Submitted to
manuscript

A Nonparametric Approach with Marginals for Modeling Consumer Choice

Yanqiu Ruan
Singapore University of Technology and Design, Singapore, yanqiu_ruan@mymail.sutd.edu.sg

Xiaobo Li
National University of Singapore, Singapore, iselix@nus.edu.sg

Karthyek Murthy
Singapore University of Technology and Design, Singapore, karthyek_murthy@sutd.edu.sg

Karthik Natarajan
Singapore University of Technology and Design, Singapore, karthik_natarajan@sutd.edu.sg

Given data on the choices made by consumers for different offer sets, a key challenge is to develop parsimonious models that describe and predict consumer choice behavior while being amenable to prescriptive tasks such as pricing and assortment optimization. The marginal distribution model (MDM) is one such model, which requires only the specification of marginal distributions of the random utilities. This paper aims to establish necessary and sufficient conditions for given choice data to be consistent with the MDM hypothesis, inspired by the utility of similar characterizations for the random utility model (RUM). This endeavor leads to an exact characterization of the set of choice probabilities that the MDM can represent. Verifying the consistency of choice data with this characterization is equivalent to solving a polynomial-sized linear program. Since the analogous verification task for RUM is computationally intractable and neither of these models subsumes the other, MDM is helpful in striking a balance between tractability and representational power. The characterization is convenient to be used with robust optimization for making data-driven sales and revenue predictions for new unseen assortments. When the choice data lacks consistency with the MDM hypothesis, finding the best-fitting MDM choice probabilities reduces to solving a mixed integer convex program. The results extend naturally to the case where the alternatives can be grouped based on the similarity of the marginal distributions of the utilities. Numerical experiments show that MDM provides better representational power and prediction accuracy than multinominal logit and significantly better computational performance than RUM.

Key words: discrete choice, nonparametric modeling, optimization, additive perturbed utility model
History:

1. Introduction

Discrete choice models have been used extensively in economics (Allenby and Ginter 1995), marketing (McFadden 1986), healthcare (de Bekker-Grob et al. 2018), transportation (Ben-Akiva and Lerman 1985), and operations management (Talluri and Van Ryzin 2004). Such models describe the observable

distribution of demand from the behavior of one or more consumers who choose their most preferred alternative from a discrete collection of alternatives.

As choice models specify the conditional probability distribution over any offer set, they are inherently high dimensional. Given data on the choices made by consumers over a limited collection of offer sets (also referred as assortments), the specification of a choice model hypothesis is essential in linking data from the observed offer sets to predictions for new offer sets for which no data is available. The classical multinomial logit choice model (MNL) derived by Luce (1959) and Plackett (1975) is among the simplest and most widely used choice model. It stipulates that the ratio of choice probabilities for any two alternatives i and j does not depend on any alternatives other than i and j. A popular model at the expressive end of the spectrum is the random utility model (RUM) which hypothesizes that the utilities of the alternatives are random variables and the consumers are utility maximizers. In settings with finite alternatives, MNL is subsumed by RUM and a generic RUM is describable by a distribution over the rankings (or preference lists) of the alternatives (Mas-Colell et al. 1995). Such a description of RUM over n alternatives requires about $n!$ parameters, and even the task of verifying whether given choice data is consistent with the RUM hypothesis is understood to be computationally intractable (see Jagabathula and Rusmevichientong 2019).

There has been a recent surge of interest in developing choice models with good representational power using machine learning techniques. Examples of such models include those proposed by Wang et al. (2020), Sifringer et al. (2020) and Aouad and Désir (2022), who utilize neural networks to fit expressive utilities within the context of MNL and RUM hypotheses. Additionally, the decision forest choice model (Chen and Mišić 2022, Chen et al. 2019) has been shown to be capable of approximating any choice data with increasing forest depth. The expressiveness of these models however comes at the cost of requiring significant amounts of data and computation to learn. Furthermore, it has been observed that the reliability of economic information obtained from deep neural network based models is compromised when the data size is small (see Wang et al. 2020). Therefore, a natural question is to examine the representational power of other choice models and to identify choice model hypotheses that offer a balance between richer representational power and tractability while allowing for robust procedures for estimation and prediction from limited data.

1.1. The choice model and the research questions

An alternative to RUM in offering a substantial generalization to MNL is the marginal distribution model (MDM) proposed by Natarajan et al. (2009). It subsumes MNL (see Mishra et al. 2014) and the well-known additive perturbed utility (APU) model treated in Fudenberg et al. (2015). Specifying an MDM choice model requires only the specification of the marginal distributions of the random utilities of the alternatives. Then the MDM choice probabilities are computed with the extremal

distribution maximizing the expected consumer utility over all joint distributions with the given collection of marginals. A precise description of MDM is provided in Section 2. A key advantage of this model is that it allows choice probabilities to be readily computed from tractable convex optimization formulations. Besides tractability, MDM has been shown to exhibit good empirical performance in various applications using real-world datasets (see Natarajan et al. 2009, Mishra et al. 2014, Ahipasaoglu et al. 2019, 2020, Yan et al. 2022, Liu et al. 2022). More recently, price optimization has been shown to be computationally tractable with MDM (see Yan et al. 2022) and a half approximation guarantee has been developed for profit-nested heuristic in assortment optimization (see Ahipasaoglu et al. 2020). The formulation of MDM has also become useful in deriving prophet inequalities for Bayesian online selection problems (see Feldman et al. 2021) and solving smoothed optimal transport formulations (see Taşkesen et al. 2022).

Although a general specification of MDM does not impose restrictions on the marginal distributions of the random utilities, the estimation of MDM from data in practice typically requires first committing to an appropriate parametric family for the marginal distributions of the utilities (an exception in Yan et al. 2022 which uses piece-wise linear marginal cumulative distribution functions). Upon fixing suitable parametric families for the marginal distributions, the respective parameters are estimated from data using a procedure like maximum-likelihood and the corresponding choice probability predictions are made using convex optimization (see, e.g., Mishra et al. 2014). Fixing the "right" parametric families can however be a tricky exercise and is prone to suffering from underfitting and overfitting issues. Just as how parametric restrictions to RUM are deemed to be restrictive (see, e.g., Farias et al. 2013), a workflow requiring prior commitment to fixed parametric families of distributions does not allow one to leverage the full modeling power offered by MDM to extract as much structural information as possible from the data.

Though one may wish to perform data-driven estimation and prediction under the MDM hypothesis, it is currently unknown how to do so given choice data across different offer sets without imposing additional parametric restrictions on the marginal distributions of utilities. Similar to RUM, does working with the entirety of MDM lead to intractable formulations? To begin with, is verifying the consistency of given choice data with MDM computationally tractable?

In order to address these questions and gain an understanding of the representational power of MDM, this paper seeks to answer the following fundamental question: *What is the structure of the observable sales data that is necessary and sufficient for consistency with the MDM hypothesis?* More formally, suppose that $N = \{1, \ldots, n\}$ is the universe of products, and we have choice data for a collection $\mathcal{S}$ of subsets of N. For each subset $S \in \mathcal{S}$, let $p_{i,S} \in [0,1]$ denote the fraction of customers who purchased product i when the assortment S was offered. Our goal is to identify necessary and sufficient conditions for the choice probability data $\boldsymbol{p}_{\mathcal{S}} = (p_{i,S} : i \in S, S \in \mathcal{S})$ to be representable by

an MDM instance. If our investigation leads to a tractable characterization, then we aim to utilize it to develop data-driven procedures that can leverage the full modeling power of the MDM hypothesis to make revenue and sales predictions, without restricting one to make parametric distributional assumptions.

1.2. Contributions

An effort towards addressing these goals leads us to the following contributions in this paper.

An exact characterization of the choice probabilities represented by MDM and its tractability. We show that the choice data $\boldsymbol{p}_{\mathcal{S}} = (p_{i,S} : i \in S, S \in \mathcal{S})$ given for a collection of assortments $\mathcal{S}$ is representable by MDM if and only if there exists a utility function $U : \mathcal{S} \to \mathbb{R}$ representing the preferences expressed across assortments in the choice data; in particular, the utility U should exhibit a strict preference for an assortment T over another assortment S containing a common product i if $p_{i,S} < p_{i,T}$, and exhibit indifference between S and T if $p_{i,S} = p_{i,T} > 0$ (see Theorem 1 for a precise statement). The existence of a utility function implies a rational preference relation (or a ranking) over assortments, and it allows us to make the following deductions regarding the tractability and representational power of MDM:

- The characterization in Theorem 1, in terms of the existence of a ranking over assortments, lends itself to be verified with a linear program whose size is polynomial in the number of products and assortments. This is in contrast to RUM which requires the existence of a distribution over the $n!$ rankings possible for n products. Thus, unlike RUM, verifying the consistency of given choice data with the MDM hypothesis can be accomplished in polynomial time.
- The collection of MDM representable choice probabilities possesses a non-zero measure when considered relative to the collection of all possible choice probabilities. Contrast this with parametric family alternatives with a fixed number of parameters such as multinomial logit or nested logit which have zero measure. Additionally, the characterization in Theorem 1 reveals that MDM and RUM do not subsume each other in terms of the choice probabilities they can represent.

A nonparametric data-driven approach to prediction and estimation. As the sales data available in practice is often inadequate to entirely specify the probabilistic behavior of the utilities, we utilize the nonparametric MDM characterization in Theorem 1 together with robust optimization as the basis for making sales and revenue predictions for any new assortment with no prior sales data. Specifically, we develop a data-driven approach which builds upon the exact MDM characterization to produce worst-case estimates of sales and revenues computed over all MDM instances that are consistent with the given choice data. This robust approach mitigates the risk of misspecification and renders end-to-end learning feasible for MDM. The characterization can also be used to develop optimistic best-case estimates, which, along with the worst-case estimates, yield prediction intervals

for sales and revenues over all MDM instances consistent with the given choice data. The procedure yields narrower intervals for sales and revenue predictions when data for more assortments become available, as is desirable for any data-driven method.

When the choice data is not fully consistent with the MDM hypothesis, we develop a "limit of MDM" formulation to quantify the degree of inconsistency. Inspired by the "limits of rationality" measure proposed in Jagabathula and Rusmevichientong (2019), we define the limit of MDM as the smallest loss that can be obtained by fitting MDM to the given choice data. A model which attains the minimum loss can be interpreted as offering the best fit, within the MDM family, to any given choice data. This best fitting MDM can be subsequently used to produce robust revenue and sales predictions as described above. Utilizing the exact MDM characterization in Theorem 1, we reduce the computation of the limit of MDM to the rank aggregation problem (Dwork et al. 2001) and show that it is NP-hard. We develop a mixed integer convex program that is applicable generally for computing the limit (see Proposition 5). We also propose algorithms whose running time are polynomial in both the number of alternatives and the size of the assortment collection if the assortment collection possesses suitable structure (see Corollary 2).

An extension to include product groupings. As an extension of the above study, we consider the case where the products can be grouped suitably based on the similarity of the marginal distribution of their utilities. Depending on the grouping of products considered, the resulting grouped MDM (G-MDM) spans the spectrum of models interpolating between the general MDM considered above and the APU model. While a general MDM does not require any product grouping, the APU model corresponds to stipulating that all products are grouped together to have the same marginal distributions for their utilities. Thus, APU is subsumed by MDM. By flexibly allowing products to be grouped based on the similarities of their utility distributions, G-MDM imparts domain knowledge to improve model estimation. We show that the G-MDM also offers a tractable characterization of the choice probabilities they represent. In turn, this characterization allows the development of data-driven prediction and estimation methods analogous to those described in Section 1.2. We supplement these with a procedure based on K-means clustering to identify the grouping information and validate its effectiveness with synthetic data experiments in Section EC.8.

Numerical insights. We present numerical results based on both synthetic and real data, which reveal MDM's utility as a computationally tractable alternative with good representational power when the assumptions underlying parametric models such as MNL are violated. Experiments employing real-world data reveal that MDM outperforms MNL in predicting revenue-based assortment rankings by an average of 27.2% and up to 50% and improves the revenue of the predicted best assortment by 9.2% on average, with a potential improvement of 22.2% over all the instances

compared to MNL. Together with the experiments using synthetic data, the numerical results reveal that fixing the marginal distribution family beforehand often leads to inconsistent sales and revenue predictions that do not lie within the nonparametric prediction intervals. These results underscore the ability of our proposed data-driven methods to mitigate misspecification risks.

The rest of this paper is organized as follows. We begin with a brief review of related literature and provide a precise description of MDM in Section 2. We derive the exact characterization of MDM representable choice probabilities in Section 3 and discuss its implications for tractability and representational power. In Sections 4 and 5, we develop data-driven methods for prediction and estimation as applications of the characterization. We discuss the case where the products can be grouped based on the similarity of the marginal distribution of their utilities in Section 6. We conclude with a discussion after presenting the results of numerical experiments in Section 7. Proofs, illustrative examples, results of additional experiments, and useful additional information on the experiments are provided in the electronic companion (EC).

2. Related literature and a description of the MDM choice model

We begin with a concise overview of studies that aim to characterize and relate choice probabilities obtainable under prominent choice model hypotheses. Additionally, we note related results on their tractability and methods for estimation when equipped with sales data.

2.1. On the characterizations available for choice models

RUM is perhaps the most popular class of models in choice modeling. RUM assumes that the utility of each alternative i in the collection of products $N = \{1, \ldots, n\}$ takes the form $u_i = \nu_i + \epsilon_i$, where $\boldsymbol{\nu} = (\nu_1, ..., \nu_n)$ and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ denote the deterministic and stochastic parts of the utilities, respectively. Assuming a joint distribution θ on the random part $\boldsymbol{\epsilon}$, the probability of choosing product i in an assortment $S \subseteq N$ is given by $p_{i,S} = \mathbb{P}_{\sim\theta}(i = \arg\max_{j \in S}\{\nu_j + \epsilon_j\})$, where the probability of ties is assumed to be 0. Here note that we do not explicitly model the outside option; instead, we treat the outside option as one of the products in N. Modeling the joint probability distribution of the random utilities with specific parametric distribution families leads to parametric subclasses of RUM such as MNL (see, e.g., McFadden 1973), generalized extreme value model (see, e.g., McFadden 1978), nested logit model (see, e.g., McFadden 1980), multinomial probit model (see, e.g., Thurstone 1927, Daganzo 1979), mixed logit model (see, e.g., McFadden and Train 2000), and the exponomial choice model (see, e.g., Alptekinoğlu and Semple 2016). Several nonparametric choice models, such as the rank list model (see, e.g., Block and Marschak 1960, Farias et al. 2013) and the Markov chain choice model (Blanchet et al. 2016), have also proved to be useful in practice. The Markov chain choice model is a special case of the rank-list model (Berbeglia 2016).

Beginning with Marschak (1960) and Block and Marschak (1960), considerable effort has been devoted in econometrics towards understanding the restrictions imposed on choice data by the RUM hypothesis. The class of RUM and the class of rank list models are shown to be equivalent in Block and Marschak (1960). Falmagne (1978) has shown that a RUM can represent a system of choice probabilities over all possible assortments if and only if the Block-Marschak conditions are met, see also Barberá and Pattanaik (1986). McFadden and Richter (1990) has shown that under certain conditions, the axiom of revealed stochastic preference provides necessary and sufficient conditions for the choice probabilities under different assortments that can be recreated by a RUM. McFadden and Train (2000) has demonstrated that any RUM can be approximated closely by a mixed logit model. McFadden (2006) adds more conditions that relate to the findings in Falmagne (1978) and McFadden and Richter (1990). However, verifying these conditions is computationally intractable when there are a large number of products. Jagabathula and Rusmevichientong (2019) has been the first to relate the hardness of the stochastically rationalizable property stipulated by RUM to the notion of choice depth. They provide examples of structured assortment collections for which verification of consistency with RUM is computationally tractable.

Using a representative agent model (RAM) constitutes another popular optimization based approach to model choice. In RAM, a single agent makes a choice on behalf of the entire population. To make her choice, the agent takes into account the expected utility while preferring some degree of diversification. More precisely, given an assortment S, the representative agent solves

$$\max \left\{ \boldsymbol{\nu}^T \boldsymbol{x} \quad C(\boldsymbol{x}) \;|\; \boldsymbol{x} \in \quad_{n\ 1},\ x_i = 0\ \forall i \notin S \right\}, \tag{1}$$

where $\quad_{n\ 1} = \{\boldsymbol{x} \in \mathbb{R}^n_+ | \sum_{i \in N} x_i = 1\}$ is the unit simplex and $C(\boldsymbol{x}) : \quad_{n\ 1} \mapsto \mathbb{R}$ is a convex perturbation function that rewards diversification. The optimal x_i value provides the fraction of the population that chooses alternative i in assortment S. Hofbauer and Sandholm (2002) has shown that all RUM can be expressed using a representative agent model under appropriate conditions on the perturbation functions $C(\boldsymbol{x})$ (see also Feng et al. 2017).

The APU model in Fudenberg et al. (2015) can be obtained as a special case of RAM in (1) by taking the additive and separable perturbation $C(\boldsymbol{x}) = \sum_i c(x_i)$, where $c(x) : [0,1] \mapsto \mathbb{R}$ is a strictly convex function. Fudenberg et al. (2015) demonstrates acyclicity and ordinal IIA property, which is a relaxation of Luce's IIA condition, as two alternative conditions that characterize the richness of APU representable choice probabilities. Equipped with these results, Fudenberg et al. (2015) argues for APU as a considerably simpler and expressive model alternative which helps go beyond RUM while requiring only the specification of the univariate convex perturbation function $c(\,)$.

2.2. The marginal distribution model and related literature

The marginal distribution model (MDM) is a semiparametric choice model that yields choice probabilities from limited information on the joint distribution of the random utilities. As in RUM, the starting point of MDM is that the utility of each alternative i in the collection of products $N = \{1,\ldots,n\}$ takes the form $u_i = \nu_i + \epsilon_i$, where $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_n)$ and $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)$ denote the deterministic and stochastic parts of the utilities. MDM only requires the specification of the marginal distributions $F_1,\ldots,F_n$ of the random variables $\epsilon_1,\ldots,\epsilon_n$, and does not impose any independence assumption among $\epsilon_1,\ldots,\epsilon_n$. To describe the model, let denote the collection of joint distributions for $\boldsymbol{\epsilon}$ with the given marginal distributions $F_1,\ldots,F_n$. For any assortment S N, the MDM considers maximization of expected consumer utility over all distributions in :

$$\sup_{\theta 2} \mathsf{E}_{\sim\ \theta}\left[\max_{i2S}\{\nu_i + \epsilon_i\}\right]. \tag{2}$$

The probability of choosing a product i from the assortment S, given by $p_{i,S} = \mathsf{P}_{\sim\ \theta}(i = \arg\max_{j2S}\{\nu_j + \epsilon_j\})$, is evaluated with the distribution θ which attains the maximum in (2). A key advantage of this model is that the choice probabilities are readily computable via convex optimization, as described in Lemma 1 below.

Assumption 1. *Each random term ϵ_i, where i 2 N, is an absolutely continuous random variable with a strictly increasing marginal distribution $F_i(\ )$ on its support and $\mathsf{E}|\epsilon_i| < 1$.*

Lemma 1. *(Natarajan et al. 2009, Mishra et al. 2014, Chen et al. 2022) Under Assumption 1, the choice probabilities for a distribution which attains the maximum in (2) is unique and is given by the optimal solution of the following strictly concave maximization problem over the simplex:*

$$\max\left\{\sum_{i2S}\nu_i x_i + \sum_{i2S}\int_{1\ x_i}^{1} F_i^{\ 1}(t)\,dt \;\middle|\; \sum_{i2S} x_i = 1, x_i \quad 0\ \forall i\ 2\ S\right\}, \tag{3}$$

with the convention that $F_i^{\ 1}(0) = \lim_{t\#0} F_i^{\ 1}(t)$ and $F_i^{\ 1}(1) = \lim_{t"1} F_i^{\ 1}(t)$.

The reformulation of MDM in (3) shows that it is a special case of the representative agent model (1), in which the perturbation function is strictly convex and separable of the form $C(\boldsymbol{x}) = \sum_i \int_{1\ x_i}^{1} F_i^{\ 1}(t)\,dt$. If the marginal distributions are identical, MDM reduces to the additive perturbed utility (APU) model described in Section 2.1. Thus by subsuming the APU model, MDM provides a probabilistic utility interpretation for APU. Given any assortment S N, a vector $(x_i : i\ 2\ S)\ 2\ \mathsf{R}^{|S|}$ maximizes (3) and yields the MDM choice probabilities $p_{i,S} = x_i$ if and only if it satisfies the following optimality conditions for (3):

$$\nu_i + F_i^{\ 1}(1 \quad x_i) \quad \lambda + \lambda_i = 0, \quad \forall i\ 2\ S,$$

$$\begin{aligned} \lambda_i x_i &= 0, \quad \forall i \in S, \\ \sum_{i \in S} x_i &= 1, \\ x_i \geq 0, \lambda_i &\geq 0, \quad \forall i \in S, \end{aligned} \tag{4}$$

where λ and λ_i are the Lagrange multipliers associated with the constraints defining the simplex. An additional assumption that $F_i^{-1}(1) = +\infty$ is often made (see, e.g., Mishra et al. 2014) to guarantee strictly positive choice probabilities. However, in real datasets, one or more alternatives offered in an assortment might never be chosen by consumers. In this paper, we allow for this possibility by permitting the right end point of the support, $F_i^{-1}(1)$, to be possibly finite or infinite.

2.3. Related estimation approaches

Maximum likelihood estimation is the widely used method for parameter estimation when working with parametric subclasses of RUM and MDM. One may refer Manski and McFadden (1981), Manski and Lerman (1977), Cosslett (1981) and Mishra et al. (2014) for illustrations on how one may use maximum likelihood to estimate parametric models including the MNL model, the multinomial probit model, and the generalized extreme value model from choice data, after making specific distributional assumptions under the RUM and MDM modeling paradigms. Farias et al. (2013) offers a major departure by developing a nonparametric approach which makes revenue predictions under the RUM hypothesis without imposing restrictive parametric distributional assumptions. In particular, Farias et al. (2013) showcases the efficacy of using the worst-case expected revenue over the collection of RUM models that are consistent with the available sales data. Sturt (2021) presents an account of when these robust RUM revenue predictions can serve as a suitable basis for assortment optimization. Given an opportunity to collect data by performing pricing experiments, Yan et al. (2022) and Liu et al. (2022) use a nonparametric approach to identify the MDM objective in (3), upto a constant shift, by performing sufficient pricing experiments. Our approach for revenue predictions, which is described in Section 1.2, follows the same philosophy as Farias et al. (2013), but with the novelty of using the MDM characterization to make predictions that are consistent with the MDM hypothesis.

3. An exact characterization for MDM and its implications

In this section, we first develop necessary and sufficient conditions for the choice probabilities given by a collection of assortments that are representable by MDM. We follow this up with a discussion of its implications for tractability and representational power.

3.1. A tractable characterization for MDM

We begin by recalling that $N = \{1, \ldots, n\}$ denotes the universe of the products (or alternatives) and $\mathcal{S}$ denotes the collection of subsets of N for which choice data is available. Each $S \in \mathcal{S}$ is an assortment

of the products presented to the consumers. For each assortment $S \in \mathcal{S}$, let $p_{i,S}$ be the fraction of population who choose product $i \in S$. For any assortment collection $\mathcal{S}$, let $\mathcal{I}_{\mathcal{S}}$ denote the collection of all product-assortment pairs (i,S) with $i \in S$ and $S \in \mathcal{S}$. Then the observed choice probability collection $\boldsymbol{p}_{\mathcal{S}} = (p_{i,S} : i \in S, S \in \mathcal{S})$ is a non-negative vector in $\mathbb{R}^{|\mathcal{I}_{\mathcal{S}}|}$ and satisfies $\sum_{i \in S} p_{i,S} = 1$ for every $S \in \mathcal{S}$. We are interested in identifying necessary and sufficient conditions on the observable sales data $\boldsymbol{p}_{\mathcal{S}}$ which make it consistent with the MDM hypothesis. A natural follow-up question is: Can these conditions be verified in polynomial time? The following theorem provides an affirmative answer to these questions.

