stoye (2018) for an extension of the McFadden and Richter (1990) with endogenous budgets sets.

for every choice y, covariate x, and choice set $D \in \mathcal{D}_x$. F^{RUM} is a conditional probability mass function that captures the random utility maximization part of DMs choices. Thus, in terms of random choices, Assumption 3 implies that

$$\{\mathbf{y}_t\}_{t=\mathcal{T}}|(\mathbf{D} = D, \mathbf{x} = x) \sim \text{i.i.d. } F^{\text{RUM}}(\cdot|D, x). \tag{2}$$

Because of the independence assumption and finiteness of the choice set, we can rewrite the conditional distribution over observed choices at any $t \in \mathcal{T}$ as a finite mixture model:

$$\Pr(\mathbf{y}_t = y|\mathbf{x} = x) = \sum_{D \in \mathcal{D}_x} m(D|x) F^{\text{RUM}}(y|D, x)$$

for all x and y. Using the data on choices and covariates, the researcher is interested in recovering the conditional distribution over choice sets captured by m and the random utility maximization aspects of the model captured by F^{RUM}.

2.2. Nestedness of Choice Sets

Given a grand choice set Y, the biggest possible support for $\mathbf{D}$ is $2^Y \setminus \{\emptyset\}$. Since we may only have data on choices from Y, then the information contained in $|Y|$ outcomes, where $|Y|$ denotes the cardinality of Y, is not enough to identify the distribution on $\left|2^Y \setminus \{\emptyset\}\right|$ points. That is why in order to pin down the distribution of choice sets we need to restrict the support of $\mathbf{D}$.

Assumption 4 (Nestedness)**.** *For every $x \in X$, the conditional support of the random choice set, $\mathcal{D}_x$, is a collection of nested sets. That is, for all $x \in X$, $\mathcal{D}_x = \{D_k\}_{k=1}^{d_{D,x}}$ such that $D_{k-1} \subseteq D_k$ for $k = 2, \ldots, d_{D,x}$.*

Note that we do not assume that the identity or the number of support points $d_{D,x} = |\mathcal{D}_x|$ is known or fixed. Distinct (in terms of covariates) DMs may be endowed with completely different distributions over different choice sets.

The following examples demonstrate the applicability of our setup to different choice problems.

Example 1 (Multinomial choice)**.** Suppose that preferences over a set of products are captured by

$$\mathbf{u}_t(y) = v(y) + \mathbf{e}_{y,t},$$

where $\{\mathbf{e}_{y,t}\}_{y\in Y,t\in\mathcal{T}}$ are taste shocks and $v(y)$ is alternative y mean utility. Each alternative y is characterized by some unobservable attributes such as price or quality. The DM forms her choice set by ranking the alternatives according to one of the attributes and given a realization of a random threshold. She considers only those alternatives that are below the threshold. Formally, if $\{\mathbf{e}_{y,t}\}_{y\in Y,t\in\mathcal{T}}$ are i.i.d. across time and are independent of $\mathbf{x}$ and the random threshold, then Assumptions 2 and 3 are satisfied. Assumption 4 is satisfied because we assume that the consideration set is formed by ranking an attribute of alternatives. The ranking over attributes creates a ranking over alternatives such that we can index the alternatives from the lowest value of the attribute to the biggest: $Y = \{1, 2, \dots, d_y\}$. The DM only considers those alternatives whose index is above a realization of a random threshold $\boldsymbol{\tau}$. Hence, for each pair of realization of $\boldsymbol{\tau}$, τ and τ', the consideration sets are such that

$$\tau \leq \tau' \iff \{y \in Y : y \geq \tau'\} \subseteq \{y \in Y : y \geq \tau\}.$$

Example 2 (Budgets)**.** Consider a DM that, given a price vector $\mathbf{x}$ and an unobserved by the analyst income $\mathbf{e} > 0$, faces a budget $\mathbf{D} = \{y \in \mathbb{N}^{d_y} \;:\; \mathbf{x}^{\mathsf{T}} y + \boldsymbol{\eta} \leq \mathbf{e}\}$, where $\boldsymbol{\eta}$ is unobserved expenditures on divisible goods. She maximizes her utility $\mathbf{u}_t$ over $\mathbf{D}$. If after conditioning on observed prices $\mathbf{x}$ the preferences over observed choices $\{\mathbf{u}_t\}_{t=\in\mathcal{T}}$ are i.i.d. across time and are independent from $\mathbf{e} - \boldsymbol{\eta}$, then Assumptions 2 and 3 are satisfied. Assumption 4 in this example is satisfied since

$$e \leq e' \iff \{y \in \mathbb{N}^{d_y} \,:\, x^\intercal y \leq e\} \subseteq \{y \in \mathbb{N}^{d_y} \,:\, x^\intercal y \leq e'\}.$$

Example 3 (Different Stores)**.** In this example, we use t to index different DMs. Consider a population of locally isolated markets (say small towns). Every market has a store, and markets are different in terms of the size of the stores they have. Let $\mathbf{D}$ represent the set of varieties one may find in a market. Assume that there are at least T DMs in every market. At the market $\mathbf{D} = D$ every DM t is endowed with a utility function $\mathbf{u}_t$ and picks the best alternative from D. If the preferences of DMs are independent of each other and from the size of the store they go to (captured by $\mathbf{D}$), then Assumptions 2 and 3 are satisfied. Assumption 4 is satisfied if we assume that stores can be ranked according to their size: $D_1 \subseteq D_2$ means that a bigger store with the menu of varieties D_2 sells everything that a smaller store with the choice set D_1 does and maybe more.

Example 4 (Distance)**.** Consider a DM who picks a restaurant at different time periods $t \in \mathcal{T}$. At every t the preferences of the DM over restaurants are captured by $\mathbf{u}_t$. 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 $\mathbf{D}$ captures the set of restaurants the DM can get to. If the type of transportation and preferences are independent, and preferences are i.i.d. across time, then Assumptions 2 and 3 are satisfied. Assumption 4 is satisfied in this example 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. More precisely, given different costs of transportation the same distance will be attainable for some consumers but not for others. But the cheaper the cost of transportation the farther these consumers can travel.

3. Identification

In this section, we establish identification of the distribution of the latent choice sets m and the conditional distribution of choices F^{RUM} from observed data on choices and covariates. First, we need an additional assumption.

Assumption 5 (Full Support)**.** *For every $x \in X$, $D \in \mathcal{D}_x$, and $y \in D$,*

$$F^{\mathrm{RUM}}(y|D, x) > 0.$$

Assumption 5 is a standard full support assumptions in discrete choice literature: every alternative in every choice set is chosen with positive probability. McFadden (1973) pointed out that in finite samples, Assumption 5 is not testable, since zero market shares are not distinguishable from arbitrarily small but positive market shares. Additionally, when an alternative is never observed in the data this may happen either because it enters no one choice set or there is always another alternative that is better. Assumption 5 excludes such cases.

We are ready to state our main result.

Theorem 1. *Suppose Assumptions 1-5 hold and $T \geq 3$, then $m(\cdot|x)$ and $F^{\mathrm{RUM}}(\cdot|D, x)$ are constructively identified for all $x \in X$ and $D \in \mathcal{D}_x$.*

Theorem 1 is remarkable because it recovers jointly and nonparametrically the distribution of choice sets and the conditional distribution of choices. To the best of our knowledge no other work on this topic achieves this.[8]

In one of the steps of the proof of Theorem 1 we use the eigendecomposition argument of Hu (2008) and Hu et al. (2013). That is why we need to observe the choices of the same individual at least three times. However, we do not need to impose

[8] Formally, for our identification result Assumptions 2 and 3 can be replaced by condition (2). So we can allow preferences to be correlated with choice sets. However, as we discussed before, to do robust welfare analysis, we need to assume independence between preferences and choice sets.

any monotonicity restrictions on F^{RUM}. Also note that we provide conditions on primitives of our model for the uniqueness of the eigendecomposition step, combining the nestedness of the choice sets and the implications of random utility behavior. Another difference from Hu (2008) and Hu et al. (2013) is that we do not need to know the number of possible choice sets $d_{D,x}$.

We conclude this section by emphasizing that our constructive identification strategy leads to a "plug-in"-type estimation strategy. For example, the cardinality of $\mathcal{D}_x$ satisfies

$$d_{D,x} = \operatorname{rank}\left(\left[\mathbb{P}\left(\mathbf{y}_1 = i,\ \mathbf{y}_2 = j|\mathbf{x} = x\right)\right]_{i,j\in Y}\right)$$

for all x.[9] So, a consistent nonparametric estimator of the joint distribution of choices delivers a consistent estimator of the number of possible choice sets. For more details on the estimation of the objects of interest see the proof of Theorem 1.

3.1. Full Choice Set Variation and Identification of F^{RUM}

Theorem 1 does not identify $F^{\mathrm{RUM}}(\cdot|D,x)$ for all possible choice sets (i.e., for all $D \in 2^Y \setminus \{\emptyset\}$). Instead, for a fixed value of the covariate x it identifies F^{RUM} on the conditional support of the nested choice sets $\mathcal{D}_x$.

In order to recover the random utility distribution over all possible choice sets and to be able to uncover all possible substitution patterns, we can strengthen our assumptions. We highlight that full choice set variation allows us to recover all possible ordinal information about the distribution of preferences in a fully nonparametric fashion. We underline that Theorem 1 does not use any variation in observed covariates x. In this section we use additional information contained in x to learn more about choices of DMs.

Suppose that the vector of covariates x can be partitioned into $z \in Z$ and $w \in W$. Let Z_w denote the conditional support of z conditional on $\mathbf{w} = w$.

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

Assumption 6 (Excluded covariate)**.** *$F^{\mathrm{RUM}}(\cdot|D,x) = F^{\mathrm{RUM}}(\cdot|D,x')$ for all $x \neq x'$ with $w = w'$, for all D.*

Assumption 6 implies that there are covariates that can serve as exclusion restrictions: changes in the value of these covariates generates variation in choice sets, but does not affect the distribution of preferences. In many multinomial choice environments prices of goods can be taken as excluded covariates, since increasing prices shrink the set of feasible goods while not affecting preferences of DMs. In Example 1 one can use the distance to the store as an excluded covariate.