Theorem 1 (A tractable characterization for MDM). *Under Assumption 1, a choice probability collection $\boldsymbol{p}_{\mathcal{S}}$ is representable by an MDM if and only if there exists a function $\lambda : \mathcal{S} \to \mathbb{R}$ such that for any two assortments $S, T \in \mathcal{S}$ containing a common product $i \in N$,*

$$\begin{aligned} &\lambda(S) > \lambda(T) \quad \text{if} \quad p_{i,S} < p_{i,T}, \\ &\lambda(S) = \lambda(T) \quad \text{if} \quad p_{i,S} = p_{i,T} \neq 0. \end{aligned} \tag{5}$$

As a result, checking whether the given choice data $\boldsymbol{p}_{\mathcal{S}}$ satisfies the MDM hypothesis can be accomplished by solving a linear program with $O(|\mathcal{S}|)$ continuous variables and $O(n|\mathcal{S}|)$ constraints.

A proof for Theorem 1 is given immediately following this discussion. Letting $U(S) = \lambda(S)$ for $S \in \mathcal{S}$, one may understand the characterization in Theorem 1 as follows: The choice data $\boldsymbol{p}_{\mathcal{S}} = (p_{i,S} : i \in S, S \in \mathcal{S})$ given for a collection of assortments $\mathcal{S}$ is representable by MDM if and only if there exists a utility function $U : \mathcal{S} \to \mathbb{R}$ satisfying the following two conditions: (i) By assigning $U(S) < U(T)$, the utility U should exhibit a strict preference for an assortment T over another assortment S containing a common product i whenever $p_{i,S} < p_{i,T}$; and (ii) by assigning $U(S) = U(T)$, it should exhibit indifference between S and T whenever $p_{i,S} = p_{i,T} \neq 0$.

Since utility functions represent rational preference relationships (or rankings), one may equivalently understand the conditions in Theorem 1 as stipulating the existence of a preference relation over the assortment collection $\mathcal{S}$ which is consistent with the partial preferences observed in the choice data. This characterization for MDM, in terms of the existence of a consistent ranking over assortments, is in contrast to RUM which requires the existence of a probability distribution over the $n!$ rankings possible for n products. This is the key reason, why unlike RUM, verifying the consistency of given choice data with the MDM hypothesis can be done in polynomial time.

Proof. Necessity of (5): Suppose $\boldsymbol{p}_{\mathcal{S}}$ is MDM-representable. Then there exist marginal distributions $\{F_i : i \in N\}$ and deterministic utilities $\{\nu_i : i \in N\}$ such that for any assortment $S \in \mathcal{S}$, the given choice probability vector $(p_{i,S} : i \in S)$ and the respective Lagrange multipliers $\lambda_S, \{\lambda_{i,S} : i \in S\}$ are

obtainable by solving the optimality conditions (4). That is, there exist $\{\lambda_S, \lambda_{i,S} : i \in S, S \in \mathcal{S}\}$ for some fixed choice of $\{F_i : i \in N\}$ and $\{\nu_i : i \in N\}$ such that

$$\nu_i + F_i^{-1}(1 - p_{i,S}) \quad \lambda_S + \lambda_{i,S} = 0 \quad \forall (i,S) \in \mathcal{I}_{\mathcal{S}}, \tag{6}$$

$$\lambda_{i,S}\, p_{i,S} = 0 \quad \forall (i,S) \in \mathcal{I}_{\mathcal{S}}. \tag{7}$$

For each product $i \in N$ and any two assortments $S, T \in \mathcal{S}$ containing i as a common product,

$$\lambda_S \quad \nu_i = \lambda_{i,S} + F_i^{-1}(1 - p_{i,S}) \quad \text{and} \quad \lambda_T \quad \nu_i = \lambda_{i,T} + F_i^{-1}(1 - p_{i,T}).$$

If $p_{i,S} < p_{i,T}$, then $\lambda_{i,S} \quad 0$ and $\lambda_{i,T} = 0$ because of the complementary slackness condition (7). Since $F_i^{-1}(1 - p)$ is a strictly decreasing function over $p \in [0,1]$, by (6), we obtain:

$$\lambda_S \quad \nu_i \quad F_i^{-1}(1 - p_{i,S}) > F_i^{-1}(1 - p_{i,T}) = \lambda_T \quad \nu_i.$$

Subtracting ν_i on both sides, we obtain that the Lagrange multipliers should satisfy $\lambda_S > \lambda_T$. If on the other hand $p_{i,S} = p_{i,T} \neq 0$, we have $\lambda_{i,S} = \lambda_{i,T} = 0$ from the optimality conditions. Then $\lambda_S \quad \nu_i = F_i^{-1}(1 - p_{i,S}) = F_i^{-1}(1 - p_{i,T}) = \lambda_T \quad \nu_i$. Again, subtracting ν_i on both sides, we obtain that the Lagrange multipliers should satisfy $\lambda_S = \lambda_T$. Thus, setting $\lambda(S) = \lambda_S$ for all $S \in \mathcal{S}$, we see that there exists a function $\lambda : \mathcal{S} \to \mathbb{R}$ satisfying (5).

Sufficiency of (5): Given $\boldsymbol{p}_{\mathcal{S}}$ and $\lambda : \mathcal{S} \to \mathbb{R}$ such that (5) holds for all $(i,S), (i,T) \in \mathcal{I}_{\mathcal{S}}$, we next exhibit a construction of marginal distributions $(F_i : i \in N)$ and utilities $(\nu_i : i \in N)$ for MDM. This construction will be such that it yields the given $(p_{i,S} : i \in S)$ as the corresponding choice probabilities from the optimality conditions in (4), for any assortment $S \in \mathcal{S}$.

For any product $i \in N$, let $\mathcal{S}_i = \{S \in \mathcal{S} : i \in S\}$ denote the subcollection of assortments $S \in \mathcal{S}$ which contain the product i and let $m_i = |\mathcal{S}_i|$. Further, let l_i denote the number of assortments containing product i for which $p_{i,S} > 0$. Here $l_i = m_i$ when the choice probabilities $\{p_{i,S} : S \in \mathcal{S}_i\}$ are all non-zero. Equipped with this notation, we construct the marginal distribution $F_i(\,)$ for any product $i \in N$ as follows:

(a) Consider any ordering $(S_1, S_2 \ldots, S_{l_i}, S_{l_i+1}, \ldots, S_{m_i})$ over the assortments in $\mathcal{S}_i$ for which $\lambda(S_1) \quad \lambda(S_2) \quad \ldots \quad \lambda(S_{l_i}) < \lambda(S_{l_i+1}) \quad \lambda(S_{l_i+2}) \quad \ldots \quad \lambda(S_{m_i})$. With l_i defined as the number of assortments in $\mathcal{S}_i$ for which $p_{i,S} > 0$, note that it is necessary to have $\lambda(S_{l_i}) < \lambda(S_{l_i+1})$ whenever $l_i < m_i$. This follows from the observations that $\lambda(\,)$ satisfies (5) and $p_{i,S_{l_i}} > 0 = p_{i,S_{l_i+1}}$. Further, due to the conditions in (5), the choice probabilities $(p_{i,S} : S \in \mathcal{S}_i)$ must necessarily satisfy the ordering $p_{i,S_1} \quad p_{i,S_2} \quad \ldots \quad p_{i,S_{l_i}} > 0$ and $p_{i,S_{l_i+1}} = p_{i,S_{l_i+2}} = \ldots, p_{i,S_{m_i}} = 0$.

(b) Construct the cumulative distribution function $F_i(\,)$ by first setting $F_i(\lambda(S_k)) = 1 \quad p_{i,S_k}$ for $k = 1, \quad , l_i$. With this assignment, we complete the construction of the distribution F_i in between these points by connecting them with line segments as follows: For any two consecutive assortments S_k and S_{k+1} in the ordering satisfying $\lambda(S_k) < \lambda(S_{k+1})$, connect the respective points $(\lambda(S_k), 1 \quad p_{i,S_k})$ and $(\lambda(S_{k+1}), 1 \quad p_{i,S_{k+1}})$ with a line segment (see Figure 1). For $k \quad l_i$, note that if the consecutive assortments S_k and S_{k+1} are such that $\lambda(S_k) = \lambda(S_{k+1})$, then the corresponding points $(\lambda(S_k), 1 \quad p_{i,S_k})$ and $(\lambda(S_{k+1}), 1 \quad p_{i,S_{k+1}})$ coincide and there is no need to connect them. Further note that $p_{i,S_k} > p_{i,S_{k+1}}$ when $\lambda(S_k) < \lambda(S_{k+1})$, because of (5), and hence the cumulative distribution function F_i is strictly increasing in the interval $[\lambda(S_1), \lambda(S_{l_i})]$.

(c) Lastly we construct the tails of the distribution F_i as follows: For the right tail, connect the points $(\lambda(S_{l_i}), 1 \quad p_{i,S_{l_i}})$ and $(\lambda(S_{l_i+1}), 1)$ with a line segment if $l_i < m_i$. We then have $F_i(x) = 1$ for every $x \quad \lambda(S_{l_i+1})$ and therefore $F_i^{-1}(1) = \lambda(S_{l_i+1})$. If $l_i = m_i$, connect the points $(\lambda(S_{l_i}), 1 \quad p_{i,S_{l_i}})$ and $(\lambda(S_{l_i}) + \delta, 1)$ by choosing any arbitrary $\delta > 0$ (see Figure 1). In this case, we will have $F_i(x) = 1$ for every $x \quad \lambda(S_{l_i}) + \delta$. For the left tail, if $p_{i,S_1} = 1$, then we have $F_i(x) = 0$ for every $x \quad \lambda(S_1)$. Both the cumulative distribution functions drawn in Figure 1 illustrate this case. On the other hand, if $p_{i,S_1} < 1$, we use a line segment to connect $(\lambda(S_1), 1 \quad p_{i,S_1})$ and $(\lambda(S_1) \quad \delta, 0)$ by choosing an arbitrary $\delta > 0$. In this case, $F_i(x) = 0$ for every $x \quad \lambda(S_1) \quad \delta$.

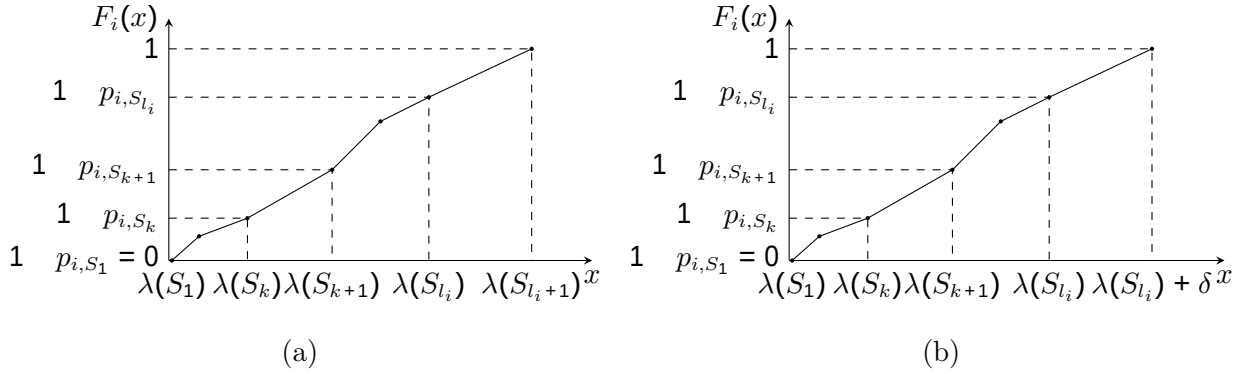


Figure 1 An illustration of the construction of the marginal distribution F_i when: (a) there is an assortment S for which $p_{i,S} = 0$ (the case where $l_i < m_i$) and (b) $p_{i,S} > 0$ for all assortments with product i (the case where $l_i = m_i$).

The above construction gives marginal distribution functions $(F_i : i \; 2 \; N)$ which are absolutely continuous and strictly increasing within its support. We next show that the constructed marginal distributions yield the given choice probabilities $(p_{i,S} : i \; 2 \; S)$, for any assortment $S \; 2 \; S$, when they are used in the optimality conditions (4) together with the assignment $\nu_i = 0$, for $i \; 2 \; N$. In other words, given $\boldsymbol{p}_S$, we next verify that

$$F_i^{-1}(1 \quad p_{i,S}) \quad \lambda(S) + \lambda_{i,S} = 0, \quad \lambda_{i,S} \; p_{i,S} = 0, \quad \text{and} \quad \lambda_{i,S} \quad 0, \quad 8(i,S) \; 2 \; I_S.$$

For any $(i,S) \in \mathcal{I}_{\mathcal{S}}$ with $p_{i,S} > 0$, we have from the construction of F_i that $F_i(\lambda(S)) = 1 - p_{i,S}$. Then for such $p_{i,S}$, we see that the optimality condition $F_i^{-1}(1 - p_{i,S}) - \lambda(S) + \lambda_{i,S} = 0$ readily holds since the optimality conditions also stipulate that $\lambda_{i,S} = 0$ when $p_{i,S} > 0$.

For any $(i,S) \in \mathcal{I}_{\mathcal{S}}$ such that $p_{i,S} = 0$, we have from Steps (a) and (c) of the above construction that $\lambda(S) - \lambda(S_{l_i+1}) = F_i^{-1}(1) = F_i^{-1}(1 - p_{i,S})$. Then if we take $\lambda_{i,S} = \lambda(S) - \lambda(S_{l_i+1})$, we again readily have $F_i^{-1}(1 - p_{i,S}) - \lambda(S) + \lambda_{i,S} = 0$. This completes the verification that for any choice data $\boldsymbol{p}_{\mathcal{S}}$ satisfying (5), there exists marginal distributions $\{F_i : i \in N\}$ and deterministic utilities $\{\nu_i : i \in N\}$ which yield $\boldsymbol{p}_{\mathcal{S}}$ as the MDM choice probabilities.

Lastly, checking whether the conditions in (5) are satisfied for given choice data $\boldsymbol{p}_{\mathcal{S}}$ is equivalent to testing if there exists an assignment for variables $(\lambda_S : S \in \mathcal{S})$ and $\epsilon > 0$ such that,

$$\lambda_S - \lambda_T + \epsilon \quad \text{if} \quad p_{i,S} < p_{i,T},$$
$$\lambda_S = \lambda_T \quad \text{if} \quad p_{i,S} = p_{i,T} \neq 0,$$

for all $(i,S),(i,T) \in \mathcal{I}_{\mathcal{S}}$. This is possible in polynomial time by solving a linear program where the above conditions are formulated as constraints and maximizing ϵ. This linear program involves $|\mathcal{S}|$ variables for $(\lambda_S : S \in \mathcal{S})$ and one variable for ϵ, and at most $n|\mathcal{S}|$ constraints.

3.2. On the representational power of MDM

For any assortment collection $\mathcal{S}$, let $P_{\text{mdm}}(\mathcal{S})$ denote the collection of choice probabilities for the assortments in $\mathcal{S}$ which are representable by any MDM choice model. We use the notation $\lambda(S)$ and λ_S interchangeably here onwards. Due to the characterization in Theorem 1, we have the following succinct description for MDM: $P_{\text{mdm}}(\mathcal{S}) = \text{Proj}_{\boldsymbol{x}}(\Lambda_{\mathcal{S}}) := \{\boldsymbol{x} : (\boldsymbol{x},\boldsymbol{\lambda}) \in \Lambda_{\mathcal{S}}\}$, where $\Lambda_{\mathcal{S}}$ is defined as

$$\Lambda_{\mathcal{S}} = \Big\{(\boldsymbol{x},\boldsymbol{\lambda}) \in \mathbb{R}^{|\mathcal{I}_{\mathcal{S}}|} \times \mathbb{R}^{|\mathcal{S}|} : x_{i,S} \geq 0, \forall (i,S) \in \mathcal{I}_{\mathcal{S}}, \sum_{i \in S} x_{i,S} = 1, \forall S \in \mathcal{S}, \quad (8)$$
$$\lambda_S > \lambda_T \text{ if } x_{i,S} < x_{i,T},\ \lambda_S = \lambda_T \text{ if } x_{i,S} = x_{i,T} \neq 0, \forall (i,S),(i,T) \in \mathcal{I}_{\mathcal{S}}\Big\},$$

for any assortment collection $\mathcal{S}$. One may understand the set $\Lambda_{\mathcal{S}}$ as the collection of MDM choice probabilities augmented with the disutilities $\lambda(\cdot)$ over the assortments. In Theorems 2 and 3 below, we seek to use the characterization in Theorem 1 to understand the representation power of MDM.

Theorem 2. *For any assortment collection $\mathcal{S}$, the collection of choice probabilities represented by MDM has a positive measure. Specifically, $\mu\big(P_{\text{mdm}}(\mathcal{S})\big) > 0$, where μ is the Lebesgue measure on $\prod_{S \subseteq \mathcal{S}} \Delta_S$ and Δ_S denotes the probability simplex $\Delta_S = \{(x_{i,S} : i \in S) : x_{i,S} \geq 0, \forall i \in S, \sum_{i \in S} x_{i,S} = 1\}$.*

Theorem 2 brings out the contrast with the representation power of parametric choice models such as MNL and nested logit. The choice probabilities represented by these parametric alternatives possess zero Lebesgue measure due to the restrictions imposed by the IIA property overall or within

the nests. In Lemma 2 below, we observe that the choice probabilities modeled by MDM are *regular* in the sense that the probability of choosing a specific product $i \in S$ cannot increase if S is enlarged.

Lemma 2. *Suppose that the choice probability collection $\boldsymbol{p}_S \in P_{\text{mdm}}(\mathcal{S})$. Then for any two assortments $S, T \in \mathcal{S}$, we have*

a) $\boldsymbol{p}_S$ satisfies the regularity property, that is, $p_{i,S} \geq p_{i,T}$ if $i \in S$ and $S \subseteq T$; and

b) $p_{i,S} \geq p_{i,T}$ if $p_{j,S} < p_{j,T}$ and $i, j \in S \setminus T$.

Utilizing the MDM characterization in Theorem 1, Theorem 3 below shows that MDM and RUM do not subsume each other generally. It also reveals that RUM and MDM have equivalent representational power as the class of regular choice models, when the assortments collection $\mathcal{S}$ has a special structure, like nested or laminar collections that are frequently encountered in inventory and revenue management applications. To state Theorem 3, we require the following definitions.

An assortment collection $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ is said to be *nested* if $S_1 \subseteq S_2 \subseteq \ldots \subseteq S_m$ for some indexing of the assortments. In other words, the smaller sets are always contained in the larger sets. An assortment collection $\mathcal{S}$ is said to be *laminar* if for any two distinct sets $S, T \in \mathcal{S}$, either $S \subseteq T$, or $T \subseteq S$, or $S \setminus T = \emptyset$. Equivalently, any two sets are either disjoint or related by containment. For any assortment collection $\mathcal{S}$, let $P_{\text{rum}}(\mathcal{S})$ and $P_{\text{reg}}(\mathcal{S})$ denote the collection of choice probabilities over the assortments in $\mathcal{S}$ which are representable by RUM and the class of regular choice models, respectively. Here regular choice models denote the broad collection of all choice models which satisfy the aforementioned regularity property in Part (a) of Lemma 2. For any assortment collection $\mathcal{S}$, it is well-known that $P_{\text{rum}}(\mathcal{S}) \subseteq P_{\text{reg}}(\mathcal{S})$ (see, e.g., Berbeglia and Joret 2017) and we have from Lemma 2 that $P_{\text{mdm}}(\mathcal{S}) \subseteq P_{\text{reg}}(\mathcal{S})$. From the first condition in (5), we observe that the collection $P_{\text{mdm}}(\mathcal{S})$ is not necessarily a closed set. We use $\text{closure}\big(P_{\text{mdm}}(\mathcal{S})\big)$ to denote the closure of $P_{\text{mdm}}(\mathcal{S})$.

Theorem 3 (Relationship between MDM and RUM). *Suppose that $\mathcal{S}$ is a collection of assortments formed over n products. Then the following hold:*

a) When $n = 2$, the choice probabilities represented by RUM and MDM coincide; when $n = 3$, the collection of choice probabilities represented by MDM is subsumed by that of RUM; and when $n \geq 4$, there exist choice probabilities over $\mathcal{S}$ that can be represented by both RUM and MDM and neither models subsume the other: specifically, there exist assortment collections $\mathcal{S}$ such that $P_{\text{mdm}}(\mathcal{S}) \not\subseteq P_{\text{rum}}(\mathcal{S})$ and $P_{\text{rum}}(\mathcal{S}) \not\subseteq P_{\text{mdm}}(\mathcal{S})$.

b) If the assortment collection $\mathcal{S}$ is either nested or laminar, then the corresponding choice probabilities over $\mathcal{S}$ represented by MDM, RUM, and the class of regular models enjoy the following equivalence regardless of n: $P_{\text{rum}}(\mathcal{S}) = \text{closure}\big(P_{\text{mdm}}(\mathcal{S})\big) = P_{\text{reg}}(\mathcal{S})$.

4. A nonparametric approach towards prediction for new assortments

As an application of the exact characterization derived in Theorem 1, we first develop a nonparametric data-driven approach for making revenue or sales predictions for new assortments with no prior sales data. The key idea behind the proposed nonparametric approach is as follows. To predict the revenue or sales for a new assortment, we consider the collection of all MDM choice models which are consistent with the observed sales data and offer the worst-case expected revenue over this collection as an estimate for the revenue or sales. Thus, robust optimization serves as the basis in our approach for allowing data to select a suitable model based on the prediction task at hand.

4.1. A robust optimization formulation for sales and revenue predictions

As in the previous sections, let $N = \{1, \ldots, n\}$ denote the universe of products, $\mathcal{S}$ denote the collection of assortments for which historical choice data, denoted by $\boldsymbol{p}_{\mathcal{S}}$, is available. Suppose that we wish to make sales or revenue predictions for a new assortment $A \notin \mathcal{S}$. Utilizing MDM hypothesis to make predictions is most sensible when the given choice data exhibits the MDM demand characteristics identified in Theorem 1. Therefore, we begin with the assumption that the choice data $\boldsymbol{p}_{\mathcal{S}}$ is MDM-representable. For $i \in N$, let $r_i \in (0, 1)$ denote the revenue obtained by selling one unit of the product i. The collection of all MDM choice probability vectors $\boldsymbol{x}_A = (x_{i,A} : i \in A)$ for the new assortment A which are consistent with the observed choice data $\boldsymbol{p}_{\mathcal{S}}$ is given by,

$$\mathcal{U}_A := \{\boldsymbol{x}_A : (\boldsymbol{p}_{\mathcal{S}}, \boldsymbol{x}_A) \in P_{\text{mdm}}(\mathcal{S}^0)\},$$

where $\mathcal{S}^0 = \mathcal{S} \cup \{A\}$. Here, as before, $P_{\text{mdm}}(\mathcal{S}^0)$ denotes the collection of all MDM representable choice probabilities over the assortment collection $\mathcal{S}^0$. The worst-case expected revenue over the consistent collection $\mathcal{U}_A$ is then defined as

$$\underline{r}(A) := \inf_{x_A \in \mathcal{U}_A} \sum_{i \in A} r_i x_{i,A}. \tag{9}$$

Due to the exact characterization in Theorem 1, we obtain the following reformulation for $\underline{r}(A)$.