Define

$$\mathcal{B}_w = \bigcup_{z \in Z_w} \mathcal{D}_{(z^\intercal, w^\intercal)^\intercal}$$

for every $w \in W$. Note that when w is fixed the preference distribution is also fixed. Thus, $\mathcal{B}_w$ contains all possible choice sets that DMs may face. The elements of $\mathcal{B}_w$ are not necessary nested since Assumption 4 needs to be satisfied after conditioning on both excluded and nonexcluded covariates. The variation in $\mathbf{z}$ can generate a substantial variation in the support of the random choice set $\mathcal{D}_x$ (See Figure 1). For example, $\mathcal{B}_w$ can be equal to $2^Y \setminus \{\emptyset\}$.

Corollary 1. *Under conditions of Theorem 1, if*

$$\mathcal{B}_w = 2^Y \setminus \{\emptyset\},$$

then $F^{\mathrm{RUM}}(\cdot|D,x)$ is identified for all $x \in X$ and $D \in 2^Y \setminus \{\emptyset\}$.

4. Welfare

In this section, we show how the choice set variation can help in welfare analysis. First, we establish sharp bounds on the fraction of individuals that are better off

$$\begin{tikzcd}
\mathcal{D}_{x_1} & \{a\} \arrow[r, leftrightarrow, dotted] \arrow[dr, leftrightarrow] & \{a,b\} \arrow[r, leftrightarrow, dotted] & \{a,b,c\} \\
\mathcal{D}_{x_2} & \{b\} \arrow[r, leftrightarrow, dotted] & \{a,b\} \arrow[dr, leftrightarrow] & \\
\mathcal{D}_{x_3} & & \{b,c\} \arrow[r, leftrightarrow, dotted] & \{a,b,c\}
\end{tikzcd}$$

Figure 1 – Excluded covariates generate choice set variation Without exclusion restrictions only comparisons within $\mathcal{D}_{x_i}$ are possible (e.g. dotted arrows). With exclusion restrictions we can also compare the distribution of choices among sets in different layers (e.g. solid arrows)

when moving from one choice set situation to another one. Second, we show how to identify the ranking of the mean utilities in multinomial choice models with additive random utility.

4.1. Ordinal Welfare Analysis

We can use our framework to make welfare comparisons. For a fixed $w \in W$ we study how welfare changes when DMs move from a choice set $B_0 \subseteq \mathcal{B}_w$ to a new choice set $B_1 \subseteq \mathcal{B}_w$. The welfare object of interest is the fraction of DMs that are better off or the fraction of DMs that are worse off in the new situation. This is a purely ordinal nonparametric comparison that only uses choice set variation. We will present the computation of welfare for the fraction of DMs that are better off, and the case of the fraction of DMs that are worse off is completely symmetric and therefore omitted.

First, let $\overline{U}$ be the set of (normalized) utility functions for which there is an isomorphism with the set of all linear orders (i.e., strict rational preferences) on Y. Namely, $\overline{U}$ contains one utility representation of each strict ranking or linear order. We define A as the matrix with rows indexed by the pairs (y, B) where $y \in B$ and $B \in \mathcal{B}_w$, and columns indexed by the utilities $u \in \overline{U}$, with entry (i, j) equal to 1, if $u(y) > u(y')$ for all $y' \in B \setminus \{y\}$, and zero otherwise. We collect all the information in $F^{\mathrm{RUM}}(\cdot|\cdot, x)$ into a vector $F_x^{\mathrm{RUM}} = (F^{\mathrm{RUM}}(y|B, x))_{y \in B, B \in \mathcal{B}_w}$. Let $\Delta(\bar{U})$ denote the simplex on $\bar{U}$. Then $\pi \in \Delta(\bar{U})$ can generate F_x^{RUM} if and only if

$$A\pi = F_x^{\mathrm{RUM}}. \tag{3}$$

That is, equation (3) fully characterizes the identified set for the distribution of preferences – the set of the preference distributions that can generate the observed data (McFadden and Richter (1990)).

Second, when computing the fraction of individuals that are better off when moving from a choice set situation B_0 to a new choice set situation B_1, we want to compare any alternative in B_1 to all alternatives in B_0. For any $B \in \mathcal{B}_w$ and $y \in B$ define $\mathbf{1}_B^y$ as the vector of length $|\overline{U}|$ with the ith entry equal to 1 if the corresponding ith utility in $\overline{U}$ is such that $u_i(y) > u_i(y')$ for all $y' \in B \setminus \{y\}$. Otherwise, the ith entry is equal to 0. That is, $\mathbf{1}_B^y$ indicates whether y is the best alternative in B according to different preference orders. Then, we can compute the upper bound and lower bounds of the fraction of DMs that are better off due to the presence of $y \in B_1$ with respect to the original situation B_0. Indeed, if $\pi \in \Delta(\bar{U})$ is a distribution over preference orders that can generate the data, then $\pi^\mathsf{T}\mathbf{1}_{B_0 \cup B_1}^y$ is the fraction of DMs that are better of from choosing y that was not available before. Formally, for a given $y \in B_1$, the upper bound is given by:

$$\begin{aligned} \overline{\rho}_y &= \max_{\pi \in \Delta(\overline{U})} \pi^\mathsf{T}\mathbf{1}_{B_0 \cup B_1}^y, \\ \text{s.t.} \quad & A\pi = F_x^{\mathrm{RUM}}. \end{aligned}$$

The lower bound is given by:

$$\begin{aligned} \underline{\rho}_y &= \min_{\pi \in \Delta(\overline{U})} \pi^\mathsf{T}\mathbf{1}_{B_0 \cup B_1}^y, \\ \text{s.t.} \quad & A\pi = F_x^{\mathrm{RUM}}. \end{aligned}$$

To compute the total fraction of individuals that are better off in B_1 with respect to B_0, we need to aggregate the upper and lower bounds with respect to all $y \in B_1$. Namely, the total fraction of individuals that are better off in B_1, ρ, can take values in

$$\rho \in \left[\sum_{y \in D_1} \underline{\rho}_y, \sum_{y \in D_1} \overline{\rho}_y \right].$$

For the special case when $D_0 \subset D_1$, then $\rho = \sum_{a \in D_1 \setminus D_0} F_x^{\mathrm{RUM}}(a|D_1)$, because this is just the fraction of DMs for which the new objects are at the top of the expanded menu. In this case the lower and the upper bounds coincide and ρ is uniquely determined even if we are not able to identify the underlying preference distribution π.

Note that the above bounds on ρ are sharp since Equation (3) fully characterizes the identified set for the distribution of preferences. In fact, ρ is uniquely identified with full choice set variation.

Theorem 2. *Under conditions of Theorem 1, if*

$$\mathcal{B}_w = 2^Y \setminus \{\emptyset\},$$

then

$$\rho = \sum_{y \in D_1} \underline{\rho}_y = \sum_{y \in D_1} \overline{\rho}_y.$$

We finish this subsection by remarking that our ordinal approach puts no restriction on the random utility distribution while being computationally superior to approaches based on the mixed logit literature (McFadden and Train (2000)). In contrast to the mixed logit approach, we do not need to integrate over (unknown) distributions of parameters of high order polynomial approximations of the utility function. Instead, by taking a purely ordinal approach and exploiting the fact that for any finite choice set Y there are at most $|Y|!$ linear orders, our welfare analysis requires solving a linear program.[10]

4.2. Average Welfare in Multinomial Choice with Additive Random Utility

In this section we show how choice set variation can help to make welfare analysis in multinomial choice settings. Following the Example 1, suppose that DMs preferences over alternatives in Y are captured by

$$\mathbf{u}_t(y) = v(y, \mathbf{x}) + \mathbf{e}_{y,t},$$

where $\{\mathbf{e}_{y,t}\}_{y \in Y, t \in \mathcal{T}}$ are mean zero taste shocks and $v(y, x)$ is the mean (average) utility of alternative y. Suppose that for any $y, y' \in Y$

$$v(y, \mathbf{x}) \neq v(y', \mathbf{x}) \text{ a.s..}$$

Manski (1975) shows that in the case where all DMs face the same menu and it is the grand choice set (i.e., $\mathcal{D}_x = \{Y\}$) if, conditional on $\mathbf{x} = x$, the taste shock $\mathbf{e}_{y,t}$ has support equal to $\mathbb{R}$, and has an absolutely continuous, independent, and identical distribution for all alternatives, then the ranking over observed frequencies of choices uniquely identifies the ranking over the set of the mean utilities. That is,

$$\mathbb{P}\left(\mathbf{y}_t = y | \mathbf{x} = x\right) > \mathbb{P}\left(\mathbf{y}_t = y' | \mathbf{x} = x\right) \iff v(y, x) > v(y', x)$$

for all $y, y' \in Y$ and $x \in X$. The following example demonstrates that this result does not hold anymore if the actual choice set is unobserved.

[10] $n!$ denotes the factorial of n.

Example 5. Suppose $Y = \{1, 2\}$ and $v(2) > v(1)$. Suppose that

$$\mathbb{P}(\mathbf{D} = D) = \begin{cases} 0.9, & \text{if } D = \{1\}, \\ 0.1, & \text{if } D = \{1, 2\}. \end{cases}$$

Then the analyst will observe option 1 with probability greater than 0.9 (1 is always chosen when $D = \{1\}$ and sometimes is chosen when $D = \{1, 2\}$) and would incorrectly conclude that $v(1) > v(2)$.

Our framework can deliver the identification of the mean utility ranking even if the choice set is unobserved.

Proposition 1. *Assume that conditions of Theorem 1 are satisfied. For every x, if conditional on $\mathbf{x} = x$ (i) $\mathbf{e}_{y,t}$ has support equal to $\mathbb{R}$ and an absolutely continuous, independent, and identical distribution for all $y \in Y$; (ii) for all $y, y' \in Y$ there exists $D \in \mathcal{B}_w$ such that $y, y' \in D$, then the ranking over $\{v(y, x)\}_{y \in Y}$ is uniquely identified for all x.*

Goeree et al. (2005) provide a sufficient condition that is weaker than independence of taste shocks: interchangeability.[11] In contrast to the independence condition, the interchangeability condition allows for correlations between taste shocks. However, it still may be restrictive for large choice sets since interchangeability implies that the correlation between two taste shocks is bounded from below:

$$\operatorname{Corr}(\mathbf{e}_{y,t}, \mathbf{e}_{y',t}) \geq -\frac{1}{|Y| - 1}.$$

Rich choice set variation can solve this problem as the following proposition demonstrates.

[11] A collection $\{\mathbf{e}_{y,t}\}_{y \in Y}$ is interchangeable (exchangeable) if for any permutation of the indices $\{\sigma(y)\}_{y \in Y}$, the joint probability distribution of $\{\mathbf{e}_{\sigma(y),t}\}_{y \in Y}$ is the same as the joint probability distribution of the original sequence $\{\mathbf{e}_{y,t}\}_{y \in Y}$.