Proposition 1. *Suppose that the given choice data collection $\boldsymbol{p}_{\mathcal{S}} \in P_{\text{mdm}}(\mathcal{S})$ and $A \notin N$ is an assortment not in $\mathcal{S}$. Then the worst-case expected revenue $\underline{r}(A)$ equals,*

$$\begin{aligned}
&\min_{x_A,\ \lambda \in \mathbb{R}^{|\mathcal{S}^0|}} \sum_{i \in A} r_i x_{i,A} \\
&\text{s.t.} \quad x_{i,A} \geq p_{i,S} \quad \text{if} \quad \lambda_A \geq \lambda_S, \qquad \forall i \in A,\ (i,S) \in I_{\mathcal{S}}, && \text{(10a)}\\
&\qquad\ x_{i,A} \leq p_{i,S} \quad \text{if} \quad \lambda_A \geq \lambda_S, \qquad \forall i \in A,\ (i,S) \in I_{\mathcal{S}}, && \text{(10b)}\\
&\qquad \sum_{i \in A} x_{i,A} = 1, \\
&\qquad x_{i,A} \geq 0, \qquad \forall i \in A, \\
&\qquad \lambda_S > \lambda_T \quad \text{if} \quad p_{i,S} < p_{i,T}, \qquad \forall (i,S), (i,T) \in I_{\mathcal{S}}, \\
&\qquad \lambda_S = \lambda_T \quad \text{if} \quad p_{i,S} = p_{i,T} \neq 0, \quad \forall (i,S), (i,T) \in I_{\mathcal{S}}.
\end{aligned}$$

One may also obtain worst-case sales predictions for a product i, when offered within assortment A, by letting $r_i = 1$ and $r_j = 0$ for $j \neq i$ in the objective in the reformulation in Proposition 1. Similarly, replacing the minimization in this reformulation with a maximization yields an best-case (optimistic) revenue estimate $r(A)$.

Observe when choice data is available for a richer assortment collection, it leads to less-conservative estimate for $\underline{r}(A)$ and a narrower interval $[r(A), \underline{r}(A)]$ as plausible values for revenue estimates which are consistent with the given data and the MDM hypothesis. This is because the number of constraints in the constraint collections (10a)-(10b) is larger when the assortment collection S for which choice data is available is made richer. In other words, $I_{S_1} \quad I_{S_2}$ when $S_1 \quad S_2$ and therefore the resulting $P_{\text{mdm}}(S_2 [fA)$ is nested within $P_{\text{mdm}}(S_1 [fA)$.

4.2. A mixed integer linear formulation for the worst case expected revenue

As a generally applicable approach for evaluating the worst-case revenue $\underline{r}(A)$, one may model the "if" conditions in (10a) - (10b) via additional binary variables $(\delta_S^+, \delta_S : S \in S)$ to obtain the mixed-integer linear reformulation with $O(n|S|)$ binary variables, $O(n + |S|)$ continuous variables, and $O(n|S|)$ constraints as follows.

Proposition 2. *Suppose that the assumptions in Proposition 1 are satisfied. Then for any* $0 < \epsilon < 1/(2|S|)$, *the worst-case expected revenue* $\underline{r}(A)$ *equals the value of the following mixed integer linear program:*

$$\min_{x_A, \ , \ , \ ^+, \ } \quad \sum_{i \in A} r_i x_{i,A}$$

$$\text{s.t.} \quad \delta_{A,S} \quad \lambda_A \quad \lambda_S \quad 1 \quad (1+\epsilon)\delta_{A,S}, \qquad \forall i \in A, (i,S) \in I_S, \tag{11a}$$

$$\delta_{S,A} \quad \lambda_S \quad \lambda_A \quad 1 \quad (1+\epsilon)\delta_{S,A}, \qquad \forall i \in A, (i,S) \in I_S, \tag{11b}$$

$$\delta_{A,S} \quad 1 \quad x_{i,A} \quad p_{i,S} \quad 1 \quad \delta_{S,A}, \qquad \forall i \in A, (i,S) \in I_S, \tag{11c}$$

$$(\delta_{A,S} + \delta_{S,A}) \quad x_{i,A} \quad p_{i,S} \quad \delta_{A,S} + \delta_{S,A}, \quad \forall i \in A, (i,S) \in I_S, \tag{11d}$$

$$\lambda_S \quad \lambda_T \quad \epsilon, \qquad \forall (i,S),(i,T) \in I_S \ \text{ s.t. } \ p_{i,S} < p_{i,T},$$

$$\lambda_S \quad \lambda_T \quad = \quad 0, \qquad \forall (i,S),(i,T) \in I_S \ \text{ s.t. } \ p_{i,S} = p_{i,T} \neq 0,$$

$$\sum_{i \in A} x_{i,A} \quad = \quad 1,$$

$$0 \quad \lambda_A \quad 1, \quad x_{i,A} \quad 0, \ \forall i \in A, \quad 0 \quad \lambda_S \quad 1, \quad \delta_{A,S}, \delta_{S,A} \in \{0,1\}, \ \forall S \in S.$$

Likewise, the optimistic expected revenue $r(A) := \sup_{x_A \in U_A} \sum_{i \in A} r_i x_{i,A}$ *equals the optimal value obtained by maximizing over the constraints in the above mixed integer linear program.*

4.3. Polynomial-time algorithms for prediction with structured collections

Besides the generally applicable mixed integer convex program in Proposition 2, we develop an alternative solution approach leveraging special structures, such as nested or laminar structures, in the assortment collection in order to evaluate the worst-case revenues $\underline{r}(A)$ in polynomial time. Corollary 1 and Proposition 3 below show that computing the worst-case expected revenue can be efficient when the assortment collection $\mathcal{S}$ is structured.

Corollary 1. *When $\mathcal{S}^0$ is either nested or laminar, evaluating $\underline{r}(A)$ in (9) is equivalent to solving the following linear program with $O(n)$ continuous variables and $O(n|\mathcal{S}|)$ constraints:*

$$\begin{aligned} \min_{x_A} \quad & \sum_{i \in A} r_i x_{i,A} \\ \text{s.t.} \quad & x_{i,A} \le p_{i,S}, \ \text{s.t. } S \supseteq A, \quad \forall i \in A, (i,S) \in I_{\mathcal{S}}, \\ & x_{i,A} \ge p_{i,S}, \ \text{s.t. } A \supseteq S, \quad \forall i \in A, (i,S) \in I_{\mathcal{S}}, \\ & \sum_{i \in A} x_{i,A} = 1, \quad x_{i,A} \ge 0, \quad \forall i \in A. \end{aligned} \tag{12}$$

The result in Corollary 1 follows from the conclusion in Theorem 3 that $\text{closure}\big(P_{\text{mdm}}(\mathcal{S}^0)\big) = P_{\text{reg}}(\mathcal{S}^0)$ when $\mathcal{S}^0$ is either nested or laminar. Next, we focus on the nested structure and relax the structure assumption on A. Without loss of generality, when $\mathcal{S}$ is nested, let $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ such that $S_1 \supseteq S_2 \supseteq \cdots \supseteq S_m$. For ease of notation, let $p_{i,S_0} = 1$ and $p_{i,S_{m+1}} = 0$ for any $i \in A$.

Proposition 3. *Suppose that the assortment collection $\mathcal{S}$ is nested. Then for any given A,*

$$\begin{aligned} \underline{r}(A) = \min_{k=0,1,\ldots,|\mathcal{S}|} \quad & \boldsymbol{R}_k \\ \text{where} \quad \boldsymbol{R}_k = \min_{x_A} \quad & \sum_{i \in A} r_i x_{i,A} \\ \text{s.t.} \quad & x_{i,A} \le p_{i,S_k}, \quad \forall i \in A \setminus S_k, \\ & x_{i,A} \ge p_{i,S}, \quad \forall i \in A, (i,S) \in I_{\mathcal{S}}, S_{k+1} \supseteq S \supseteq S_k, \\ & \sum_{i \in A} x_{i,A} = 1, \quad x_{i,A} \ge 0 \quad \forall i \in A. \end{aligned} \tag{13}$$

Observe that (13) involves $|\mathcal{S}| + 1$ linear programs, each with $O(n)$ continuous variables and $O(n|\mathcal{S}|)$ constraints. Thus, both (12) and (13) are tractable.

5. Limit of MDM and the estimation of best-fitting MDM probabilities

Customer preferences captured by choice data need not always satisfy any specific choice model hypothesis perfectly. Considering choice data instances that are not MDM-representable, we next seek to quantify the limit or the cost of approximating given choice data with an MDM choice model and a procedure for identifying MDM-representable choice probabilities offering the best fit.

5.1. A limit of MDM formulation

Given choice data $\boldsymbol{p}_S$ and any $\boldsymbol{x}_S \in P_{\text{mdm}}(S)$, suppose that a loss function $\boldsymbol{x}_S \mapsto \text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S)$ measures the degree of inconsistency in approximating choice data $\boldsymbol{p}_S$ with an MDM-consistent choice probability assignment $\boldsymbol{x}_S$. We take the loss function to be non-negative, strictly convex, and satisfying the property that $\text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) = 0$ if and only if $\boldsymbol{x}_S = \boldsymbol{p}_S$. Suppose that $(w_S : S \in \mathcal{S})$ is a vector of non-negative weights over assortments in S. Then a norm-based loss such as $\sum_{S \in \mathcal{S}} w_S \|\boldsymbol{p}_S - \boldsymbol{x}_S\|$ or a Kullback-Liebler divergence based loss such as $\sum_{S \in \mathcal{S}} w_S \sum_{i \in S} p_{i,S} \log(x_{i,S}/p_{i,S})$ serve as prominent examples among the losses which satisfy these assumptions. For $S \in \mathcal{S}$, the weight w_S may be taken, for example, to be the frequency with which the offer set S has been shown to customers in the choice dataset.

We define the limit of the MDM, denoted by $L(\boldsymbol{p}_S)$, as the smallest value of $\text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S)$ attainable by fitting the observed data $\boldsymbol{p}_S$ with an MDM choice model:

$$L(\boldsymbol{p}_S) = \inf \{\text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) : \boldsymbol{x}_S \in P_{\text{mdm}}(S)\}. \tag{14}$$

As is evident from the definition above, evaluating the limit $L(\boldsymbol{p}_S)$ can be viewed as identifying a choice probability assignment $\boldsymbol{x}_S$ which is consistent with the MDM hypothesis and is about as close any MDM model can be to the observed choice data $\boldsymbol{p}_S$. Thus any $\boldsymbol{x}_S$ attaining the minimum in (14) can be seen as offering the best fit, within the MDM family, to the observed choice data. In particular, suppose we take $\text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) = \sum_{S \in \mathcal{S}} w_S \sum_{i \in S} p_{i,S} \log(x_{i,S}/p_{i,S})$ and the weight w_S, for $S \in \mathcal{S}$, to be equal to the number of observations available for an assortment S in the choice dataset. Then, as highlighted in Example 2.1 of Jagabathula and Rusmevichientong (2019), $\boldsymbol{x}_S$ is a minimizer in the limit formulation (14) if and only if it maximizes the likelihood. Thus, in this case, a solution to the limit (14) can be viewed as being obtained from *maximum likelihood estimation* in the MDM family without any parametric restrictions.

Besides this use in estimation, one may also use the limit $L(\boldsymbol{p}_S)$ as a diagnostic tool for determining how well MDM is suitable for fitting choice data and comparing it with how effective any parametric subclass is in accomplishing the same. To see this use at a conceptual level, suppose that $\bar{\boldsymbol{x}}_S$ denotes the choice probabilities obtained by fitting a parametric subclass of MDM, such as MNL (or) marginal exponential model. Then, as put forward by Jagabathula and Rusmevichientong (2019), one may view the overall loss captured by $\text{loss}(\boldsymbol{p}_S, \bar{\boldsymbol{x}}_S)$ as below:

$$\text{loss}(\boldsymbol{p}_S, \bar{\boldsymbol{x}}_S) = L(\boldsymbol{p}_S) + \{\text{loss}(\boldsymbol{p}_S, \bar{\boldsymbol{x}}_S) - L(\boldsymbol{p}_S)\},$$

where the second component $\text{loss}(\boldsymbol{p}_S, \bar{\boldsymbol{x}}_S) - L(\boldsymbol{p}_S)$ is the incremental cost that comes with employing a parametric model within the MDM family in order to approximate the choice data. If data suggests

that this incremental parametric cost is higher relative to the limit $L(\boldsymbol{p}_S)$, then one should consider a richer parametric model (or) use the general nonparametric MDM over the chosen parametric class. If, on the other hand, the loss $L(\boldsymbol{p}_S)$ due to MDM itself is large, then MDM should possibly not be considered as a suitable model for the given choice data.

Recall the characterization $P_{\text{mdm}}(S)$ as the projection $\{\boldsymbol{x} : (\boldsymbol{x}, \lambda) \in \quad _S\}$, where $\quad _S$ is defined in (8). Due to this characterization, we have the following equivalent formulation for the limit $L(\boldsymbol{p}_S)$.

Proposition 4. *Under Assumption 1, the limit $L(\boldsymbol{p}_S)$ equals*

$$\begin{array}{lll} \min\limits_{x_S,} & \sum\limits_{S \in S} \text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) & \\ \text{s.t.} & x_{i,S} \quad x_{i,T} \text{ if } \lambda_S \quad \lambda_T, & \forall (i,S),(i,T) \in I_S, \\ & \sum\limits_{i \in S} x_{i,S} = 1, & \forall S \in S, \\ & x_{i,S} \quad 0, & \forall (i,S) \in I_S. \end{array} \tag{15}$$

The set of MDM-representable choice probabilities $P_{\text{mdm}}(S)$is non-convex (see, Example EC.3). The following theorem is based on reducing a specific instance of the formulation (15) to the Kemeny optimal rank aggregation problem.

Theorem 4. *Problem* (15) *is NP-hard.*

5.2. A mixed integer convex reformulation for the limit of MDM

Proposition 5 below provides a generally applicable mixed-integer convex reformulation for (15).

Proposition 5. *Suppose that Assumption 1 is satisfied. Then for any $0 < \epsilon < 1/(2|S|)$, the limit $L(\boldsymbol{p}_S)$ equals the value of the following mixed integer convex program:*

$$\begin{array}{lll} \min\limits_{x, \ ,} & \sum\limits_{S \in S} \text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) & \\ \text{s.t.} & \delta_{S,T} \quad \lambda_S \quad \lambda_T \quad 1 \quad (1+\epsilon)\delta_{S,T}, & \forall (i,S), (i,T) \in I_S, \\ & \delta_{S,T} \quad 1 \quad x_{i,S} \quad x_{i,T} \quad 1 \quad \delta_{T,S}, & \forall (i,S), (i,T) \in I_S, \\ & (\delta_{S,T} + \delta_{T,S}) \quad x_{i,S} \quad x_{i,T} \quad \delta_{S,T} + \delta_{T,S}, & \forall (i,S), (i,T) \in I_S, \\ & \sum\limits_{i \in S} x_{i,S} = 1, \ \forall S \in S, & x_{i,S} \quad 0, \ \forall (i,S) \in I_S, \\ & 0 \quad \lambda_S \quad 1, \ \forall S \in S, & \delta_{S,T} \in \{0,1\}, \ \forall S,T \in S. \end{array} \tag{16}$$

Suppose that $(\boldsymbol{x}_S, \boldsymbol{\lambda})$ attains the minimum in (15) or equivalently in (16). Then for any new assortment $A \not\in S$, one may use the constraints in Proposition 2 to obtain the robust revenue estimate $\underline{r}(A)$ consistent with the fitted choice probabilities $\boldsymbol{x}_S$ as below:

$$\min_{x, \lambda_A, \ ^+,} \quad \sum_{i \in A} r_i x_{i,A}$$

$$
\begin{aligned}
\text{s.t.} \quad & \delta_{A,S} \quad \lambda_A \quad \lambda_S \quad 1 \quad (1+\epsilon)\delta_{A,S}, && \forall i \in A, (i,S) \in I_S, \\
& \delta_{S,A} \quad \lambda_S \quad \lambda_A \quad 1 \quad (1+\epsilon)\delta_{S,A}, && \forall i \in A, (i,S) \in I_S, \\
& \delta_{A,S} \quad 1 \quad x_{i,A} \quad x_{i,S} \quad 1 \quad \delta_{S,A}, && \forall i \in A, (i,S) \in I_S, \\
& (\delta_{A,S} + \delta_{S,A}) \quad x_{i,A} \quad x_{i,S} \quad \delta_{A,S} + \delta_{S,A}, && \forall i \in A, (i,S) \in I_S, \\
& \sum_{i \in A} x_{i,A} = 1, \quad 0 \quad \lambda_A \quad 1, \\
& x_{i,A} \quad 0, \ \forall i \in A, \quad \delta_{A,S}, \delta_{S,A} \in \{0,1\}, \ \forall S \in \mathcal{S}.
\end{aligned}
$$

Since the constraints in (15) allow $x_{i,S} = x_{i,T}$ even when the counterpart $\lambda_S \neq \lambda_T$, observe that the solution $\boldsymbol{x}_S$ can only be guaranteed to be arbitrarily close to the MDM-representable collection $P_{\text{mdm}}(S)$. Therefore if one wishes to obtain a δ-optimal MDM-representable choice probability assignment, for some $\delta > 0$, they may do so as follows: Equipped with the optimal value $L(\boldsymbol{p}_S) = \text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S)$ and an optimal $\boldsymbol{\lambda}$, a δ-optimal MDM-representable choice probability assignment $\boldsymbol{x}_S$ can be obtained by solving the following convex program:

$$
\begin{aligned}
\max_{x_S, \epsilon} \quad & \epsilon && (17) \\
\text{s.t.} \quad & \text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) \quad L(\boldsymbol{p}_S) + \delta, \\
& x_{i,S} \quad x_{i,T} + \epsilon \text{ if } \lambda_S < \lambda_T, && \forall (i,S), (i,T) \in I_S, \\
& x_{i,S} = x_{i,T} \text{ if } \lambda_S = \lambda_T, && \forall (i,S), (i,T) \in I_S, \\
& \sum_{i \in S} x_{i,S} = 1, \ \forall S \in \mathcal{S}, \quad x_{i,S} \quad 0, \ \forall (i,S) \in I_S.
\end{aligned}
$$

Due to the constraints in (17) and the characterization in Theorem 1, we have that any choice probability collection $\boldsymbol{x}_S$ obtained by solving (17) is MDM-representable. Further, it cannot be improved to offer a fit which is better by more than δ magnitude, for any arbitrary choice of $\delta > 0$, due to the constraint $\text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) \quad L(\boldsymbol{p}_S) + \delta$. Note that when loss(,) is defined in terms of the L_1-norm, the formulation (17) is a linear program with $O(n|S|)$ continuous variables and $O(n|S|^2)$ constraints and (16) is a mixed-integer linear program with $O(|S|^2)$ binary variables, $O(n|S|)$ continuous variables and $O(n|S|^2)$ constraints.

5.3. Polynomial time algorithms for special cases

Besides the mixed integer convex program in Proposition 5, we develop an alternative solution approach that seeks to evaluate the limit of MDM by searching over admissible rankings over assortments in $\mathcal{S}$. This algorithm is capable of evaluating the limit in polynomial time either if the assortment collection $\mathcal{S}$ possesses a nested or laminar structure, or, if $|\mathcal{S}|$ is fixed. In particular, Corollary 2 below shows that evaluating $L(\boldsymbol{p}_S)$ can be efficient by utilizing the Theorem 3 conclusion that $\text{closure}\big(P_{\text{mdm}}(S)\big) = P_{\text{reg}}(S)$ under nested or laminar $\mathcal{S}$. In this case, the constraints of $L(\boldsymbol{p}_S)$ in (15) can be replaced with the regularity conditions for choice probabilities over $\mathcal{S}$.

Corollary 2. *When S is nested or laminar, evaluating the limit $L(\boldsymbol{p}_S)$ in Proposition 4 reduces to the following convex program with $O(n|S|)$ continuous variables and $O(n|S|^2)$ linear constraints:*

$$\begin{aligned} \min_{x_S} \quad & \sum_{S \in \mathcal{S}} \text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) \\ \text{s.t.} \quad & x_{i,S} \quad x_{i,T} \text{ if } S \quad T, \quad \forall (i,S),(i,T) \in I_{\mathcal{S}}, \\ & \sum_{i \in S} x_{i,S} = 1, \ \forall S \in \mathcal{S}, \quad x_{i,S} \quad 0, \ \forall (i,S) \in I_{\mathcal{S}}. \end{aligned} \tag{18}$$

Besides the polynomial algorithms for the limit with special collection structures, evaluating the limit $L(\boldsymbol{p}_S)$ in Proposition 4 is polynomial in the number of alternatives n when the size of the assortment collection $|\mathcal{S}|$ is fixed. An algorithm for this case is provided in Section EC.5.

6. Grouped Marginal Distribution Model

In this section, we consider MDM choice models in which the alternatives can be grouped based on the marginal distributions of the stochastic components of the utilities.

Definition 1 (G-MDM). An MDM specified by the marginal distributions $\{F_i : i \in N\}$ and deterministic utilities $\{\nu_i : i \in N\}$ is called a grouped marginal distribution model (or G-MDM) if there exists (i) a partition $G = \{G_1, \ldots, G_K\}$ of the set of alternatives N, and (ii) a distinct marginal distribution $\hat{F}_l$ for each group G_l such that $F_i(\cdot) = \hat{F}_l(\cdot)$ for every $i \in G_l$ and $l \in \{1, \ldots, K\}$.

As is evident from Definition 1, G-MDM places additional restrictions on the distributions of the stochastic noise terms when compared to the general treatment for MDM we have been considering. Consequently, the set of choice probabilities representable by G-MDM over assortments in $\mathcal{S}$ will be a subset of $P_{\text{mdm}}(\mathcal{S})$, for any assortment collection $\mathcal{S}$. The worst-case sales and revenue estimates one arrives with G-MDM will be less conservative than a general MDM in the case of strict containment, and the utility of G-MDM lies therein. G-MDM offers an opportunity for the modeler to translate domain knowledge, if any, about the distributions of the noise terms into less conservative sales and revenue estimates for new assortments with no prior sales data.

The concept of grouping in G-MDM can be understood as organizing products based on the similarity of their marginal distributions of the noise terms. This notion is related to heteroskedasticity, a concept in economics where different random variables exhibit differing variances. In the realm of choice modeling, researchers have explicitly incorporated heteroskedasticity through models such as the heteroscedastic extreme value model (Bhat 1995), the heteroscedastic exponomial choice model (Alptekinoğlu and Semple 2021), and the marginal exponential model (Mishra et al. 2014).

To illustrate heteroskedasticity, let us consider the example of purchasing a cellphone. Suppose the choice set consists of iPhones and other phones from lesser-known brands. Since iPhones are well-known and have consistent features, the utilities associated with them are likely to have less

variability. In contrast, phones from lesser-known brands may have more extreme utility values due to varying preferences: some individuals might highly value specific features, while others may assign very low utilities to these phones.

In the context of G-MDM, a modeler can benefit by grouping all iPhones together and placing phones from lesser-known brands in a separate group. By doing so, G-MDM can strike a better trade-off compared to general MDM by avoiding overfitting, and compared to APU by avoiding misspecification. Such grouping allows G-MDM to capture the heteroskedasticity in data, leading to more accurate and robust results in choice modeling scenarios.

For notational convenience, let $g : N \to G$ be a function that maps an alternative in N to a group in G. For example, $g(i) = l$ means $i \in G_l$. Then one can write the objective function (3) in the convex formulation which yields the respective G-MDM choice probabilities as,

$$\sum_{i \in S} \nu_i x_i + \sum_{i \in S} \int_{x_i}^{1} F_{g(i)}^{-1}(t)\, dt = \sum_{i \in S} \nu_i x_i + \sum_{l=1}^{K} \sum_{i \in S: g(i)=l} \int_{x_i}^{1} \hat{F}_l^{-1}(t)\, dt. \tag{19}$$

As mentioned in Section 2, a special case of MDM in which the marginal distributions $\{F_i : i \in N\}$ are taken to be identical leads to the well-known APU model in Fudenberg et al. (2015). Similar to the general MDM considered in previous sections, we shall see that the restrictions imposed on the choice probabilities by the grouping of products can be captured by computationally tractable necessary and sufficient conditions, and one can leverage such conditions to develop nonparametric procedures for estimation and prediction.