Proposition 2. *Assume that conditions of Theorem 1 are satisfied. For every x and $y, y' \in Y$, if conditional on $\mathbf{x} = x$ (i) $\{\mathbf{e}_{y,t}, \mathbf{e}_{y',t}\}$ have support equal to $\mathbb{R}^2$ and are interchangeable; (ii) $\{y, y'\} \in \mathcal{B}_w$, then the ranking over $\{v(y, x)\}_{y \in Y}$ is uniquely identified for all x.*

Note that the pairwise interchangeability (condition (i) in Proposition 2) does not restrict correlations between taste shocks since

$$\mathrm{Corr}\left(\mathbf{e}_{y,t}, \mathbf{e}_{y',t}\right) \geq -\frac{1}{2-1} = -1.$$

Although independence implies interchangeability Proposition 2 is not a generalization of Proposition 1. Proposition 2 imposes minimal restrictions on the distribution of unobservables, but requires some DM to chose only from binary choice sets and full variation in these binary choice sets. Proposition 1 does not require existence of binary choice sets, but needs full joint independence (or interchangeability).

Since Proposition 2 uses binary choice sets, instead of interchangeability one may assume conditional median independence of $\mathbf{e}_{y,t} - \mathbf{e}_{y',t}$ for all $y, y' \in Y$. Indeed, in binary choice problems

$$\begin{aligned}\mathbb{P}\left(\mathbf{y}_t = y | \mathbf{D} = \{y, y'\}, \mathbf{x} = x\right) &= \mathbb{P}\left(v(y', \mathbf{x}) - v(y, \mathbf{x}) \leq \mathbf{e}_{y,t} - \mathbf{e}_{y',t} | \mathbf{x} = x\right) = \\ &= 1 - F_{\mathbf{e}_{y,t} - \mathbf{e}_{y',t} | \mathbf{x}}(v(y', x) - v(y, x) | x),\end{aligned}$$

where $F_{\mathbf{e}_{y,t} - \mathbf{e}_{y',t} | \mathbf{x}}(\cdot | x)$ is a conditional c.d.f. of $\mathbf{e}_{y,t} - \mathbf{e}_{y',t}$ conditional on $\mathbf{x} = x$. Hence, if

$$\mathrm{Median}(\mathbf{e}_{y,t} - \mathbf{e}_{y',t} | \mathbf{x} = x) = 0$$

for all $y, y' \in Y$ and x, then the ranking over $\{v(y, x)\}_{y \in Y}$ is uniquely identified for all x.

We conclude this section by noting that the binary menus used in Proposition 2 can also be used for identification of the distribution of random coefficients in multinomial choice models under minimal restrictions (e.g. Ichimura and Thompson (1998) and Gautier and Kitamura (2013)).

5. General Choice Set Structures

Our main results have imposed no restriction on random utility behavior. Instead, we exploited only nestedness of choice sets to ensure identification of the hidden choice sets and preference distributions. In this section, we generalize our main theorem to allow for nonnested choice sets while imposing a weak condition on the data generating process. Namely, we impose a linear independence restriction on random utility.

Assumption 7 (Linear independence)**.** *For every $x \in X$, $\{F^{\mathrm{RUM}}(\cdot|D,x)\}_{D\in\mathcal{D}_x}$ is a collection of linearly independent functions. That is, for any x and any collection of reals $\{\alpha_D\}_{D\in\mathcal{D}_x}$*

$$\forall y \in Y, \sum_{D\in\mathcal{D}_x} \alpha_D F^{\mathrm{RUM}}(y|D,x) = 0 \iff \forall D \in \mathcal{D}_x : \alpha_D = 0.$$

Assumption 7 means that none of the possible choice sets is redundant in generating the data.[12] Recall the finite mixture representation of our model:

$$\Pr(\mathbf{y}_t = y|\mathbf{x} = x) = \sum_{D\in\mathcal{D}_x} m(D|x)F^{\mathrm{RUM}}(y|D,x)$$

for all x and y. So if the distribution over some big set can be represented as the mixture of the distribution over its subsets, then these subsets are redundant and do not contain information on top of what is already contained in the distribution over the big set (see Example 7 for a numerical example). The nestedness condition we

[12]In the context of auctions a similar assumption is made in An (2017) and Luo (2018).

used before (Assumption 4) is sufficient for Assumption 7 to hold.[13]

Similar to the nestedness condition, one of the implications of Assumption 7 is a restriction on the cardinality of $\mathcal{D}_x$. Indeed, it must be that for all $x \in X$, the number of points in the support of the random choice set, $d_{D,x} = |\mathcal{D}_x|$, has to be smaller than or equal to the total number of alternatives, i.e., $d_{D,x} \leq |Y|$. We believe this is a desirable property as having more choice sets than outcomes after conditioning on observables may not be realistic nor tractable in empirical applications.

Theorem 3. *Suppose Assumptions 1-3, 5, and 7 hold and* $T \geq 3$*, then* $m(\cdot|x)$ *and* $F^{\mathrm{RUM}}(\cdot|D, x)$ *are constructively identified for all* $x \in X$ *and* $D \in \mathcal{D}_x$*.*

Theorem 3 covers many new cases of interest that greatly extend the scope of our results. Namely, we can study other structural restrictions on the support of the random choice set $\mathcal{D}_x$. Consider the following example of the environment where Assumption 7 is satisfied.

Example 6 (Consideration Sets: Categorization by Partitions)**.** DMs have a two-stages choice procedure, where they first categorize alternatives using a partition of Y. Then they maximize utility given the category they consider in the first stage. More formally, the support of the unobserved random consideration set, $\mathcal{D}_x$, is such that $\bigcup_{D \in \mathcal{D}_x} D = Y$ and $D \cap D' = \emptyset$ for any $D, D' \in \mathcal{D}_x$, $D \neq D'$. Under the full support assumption (Assumption 5), this example will satisfy Assumption 7.[14] An example of categorization is the car choice problem with choice sets being represented by brands (e.g. Nissan, Toyota, etc.) or types of the car (e.g., SUV, sedan, etc.).

In the previous example we only impose additional structure on the choice set formation process. No extra restrictions are imposed on F^{RUM}. Example 6 shows that Assumption 7 can be satisfied in a variety of empirical applications. However, as the following example demonstrates, it may also fail to hold.

[13]Because it makes the matrix $[F^{\mathrm{RUM}}(y|D, x)]_{y \in Y, D \in \mathcal{D}_x}$ triangular.

[14]Categorization as a consideration set heuristic has been studied in Aguiar (2017) and Zhang (2016).

Example 7. Suppose that $Y = \{1, 2, 3, 4\}$, $\mathcal{D}_x = \{\{1, 2\}, \{3, 4\}, \{1, 2, 3, 4\}\}$. Suppose that $F^{\mathrm{RUM}}(\cdot|D)$ is uniform for every $D \in \mathcal{D}_x$. Then Assumption 7 does not hold since

$$F^{\mathrm{RUM}}(y|\{1, 2, 3, 4\}) = \frac{1}{2}F^{\mathrm{RUM}}(y|\{1, 2\}) + \frac{1}{2}F^{\mathrm{RUM}}(y|\{3, 4\})$$

for all y. However, Assumption 7 is satisfied, if $\mathcal{D}_x = \{\{1, 2\}, \{3, 4\}\}$, $\mathcal{D}_x = \{\{1, 2\}, \{1, 2, 3, 4\}\}$, or $\mathcal{D}_x = \{\{3, 4\}, \{1, 2, 3, 4\}\}$.

Example 7 shows that, essentially, Assumption 7 requires the support of $\mathbf{D}$ to have the minimal collection of sets that can generate the data. In Example 7 the DM whose choice set is Y is observationally equivalent to the DM whose choice set is equal to $\{1, 2\}$ or to $\{3, 4\}$ with equal probabilities.

6. Conclusion

We have studied the problem of preference identification and welfare analysis when choice sets are not observable. We showed that observing three or more choices from the same latent choice set is sufficient to identify and consistently estimate jointly the distribution of choice sets and preferences, when choice sets are nested. With exclusion restrictions on covariates we can recover rich variation of the unobserved choice sets making possible to do sharp nonparametric welfare analysis. When full choice set variation is attained exact nonparametric welfare analysis is possible. Alternatively, when the analyst is willing to impose additional restrictions on random utility, namely allowing for additive random utility, then average welfare analysis can be done in a fully nonparametric fashion.

We extend our main result to more general choice sets structures. We provide a simple linear independence condition on the conditional distribution of choices that suffices for nonparametric identification of the joint distribution of choice sets and preferences.

References

Abaluck, J. and Adams, A. (2017). What do consumers consider before they choose? identification from asymmetric demand responses. Technical report, National Bureau of Economic Research.

Aguiar, V. H. (2017). Random categorization and bounded rationality. *Economics Letters*, 159:46–52.

Allen, R. and Rehbeck, J. (2019). Revealed stochastic choice with attributes. *Available at SSRN 2818041*.

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

Barseghyan, L., Molinari, F., and Thirkettle, M. (2019). Discrete choice under risk with limited consideration. *arXiv preprint arXiv:1902.06629*.

Block, H. D. and Marschak, J. (1960). Random orderings and stochastic theories of responses. *Contributions to probability and statistics*, 2:97–132.

Brady, R. L. and Rehbeck, J. (2016). Menu-Dependent Stochastic Feasibility. *Econometrica*, 84(3):1203–1223.

Crawford, G. S., Griffith, R., and Iaria, A. (2019). Demand estimation with unobserved choice set heterogeneity using sufficient sets. Technical report, CEPR Discussion Papers.

Dardanoni, V., Manzini, P., Mariotti, M., and Tyson, C. (2019). Inferring cognitive heterogeneity from aggregate choices. Technical report, Department of Economics, University of Sussex.

Falmagne, J. C. (1978). A representation theorem for finite random scale systems. *Journal of Mathematical Psychology*, 18(1):52–72.

Fox, J. T. (2007). Semiparametric estimation of multinomial discrete-choice models using a subset of choices. *The RAND Journal of Economics*, 38(4):1002–1019.

Fudenberg, D., Iijima, R., and Strzalecki, T. (2015). Stochastic choice and revealed perturbed utility. *Econometrica*, 83(6):2371–2409.

Gautier, E. and Kitamura, Y. (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica*, 81(2):581–607.

Goeree, J. K., Holt, C. A., and Palfrey, T. R. (2005). Regular quantal response equilibrium. *Experimental Economics*, 8(4):347–367.