6.1. A tractable characterization for G-MDM

Theorem 5 below reveals how the additional restrictions, which are imposed by the identicalness of marginal distributions within the product groups, manifest in the choice probability observations.

Theorem 5 (A tractable characterization for G-MDM). *Under Assumption 1, a choice probability collection $\boldsymbol{p}_S$ is representable by a G-MDM if and only if there exists a function $\lambda : S \to \mathbb{R}$ and $\nu \in \mathbb{R}^n$ such that for all $(i, S), (j, T) \in I_S$ with $g(i) = g(j)$:*

$$\begin{aligned} &\lambda(S) \quad \nu_i > \lambda(T) \quad \nu_j \quad \text{if} \quad p_{i,S} < p_{j,T}, \\ &\lambda(S) \quad \nu_i = \lambda(T) \quad \nu_j \quad \text{if} \quad p_{i,S} = p_{j,T} \neq 0. \end{aligned} \tag{20}$$

As a result, checking whether given choice data $\boldsymbol{p}_S$ satisfies the G-MDM hypothesis can be accomplished by solving a linear program with $O(n|S|)$ continuous variables and $O(n^2|S|^2)$ constraints.

It is instructive to compare the necessary and sufficient conditions in Theorem 5 with that obtained for the MDM in Theorem 1. The conditions in Theorem 1 do not require comparing choice probabilities $p_{i,S}$ and $p_{j,T}$ for $i \neq j$. With fewer constraints, the collection of choice probabilities allowed by MDM is strictly larger than any of its G-MDM counterparts in which at least two

products are grouped together. The conditions in Theorem 5 also reveal that a G-MDM with fewer groups contains strictly fewer conditions and thus provides strictly more representation power (see Example EC.4). By setting $K = 1$, we deduce Corollary 3 below for the special case where all the products are grouped together: that is, grouping $G = \{G_1\}$ and $G_1 = N$.

Corollary 3. *(A tractable characterization under identical marginals). When* $K = 1$*,* $\boldsymbol{p}_S$ *is G-MDM-representable if and only if there exists there exists a function* $\lambda : S \to \mathbb{R}$ *and* $\boldsymbol{\nu} \in \mathbb{R}^n$ *such that the conditions in* (20) *are satisfied for all* $(i, S), (j, T) \in I_S$.

The conditions in Corollary 3 are equivalent to those derived in Fudenberg et al. (2015) and describe the probabilities that can be obtained with the APU model. For any grouping G over the products in an assortment collection S, let $P_G(S)$ denote the collection of choice probabilities $\boldsymbol{p}_S$ which satisfy the G-MDM representable conditions in Theorem 5 and $P_{\text{apu}}(S)$ denote the collection of choice probabilities $\boldsymbol{p}_S$ which satisfies the APU representable conditions in Corollary 3. Analogous to Theorem 2, the set of G-MDM representable choice probabilities has a positive Lebesgue measure.

Theorem 6. *For any assortment collection* S *and a grouping* G *of the products, the collection of choice probabilities represented by G-MDM has a positive measure: Specifically,* $\mu\big(P_G(S)\big) - \mu\big(P_{\text{apu}}(S)\big) > 0$*, where* μ *is the Lebesgue measure on the probability simplexes* $\prod_{S \in \mathcal{S}} \Delta_S$.

In Theorem 6, we demonstrate that the entire APU already possesses a positive measure. This result implies that by removing the parametric assumption in the marginal distributions of MDM, even if we enforce identical marginals in MDM, the choice model exhibits a significantly greater representation power compared to many other parametric choice models. Similar to Theorem 2, the proof of Theorem 6 relies on constructing an MNL model, which forms the core around which perturbations in choice probabilities can still be represented by the APU. However, due to the APU's reduced representational capacity, constructing such an MNL model becomes more challenging. Nevertheless, we succeed in proving Lemma EC.5, which establishes a close connection to number theory and leads us to the desired property. The proof technique employed for these two theorems and the associated technical lemma is innovative and may be of independent interest.

6.2. Revenue Prediction with G-MDM

Similar to MDM, suppose that we wish to make sales or revenue predictions for a new assortment $A \notin \mathcal{S}$ with a G-MDM choice model over a grouping G. Like in Section 4, we begin the assumption that the choice data $\boldsymbol{p}_S \in P_G(S)$: that is, $\boldsymbol{p}_S$ is representable by a G-MDM. Let the collection of all G-MDM choice probability vectors $\boldsymbol{x}_A = (x_{i,A} : i \in A)$ for the new assortment A which are consistent with the observed choice data $\boldsymbol{p}_S$ be denoted by $U_G := \{\boldsymbol{x}_A : (\boldsymbol{p}_S, \boldsymbol{x}_A) \in P_G(S^0)\}$, where $S^0 = S \cup \{A\}$. The worst-case expected revenue over the consistent collection U_G is then defined

as $\underline{r}(A) := \inf_{x_A \in U_G} \sum_{i \in A} r_i x_{i,A}$. Similar to Proposition 2, we have in Proposition 6 below a mixed integer linear reformulation for evaluating $\underline{r}(A)$ under the G-MDM assumption. Recall the notations $S^0 = S \cup \{A\}$ and $I_{S^0} = \{(i,S) : i \in S, S \in S^0\}$.

Proposition 6. *For any $0 < \epsilon < 1/(2n|S|)$, the worst-case expected revenue $\underline{r}(A)$ equals the value of the following mixed integer linear program:*

$$
\begin{aligned}
\min_{x,\,\lambda,\,\nu,\,\delta} \quad & \sum_{i \in A} r_i x_{i,A} \\
\text{s.t.} \quad & x_{i,S} = p_{i,S}, && \forall (i,S) \in I_S, \\
& \delta_{i,j,S,T} \le \lambda_S + \nu_i + \lambda_T + \nu_j \le 1 + (1+\epsilon)\delta_{i,j,S,T}, && \forall (i,S),(j,T) \in I_{S^0} : g(i) = g(j), \\
& \delta_{i,j,S,T} \le 1 - x_{i,S} + x_{j,T} \le 1 + \delta_{j,i,T,S}, && \forall (i,S),(j,T) \in I_{S^0} : g(i) = g(j), \\
& (\delta_{i,j,S,T} + \delta_{j,i,T,S}) \le x_{i,S} + x_{i,T} \le \delta_{i,j,S,T} + \delta_{j,i,T,S}, && \forall (i,S),(j,T) \in I_{S^0} : g(i) = g(j), \\
& \sum_{i \in A} x_{i,A} = 1, \quad x_{i,A} \ge 0, \ \forall i \in A, \quad 0 \le \lambda_S + \nu_i \le 1, && \forall (i,S) \in I_{S^0}, \\
& \delta_{i,j,S,T} \in \{0,1\}, && \forall (i,S),\ (j,T) \in I_{S^0} : g(i) = g(j).
\end{aligned}
\tag{21}
$$

Likewise, the optimistic expected revenue $r(A) := \sup_{x_A \in U_G} \sum_{i \in A} r_i x_{i,A}$ equals the optimal value obtained by maximizing over the constraints in the above mixed integer linear program.

6.3. Limit of G-MDM and the estimation of best-fitting G-MDM probabilities

Given choice data $\boldsymbol{p}_S$ over assortments in a collection S and a strictly convex non-negative loss function satisfying the assumptions in Section 5, we define the limit of G-MDM over a grouping G as

$$
L_G(\boldsymbol{p}_S) = \inf \{\text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) : \boldsymbol{x}_S \in P_G(S)\}. \tag{22}
$$

The limit $L_G(\boldsymbol{p}_S)$ measures the cost of approximating given choice data $\boldsymbol{p}_S$ with a G-MDM equipped with grouping G. As in Section 5, a key use of the limit $L_G(\boldsymbol{p}_S)$ is that a near-optimal solution $\boldsymbol{x}_S$ to (22) delivers a fit which is about as good one can get in fitting the given choice data $\boldsymbol{p}_S$ within the $P_G(S)$ family. As a result, it can also serve as a diagnostic tool suitable for determining whether a chosen grouping of products is supported by data. To see this, observe that the cost $L_G(\boldsymbol{p}_S)$ can be decomposed as $L_G(\boldsymbol{p}_S) = L(\boldsymbol{p}_S) + \{L(\boldsymbol{p}_S) - L_G(\boldsymbol{p}_S)\}$, where the second component $L(\boldsymbol{p}_S) - L_G(\boldsymbol{p}_S)$ can be viewed as the incremental cost of the modeling choice made in combining the products into groups G. If one observes this incremental cost to be large relative to the respective MDM limit $L(\boldsymbol{p}_S)$ in this decomposition, then it is an indication lent by data that the modeler would benefit by moving beyond the grouping G; one may proceed in this case by either working with the general MDM family treated in Sections 3 - 5 (or) by refining the grouping to have finer partitions than G.

Theorem 5 allows us reformulate (22) as following in (23), which in turn leads to a mixed integer convex program in Proposition 7 below.

$$
\begin{aligned}
L_G(\boldsymbol{p}_S) = \inf_{x_S,\ ,\ ,} \quad & \sum_{S \in \mathcal{S}} \text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) \\
\text{s.t.} \quad & x_{i,S} < x_{j,T} \text{ if } \lambda_S \quad \nu_i > \lambda_T \quad \nu_j, \qquad \forall (i,S),(j,T) \in I_S : g(i) = g(j), \\
& x_{i,S} = x_{j,T} > 0 \text{ if } \lambda_S \quad \nu_i = \lambda_T \quad \nu_j, \quad \forall (i,S),(j,T) \in I_S : g(i) = g(j), \\
& \sum_{i \in S} x_{i,S} = 1, \ \forall S \in \mathcal{S}, \qquad x_{i,S} \quad 0, \ \forall (i,S) \in I_S,
\end{aligned}
\tag{23}
$$

where the decision variables are the choice probabilities $x_{i,S}$ associated with alternative-assortment pairs, the Lagrange multipliers λ_S associated with assortments, and the deterministic utilities ν_i associated with the alternatives.

Proposition 7. *Suppose that Assumption 1 is satisfied. Then for any* $0 < \epsilon < 1/(2n|\mathcal{S}|)$, *the limit* $L_G(\boldsymbol{p}_S)$ *in* (22) *equals the value of the following mixed integer convex program:*

$$
\begin{aligned}
\min_{x_S,\ ,\ ,} \quad & \sum_{S \in \mathcal{S}} \text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) \\
\text{s.t.} \quad & \delta_{i,j,S,T} \quad \lambda_S \quad \nu_i \quad \lambda_T + \nu_j \quad 1 \quad (1+\epsilon)\delta_{i,j,S,T}, \quad \forall (i,S),\ (j,T) \in I_S : g(i) = g(j), \\
& \delta_{i,j,S,T} \quad 1 \quad x_{i,S} \quad x_{j,T} \quad 1 \quad \delta_{j,i,T,S}, \qquad \forall (i,S),\ (j,T) \in I_S : g(i) = g(j), \\
& (\delta_{i,j,S,T} + \delta_{j,i,T,S}) \quad x_{i,S} \quad x_{i,T} \quad \delta_{i,j,S,T} + \delta_{j,i,T,S}, \ \forall (i,S),\ (j,T) \in I_S : g(i) = g(j), \\
& \sum_{i \in S} x_{i,S} = 1, \ \forall S \in \mathcal{S}, \quad x_{i,S} \quad 0, \quad 0 \quad \lambda_S \quad \nu_i \quad 1, \ \forall (i,S) \in I_S, \\
& \delta_{i,j,S,T} \in \{0,1\}, \qquad \forall (i,S),\ (j,T) \in I_S : g(i) = g(j).
\end{aligned}
\tag{24}
$$

Observe that the formulations in (21) and (24) are mixed integer programs involving $O(n^2|\mathcal{S}|^2)$ binary variables, $O(n|\mathcal{S}|)$ continuous variables, and $O(n^2|\mathcal{S}|^2)$ constraints.

Additional numerical experiments are provided in Section EC.8 to show that the model incorporating grouping yields narrower prediction intervals. We supplement these results in EC.8 with (i) a K-means clustering procedure to identify grouping based on the G-MDM representable conditions in Theorem 5 and (ii) a validation of the effectiveness of the grouping identification procedure.

7. Numerical Experiments

7.1. Experiment results with Synthetic Data

In Experiments 1 - 2 below, we compare the representational ability of MDM with RUM and MNL model. Experiment 3 compares the prediction performance offered by the nonparametric approach proposed in this paper with that offered by models involving parametric assumptions. Experiment 4 compares the limit of approximating choice probabilities with MDM, RUM, and MNL models. Additional useful details on the precise setup of all the experiments are furnished in EC.7.

The representation power and tractability of MDM compared to RUM and MNL. In Experiment 1, we investigate the representational power of MDM for a large number of alternatives ($n = 1000$) by randomly perturbing choice probabilities obtained from an underlying MNL model. We test for the fraction of instances that can be represented by MDM where the parameter α controls the fraction of choice probabilities that are perturbed from the MNL model (a larger value indicates more entries are modified from the underlying MNL model). While checking the representability of these models can be done by solving linear programs, RUM quickly becomes intractable as n increases. In Figure 2, we see that even with small perturbations to the choice probabilities of the underlying MNL model, none of the MNL models can represent the perturbed choice data. However, MDM which subsumes the MNL model can capture many of these instances. This shows that MDM is a much more robust model than MNL model. The runtimes for these large instances were less than 1 second. The computational requirements for RUM make it impossible to run at this scale.

In Experiment 2, we compare the representational power and computational time for MDM and RUM for a small number of alternatives. We find that both MDM and RUM show good representational power: In particular, with the collection size $|S| = 20$, round 80 percent of the instances can still be represented by MDM when 25 percent of the choice probability entries are perturbed; this drops to 60 percent when 100 percent of entries are perturbed, RUM has better representation power in these examples (see Figure 3). However, this comes at a significant run time cost even at this scale as seen in Figure 3 as compared to MDM.

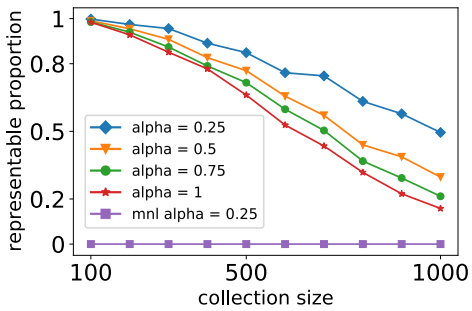
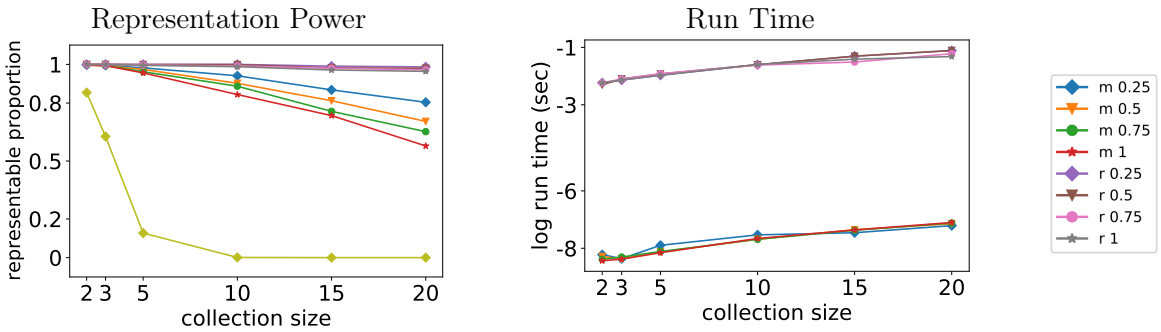


Figure 2 The representational power of MDM

Revenue and choice probability prediction with nonparametric MDM. In Experiment 3, we generate 20 random instances with a product size of 7 and a collection size $|S|$ ranging among $\{20, 40, 80\}$, using nonidentical exponential distributions for the marginal distributions to generate the underlying choice probabilities. In Figure 4, we compare the predictions offered by the following two methods: (1) computing the nonparametric MDM lower and upper bounds of revenue and choice probabilities for each instance by solving $\underline{r}(A)$ and $r(A)$; and (2) restricting the marginal



Notes. m stands for MDM, r stands for RUM, and the numbers stand for the perturbation parameters.

Figure 3 Comparison of the performance of MDM and RUM

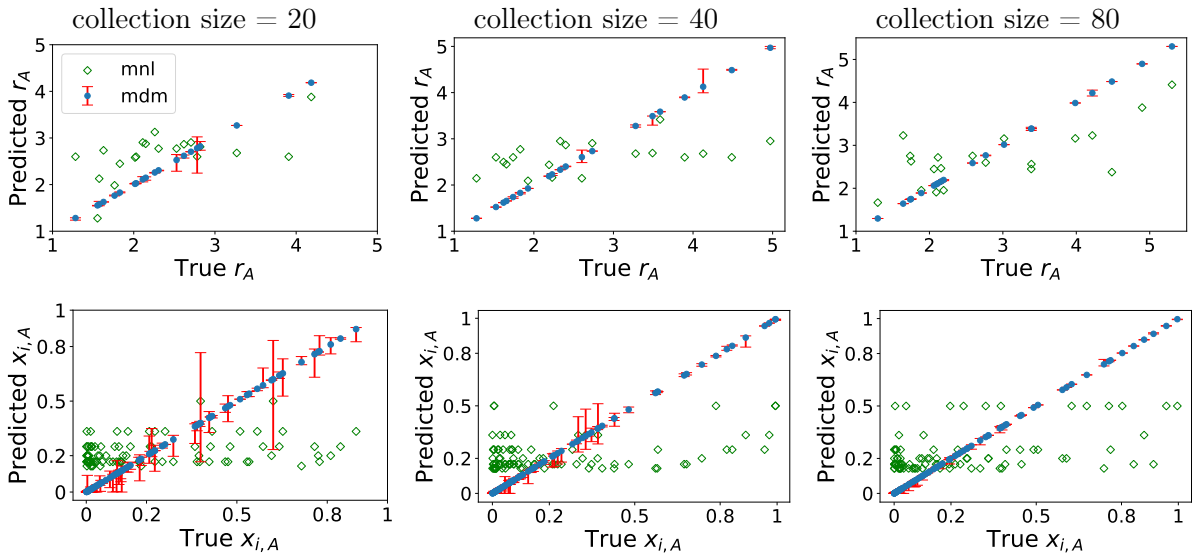
distributions for MDM to be identical exponential distributions (which leads to the underlying choice model being MNL), and estimating the preference parameters using maximum likelihood estimation (MLE); we then using the estimated MNL model to predict revenue for unseen assortments in each instance. While Figure 4 reveals the proposed nonparametric approach to be correctly predicting the true revenue or choice probabilities, the MLE of the parametric approach with mis-specified marginals is often found to lead to inaccurate predictions which are far from the truth and also far out of the nonparametric MDM prediction intervals. When more assortments are offered, the prediction under nonparametric MDM becomes more accurate while the prediction results made under the incorrect parametric model become worse. Thus, besides revealing the benefits of the proposed nonparametric data-driven approach for prediction based on MDM, Experiment 3 brings out the risks in stipulating apriori distributional assumptions on the model.

Estimation performance of MDM compared to RUM and MNL. In Experiment 4, we compare the explanatory ability of MDM, RUM and MNL models by examining the cumulative absolute deviation loss suffered in fitting them to uniformly generated choice data instances. Figure 5 reveals that nonparametric MDM and RUM models are competitive and have much higher explanatory ability than MNL with increasing collection sizes. In particular, MDM incurs about 44% lesser loss, on average, than the best-fitting MNL model for the largest collection size considered.

7.2. Experiment Results with Real-World Data

In Experiments 5-7 below, we use the dataset from JD.com (introduced in Shen et al. 2020) in order to evaluate the feasibility of representing it with an MDM, the efficacy of predictions obtained by the proposed nonparametric approach, and the explanatory ability captured by the limit formulations.

Data processing. To pre-process the data, we follow the same pre-processing steps as in Ahipasaoglu et al. (2020). The dataset includes millions of transaction records and over 3000 alternatives. Each transaction records the set of products viewed by a customer (by clicking these products



Notes. In each figure, the blue dots represent the true revenues or choice probabilities, while the red ranges represent the predicted revenue intervals or choice probability intervals with the nonparametric MDM, and the green squares represent the predicted revenues or choice probabilities using the MLE of the MNL model.

Figure 4 Comparison of prediction accuracy between MDM and MNL with randomly generated instances

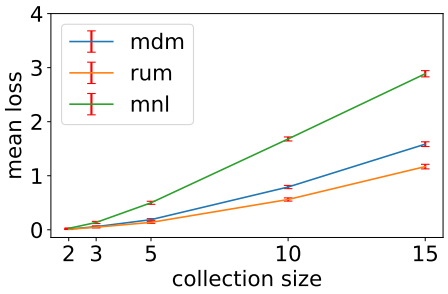


Figure 5 The limit loss comparison among MDM, RUM, and MNL

on the website) and the choice behavior of the customer (either making a purchase or leaving the system without buying anything). We assume that the set of products viewed by the customer is the offered assortment. We select the top 8 purchased products and combine the remaining products and the non-purchase option as the outside option for customers. We remove any transaction records that do not align with our model assumptions, such as multiple units of a product purchased in one record. As a result, the processed dataset contains 1784 customers and 8097 times of purchases in total. After preprocessing the data, we group the transaction records by product-assortment pairs and count the frequency of each pair. Dividing this frequency by the number of times the corresponding assortment is offered results in the empirical choice probabilities, denoted by $\boldsymbol{p}_S$. In the following discussion, we use O_S to denote the number of times an assortment S is offered.

Representational power comparison among several models. Experiment 5 compares the representation power of MDM with the MNL model and the class of regular choice models. The tested instances feature assortments which are offered at least O_S times, with O_S values ranging from 60 to 100. If we include data on the outside option, none of the models considered are found to exactly represent the data even when $O_S = 100$. By focusing on the sales data of the products offered by the firm, Table 1 shows that the nonparametric MDM and regular choice models are able to represent the choice data obtained from $O_S = 75$ and 100, whereas MNL models fail to represent any of the instances. We could not report the results for RUM here because of its intractability.

Table 1 The representability of MNL, MDM, and the class of regular choice models

O_S	# assortments offered at least O_S times	MNL	MDM	Regular Model
60	13	0	0	0
75	12	0	1	1
100	11	0	1	1

Notes. 1 denotes an instance that can be represented by the tested model, while 0 denotes the opposite.