Goeree, M. S. (2008). Limited information and advertising in the US personal computer industry. *Econometrica*, 76(5):1017–1074.

Gul, F. and Pesendorfer, W. (2006). Random expected utility. *Econometrica*, 74(1):121–146.

Hickman, W. and Mortimer, J. H. (2016). Demand estimation with availability variation. *Handbook on the Economics of Retailing and Distribution*, page 306.

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

Hu, Y., McAdams, D., and Shum, M. (2013). Identification of first-price auctions with non-separable unobserved heterogeneity. *Journal of Econometrics*, 174(2):186–193.

Ichimura, H. and Thompson, T. S. (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics*, 86(2):269–295.

Kitamura, Y. and Stoye, J. (2018). Nonparametric analysis of random utility models. *Econometrica*, 86(6):1883–1909.

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

Luo, Y. (2018). Unobserved heterogeneity in auctions under restricted stochastic dominance. Technical report.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of econometrics*, 3(3):205–228.

Manzini, P. and Mariotti, M. (2014). Stochastic Choice and Consideration Sets. *Econometrica*, 82(3):1153–1176.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142.

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

McFadden, D. and Train, K. (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics*, 15(5):447–470.

Polisson, M. and Quah, J. K.-H. (2013). Revealed preference in a discrete consumption space. *American Economic Journal: Microeconomics*, 5(1):28–34.

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

Zhang, J. (2016). Stochastic Choice with Subjective Categorization. *Available at SSRN*.

A. Proofs

A.1. Proof of Theorem 1

Fix some $x \in X$. To simplify the notation we drop the conditioning on $\mathbf{x} = x$ from the notation below. For instance, when we write $\mathbb{P}(\mathbf{y}_1 = 1)$ we mean $\mathbb{P}(\mathbf{y}_1 = 1|\mathbf{x} = x)$. Define the following matrices

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

where $\operatorname{diag}(z)$ is a diagonal matrix with vector z on the diagonal.

Step 1. In this step we will show how to identify the number of choice sets that are considered with positive probability. By the law of total probability, under the independence assumption,

$$\begin{aligned}
\mathbb{P}\left(\mathbf{y}_1 = i, \mathbf{y}_2 = j\right) &= \sum_k \mathbb{P}\left(\mathbf{y}_1 = i, \mathbf{y}_2 = j|\mathbf{D} = D_k\right)\mathbb{P}\left(\mathbf{D} = D_k\right)\\
&= \sum_k \mathbb{P}\left(\mathbf{y}_1 = i|\mathbf{D} = D_k\right)\mathbb{P}\left(\mathbf{y}_2 = j|\mathbf{D} = D_k\right)\mathbb{P}\left(\mathbf{D} = D_k\right).
\end{aligned}$$

Or in matrix notation

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

Under the nestedness assumption the maximal number of the points in the support of $\mathbf{D}$ is equal to the number of the possible outcomes. That is, $d_D \leq |Y|$.

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

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

That is, the rank of $L_{1,2}$ is equal to $d_D = |\mathcal{D}_x|$.

To show that $L_{1|D}$ and $L_{2|D}$ have full column rank note that since the choice sets are nested we can always find a set of alternatives $\{y_k^*\}_{k=1}^{d_D}$ such that

$$\begin{aligned}
&y_1^* \in D_k, \quad k = 1, 2, \ldots, d_D,\\
&y_2^* \in D_k, \quad k = 1, 2, \ldots, d_D - 1,\\
&\ldots,\\
&y_{d_D}^* \in D_k, \quad k = 1.
\end{aligned}$$

Stacking together the rows of the matrix $L_{1|D}$ that correspond to $\{y_k^*\}_{k=1}^{d_D}$, we obtain a matrix with zero entries below the anti-diagonal (the diagonal going from the lower left corner to the upper right corner).[15] Since we assume that every element of every choice set is picked with positive probability, this square matrix has full rank. Repeating the same construction with $L_{2|D}$ we can conclude that $L_{1|D}$ and $L_{2|D}$ have full column rank that is equal to d_D.

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, take a collection of alternatives in Y, $\{\tilde{y}_k\}_{k=1}^{d_D}$, such that the following observable modification

[15]The probability of picking an element that does not belong to the choice set is zero.

of $L_{1,2}$ is nonsingular (have full rank):

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

Such collection $\{\tilde{y}_k\}_{k=1}^{d_D}$ 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

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

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). For estimation purposes it is sufficient to find any $\{\tilde{y}_k\}_{k=1}^{d_D}$ that make $\tilde{L}_{1,2}$ nonsingular. Similar to $L_{1,2}$, $\tilde{L}_{1,2}$ can be easily estimated from the observed data.