Estimation performance comparison between MDM and MNL. In Experiment 6, we compare the explanatory ability of the MDM and MNL model by computing the limit loss over choice data obtained by considering assortments that are offered at least O_S times, where O_S is set to vary from 1 to 100. Using 1-norm as the loss function, the results in Table 2 show that nonparametric MDM suffers much lesser cost in approximating the choice data, and hence greater explanatory ability. We also observe that the run time of solving the limit of MDM grows when the size of the assortment collection becomes larger. We further assess the accuracy of the nonparametric MDM and

Table 2 Comparison of the estimation performance of MDM and MNL model

O_S	# assortments offered at least O_S times	MDM		MNL	
		log loss	run-time	log loss	run-time
1	134	-1.50	3600	-1.666	0.839
10	42	-3.603	3600	-1.803	0.253
20	29	-3.982	13.654	-1.893	0.193
30	24	-4.063	6.298	-1.918	0.318
40	19	-4.165	2.379	-1.945	0.149
50	15	-4.403	0.237	-2.002	0.146
60	13	-4.544	0.094	-2.028	0.114
75	12	-4.633	0.060	-2.045	0.120
100	11	-4.622	0.046	-2.039	0.112

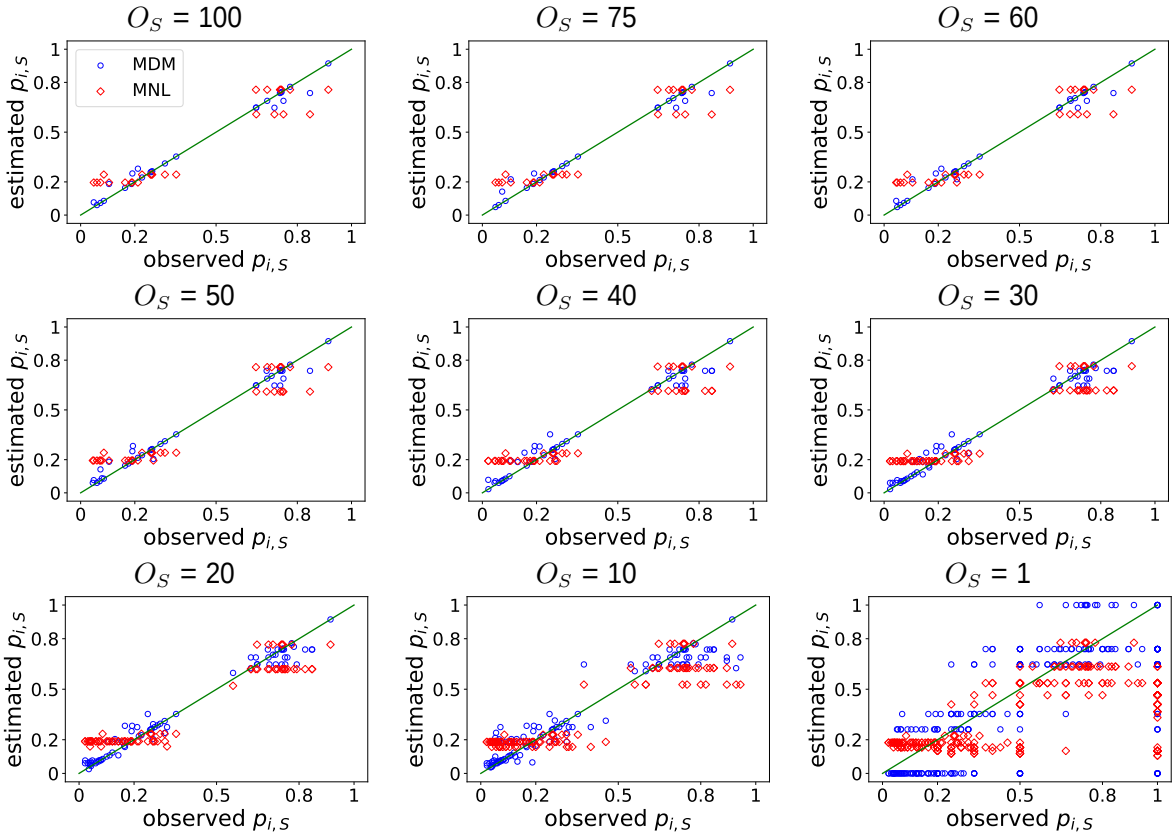
MNL models by comparing the observed (true) and estimated choice probabilities via scatter plots.

Figure 6 shows these scatter plots, where each point represents an observed and estimated choice probability pair. The horizontal axis shows the observed probability and the vertical axis shows the estimated probability. The closer the points are to the 45-degree line segment (green segment in Figure 6), the better the estimation accuracy. The scatter plots reveal the following findings:

(i) When $O_S = \{50, 60, 75, 100\}$, MDM is seen to correctly estimates most data points due to its proximity to the 45^o line while most points from MNL estimation are still away from the 45^o line.

(ii) When $O_S = \{10, 20, 30, 40\}$, although both MDM and MNL model are limited in their abilities to exactly represent the choice data, MDM shows much better estimation accuracy than MNL model with most points by MDM being much closer to the 45^o line than the MNL model.

(iii) In the noisy environment where many assortments are just shown once (corresponding to $O_S = 1$), both MDM and MNL fail understandably with most points falling away from the 45^o line.



Notes. In each plot, each point corresponds to the coordinate (true choice probability, estimated choice probability) of each observation of the processed data with O_S. The green line is the 45^o line. The blue dots represent the estimation results with the nonparametric MDM and the red squares represent the estimation results with the MNL model.

Figure 6 Scatter plots to compare the estimation accuracy of MDM and MNL

Prediction performance comparison between MDM and MNL. Experiment 7 evaluates the predictive-cum-prescriptive abilities of the nonparametric MDM and the MNL model by comparing their accuracies in identifying (i) a ranking over unseen test assortments based on their expected revenues, and (ii) the average revenue of the assortment identified to offer the largest revenue in the test set. Considering assortments that are shown at least O_S times (with O_S taken to vary from 20 to 50), we report the average out-of-sample performance over instances generated by randomly picking $\max\{2, \beta 0.2|S|\}$ fraction of the assortments to be the test set and the remaining to be the training set. For MNL, we use the Maximum Likelihood Estimator (MLE) obtained from training data to estimate choice probabilities for the test assortments and use them subsequently to rank the test assortments in a decreasing order of expected revenues. For MDM, we compute the robust revenue $\underline{r}(A)$ and the optimistic revenue $r(A)$ and record the corresponding choice probabilities for each tested assortment A in the test set if the training data can be represented by MDM. If we find the training data to be not exactly representable by MDM, we solve the limit of MDM (Problem (16)) with the training data and use the choice probabilities yielded by solving (17) to proceed as before with ranking the assortments in a decreasing order of expected revenues.

For comparing the quality of rankings offered by the MDM and the MNL model, we take the well-known Kendall Tau distance (see Definition EC.2) as a natural metric for evaluating the closeness of the predicted ranking with the ground truth hidden from training. For both models, we also compare the true revenues of the assortments which are predicted to rank at the top. The average of these out-of-sample performance metrics across randomly generated train-test splits are reported in Table 3. The results in Table 3 show that the optimistic prediction results of nonparametric MDM outperform the MNL model, yielding uniformly lower average Kendall tau distances and higher average revenues for the predicted best assortments across all scenarios. Similarly, the robust prediction results of nonparametric MDM outperform the MNL models in most scenarios, except for instances with $O_S = 30$ in terms of average Kendall tau distances and instances with $O_S = 50$ in terms of average revenue predictions. Figure 7 illustrates that the nonparametric MDM approach predicts the true revenue more accurately than MNL, as the predicted intervals include the true revenue or are closer to it.

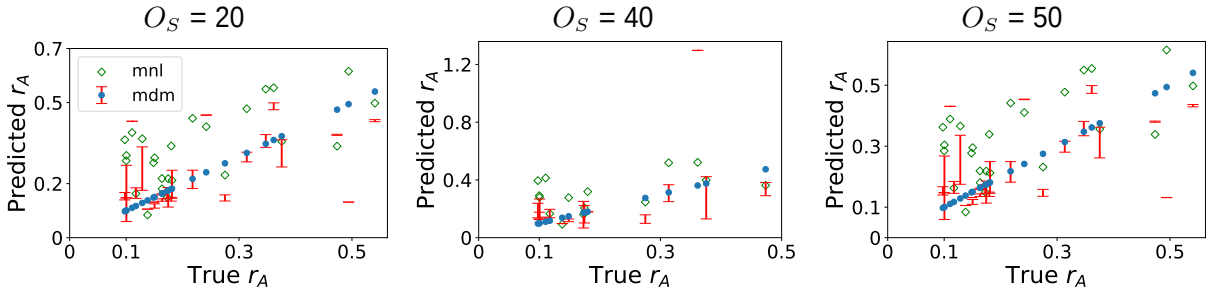
8. Conclusions

We identify the first-known sufficient and necessary conditions to verify whether a set of given choice data can be represented by MDM. Besides being verifiable in polynomial time, these representable conditions lead to a mixed integer linear program to predict the revenue or choice probability of alternatives for an unseen assortment and a mixed integer convex program that can give the closest fitting MDM choice probabilities to the given dataset. We extend the representable conditions,

Table 3 The prediction performance of MDM and MNL

O_S	resulting jSj	#test assortments	Average Kendall Tau Distance			Average Revenue of the Predicted Best Assortments		
			MNL	MDM_LB	MDM_UB	MNL	MDM_LB	MDM_UB
20	29	23	4.9	3.1	3.9	0.385	0.416	0.422
30	24	20	3.0	3.1	2.5	0.389	0.420	0.403
40	19	15	1.8	1.2	1.1	0.225	0.271	0.275
50	15	12	1.2	0.9	0.6	0.257	0.249	0.271

Notes. MDM_LB represents the results by solving $\underline{r}(A)$ while MDM_UB represents the results by solving $r(A)$.



Notes. In each figure, the blue dots represent the true revenues, while the red ranges represent the predicted revenue intervals with the nonparametric MDM, & the green squares represent the predicted revenues using the MNL model.

Figure 7 Revenue predictions vs. true revenue and for the nonparametric MDM and the MNL model

prediction, and limit formulations to the novel grouped-MDM, which allows the flexibility of grouping the alternatives based on distribution of the random utilities. Our numerical experiments demonstrate that checking the representability of MDM is computationally efficient compared to RUM, and MDM provides better representation power, estimation, and prediction performance than MNL.

Acknowledgments

The authors gratefully acknowledge support from the Ministry of Education Academic Research Fund of Singapore under grants MOE2019-T2-2-163 and T2MOE1906.

References

S. Ahipasaoglu, X. Li, Z. Sun, and Y. Yuan. A unified analysis for assortment planning with marginal distributions. *Available at SSRN: https://ssrn.com/abstract=3638783*, 2020.

S. D. Ahipasaoglu, U. Arikan, and K. Natarajan. Distributionally robust markovian tra□c equilibrium. *Transportation Science*, 53(6):1546–1562, 2019.

G. M. Allenby and J. L. Ginter. Using extremes to design products and segment markets. *Journal of Marketing Research*, 32(4):392–403, 1995.

A. Alptekinoğlu and J. H. Semple. The exponomial choice model: A new alternative for assortment and price optimization. *Operations Research*, 64(1):79–93, 2016.

A. Alptekinoğlu and J. H. Semple. Heteroscedastic exponomial choice. *Operations Research*, 69(3):841–858, 2021.

A. Aouad and A. Désir. Representing random utility choice models with neural networks. *arXiv preprint arXiv:2207.12877*, 2022.

S. Barberá and P. K. Pattanaik. Falmagne and the rationalizability of stochastic choices in terms of random orderings. *Econometrica: Journal of the Econometric Society*, pages 707–715, 1986.

M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, 1985.

G. Berbeglia. Discrete choice models based on random walks. *Operations Research Letters*, 44(2):234–237, 2016.

G. Berbeglia and G. Joret. Assortment optimisation under a general discrete choice model: A tight analysis of revenue-ordered assortments. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, page 345–346. ACM, 2017. ISBN 9781450345279.

C. R. Bhat. A heteroscedastic extreme value model of intercity travel mode choice. *Transportation Research Part B: Methodological*, 29(6):471–483, 1995.

J. Blanchet, G. Gallego, and V. Goyal. A markov chain approximation to choice modeling. *Operations Research*, 64(4):886–905, 2016.

H. D. Block and J. Marschak. Random orderings and stochastic theories of response. *In: Economic Information, Decision, and Prediction. Theory and Decision Library*, 7-1, 1960.

L. Chen, W. Ma, K. Natarajan, D. Simchi-Levi, and Z. Yan. Distributionally robust linear and discrete optimization with marginals. *Operations Research*, 70(3):1822–1834, 2022.

N. Chen, G. Gallego, and Z. Tang. The use of binary choice forests to model and estimate discrete choices. *arXiv preprint arXiv:1908.01109*, 2019.

Y.-C. Chen and V. V. Mišić. Decision forest: A nonparametric approach to modeling irrational choice. *Management Science*, 2022.

S. R. Cosslett. Maximum likelihood estimator for choice-based samples. *Econometrica: Journal of the Econometric Society*, pages 1289–1316, 1981.

C. Daganzo. *Multinomial probit: the theory and its application to demand forecasting*. Academic Press, New York, 1979.

E. W. de Bekker-Grob, J. Veldwijk, M. Jonker, B. Donkers, J. Huisman, S. Buis, J. Swait, E. Lancsar, C. L. Witteman, G. Bonsel, et al. The impact of vaccination and patient characteristics on influenza vaccination uptake of elderly people: a discrete choice experiment. *Vaccine*, 36(11):1467–1476, 2018.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

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

V. F. Farias, S. Jagabathula, and D. Shah. A nonparametric approach to modeling choice with limited data. *Management science*, 59(2):305–322, 2013.

M. Feldman, O. Svensson, and R. Zenklusen. Online contention resolution schemes with applications to bayesian selection problems. *SIAM Journal on Computing*, 50(2):255–300, 2021.

G. Feng, X. Li, and Z. Wang. Technical note—on the relation between several discrete choice models. *Operations Research*, 65(6):1429–1731, 2017.

S. Fiorini. A short proof of a theorem of falmagne. *Journal of Mathematical Psychology*, 48(1):80–82, 2004.

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

J. Hofbauer and W. H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70 (6):2265–2294, 2002.

S. Jagabathula and P. Rusmevichientong. The limit of rationality in choice modeling: Formulation, computation, and implications. *Management Science*, 65(5):2196–2215, 2019.

C. Liu, M. Liu, H. Sun, and C.-P. Teo. Product and ancillary pricing optimization: Market share analytics via perturbed utility model. *Available at SSRN 4095769*, 2022.

R. D. Luce. *Individual choice behavior*. John Wiley, 1959.

C. F. Manski and S. R. Lerman. The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, pages 1977–1988, 1977.

C. F. Manski and D. McFadden. Alternative estimators and sample designs for discrete choice analysis. *Structural analysis of discrete data with econometric applications*, 2:2–50, 1981.

J. Marschak. *Binary-Choice Constraints and Random Utility Indicators*, pages 218–239. Springer Netherlands, Dordrecht, 1960. ISBN 978-94-010-9276-0. doi: 10.1007/978-94-010-9276-0_9. URL https://doi.org/10.1007/978-94-010-9276-0_9.

A. Mas-Colell, M. D. Whinston, J. R. Green, et al. *Microeconomic theory*, volume 1. Oxford university press New York, 1995.

D. McFadden. Conditional logit analysis of qualitative choice behaviour. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press New York, New York, NY, USA, 1973.

D. McFadden. Modelling the choice of residential location. *Spatial Interaction Theory and Residential Location*, pages 75–96, 1978.

D. McFadden. Econometric models for probabilistic choice among products. *Journal of Business*, pages S13–S29, 1980.

D. McFadden. The choice theory approach to market research. *Marketing science*, 5(4):275–297, 1986.

D. McFadden and M. K. Richter. Stochastic rationality and revealed stochastic preference. preferences, uncertainty, and optimality, essays in honor of leo hurwicz, 1990.

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

D. L. McFadden. Revealed stochastic preference: a synthesis. In *Rationality and Equilibrium*, pages 1–20. Springer, 2006.

V. K. Mishra, K. Natarajan, D. Padmanabhan, C.-P. Teo, and X. Li. On theoretical and empirical aspects of marginal distribution choice models. *Management Science*, 60(6):1511–1531, 2014.

K. Natarajan, M. Song, and C.-P. Teo. Persistency model and its applications in choice modeling. *Management Science*, 55(3):453–469, 2009.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2346567.

A. Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

M. Shen, C. S. Tang, D. Wu, R. Yuan, and W. Zhou. Jd. com: Transaction-level data for the 2020 msom data driven research challenge. *Manufacturing & Service Operations Management*, 2020.

B. Sifringer, V. Lurkin, and A. Alahi. Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*, 140:236–261, 2020.

B. Sturt. The value of robust assortment optimization under ranking-based choice models. *arXiv preprint arXiv:2112.05010*, 2021.

K. Talluri and G. Van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004.

B. Taşkesen, S. Shafieezadeh-Abadeh, and D. Kuhn. Semi-discrete optimal transport: Hardness, regularization and numerical solution. *Mathematical Programming*, pages 1–74, 2022.

L. L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

S. Wang, Q. Wang, and J. Zhao. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. *Transportation Research Part C*, 118, 2020.

Z. Yan, K. Natarajan, C.-P. Teo, and C. Cheng. A representative consumer model in data-driven multi-product pricing optimization. *Management Science*, 68(8):5798–5827, 2022.

E-companion

The E-companion is organized as follows. Sections EC.1 to EC.4 present the proofs of the results, respectively, in Sections 3.2, 4, 5 and 6. Section EC.5 provides an algorithm to evaluate the limit of MDM in (15). Section EC.6 collects and presents all the illustrative examples mentioned in the paper. Section EC.7 details the implementation of the experiments. Section EC.8 and Section EC.9 provide additional experiment results and the implementation details of G-MDM.

EC.1. Proofs of the Results in Section 3.2

EC.1.1. Proof of Theorem 2

To prove Theorem 2, we first present a lemma that can construct special cases of choice probabilities obtained from the MNL model.

Lemma EC.1. *For any fixed n, let $S,T \subseteq N$ with $|S|,|T| \geq 2$. Let $i \in S \setminus T$. Then there exists a set of positive integers $x_1,\ldots,x_n$ such that*

$$\frac{x_i}{\sum_{k\in S} x_k} > \frac{x_i}{\sum_{k\in T} x_k}, \tag{EC.1}$$

as long as $S \not\subseteq T$.

Proof. We prove Lemma EC.1 holds for a set of positive integers such that $x_k = 2^k$ when $k \geq 1$. When $S \not\subseteq T$, to show $\frac{x_i}{\sum_{k\in S} x_k} > \frac{x_i}{\sum_{k\in T} x_k}$, it's equivalent to show $\sum_{k\in S} x_k > \sum_{k\in T} x_k$. Given that any subset can be viewed as a binary number, the intuitive understanding of its uniqueness becomes evident. However, for the sake of comprehensiveness and rigor, we shall furnish a formal proof as follows.

It's obvious that $\sum_{k\in S} x_k > \sum_{k\in T} x_k$ when $S \supset T$ or $T \supset S$. Next, we prove that $\sum_{k\in S} x_k > \sum_{k\in T} x_k$ when $S \not\supset T$ or $T \not\supset S$. Let $k_1 = \arg\max_{k\in S\setminus T} x_k$ and $k_2 = \arg\max_{k\in T\setminus S} x_k$. Without loss of generality, let $k_1 > k_2$. Then, we have $k_1 \geq k_2 \geq 1$. We have

$$\sum_{k\in S} x_k \geq \sum_{k\in T} x_k \geq x_{k_1} \geq \sum_{k\in T\setminus S} x_k \geq x_{k_1} \geq \sum_{i=1}^{k_2} x_i = 2^{k_1} \geq \frac{2\cdot(1-2^{k_2})}{1-2} = 2^{k_1} \geq 2^{k_2+1}+2 > 0.$$

The first inequity is due to $k_1 \in S\setminus T$, and the second inequity is due to $T\setminus S \subseteq \{1,\ldots,k_2\}$. The first equality is due to the formula of the sum of geometric series. This completes the proof.

Recall that the probability of choosing product i in assortment S under an MNL model is $p_{i,S} = \frac{e^{\nu_i}}{\sum_{j\in S} e^{\nu_j}}$. Then, there exists $\nu_i = \ln x_i$ for all $i \in N$, such that the instance in Lemma EC.1 is an MNL instance. Equipped with Lemma EC.1, we prove Theorem 2 as follows:

Proof. Let $\boldsymbol{p}_S$ be an instance following the manners in Lemma EC.1. Since MNL is a special case of MDM, $\boldsymbol{p}_S$ is MDM-representable. We next show that any instance $\boldsymbol{p}'_S$ that lies in the ball centered at $\boldsymbol{p}_S$ with the radius $\epsilon > 0$ is an MDM instance. Let $0 < \epsilon < \min\{p_{i,S} \quad p_{i,T}, \forall (i,S),(i,T) \in I_S$. We perturb $\boldsymbol{p}_S$ to be $\boldsymbol{p}'_S$ by letting $p'_{i,S} = p_{i,S} + \epsilon$ and $p'_{j,S} = p_{j,S} \quad \epsilon$ by arbitrarily choosing $(i,S),(j,S) \in I_S$, and keeping other entries of $\boldsymbol{p}'_S$ the same as $\boldsymbol{p}_S$. Since $\boldsymbol{p}_S$ is MDM-representable, we have

$$\lambda_S > \lambda_T \text{ if } p_{i,S} < p_{i,T} \ \forall (i,S),(i,T) \in I_S, \text{ and } \lambda_T > \lambda_S \text{ if } p_{j,T} < p_{j,S} \ \forall (j,S),(j,T) \in I_S.$$

Since $\epsilon < \min\{p_{i,S} \quad p_{i,T}, \forall (i,S),(i,T) \in I_S$, we have

$$\lambda_S > \lambda_T \text{ if } p_{i,S} + \epsilon < p_{i,T} \ \forall (i,S),(i,T) \in I_S, \text{ and } \lambda_T > \lambda_S \text{ if } p_{j,T} < p_{j,S} \quad \epsilon \ \forall (j,S),(j,T) \in I_S.$$

By the construction of $\boldsymbol{p}'_S$, equivalently, we have

$$\lambda_S > \lambda_T \text{ if } p'_{i,S} < p'_{i,T} \ \forall (i,S),(i,T) \in I_S, \text{ and } \lambda_T > \lambda_S \text{ if } p'_{j,T} < p'_{j,S} \ \forall (j,S),(j,T) \in I_S.$$

Thus, $\boldsymbol{p}'_S$ is MDM-representable.

EC.1.2. Proof of Lemma 2

Proof. We first prove a) in Lemma 2 by contradiction. From Theorem 1, if there exists some alternative i such that $p_{i,S} > p_{i,S \setminus T}$, then we have $\lambda_S < \lambda_{S \setminus T}$, which is equivalent to $\lambda_S < \lambda_{S \setminus T}$ for all $j \in S \setminus T$. This implies $\lambda_S \quad \lambda_{S \setminus T}$ which gives $p_{j,S} \quad p_{j,S \setminus T}$ for all $(j,S),(j,S \setminus T) \in I_S$. Since $\sum_{j \in S} p_{j,S} = 1$, we get $\sum_{j \in S \setminus T} p_{j,S \setminus T} < 1$ contradicting the condition $\sum_{j \in S \setminus T} p_{j,S \setminus T} = 1$. For b), from Theorem 1, if $i, j \in S \setminus T$, we have $p_{j,S} < p_{j,T} \Rightarrow \lambda_S > \lambda_T \Rightarrow \lambda_S \quad \lambda_T \Rightarrow p_{i,S} \quad p_{i,T}$.

EC.1.3. Proof of Theorem 3

Proof. We prove *a*) of Theorem 3 first. We use the following notations for the rank list model since any RUM can be described by a rank list model (see, e.g., Block and Marschak 1960). Let $\quad_n$ denote the set of all permutations of n alternatives. Each element $\sigma \in \quad_n$ denotes a ranking of n alternatives. For instance, $\sigma = \{1 \quad 2 \quad 3\}$ means alternative 1 is more preferred than alternative 2 which is more preferred than alternative 3. The probability of each ranking is $P(\sigma)$ and $\sum_{\sigma \in \quad_n} P(\sigma) = 1$. We prove the result case by case.