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

$$\begin{aligned}
\tilde{L}_{1,D} &= \left[\mathbb{P}\left(\mathbf{y}_1 = \tilde{y}_i, \mathbf{D} = D_k\right)\right]_{i,k\in\{1,\ldots,d_D\}}, \\
\tilde{L}_{2,1,y} &= \left[\mathbb{P}\left(\mathbf{y}_2 = \tilde{y}_i, \mathbf{y}_1 = \tilde{y}_j, \mathbf{y}_3 = y\right)\right]_{i,j\in\{1,\ldots,d_D\}}, \\
A_{y|D} &= \operatorname{diag}\left(\left(\mathbb{P}\left(\mathbf{y}_3 = y | \mathbf{D} = D_k\right)\right)_{k\in\{1,\ldots,d_D\}}\right) = \operatorname{diag}\left(\left(F^{\mathrm{RUM}}(y|D_k)\right)_{k\in\{1,\ldots,d_D\}}\right).
\end{aligned}$$

By the law of total probability, under the independence assumption,

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

Hence, in matrix notation we get

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

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

$$\tilde{L}_{1,D}^{\intercal} = \tilde{L}_{2|D}^{-1}\tilde{L}_{1,2}^{\intercal}. \tag{4}$$

Similarly to the previous calculations, under the independence assumption, we have

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

Combining the latter with equation (4) we get the following eigenvector-eigenvalue decomposition of $R_y = \tilde{L}_{2,1,y}\left(\tilde{L}_{1,2}^{\intercal}\right)^{-1}$

$$R_y = \tilde{L}_{2|D}A_{y|D}\tilde{L}_{2|D}^{-1}. \tag{5}$$

Step 4. Note that in the decomposition (5) 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). For estimation purposes any normalization of the eigenvectors suffices (e.g., one can take eigenvectors that have the norm equal to 1). In contrast to the existing results (e.g., Hu et al. (2013)) we will 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, $\{\eta_k\}_{k=1}^{d_D}$. We do not know which η corresponds to which choice set. Hence, we need to construct a map $h : \{\eta_k\}_{k=1}^{d_D} \to 2^{y} \setminus \{\emptyset\}$ that assigns an eigenvector to a choice set. Note that

$$R_y\eta_k = \lambda_{y,k}\eta_k = F^{\mathrm{RUM}}(y|D_k)\eta_k$$

Hence,

$$R_y \eta_k = 0 \iff y \notin h(\eta_k).$$

As a result, by checking every y and η_k we can assign choice sets to every eigenvector, thus we can identify $\mathcal{D}_x$. Moreover, since $\lambda_{y,k} = F^{\text{RUM}}(y|D_k)$ we also identify F^{RUM} and $L_{2|D}$ since

$$L_{2|D} = \left[F^{\text{RUM}}(i|D_k)\right]_{i\in Y, k=1,\dots,d_D}.$$

Step 5. Finally, let $m = (m(D))_{D\in\mathcal{D}_x}$, then

$$m = L_{1,D}^{\intercal}\iota,$$

where ι is the vector of ones. Hence,

$$L_{1,2}^{\intercal}\iota = L_{2|D}L_{1,D}^{\intercal}\iota = L_{2|D}m.$$

Since $L_{1,2}$ is observed (can be consistently estimated), and $L_{2|D}$ is constructively identified (also can be estimated) and has full column rank, we also identify and can consistently estimate the distribution over choice sets.

A.2. Proof of Theorem 2

Under conditions of Theorem 1, if $\mathcal{B}_w = 2^Y \setminus \{\emptyset\}$, it has to be that

$$F^{\text{RUM}}(y|D,x) = \sum_{u\in\overline{U}} \pi(u)\mathbb{1}\left(u(y) > u(y'),\ \forall y' \in Y \setminus \{y\}\right),$$

where π satisfies $A\pi = F^{\text{RUM}}$. This implies that the probability of being the top element according to the distribution over utilities supported on $\overline{U}$, is equal to the identified F^{RUM}. This makes this quantity unique when $\mathcal{B}_w = 2^Y \setminus \{\emptyset\}$.

A.3. Proof of Proposition 1

Take any $y, y' \in Y$. Since there exists $D \in \mathcal{B}_w$ such that $y, y' \in D$, then by Lemma 1 in Fox (2007)

$$\mathbb{P}(\mathbf{y}_t = y | \mathbf{D} = D, \mathbf{x} = x) > \mathbb{P}(\mathbf{y}_t = y' | \mathbf{D} = D, \mathbf{x} = x) \iff v(y, x) > v(y', x).$$

The ranking between y and y' then follows from the fact that $F^{\mathrm{RUM}}(\cdot|D, x) = \mathbb{P}(\mathbf{y}_t = \cdot|\mathbf{D} = D, \mathbf{x} = x)$ is identified. Since we can recover all binary comparison, we can recover the whole ranking.

A.4. Proof of Theorem 3

Fix some $x \in X$. Similar to the proof of Theorem 1 to simplify the notation we drop the conditioning on $\mathbf{x} = x$ from the notation below.

Note that Theorem 3 is a generalization of Theorem 1 since instead of nestedness of choice sets we assume that $\{F^{\mathrm{RUM}}(\cdot|D)\}_{D \in \mathcal{D}_x}$ are linearly independent. In the proof of Theorem 1 the nestedness assumption was used only once in Step 1. It was used to prove that matrices $L_{1|D}$ and $L_{2|D}$ have full column rank. Note that

$$L_{1|D} = [\mathbb{P}(\mathbf{y}_1 = i | \mathbf{D} = D_k)]_{i \in Y, k=1,\dots,d_D} = \left[F^{\mathrm{RUM}}(i|D_k)\right]_{i \in Y, k \in \{1,\dots,d_D\}}.$$

By Assumption 7 the columns of $L_{1|D}$ are linearly independent, and, hence, $L_{1|D}$ has full column rank (similar logic applies to $L_{2|D}$). The rest of the proof follows from the proof of Theorem 1.