1. $n = 2$: Here $P_{\text{mdm}}(S) = P_{\text{rum}}(S)$. This is straightforward since all probabilities that satisfy $0 \quad p_{1,\{1,2\}} \quad p_{1,\{1\}} = 1$, and $0 \quad p_{2,\{1,2\}} \quad p_{2,\{2\}} = 1$ where $p_{1,\{1,2\}} + p_{2,\{1,2\}} = 1$, are representable by both models.
2. $n = 3$: Lemma 2 implies that $P_{\text{mdm}}(S) \quad P_{\text{reg}}(S)$. $P_{\text{rum}}(S) = P_{\text{reg}}(S)$ since

$$P(\{1 \quad 2 \quad 3\}) = p_{2,\{2,3\}} \quad p_{2,\{1,2,3\}} \quad 0 \text{ and } P(\{1 \quad 3 \quad 2\}) = p_{3,\{2,3\}} \quad p_{3,\{1,2,3\}} \quad 0,$$
$$P(\{2 \quad 1 \quad 3\}) = p_{1,\{1,3\}} \quad p_{1,\{1,2,3\}} \quad 0 \text{ and } P(\{2 \quad 3 \quad 1\}) = p_{3,\{1,3\}} \quad p_{3,\{1,2,3\}} \quad 0,$$
$$P(\{3 \quad 1 \quad 2\}) = p_{1,\{1,2\}} \quad p_{1,\{1,2,3\}} \quad 0 \text{ and } P(\{3 \quad 2 \quad 1\}) = p_{2,\{1,2\}} \quad p_{2,\{1,2,3\}} \quad 0,$$

where $\sum_{\sigma 2 \quad n} P(\sigma) = 3 \quad 2 = 1$. We next show that $P_{\mathsf{mdm}}(S) \quad P_{\mathsf{rum}}(S)$ for $n = 3$ by giving an example of choice probabilities with $S = f f1, 2, 3g, f1, 2g, f1, 3g, f2, 3gg$ in Table EC.1 that can be represented by RUM but not by MDM.

Table EC.1 Choice probabilities that cannot be represented by MDM for $n = 3$.

Alternative	A={1,2,3}	B={1,2}	C={1,3}	D={2,3}
1	1/3	5/9	4/9	-
2	1/3	4/9	-	5/9
3	1/3	-	5/9	4/9

This collection of choice probabilities $\boldsymbol{p}_S$ cannot be represented by MDM because $p_{1,B} > p_{1,C}$, $p_{2,D} > p_{2,B}$, $p_{3,C} > p_{3,D}$ implies $\lambda_B < \lambda_C$, $\lambda_D < \lambda_B$ and $\lambda_C < \lambda_D$. This gives $\lambda_D < \lambda_B < \lambda_C < \lambda_D$ which is inconsistent. So, $\boldsymbol{p}_S$ in Table EC.1 cannot be represented by MDM. On the other hand, it is straightforward to check that by setting the ranking probabilities for RUM as follows: $P(f1 \quad 2 \quad 3g) = 2/9$, $P(f1 \quad 3 \quad 2g) = 1/9$, $P(f2 \quad 1 \quad 3g) = 1/9$, $P(f2 \quad 3 \quad 1g) = 2/9$, $P(f3 \quad 1 \quad 2g) = 2/9$, $P(f3 \quad 2 \quad 1g) = 1/9$, we obtain the choice probabilities in Table EC.1. This implies $\boldsymbol{p}_S$ in table EC.1 can be represented by RUM but not MDM.

3. $n \quad 4$: We show $P_{\mathsf{mdm}}(S)\ 6\ P_{\mathsf{rum}}(S)$ and $P_{\mathsf{rum}}(S)\ 6\ P_{\mathsf{mdm}}(S)$ by providing two examples: (1) $\boldsymbol{p}_S$ can be represented by RUM but not MDM and (2) $\boldsymbol{p}_S$ can be represented by MDM but not RUM when $S = f f1, 2, 3, 4g, f1, 2, 3g, f1, 2, 4g, f1, 2gg$. The examples are provided for $n = 4$. For larger n, we can add the alternatives in the assortments and set the choice probabilities for these added alternatives to be zero. Firstly, we observe that the multinomial logit choice probabilities can be obtained from both RUM and MDM. This follows from using independent and identically distributed Gumbel distributions for RUM (see, e.g., Ben-Akiva and Lerman 1985) and exponential distributions for MDM (see, e.g., Mishra et al. 2014). Hence the intersection between the two sets is nonempty for any n. Next consider the choice probabilities in Table EC.2. This can be

Table EC.2 Choice probabilities can be represented by MDM but not by RUM for $n = 4$.

Alternative	A={1,2,3,4}	B={1,2,3}	C={1,2,4}	D={1,2}
1	3/20	7/20	2/8	1/2
2	3/20	2/8	7/20	1/2
3	7/20	2/5	-	-
4	7/20	-	2/5	-

recreated by RUM using the distribution over the ranking as follows:

$P(f1\ 2\ 3\ 4g) = 1/40$ $P(f1\ 2\ 4\ 3g) = 1/40$ $P(f1\ 3\ 2\ 4g) = 1/40$
$P(f1\ 3\ 4\ 2g) = 1/40$ $P(f1\ 4\ 2\ 3g) = 1/40$ $P(f1\ 4\ 3\ 2g) = 1/40$
$P(f2\ 1\ 3\ 4g) = 1/40$ $P(f2\ 1\ 4\ 3g) = 1/40$ $P(f2\ 3\ 1\ 4g) = 1/40$
$P(f2\ 3\ 4\ 1g) = 1/40$ $P(f2\ 4\ 1\ 3g) = 1/40$ $P(f2\ 4\ 3\ 1g) = 1/40$
$P(f3\ 1\ 2\ 4g) = 1/20$ $P(f3\ 1\ 4\ 2g) = 1/20$ $P(f3\ 2\ 1\ 4g) = 1/10$
$P(f3\ 2\ 4\ 1g) = 1/10$ $P(f3\ 4\ 1\ 2g) = 1/40$ $P(f3\ 4\ 2\ 1g) = 1/40$
$P(f4\ 1\ 3\ 2g) = 1/10$ $P(f4\ 1\ 3\ 2g) = 1/10$ $P(f4\ 2\ 1\ 3g) = 1/20$
$P(f4\ 2\ 3\ 1g) = 1/20$ $P(f4\ 3\ 1\ 2g) = 1/40$ $P(f4\ 3\ 2\ 1g) = 1/40$

Now $p_{1,B} > p_{1,C}$ implies $\lambda_B < \lambda_C$ and $p_{2,B} > p_{2,C}$ implies $\lambda_B > \lambda_C$. Hence $\boldsymbol{p}_S$ in Table EC.2 is not representable by MDM. Next consider the choice probabilities in Table EC.3. Here $p_{1,A} < p_{1,B} =$

Table EC.3 Choice probabilities can be represented by MDM but not by RUM for $n = 4$.

Alternative	A={1,2,3,4}	B={1,2,3}	C={1,2,4}	D={1,2}
1	0.1	0.2	0.2	0.25
2	0.2	0.25	0.25	0.75
3	0.2	0.55	-	-
4	0.5	-	0.55	-

$p_{1,C} < p_{1,D}$ implies $\lambda_A > \lambda_B = \lambda_C > \lambda_D$, and $p_{2,A} < p_{2,B} = p_{2,C} < p_{2,D}$ implies $\lambda_A > \lambda_B = \lambda_C > \lambda_D$, and $p_{3,A} < p_{3,B}$ implies $\lambda_A > \lambda_B$, and $p_{4,A} < p_{4,C}$ implies $\lambda_A > \lambda_C$. So we have $\lambda_A > \lambda_B = \lambda_C > \lambda_D$ which is easy to enforce and so $\boldsymbol{p}_S$ can be represented by MDM. A necessary condition for $\boldsymbol{p}_S$ to be representable by RUM are the Block-Marshak conditions provided in Block and Marschak (1960) (also see Theorem 1 in Fiorini 2004). If the choice probabilities are representable by RUM, one of these conditions is given by $p_{1,A} + p_{1,D}$ $p_{1,B} + p_{1,C}$. Here $p_{1,A} + p_{1,D} = 0.1 + 0.25 = 0.35 <$ $0.4 = 0.2 + 0.2 = p_{1,B} + p_{1,C}$. So, $\boldsymbol{p}_S$ is not representable by RUM.

We then prove b) of Theorem 3. We know that $P_{\mathsf{mdm}}(S) = P_{\mathsf{rum}}(S) = P_{\mathsf{reg}}(S)$ when $n = 2$ and $P_{\mathsf{rum}}(S)$ $P_{\mathsf{reg}}(S)$ and closure$\big(P_{\mathsf{mdm}}(S)\big)$ $P_{\mathsf{reg}}(S)$ for any S. To show b), we just need to show $P_{\mathsf{reg}}(S)$ $P_{\mathsf{rum}}(S)$ and $P_{\mathsf{reg}}(S)$ closure$\big(P_{\mathsf{mdm}}(S)\big)$ when S is nested or laminar. Under a nested collection $S = \{S_1, S_2, \ldots, S_m\}$ with S_1 S_2 $\ldots$ S_m, we have

$$P_{\mathsf{reg}}(S) = \left\{\boldsymbol{x} \in \mathbb{R}^{I_S} : x_{i,S} \quad 0, \forall (i,S) \in I_S, \sum_i x_{i,S} = 1, \forall S \in S, x_{i,S_k} \quad x_{i,S_j}\ \forall j,k \in [m], j < k, i \in S_i\right\}.$$

Under a laminar collection, we have

$$P_{\mathsf{reg}}(S) = \left\{\boldsymbol{x} \in \mathbb{R}^{I_S} : x_{i,S} \quad 0, \forall (i,S) \in I_S, \sum_i x_{i,S} = 1, \forall S \in S, x_{i,S} \quad x_{i,T}\ \forall T \quad S, (i,S), (i,T) \in I_S\right\}.$$

Next, we show $P_{\mathsf{reg}}(S)$ closure$\big(P_{\mathsf{mdm}}(S)\big)$ with a nested or laminar S by showing that for any $\boldsymbol{p}_S \in P_{\mathsf{reg}}(S)$, we can construct $\boldsymbol{\lambda}_S$ such that $(\boldsymbol{p}_S, \boldsymbol{\lambda}_S) \in$, where

$$:= \Big\{(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathsf{R}^{I_S} \quad \mathsf{R}^S : x_{i,S} \quad 0, \forall (i,S) \in I_S, \sum_i x_{i,S} = 1, \forall S \in S,$$
$$\lambda_S \quad \lambda_T \text{ if } x_{i,S} \quad x_{i,T}, \forall (i,S), (i,T) \in I_S\Big\}.$$

Suppose that $S = \{S_1, S_2, \quad, S_m\}$ is a nested collection with $S_1 \quad S_2 \quad \ldots \quad S_m$. Then we take any $\boldsymbol{\lambda}_S$ satisfying $\lambda_{S_1} \quad \lambda_{S_2} \quad \lambda_{S_m}$. Since $p_{i,S_j} \quad p_{i,S_k}$ for any $j \quad k$ (due to $\boldsymbol{p}_S \in P_{\mathsf{reg}}(S)$), the resulting $(\boldsymbol{p}_S, \boldsymbol{\lambda}_S) \in$. As a result, $\boldsymbol{p}_S \in$ closure$\big(P_{\mathsf{mdm}}(S)\big)$.

If S is laminar, we take any $\boldsymbol{\lambda}_S$ such that $\lambda_S \quad \lambda_T$ if $S, T \in S$ with $S \quad T$. Since $p_{i,S} \quad p_{i,T}$ due to the regularity $\boldsymbol{p}_S \in P_{\mathsf{reg}}(S)$, the resulting $(\boldsymbol{p}_S, \boldsymbol{\lambda}_S) \in$. Hence $\boldsymbol{p}_S \in$ closure$\big(P_{\mathsf{mdm}}(S)\big)$ and $P_{\mathsf{reg}}(S) =$ closure$\big(P_{\mathsf{mdm}}(S)\big)$ with a nested or laminar S.

Given a nested or laminar collection S, to show $P_{\mathsf{reg}}(S) \quad P_{\mathsf{rum}}(S)$, we next show that, for any $\boldsymbol{p}_S \in P_{\mathsf{reg}}(S)$, there exists a probability distribution $(P(\sigma) : \sigma \in \quad _n)$ such that $\boldsymbol{p}_S \in P_{\mathsf{rum}}(S)$.

(1) For a nested collection S, we prove $P_{\mathsf{reg}}(S) \quad P_{\mathsf{rum}}(S)$. Without loss of generality, let $S = \{S_1, S_2, \ldots, S_m\}$ be $S_k = \{1, \ldots, k\}$ for $k = 1, \quad, m$. Next, for any $\boldsymbol{p}_S \in P_{\mathsf{reg}}(S)$, we prove the existence of a probability distribution $(P(\sigma) : \sigma \in \quad _m)$ such that $\boldsymbol{p}_S \in P_{\mathsf{rum}}(S)$ from the point of view of polyhedral combinations. To show the existence of a probability distribution $(P(\sigma) : \sigma \in \quad _m)$ for $\boldsymbol{p}_S \in P_{\mathsf{rum}}(S)$ is equivalent to showing $\boldsymbol{p}_S$ lies in the multiple choice polytope characterized as,

$$\text{convex hull of } \{(\mathsf{I}[\sigma, i, S] : i \in S, S \in S) \in \{0,1\}^{\sum_{S \in S} j_S j} : \sigma \in \quad _m\},$$

where $\mathsf{I}[\sigma, i, S] = 1$ if and only if $i = \arg\min_{j \in S} \sigma(j)$ (see Section 3 in Fiorini 2004 and Lemma 2.5 of Jagabathula and Rusmevichientong 2019).

Now, we show that $\boldsymbol{p}_S$ lies in the multiple choice polytope via a graph representation of the multiple choice polytope following the steps in Section 3 in Fiorini (2004). Let $\boldsymbol{D} = (N_0, A)$ be a simple, acyclic directed graph, and let $m+1$ be the source node and 0 be the sink node of $\boldsymbol{D}$, where $N_0 = \{1, \quad, m\} \cup \{m+1, 0\}$. We encode each $m+1 \quad 0$ directed path of its arc set A in $\boldsymbol{D}$ by means of the indicator characteristic vector in the set $\{(\mathsf{I}[\sigma, i, S] : i \in S, S \in S) \in \{0,1\}^{\sum_{S \in S} j_S j} : \sigma \in \quad _m\} \subseteq \mathsf{R}^A$, which we denote r . The convex hull of the vectors r , for a $m+1 \quad 0$ directed path in $\boldsymbol{D}$, is referred to as the $m+1 \quad 0$ directed path polytope of $\boldsymbol{D}$. For a node v of $\boldsymbol{D}$, let $\delta^-(v) = \{(w,v) : w \in N_0, (w,v) \in A\}$ represent the nodes incoming to node v and $\delta^+(v) = \{(v,w) : w \in N_0, (v,w) \in A\}$ represent the nodes outgoing from node v. For $B \quad A$, let $r(B) = \sum\{r(v,w) : (w,v) \in B\}$. Let M be the matrix whose rows are indexed by modes of $\boldsymbol{D}$ such that the entry corresponding to node v and arc a equals to 1 if a enters v, and 1 if a

leaves 0, and 0 else. It's well known that M is totally unimodular (Schrijver 1998). This implies that the polyhedron $\{r \in \mathbb{R}^A : Mr = d, r \geq 0\}$ has all its vertices integer for every integral vector $d \in \mathbb{R}^A$. Assume that $\delta(m+1) = \delta(0)^+ = \emptyset$.

Lemma EC.2 (Theorem 2 in Fiorini 2004). *A point $r \in \mathbb{R}^A$ belongs to the $m+1$ – 0 directed path polytope $\boldsymbol{D}$ if and only if*

$$r(\delta^-(v)) - r(\delta^+(v)) = 0, \quad \forall v \in N_0 \setminus \{m+1, 0\}, \tag{EC.2}$$

$$r(\delta^-(0)) = 1, \tag{EC.3}$$

$$r(v,w) \geq 0, \quad \forall (w,v) \in A. \tag{EC.4}$$

In network flows, (EC.2)-(EC.4) *defines a flow of value 1 in the network $\boldsymbol{D} = (N_0, A)$, with source node $m+1$ and sink node* 0.

By Lemma EC.2, to show $\boldsymbol{p}_S$ lies in the multiple choice polytope, we need to show $(r(w,v) : (w,v) \in A)$ based on $\boldsymbol{p}_S$ satisfying (EC.2) – (EC.4) in Lemma EC.2. We demonstrate $(r(w,v) : (w,v) \in A)$ under $\boldsymbol{p}_S$ as follows:

$$\begin{aligned}
r(m+1, j) &= \sum_{\sigma \in \Pi_m : \arg\min_{i \in S_m} \sigma(i) = j} P(\sigma) = p_{j,S_m}, \quad \forall j = 1, \ldots, m, \\
r(i,j) &= \sum_{\sigma \in \Pi_m : \arg\min_{k \in S_{i-1}} \sigma(k) = j} P(\sigma) - \sum_{\sigma \in \Pi_m : \arg\min_{k \in S_i} \sigma(k) = j} P(\sigma) \\
&= p_{j,S_{i-1}} - p_{j,S_i}, \quad \forall i = 2, \ldots, m, \ 1 \leq j \leq i-1, \\
r(1,0) &= 1.
\end{aligned}$$

We provide Figure EC.1 to illustrate $(r(w,v) : (w,v) \in A)$ in the graph $\boldsymbol{D}$.

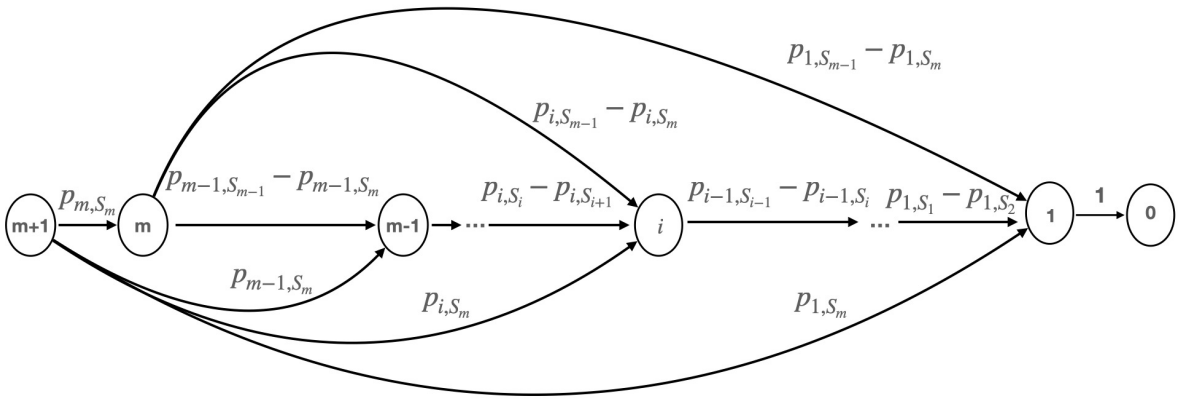


Figure EC.1 An illustration of $\boldsymbol{D}$ given $\boldsymbol{p}_S \in P_{\text{reg}}(S)$ with a nested collection S

Next, we verify such $(r(w,v) : (w,v) \in A)$ satisfies (EC.2)-(EC.4).

For (EC.2), define $p_{j,S_{m+1}} = 0$ for all j. For a node $j \in \{2, \ldots, m\}$,

$$\begin{aligned} r(\delta^+(j)) &= \sum_{k=j+1}^{m+1} r(k,j) \\ &= \sum_{k=j}^{m} p_{j,S_k} - p_{j,S_{k+1}} = p_{j,S_j} - p_{j,S_{j+1}} + p_{j,S_{j+1}} - p_{j,S_{j+2}} + \cdots + p_{j,S_m} - p_{j,S_{m+1}} \\ &= p_{j,S_j} - p_{j,S_{m+1}} = p_{j,S_j}. \\ r(\delta^+(j)) &= \sum_{k=1}^{j-1} p_{k,S_{j-1}} - p_{k,S_j} \\ &= p_{1,S_{j-1}} - p_{1,S_j} + p_{2,S_{j-1}} - p_{2,S_j} + \cdots + p_{j-1,S_{j-1}} - p_{j-1,S_j} \\ &= \sum_{k=1}^{j-1} p_{k,S_{j-1}} - \sum_{k=1}^{j-1} p_{k,S_j} = 1 - \sum_{k=1}^{j-1} p_{k,S_j} = p_{j,S_j}. \end{aligned}$$

For the node 1,

$$\begin{aligned} r(\delta^+(1)) &= \sum_{k=2}^{m+1} r(k,1) \\ &= p_{1,S_1} - p_{1,S_2} + p_{1,S_2} - p_{1,S_3} + \cdots + p_{1,S_m} - p_{1,S_{m+1}} \\ &= p_{1,S_1} - p_{1,S_{m+1}} = 1 - 0 = 1 = r(\delta^+(1)). \end{aligned}$$

Therefore, $r(\delta^+(j)) = r(\delta^+(j))$ for $j \in \{1,2,\ldots,m\}$. (EC.2) is satisfied by $(r(w,v) : (w,v) \in A)$.

For (EC.3), $r(\delta^+(0)) = r(1,0) = 1$.

For (EC.4), $r(m+1,j) = p_{j,S_m} \geq 0$, $\forall j = 1, \ldots, m$ because of the nonnegativity of choice probabilities. We have $r(i,j) = p_{j,S_{i-1}} - p_{j,S_i} \geq 0$, $\forall i = 2, \ldots, m$, $1 \leq j \leq i-1$ since p_S satisfies regularity,i.e., $\boldsymbol{p}_S \in P_{\text{reg}}(S)$. Further $r(1,0) = 1 > 0$.

Thus the assignment $(r(w,v) : (w,v) \in A)$ satisfies (EC.2)-(EC.4) in Lemma EC.2. Therefore, for any $\boldsymbol{p}_S \in P_{\text{reg}}(S)$, there exists a probability distribution $(P(\sigma) : \sigma \in \Sigma_m)$ such that $\boldsymbol{p}_S \in P_{\text{rum}}(S)$. This implies $P_{\text{reg}}(S) \subseteq P_{\text{rum}}(S)$ under the nested collection S.

(2) We prove $P_{\text{reg}}(S) \subseteq P_{\text{rum}}(S)$ under the laminar collection. By the definition of the laminar collection, we know that for $S, T \in S$, either $S \subseteq T$, or $T \subseteq S$, or $S \setminus T = \emptyset$. Then, it suffices to construct a distribution $P(\cdot)$ for $\hat{S} \supseteq S$ such that $\hat{S}$ is a nested collection. Following the proof in (1), we have $P_{\text{reg}}(S) = P_{\text{rum}}(S)$ under a laminar collection.

EC.2. Proofs of the Results in Section 4

EC.2.1. Proof of Proposition 1

Proof. Due to Theorem 1, we have $P_{\text{mdm}}(S^0) = \text{Proj}_x(\Lambda_{S^0})$, where the lifted set Λ_{S^0} equals

$$\Big\{(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathbb{R}^{I_{S^0}} \times \mathbb{R}^{S^0} : x_{i,S} \geq 0,\ \sum_i x_{i,S} = 1, \forall S \in S^0,$$
$$\lambda_S > \lambda_T \text{ if } x_{i,S} < x_{i,T},\ \lambda_S = \lambda_T \text{ if } x_{i,S} = x_{i,T} > 0,\ \forall (i,S), (i,T) \in I_{S^0}\Big\},$$

following the definition in (8). Since $U_A := \{\boldsymbol{x}_A : (\boldsymbol{x}_S, \boldsymbol{x}_A, \boldsymbol{\lambda}) \in \quad^{S^0},\ \boldsymbol{x}_S = \boldsymbol{p}_S\}$, the non-numbered constraints in the formulation in Proposition 1 are obtained by replacing $\boldsymbol{x}_S = \boldsymbol{p}_S$ in the above description of the lifted set $\quad^{S^0}$. For deducing the remaining constraints (10a) - (10b), we proceed as follows: Consider any $(i,S) \in I_S$. From the description of $\quad^{S^0}$, observe that an assignment for $x_{i,A}, \lambda_A, \lambda_S$ in any $(\boldsymbol{x}, \boldsymbol{\lambda}) \in \quad^{0}_{S}$ satisfying $\boldsymbol{x}_S = \boldsymbol{p}_S$ necessarily satisfies one of the following four cases:

In *Case 1,* we have $\lambda_A > \lambda_S$ and $x_{i,A} < p_{i,S}$: If λ_A, λ_S is such that $\lambda_A > \lambda_S$, this informs the restriction $\{x_{i,A} : x_{i,A} < p_{i,S}\}$ on the values $x_{i,A}$ can take. The closure of this restricted collection $\{x_{i,A} : x_{i,A} < p_{i,S}\}$ equals $\{x_{i,A} : x_{i,A} \quad p_{i,S}\}$.

In *Case 2,* we have $\lambda_A < \lambda_S$ and $x_{i,A} > p_{i,S}$: If λ_A, λ_S is such that $\lambda_A < \lambda_S$, the closure of the corresponding restriction $\{x_{i,A} : x_{i,A} > p_{i,S}\}$ equals $\{x_{i,A} : x_{i,A} \quad p_{i,S}\}$.

In *Case 3,* we have $\lambda_A = \lambda_S$ and $x_{i,A} = p_{i,S} \neq 0$: When λ_A, λ_S is such that $\lambda_A = \lambda_S$ and $p_{i,S} \neq 0$, the corresponding restriction on the values of $x_{i,A}$ is given by the closed set $\{x_{i,A} : x_{i,A} = p_{i,S}\}$.

Finally, in *Case 4,* we have λ_A, λ_S unconstrained and $x_{i,A} = p_{i,S} = 0$: Like in Case 3, the restriction on the values of $x_{i,A}$ corresponding to this case equals $\{x_{i,A} : x_{i,A} = 0\}$. The relationship between $x_{i,A}, p_{i,S}, \lambda_A, \lambda_S$ in this case is any one of the following sub-cases: Case (4a) $\lambda_A > \lambda_S$ and $0 = x_{i,A}$ $p_{i,S} = 0$, or Case (4b) $\lambda_A < \lambda_S$ and $0 = x_{i,A} \quad p_{i,S} = 0$, or Case (4c) $\lambda_A = \lambda_S$ and $0 = x_{i,A} = p_{i,S} = 0$.

Combining the observations in the cases (1) & (4a), (2) & (4b), and (3) & (4c), we obtain that the closure of U_A equals the collection of probability vectors $\boldsymbol{x}_A = (x_{i,A} : i \in A)$ for which there exists a function $\lambda : S^0 \to \mathbb{R}$ such that

$$\begin{aligned}
x_{i,A} \quad p_{i,S} \quad &\text{if } \lambda_A > \lambda_S, \quad \forall i \in A,\ (i,S) \in I_S, \\
x_{i,A} \quad p_{i,S} \quad &\text{if } \lambda_A < \lambda_S, \quad \forall i \in A,\ (i,S) \in I_S, \text{ and} \\
x_{i,A} = p_{i,S} \quad &\text{if } \lambda_A = \lambda_S, \quad \forall i \in A,\ (i,S) \in I_S,
\end{aligned}$$

in addition to satisfying $\lambda_S > \lambda_T$ if $p_{i,S} < p_{i,T}$ and $\lambda_S = \lambda_T$ if $p_{i,S} = p_{i,T} \neq 0$, for all $(i,S), (i,T) \in I_S$. The constraints in the formulation in Proposition 1 exactly specify these conditions describing the closure of U_A. Observe that the objective in $\inf\{\sum_{i \in A} r_i x_{i,A} : \boldsymbol{x}_A \in U_A\}$ is continuous as a function of $\boldsymbol{x}_A$. Therefore, $\inf\{\sum_{i \in A} r_i x_{i,A} : \boldsymbol{x}_A \in U_A\} = \min\{\sum_{i \in A} r_i x_{i,A} : \boldsymbol{x}_A \in \text{closure}(U_A)\}$.

EC.2.2. Proof of Proposition 2

Proof. Recall the notation $S^0 = S \cup \{A\}$. Observe that the variables $(\lambda_S : S \in S^0)$ influence the value of the formulation in Proposition 1 only via the sign of $\lambda_S \quad \lambda_T$, for any pair of variables λ_S, λ_T from the collection $(\lambda_S : S \in S^0)$. Therefore the optimal value of this optimization formulation is not affected by the presence of the following additional constraints: $0 \quad \lambda_S \quad 1$ for all $S \in S^0$, and

$$\lambda_S \quad \lambda_T \quad \epsilon \quad \text{if} \quad \lambda_S > \lambda_T, \qquad \forall (i,S), (i,T) \in I_{S^0},$$

for some suitably small value of $\epsilon > 0$. Indeed, this is because the signs of the differences $\{\lambda_S - \lambda_T : S,T \in \mathcal{S}^0\}$ are not affected by these additional constraints. Taking ϵ to be smaller than $1/(2|\mathcal{S}|)$, for example, ensures that there is a feasible assignment for $(\lambda_S : S \in \mathcal{S}^0)$ within the interval $[0,1]$ even if all these variables take distinct values.

Let F denote the feasible values for the variables $(\lambda_S : S \in \mathcal{S}^0), (x_{i,A} : i \in A)$ satisfying the constraints introduced in the above paragraph besides those in the formulation in Proposition 1. Equipped with this feasible region F, we have the following deductions from (11a) - (11c) for $(\lambda_S : S \in \mathcal{S}^0), (x_{i,A} : i \in A)$ in F: For every $i \in A$ and any $S \in \mathcal{S}$ containing i,

(i) we have $\lambda_A < \lambda_S$ if and only if $\delta_{A,S} = 1$ and $\delta_{S,A} = 0$, due to the constraints (11a) and (11b); in this case, we have from (11c) that $p_{i,S} \le x_{i,A} \le 1$;

(ii) likewise, we have $\lambda_A > \lambda_S$ if and only if $\delta_{A,S} = 0$ and $\delta_{S,A} = 1$, due to the constraints (11a) and (11b); in this case, we have from (11c) that $0 \le x_{i,A} \le p_{i,S}$.

(iii) finally, $\lambda_A = \lambda_S$ if and only if $\delta_{A,S} = 0$ and $\delta_{S,A} = 1$; here we have from (11d) that $x_{i,A} = p_{i,S}$.

Thus the binary variables $\{\delta_{A,S}, \delta_{S,A} : S \in \mathcal{S}\}$ suitably model the constraints collection (10a) - (10b) and provide an equivalent reformulation in terms of the constraints (11a) - (11d). Therefore the optimal value of the formulations in Propositions 1 and 2 are identical.

EC.2.3. Proof of Corollary 1

Proof. When $\mathcal{S}^0$ is either nested or laminar, from Theorem 3, we know that $P_{\mathrm{mdm}}(\mathcal{S}^0) = P_{\mathrm{reg}}(\mathcal{S}^0)$. So, we can solve the worst-case expected revenue in (9) with the representable conditions of the regular model which are $x_{i,A} \le p_{i,S}$ if $S \subseteq A$ and $x_{i,A} \le p_{i,S}$ if $A \subseteq S$ for all $i \in A$ and $(i,S) \in \mathcal{I}_{\mathcal{S}}$. This is a linear program with $O(n)$ continuous variables and $O(n|\mathcal{S}|)$ constraints.

EC.2.4. Proof of Proposition 3

Proof. Suppose $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$ is nested as in $S_1 \subseteq S_2 \subseteq \cdots \subseteq S_m$. We have, from the regularity of MDM in Lemma 2, that $p_{i,S_1} \le p_{i,S_2} \le \cdots \le p_{i,S_m}$, for all $i \in S_1$. Recall that $\mathrm{closure}\big(P_{\mathrm{mdm}}(\mathcal{S})\big) = \mathrm{Proj}_{\boldsymbol{x}}(\mathcal{Q}^0_{\mathcal{S}})$ where $\mathcal{Q}^0_{\mathcal{S}}$ is defined as

$$\mathcal{Q}^0_{\mathcal{S}} = \Big\{(\boldsymbol{x},\boldsymbol{\lambda}) \in \mathbb{R}^{\mathcal{I}_{\mathcal{S}}} \times \mathbb{R}^{\mathcal{S}} : x_{i,S} \ge 0, \forall (i,S) \in \mathcal{I}_{\mathcal{S}}, \sum_i x_{i,S} = 1, \forall S \in \mathcal{S}, \\ \lambda_S \ge \lambda_T \text{ if } x_{i,S} \ge x_{i,T}, \forall (i,S),(i,T) \in \mathcal{I}_{\mathcal{S}}\Big\}. \tag{EC.5}$$

Thus, for the given nested $\mathcal{S}$, we have $\lambda_{S_m} \le \lambda_{S_k} \le \lambda_{S_{k-1}} \le \lambda_{S_1}$ for any $\boldsymbol{\lambda}$ such that $(\boldsymbol{x},\boldsymbol{\lambda}) \in \mathcal{Q}^0_{\mathcal{S}}$. For ease of notation, let $\lambda_{S_0} = +1$ and $\lambda_{S_{m+1}} = -1$. Then for any given A, the corresponding λ_A must satisfy $\lambda_{S_{k+1}} \le \lambda_A \le \lambda_{S_k}$ for some $k \in \{0,1,\ldots,m\}$.

From the viewpoint of $\lambda_A \le \lambda_{S_k} \le \cdots \le \lambda_{S_0}$, we deduce the following constraints on $x_{i,A}$: For any $i \in A \setminus S_k$, we have the respective MDM feasibility constraints $x_{i,A} \le p_{i,S_k} \le \cdots \le p_{i,S_j}$ for all

$j \quad k$ such that $i \in S_j$. These constraints can be equivalently summarized by $x_{i,A} \quad p_{i,S_k}$, and this comprises the first set of constraints for evaluating $\boldsymbol{R}_k$ in (13).

From the viewpoint of $\lambda_{S_{m+1}} \qquad \lambda_{S_{k+1}} \quad \lambda_A$, we deduce the following constraints on $x_{i,A}$: For any $i \in A, (i,S) \in I_S, S_{k+1} \quad S$, we have the respective MDM feasibility constraint $x_{i,A} \quad p_{i,S}$, which comprise the second set of constraints for evaluating $\boldsymbol{R}_k$ in (13).

EC.3. Proofs of the Results in Section 5

EC.3.1. Proof of Proposition 4

Proof. Due to Theorem 1, we have $P_{\text{mdm}}(S) = \text{Proj}_x(\ _S)$, following the definition in (8). One can argue the closure of $P_{\text{mdm}}(S)$ by the similar arguments of the proof of Proposition 1. Consider any $(i,S) \in I_S$. From the description of $\ _S$, observe that an assignment for $x_{i,S}, x_{i,T}, \lambda_S, \lambda_T$ in any $(\boldsymbol{x}, \boldsymbol{\lambda}) \in \ _S$ necessarily satisfies one of the following four cases:

In *Case 1,* we have $\lambda_S < \lambda_T$ and $x_{i,S} > x_{i,T}$: If λ_S, λ_T is such that $\lambda_S < \lambda_T$, the closure of the corresponding restriction $\{x_{i,S}, x_{i,T} : x_{i,S} > x_{i,T}\}$ equals $\{x_{i,S}, x_{i,T} : x_{i,S} \quad x_{i,T}\}$.

In *Case 2,* we have $\lambda_S > \lambda_T$ and $x_{i,S} < x_{i,T}$: If λ_S, λ_T is such that $\lambda_S > \lambda_T$, the closure of the corresponding restriction $\{x_{i,S}, x_{i,T} : x_{i,S} < x_{i,T}\}$ equals $\{x_{i,S}, x_{i,T} : x_{i,S} \quad x_{i,T}\}$.

In *Case 3,* we have $\lambda_S = \lambda_T$ and $x_{i,S} = x_{i,T} > 0$: When λ_S, λ_T is such that $\lambda_S = \lambda_T$ and $x_{i,S} = x_{i,T} > 0$, the corresponding restriction on the values of $x_{i,S}, x_{i,T}$ is given by the closed set $\{x_{i,S}, x_{i,T} : x_{i,S} = x_{i,T} > 0\}$.

Finally, in *Case 4,* we have λ_S, λ_T unconstrained and $x_{i,S} = x_{i,T} = 0$: Like in Case 3, the restriction on the values of $x_{i,S}, x_{i,T}$ corresponding to this case equals $x_{i,S} = x_{i,T} = 0$. The relationship between $x_{i,S}, x_{i,T}, \lambda_S, \lambda_T$ in this case is any one of the following sub-cases: Case (4a) $\lambda_S > \lambda_T$ and $0 = x_{i,S}$ $x_{i,S} = 0$, or Case (4b) $\lambda_S < \lambda_T$ and $0 = x_{i,S} \quad x_{i,T} = 0$, or Case (4c) $\lambda_S = \lambda_T$ and $0 = x_{i,S} = x_{i,T} = 0$.

Combining the observations in the cases (1) & (4a), (2) & (4b), and (3) & (4c), we obtain that the closure of $P_{\text{mdm}}(S)$ equals the collection of probability vectors $\boldsymbol{x}$ for which there exists a function $\lambda : S \ ! \ \mathbb{R}$ such that

$$
\begin{aligned}
x_{i,S} \qquad x_{i,T} \quad &\text{if} \quad \lambda_S > \lambda_T, \quad \forall (i,S),\ (i,T) \in I_S, \\
x_{i,S} \qquad x_{i,T} \quad &\text{if} \quad \lambda_S < \lambda_T, \quad \forall (i,S),\ (i,T) \in I_S, \\
x_{i,S} \ = \ x_{i,T} \quad &\text{if} \quad \lambda_S = \lambda_T, \quad \forall (i,S),\ (i,T) \in I_S.
\end{aligned}
$$

The constraints in the formulation in Proposition 4 exactly specify these conditions describing the closure of $P_{\text{mdm}}(S)$. Observe that the objective in $\inf\{\text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) : \boldsymbol{x}_S \in P_{\text{mdm}}(S)\}$ is continuous as a function of $\boldsymbol{x}$. Therefore, $\inf\{\text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) : \boldsymbol{x}_S \in P_{\text{mdm}}(S)\} = \min\{\text{loss}(\boldsymbol{p}_S, \boldsymbol{x}_S) : \boldsymbol{x}_S \in \text{closure}(P_{\text{mdm}}(S))\}$.

EC.3.2. Proof of Theorem 4

Before we formally prove Theorem 4, we first show the following preparatory material. Problem (15) can be solved by the following optimization problem:

$$\inf \quad \text{loss}(\boldsymbol{x}_S(\boldsymbol{\lambda}), \boldsymbol{p}_S) \tag{EC.6}$$

$$\text{s.t.} \quad \boldsymbol{x}_S(\boldsymbol{\lambda}) \in \arg\inf_{x_S(\,):(x_S(\,),\,) \in S} \text{loss}(\boldsymbol{x}_S(\boldsymbol{\lambda}), \boldsymbol{p}_S | \boldsymbol{\lambda}), \tag{EC.7}$$

where $\boldsymbol{x}_S(\boldsymbol{\lambda})$ can be interpreted as a collection of MDM-representable choice probabilities given S and $\boldsymbol{\lambda}$. Next, we focus on the sub-problem (EC.7). Let $[m] = \{1, 2, \ldots, m\}$ and $[k] = \{1, 2, \ldots, k\}$ for some positive integer m and k.

Assumption EC.1. *Consider $S = \{S_1, S_2, \ldots, S_m\}$ and $\boldsymbol{p}_S$ as a $n \times m$ matrix with n rows and m columns, satisfying the following properties:*

- *$n = k + m$ with k as a positive integer;*
- *For each $l \in [k]$, there are exactly two elements in row l of $\boldsymbol{p}_S$ and $\kappa = |p_{l,S_i} - p_{l,S_j}| < \frac{1}{2m}$ is a positive constant for $i, j \in [m]$ with $l \in S_i \setminus S_j$;*
- *For each $i \in [m]$, there is exactly one element p_{k+i,S_i} in row $k + i$ and $p_{k+i,S_i} = 1 - \sum_{j=1}^{k} p_{j,S_i}$ and $p_{k+i,mS_i} > \frac{km\kappa}{2}$.*

Lemma EC.3. *Given $\boldsymbol{p}_S$ that satisfies Assumption EC.1, the sub-problem* (EC.7) *with 1-norm objective function has a closed-form optimal objective value $\sum_{l=1,i,j \in [m]: l \in S_i \setminus S_j}^{k} 2|p_{l,S_i} - p_{l,S_j}| \cdot |(p_{l,S_i} - p_{l,S_j})(\lambda_{S_i} - \lambda_{S_j}) < 0|$.*

Intuitively, $\boldsymbol{p}_S$ has exactly one product-assortment pair for product l with $l \in [k]$ and exactly one element for product $k + i$ with $i \in [m]$ which corresponds to the number of assortments. For product l with $l \in [k]$ the indicator takes value 1 if and only if the choice probabilities of the product-assortment pair violate the MDM-representable conditions, the minimum loss to make this pair to be MDM-representable under 1-norm loss is $|p_{l,S_i} - p_{l,S_j}|$, which causes the violation of the normalization constraint of the assortments S_i and S_j. To satisfy the normalization conditions of assortment S_i and S_j, the least loss is also $|p_{l,S_i} - p_{l,S_j}|$. Next, we formally prove Lemma EC.3 with three steps: (1) reformulate the sub-problem (EC.7) to a linear program; (2) construct a primal feasible solution $\boldsymbol{x}(\boldsymbol{\lambda})$ such that the desired optimal objective value is achieved, which can be served as an upper bound to (EC.7); (3) derive the dual for the problem and construct a dual feasible solution such that the desired optimal objective value is achieved.

Proof. Step (1): Let f_{sub} denote the optimal value of (EC.7). Then, we reformulate the sub-problem (EC.7) as the following problem:

$$\begin{aligned} f_{\text{sub}} = \min_{x_S} \quad & \sum_{l=1:i,j\in[m],l\in S_i\setminus S_j}^{k} (|x_{l,S_i} - p_{l,S_i}| + |x_{l,S_j} - p_{l,S_j}|) + \sum_{i=1}^{m} |x_{k+i,S_i} - p_{k+i,S_i}| \\ \text{s.t.} \quad & x_{l,S_i} \le x_{l,S_j} \text{ if } \lambda_{S_i} \le \lambda_{S_j} \quad \forall l \in [k], i,j, \in [m],\ l \in S_i \setminus S_j, \\ & \sum_{i\in S} x_{i,S} = 1, \quad \forall S \in \mathcal{S}, \\ & x_{i,S} \ge 0, \quad \forall (i,S) \in \mathcal{I}_S. \end{aligned} \tag{EC.8}$$

We reformulate Problem (EC.8) as the following linear program by introducing a new variable z_S.

$$\begin{aligned} \min_{x_S, z_S} \quad & \sum_{l=1,i,j\in[m],l\in S_i\setminus S_j}^{k} (z_{l,S_i} + z_{l,S_j}) + \sum_{i=1}^{m} z_{k+i,S_i} \\ \text{s.t.} \quad & z_{l,S_i} \ge x_{l,S_i} - p_{l,S_i}, \quad z_{l,S_j} \ge x_{l,S_j} - p_{l,S_j}, \quad \forall l \in [k], i,j \in [m], l \in S_i \setminus S_j, \\ & z_{l,S_i} + x_{l,S_i} \ge p_{l,S_i}, \quad z_{l,S_j} + x_{l,S_j} \ge p_{l,S_j}, \quad \forall l \in [k], i,j \in [m], l \in S_i \setminus S_j, \\ & x_{l,S} \le x_{l,S_j} \text{ if } \lambda_{S_i} \le \lambda_{S_j} \quad \forall l \in [k], i,j \in [m], l \in S_i \setminus S_j, \\ & z_{k+1,S_i} \ge x_{k+i,S_i} - p_{k+i,S_i}, \quad \forall i \in [m], \\ & z_{k+1,S_i} + x_{k+i,S_i} \ge p_{k+i,S_i}, \quad \forall i \in [m], \\ & \sum_{i\in S} x_{i,S} = 1, \quad \forall S \in \mathcal{S}, \\ & x_{i,S}, z_{i,S} \ge 0, \quad \forall (i,S) \in \mathcal{I}_S. \end{aligned} \tag{EC.9}$$

Step (2): Given any $\boldsymbol{\lambda}$, construct a solution $(\boldsymbol{x}_S, \boldsymbol{z}_S)$ as follows. For $l \in [k]$,$i,j \in [m]$ such that $l \in S_i \setminus S_j$:

(2a) if $\mathbf{If}(p_{l,S_i} \le p_{l,S_j})(\lambda_{S_i} \le \lambda_{S_j}) \ge 0\mathbf{g} = 0$, let

$$x_{l,S_i} = p_{l,S_i},\ x_{l,S_j} = p_{l,S_j},\ x_{k+i,S_i} = p_{k+i,S_i},\ x_{k+j,S_j} = p_{k+j,S_j},\ x_{l,S_i} = z_{l,S_j} = 0;$$

(2b) if $\mathbf{If}(p_{l,S_i} \le p_{l,S_j})(\lambda_{S_i} \le \lambda_{S_j}) \ge 0\mathbf{g} = 1$, let

$$x_{l,S_i} = x_{l,S_j} = \frac{p_{l,S_i} + p_{l,S_j}}{2},\ z_{l,S_i} = z_{l,S_j} = \frac{|p_{l,S_i} - p_{l,S_j}|}{2}.$$

For $i \in [m]$, let

$$x_{k+i,S_i} = p_{k+i,S_i} + \sum_{l=1,i,j,\in[m],l\in S_i\setminus S_j}^{k} \mathbf{If}(p_{l,S_i} \le p_{l,S_j})(\lambda_{S_i} \le \lambda_{S_j}) \ge 0\mathbf{g} \frac{\operatorname{sgn}(p_{l,S_i} - p_{l,S_j})|p_{l,S_i} - p_{l,S_j}|}{2},$$

$$z_{k+i,S_i} = \sum_{l=1,i,j,\in[m],l\in S_i\setminus S_j}^{k} \mathbf{If}(p_{l,S_i} \le p_{l,S_j})(\lambda_{S_i} \le \lambda_{S_j}) \ge 0\mathbf{g} \frac{\operatorname{sgn}(p_{l,S_i} - p_{l,S_j})|p_{l,S_i} - p_{l,S_j}|}{2},$$

where $\mathrm{sgn}(x) := \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x \leq 0. \end{cases}$ Then, the objective value of Problem (EC.9) under the constructed solution $(\boldsymbol{x}_S, \boldsymbol{z}_S)$ is $\sum_{l=1,i,j\in[m]:l\in S_i\setminus S_j}^{k} \{p_{l,S_i} - p_{l,S_j} \mid f(p_{l,S_i} - p_{l,S_j})(\lambda_{S_i} - \lambda_{S_j}) < 0\}$. This implies that $f_{\mathrm{sub}} \leq \sum_{l=1,i,j\in[m]:l\in S_i\setminus S_j}^{k} \{p_{l,S_i} - p_{l,S_j} \mid f(p_{l,S_i} - p_{l,S_j})(\lambda_{S_i} - \lambda_{S_j}) < 0\}$.

Step (3): We derive the dual of (EC.9) as follows. For $l \in [k], i, j \in [m]$ such that $l \in S_i \setminus S_j$, we introduce the following variables: $\alpha_{l,i}, \beta_{l,i}, \alpha_{l,j}, \beta_{l,j}, u_{l,i,j} \geq 0$. For $i \in [m]$, we introduce the following variables: $\alpha_{k+i,i}, \beta_{k+i,i} \geq 0, \eta_i$. The dual problem of (EC.9) is given as:

$$
\begin{aligned}
\max_{\alpha,\beta,u,\eta} \quad & \sum_{l=1,i,j\in[m],l\in S_i\setminus S_j}^{k} [p_{l,S_i}(\beta_{l,i} - \alpha_{l,i}) + p_{l,S_j}(\beta_{l,j} - \alpha_{l,j})] + \sum_{i=1}^{m} [p_{k+i,S_i}(\beta_{k+i,i} - \alpha_{k+i,i}) - \eta_i] \\
\text{s.t.} \quad & \alpha_{l,i} + \beta_{l,i} \geq 1, \quad \alpha_{l,j} + \beta_{l,j} \geq 1, \quad \forall l \in [k], i, j \in [m], l \in S_i \setminus S_j, \\
& \alpha_{l,i} + \beta_{l,i} + u_{l,i,j} \geq 0, \quad \alpha_{l,j} + \beta_{l,j} - u_{l,i,j} \geq 0, \quad \forall l \in [k], i, j \in [m], l \in S_i \setminus S_j, \\
& \alpha_{k+i,i} + \beta_{k+i,i} \geq 1, \quad \alpha_{k+i,i} + \beta_{k+i,i} \geq 0, \quad \forall i \in [m], \\
& \alpha_{l,i}, \beta_{l,i}, \alpha_{l,j}, \beta_{l,j}, u_{l,i,j} \geq 0, \quad \forall l \in [k], i, j \in [m], l \in S_i \setminus S_j, \\
& \alpha_{k+i,i}, \beta_{k+i,i} \geq 0, \quad \forall i \in [m].
\end{aligned}
\tag{EC.10}
$$

Construct a dual solution for (EC.10) as follows: For $l \in [k], i, j \in [m], l \in S_i \setminus S_j$,

(3a) if $f(p_{l,S_i} - p_{l,S_j})(\lambda_{S_i} - \lambda_{S_j}) \geq 0$, let $\alpha_{l,i} = \beta_{l,i} = \alpha_{l,j} = \beta_{l,j} = u_{l,i,j} = 0$.

(3b) if $f(p_{l,S_i} - p_{l,S_j})(\lambda_{S_i} - \lambda_{S_j}) < 0$, let $\alpha_{l,i} = 1$, $\beta_{l,i} = 0$, $\alpha_{l,j} = 0$, $\beta_{l,j} = 1$, $u_{l,i,j} = 1$.

For $i \in [m]$, let $\eta_i = 0$ and $\alpha_{k+i,i} = 0$, $\beta_{k+i,i} = \sum_{l=1}^{k} \sum_{j\in[m]:l\in S_i\setminus S_j} \frac{|p_{l,S_i} - p_{l,S_j}| \cdot f(p_{l,S_i} - p_{l,S_j})(\lambda_{S_i} - \lambda_{S_j}) < 0}{2 p_{k+i,S_i}}$.

By weak duality, we have $f_{\mathrm{sub}} \geq \sum_{l=1,i,j\in[m]:l\in S_i\setminus S_j}^{k} \{p_{l,S_i} - p_{l,S_j} \mid f(p_{l,S_i} - p_{l,S_j})(\lambda_{S_i} - \lambda_{S_j}) < 0\}$.

Summing up, we have $f_{\mathrm{sub}} = \sum_{l=1,i,j\in[m]:l\in S_i\setminus S_j}^{k} \{p_{l,S_i} - p_{l,S_j} \mid f(p_{l,S_i} - p_{l,S_j})(\lambda_{S_i} - \lambda_{S_j}) < 0\}$.

Before we show the hardness of Problem (15), we provide the following three definitions (see Dwork et al. (2001)) to describe the Kemeny optimal aggregation problem.

Definition EC.1 (Full lists and partial lists). Let $M = \{1, \ldots, m\}$ be a finite set of alternatives, called universe. A ranking over M is an ordered list. If the ranking τ contains all the elements in M, then it is called a full list (ranking). If the ranking τ contains a subset of element from the universe M, then it is called a partial list (ranking).

Definition EC.2 (Kendall-tau distance (K-distance)). The K-distance, denoted as $K(\sigma, \tau)$, is the number of pairs $i, j \in M$ such that $\sigma(i) < \sigma(j)$ but $\tau(i) > \tau(j)$ where $\sigma(i)$ stands for the position of i in σ and similar explanations are applied for $\sigma(j), \tau(i)$ and $\tau(j)$. Note that the pair (i, j) has contribution to the K-distance only if both i, j appear in both lists σ, τ.

Definition EC.3 (SK, Kemeny optimal). For a collection of partial lists $\tau_1, \ldots, \tau_k$ and a full list π, we denote

$$SK(\pi, \tau_1, \ldots, \tau_k) = \sum_{i=1}^{k} K(\pi, \tau_i).$$

We say a permutation σ is a Kemeny optimal aggregation of $\tau_1, \ldots, \tau_k$ if it minimizes $SK(\pi, \tau_1, \ldots, \tau_k)$ over all permutations π.

Lemma EC.4 (see Dwork et al. (2001)). *Finding Kemeny optimal solution for partial lists of length 2 is exactly the same problem as finding a minimum feedback arc set, and hence is NP-hard.*

Now, we are ready to prove the hardness of Problem (15).

Proof. To show Problem (15) is NP-hard, it suffices to show some instances of this problem is NP-hard. We show that Kemeny optimal aggregation of length 2 can be reduced to Problem (15).

The decision version of Kemeny optimal aggregation with a collection of partial lists of all length 2 is stated as follows:

INSTANCE: A finite set M with $|M| = m$, a collection of partial lists $\tau_1, \ldots, \tau_k$ with $|\tau_i| = 2$ for $i = 1, \ldots, k$, an upper bound on the loss L.

QUESTION:Is there a full list π, such that $\sum_{i=1}^{k} K(\pi, \tau_i) \quad L$?

The decision version of the limit problem of MDM in Problem (15) is stated as follows:

INSTANCE: A finite set N with $|N| = n$, a collection of assortments S with $|S| = m$ and $S \quad N$ for all $S \in S$, the observed choice probabilities $\boldsymbol{p}_S = (p_{i,S} : i \in S, S \in S)$ with $\sum_{i \in S} p_{i,S} = 1$ for all $S \in S$, an upper bound on the loss L^0.

QUESTION: Is there a solution $(\boldsymbol{x}_S, \boldsymbol{\lambda})$ to Problem (15) such that loss$(\boldsymbol{x}_S, \boldsymbol{p}_S) \quad L^0$?

We then will reduce the Kemeny optimal aggregation problem to Problem (15). Given any instance of Kemeny optimal aggregation problem with partial lists all of length 2, we can construct an instance of Problem (15) as follows.

(1) Let the collection of assortments $S = \{S_1, S_2, \ldots, S_m\}$ with $|S| = m$ and the set of alternatives (products), $N = \{1, \ldots, k, k+1, \ldots, k+m\}$ with $|N| = n = k + m$. Given the observed choice data $\boldsymbol{p}_S$, consider $\boldsymbol{p}_S$ as a $n \quad m$ matrix with n rows and m columns.

(2) The values of the entries in $\boldsymbol{p}_S$ are set in the following manner. For each $l \in \{1, ..., k\}$, suppose $\tau_l = \{i, j\}$ with $\tau_l(i) < \tau_l(j)$, then we set $p_{l,S_i} = \frac{1}{3^{\ k}}$ and $p_{l,S_j} = \frac{2}{3^{\ k}}$. It's easy to see that for each S_i with $1 \quad i \quad m$, $0 < \sum_{j=1}^{k} p_{j,S_i} < 1$.

(3) For each S_i with $1 \quad i \quad m$, let $p_{k+i,S_i} = 1 \quad \sum_{j=1}^{k} p_{j,S_i}$.

(4) Set other entries of $\boldsymbol{p}_S$ as zero.

(5) Set the loss function in Problem (15) to be 1-norm loss.

We give Example EC.1 and Example EC.2 to illustrate the above instance construction.

Example EC.1. Given an instance of Kemeny optimal aggregation with $M = \{1,2,3\}$ and $\tau_1 = (1 \quad 2), \tau_2 = (1 \quad 3), \tau_3 = (2 \quad 3)$, we construct an instance for Problem (15) with $\boldsymbol{p}_S$ as shown in Table EC.4.

Table EC.4 An example of a representable instance construction

alternative	$S_1 = \{1,2,4\}$	$S_2 = \{1,3,5\}$	$S_3 = \{2,3,6\}$
1	1/9	2/9	-
2	1/9	-	2/9
3	-	1/9	2/9
4	7/9	-	-
5	-	6/9	-
6	-	-	5/9

Example EC.2. Given an instance of Kemeny optimal aggregation with $M = \{1,2,3\}$ and $\tau_1 = (1 \quad 2), \tau_2 = (3 \quad 1), \tau_3 = (2 \quad 3)$, we construct an instance for Problem (15) with $\boldsymbol{p}_S$ as shown in Table EC.5.

Table EC.5 An example of an infeasible instance construction

alternative	$S_1 = \{1,2,4\}$	$S_2 = \{1,3,5\}$	$S_3 = \{2,3,6\}$
1	1/9	2/9	-
2	2/9	-	1/9
3	-	1/9	2/9
4	6/9	-	-
5	-	6/9	-
6	-	-	6/9

In Example EC.1, both the Kemeny optimal aggregation and the limit of MDM instances are feasible to their problem respectively. The optimal solution to the Kemeny optimal aggregation is $\pi = (1 \quad 2 \quad 3)$ and one of the optimal solutions to the limit of MDM is $\boldsymbol{x}_S = \boldsymbol{p}_S$ and $\lambda_{S_1} = 3$, $\lambda_{S_2} = 2$ and $\lambda_{S_3} = 1$. Both instances obtain 0 loss.

In Example EC.2, both the Kemeny optimal aggregation and the limit of MDM instance are infeasible to their problem respectively. It's trivial to see that the optimal solution to the Kemeny optimal aggregation is one of $\{(1 \quad 2 \quad 3), (2 \quad 3 \quad 1), (3 \quad 1 \quad 2)\}$. Each of such solutions obtains $SK = 1$. Given Lemma EC.3, one may make a guess for one of the optimal solutions to the limit of MDM in Example EC.2 to be $x_{2,S_1} = \frac{1}{6}$, $x_{2,S_3} = \frac{1}{6}$, $x_{4,S_1} = \frac{13}{18}$, $x_{6,S_3} = \frac{11}{18}$ and the optimal loss to be $\frac{2}{9}$. We will show that this guess is true.

Recall that Problem (15) is equivalent to Problem (EC.6). We next show that for any fixed $\boldsymbol{\lambda}$ in Problem (EC.6), then the sub-problem (EC.7) with optimal $\boldsymbol{x}_S(\boldsymbol{\lambda})$ and $f_{\text{sub}}(\boldsymbol{\lambda})$ under $\boldsymbol{\lambda}$, there exists π such that $\lambda_{S_{\pi\ 1(1)}} \quad \lambda_{S_{\pi\ 1(m)}}$ for the Kemeny optimal aggregation and $SK(\pi) = \frac{3\ k}{2} f_{\text{sub}}(\boldsymbol{\lambda})$.

By Lemma EC.3, we have

$$
\begin{aligned}
& f_{\text{sub}}(\boldsymbol{\lambda}) = \text{loss}(\boldsymbol{x}_S(\boldsymbol{\lambda}), \boldsymbol{p}_S) \\
&= \sum_{l=1, i,j \in [m]: l \in S_i \setminus S_j}^{k} \left(|x_{l,S_i} - p_{l,S_i}| + |x_{l,S_j} - p_{l,S_j}| \right) + \sum_{i=1}^{m} |x_{k+i,S_i} - p_{k+i,S_i}| \qquad (1) \\
&= \sum_{l=1, i,j \in [m]: l \in S_i \setminus S_j}^{k} 2|p_{l,S_i} - p_{l,S_j}| \mathbf{1}\{(p_{l,S_i} - p_{l,S_j})(\lambda_{S_i} - \lambda_{S_j}) < 0\} \qquad (2) \\
&= \frac{2}{3^k} \sum_{l=1, i,j \in [m]: l \in S_i \setminus S_j}^{k} \mathbf{1}\{(\pi(i) - \pi(j))(\tau_l(i) - \tau_l(j)) < 0\} \qquad (3) \\
&= \frac{2}{3^k} \sum_{l=1}^{k} K(\pi, \tau_l) \\
&= \frac{2}{3^k} SK(\pi).
\end{aligned}
$$

Equation (1) is due to the construction of $\boldsymbol{p}_S$. Equation (2) holds because of $|p_{l,S_i} - p_{l,S_j}| = \frac{1}{3^k}$ and the closed form objective value in Lemma EC.3. The argument for Equation (3) is as follows: For $l \in [k]$, $i, j \in [m] : l \in S_i \setminus S_j$, by instance construction, we have

$$p_{l,S_i} < p_{l,S_j} \text{ if } \tau_l(i) < \tau_l(j).$$

From the relation between π and $\boldsymbol{\lambda}$, we have $\lambda_{\pi^{-1}(1)} > \ldots > \lambda_{\pi^{-1}(m)}$. Then, we have

$$\pi(i) < \pi(j) \text{ if } \lambda_{S_i} - \lambda_{S_j}.$$

The above two inequities imply that

$$\mathbf{1}\{(\lambda_{S_i} - \lambda_{S_j})(p_{l,S_i} - p_{l,S_j}) > 0\} = \mathbf{1}\{(\pi(i) - \pi(j))(\tau_l(i) - \tau_l(j)) < 0\}.$$

Setting $L = \frac{3k}{2} L'$. The decision problem of the limit of MDM asks is there $(\boldsymbol{x}_S, \boldsymbol{\lambda})$ such that $f_{\text{limit}} - L'$ is equivalent to the decision problem of Kemeny optimal aggregation is there a full ranking π such that $f_{\text{kemeny}} - L$.

EC.3.3. Proof of Proposition 5

Proof. Observe that the variables $(\lambda_S : S \in \mathcal{S})$ influence the value of the formulation in Proposition 5 only via the sign of $\lambda_S - \lambda_T$, for any pair of variables λ_S, λ_T from the collection $(\lambda_S : S \in \mathcal{S})$. Therefore the optimal value of this optimization formulation is not affected by the presence of the following additional constraints: $0 - \lambda_S - 1$ for all $S \in \mathcal{S}$. Indeed, this is because the signs of the differences $\{\lambda_S - \lambda_T : S, T \in \mathcal{S}\}$ are not affected by these additional constraints. Taking ϵ to be smaller than $1/(2|\mathcal{S}|)$, for example, ensures that there is a feasible assignment for $(\lambda_S : S \in \mathcal{S})$ within the interval $[0, 1]$ even if all these variables take distinct values.

Let F denote the feasible values for the variables $(\lambda_S : S \in \mathcal{S}), (x_{i,S} : (i,S) \in \mathcal{I}_{\mathcal{S}})$ satisfying the constraints introduced in the above paragraph besides those in the formulation in Proposition 4. Equipped with this feasible region F, we have the following deductions for $(\lambda_S : S \in \mathcal{S}), (x_{i,S} : (i,S) \in \mathcal{I}_{\mathcal{S}})$ in F: For any $S, T \in \mathcal{S}$ containing i,

(i) we have $\lambda_S < \lambda_T$ if and only if $\delta_{S,T} = 1$ and $\delta_{T,S} = 0$, due to the first set of constraints of (16); in this case, we have from the second and fourth set of constraints of (16) that $0 \quad x_{i,T} \quad x_{i,S} \quad 1$;

(ii) likewise, we have $\lambda_S > \lambda_T$ if and only if $\delta_{S,T} = 0$ and $\delta_{T,S} = 1$, due to the first set of constraints; in this case, we have from the second and fourth set of constraints of (16) that $0 \quad x_{i,S} \quad x_{i,T} \quad 1$.

(iii) finally, $\lambda_S = \lambda_T$ if and only if $\delta_{S,T} = 0$ and $\delta_{T,S} = 0$; here we have from the third set of constriants of (16) that $x_{i,S} = x_{i,T}$.

Thus the binary variables $\{\delta_{S,T} : S, T \in \mathcal{S}\}$ suitably model the first set of constraints of (15) and provide an equivalent reformulation in terms of the constraints. Therefore the optimal value of the formulations in Propositions 4 and 5 are identical.

EC.4. Proofs of the Results in Section 6

EC.4.1. Proof of Theorem 5

Proof. necessity of (20). Suppose $\boldsymbol{p}_{\mathcal{S}}$ is G-MDM-representable. Then there exist marginal distributions $\{F_i : i \in N\}$ and deterministic utilities $\{\nu_i : i \in N\}$ such that for any assortment $S \in \mathcal{S}$, the given choice probability vector $(p_{i,S} : i \in S)$ and the respective Lagrange multipliers $\{\lambda_S, \lambda_{i,S} : i \in S\}$ are obtainable by solving (4). That is, there exist $\{\lambda_S, \lambda_{i,S} : i \in S\}$ for some fixed choice of $\{F_i : i \in N\}$ and $\{\nu_i : i \in N\}$, such that

$$\nu_i + F_i^{-1}(1 \quad p_{i,S}) \quad \lambda_S + \lambda_{i,S} = 0 \quad \forall (i,S) \in \mathcal{I}_{\mathcal{S}}, \tag{EC.11}$$

$$\lambda_{i,S}\, p_{i,S} = 0 \quad \forall (i,S) \in \mathcal{I}_{\mathcal{S}}. \tag{EC.12}$$

For any group l, alternatives $i, j \in G_l$, assortments S, T with $i, j \in S \setminus T$,

$$\lambda_S \quad \nu_i = \lambda_{i,S} + F_l^{-1}(1 \quad p_{i,S}) \text{ and } \lambda_T \quad \nu_j = \lambda_{j,T} + F_l^{-1}(1 \quad p_{j,T}).$$

If $p_{i,S} < p_{i,T}$, then $\lambda_{i,S} \quad 0$ and $\lambda_{i,T} = 0$ because of the complementary slackness condition (EC.12). Since $F_l^{-1}(1 \quad p)$ is a strictly decreasing function over $p \in [0,1]$, by (EC.11), we obtain:

$$\lambda_S \quad \nu_i \qquad F_l^{-1}(1 \quad p_{i,S}) > F_l^{-1}(1 \quad p_{i,T}) = \lambda_T \quad \nu_i.$$

If on the other hand $p_{i,S} = p_{i,T} \neq 0$, we have $\lambda_{i,S} = \lambda_{i,T} = 0$ from the optimality conditions. Then $\lambda_S \quad \nu_i = F_l^{-1}(1 \quad p_{i,S}) = F_l^{-1}(1 \quad p_{i,T}) = \lambda_T \quad \nu_i$. Thus, setting $\lambda(S) = \lambda_S$ for all $S \in \mathcal{S}$, we see that there exists a function $\lambda : \mathcal{S} \to \mathbb{R}$ and $(\nu_i : i \in N)$ satisfying (20).

su ciency of (20). Given $\boldsymbol{p}_S$ and $\lambda : S \to \mathbb{R}$, $\{v_i : i \in N\}$ such that (20) holds for all $(i,S),(j,T) \in \mathcal{I}_S$ with $g(i) = g(j) = l$, we next exhibit a construction of marginal distributions $(F_l : l = 1, \ldots, K)$ for G-MDM. This construction will be such that it yields the given $(p_{i,S} : i \in S)$ as the corresponding choice probabilities from the optimality conditions in (4), for any assortment $S \in \mathcal{S}$.

Towards this, let $z_{i,S} = \lambda_S \quad \nu_i$ for all $(i,S) \in \mathcal{I}_S$. For a given group l, let $\boldsymbol{z}^l_S = (z^l_{i,S} : (i,S) \in \mathcal{I}, i \in G_l)$ with $|\boldsymbol{z}^l_S| = m_l$. Further, let h_l denote the number of product and assortment pairs in the group l for which $p_{i,S} > 0$ with $g(i) = l$. See that $h_l = m_l$, with the equality holding only when the choice probabilities $\boldsymbol{p}^l_S = \{p_{i,S} : (i,S) \in \mathcal{S}, g(i) = l\}$ are all non-zero. Equipped with this notation, we construct the marginal distribution $F_l(\,)$ for any group l as follows:

(a) Consider any ordering $(\boldsymbol{i},\boldsymbol{S})^l = ((i_1,S_1),(i_2,S_2) \quad ,(i_{h_l},S_{h_l}),(i_{h_l+1},S_{h_l+1}), \quad ,(i_{m_l},S_{m_l}))$ over the product and assortment pairs in G_l for which $z^l_{i_1,S_1} \quad z^l_{i_2,S_2} \quad z^l_{i_{h_l},S_{h_l}} < z^l_{i_{h_l+1},S_{h_l+1}} \quad z^l_{i_{m_l},S_{m_l}}$. With h_l defined as the number of product and assortment pairs in group l with $p_{i,S} > 0$, note that it is necessary to have $z^l_{i_{h_l},S_{h_l}} < z^l_{i_{h_l+1},S_{h_l+1}}$ whenever $h_l < m_l$. This follows from the observations that $\boldsymbol{z}^l$ satisfies (20) and $p^l_{i_{h_l},S_{h_l}} > 0 = p^l_{i_{h_l+1},S_{h_l+1}}$. Further, due to the conditions in (20), the choice probabilities $(p_{i,S} : (i,S) \in \mathcal{I}, g(i) = l)$ must necessarily satisfy the ordering $p^l_{i_1,S_1} \quad p^l_{i_2,S_2} \quad p^l_{i_{h_l},S_{h_l}} > p^l_{i_{h_l+1},S_{h_l+1}} \quad p^l_{i_{m_l},S_{m_l}}$.

(b) Construct the cumulative distribution function $F_l(\,)$ by first setting $F_l(z_{i_k,S_k}) = 1 \quad p^l_{i_k,S_k}$ for $k = 1, \quad ,h_l$. With this assignment, we complete the construction of the distribution F_l in between these points by connecting them with line segments as follows: For any two consecutive product and assortment pairs (i_k,S_k) and (i_{k+1},S_{k+1}) in the ordering satisfying $z^l_{i_k,S_k} < z^l_{i_{k+1},S_{k+1}}$, connect the respective points $(z^l_{i_k,S_k}, 1 \quad p^l_{i_k,S_k})$ and $(z^l_{i_{k+1},S_{k+1}}, 1 \quad p^l_{i_{k+1},S_{k+1}})$ with a line segment (see Figure EC.2). For $k \quad h_l$, note that if the product and assortment pairs (i_k,S_k) and (i_{k+1},S_{k+1}) are such that $z^l_{i_k,S_k} = z^l_{i_{k+1},S_{k+1}}$, then the corresponding points $(z^l_{i_k,S_k}, 1 \quad p^l_{i_k,S_k})$ and $(z^l_{i_{k+1},S_{k+1}}, 1 \quad p^l_{i_{k+1},S_{k+1}})$ coincide and there is no need to connect them. Note that $p^l_{i_k,S_k} > p^l_{i_{k+1},S_{k+1}}$ when $z^l_{i_k,S_k} < z^l_{i_{k+1},S_{k+1}}$, because of (20), and hence the cumulative distribution function F_l is strictly increasing in the interval $[z^l_{i_k,S_k}, z^l_{i_{k+1},S_{k+1}}]$.

(c) Lastly we construct the tails of the distribution F_l as follows: For the right tail, connect the points $(z^l_{i_{h_l},S_{h_l}}, 1 \quad p^l_{i_{h_l},S_{h_l}})$ and $(z^l_{i_{h_l+1},S_{h_l+1}}, 1)$ with a line segment if $h_l < m_l$. We then have $F_l(x) = 1$ for every $x \quad z^l_{i_{h_l},S_{h_l}}$ and therefore $F_l{}^{-1}(1) = z^l_{i_{h_l+1},S_{h_l+1}}$. If $h_l = m_l$, connect the points $(z^l_{i_{h_l},S_{h_l}}, 1 \quad p^l_{i_{h_l},S_{h_l}})$ and $(z^l_{i_{h_l},S_{h_l}} + \delta, 1)$ by choosing any aribtrary $\delta > 0$ (see Figure EC.2). In this case, we will have $F_l(x) = 1$ for every $x \quad z^l_{i_{h_l},S_{h_l}} + \delta$. For the left tail, if $p_{i,S_1} = 1$, then we have $F_l(x) = 0$ for every $x \quad z^l_{i_1,S_1}$. Both the cumulative distribution functions drawn in Figure EC.2 illustrate this case. On the other hand, if $p_{i,S_1} < 1$, we use a line segment to connect $(z^l_{i_1,S_1}, 1 \quad p_{i,S_1})$ and $(z^l_{i_1,S_1} \quad \delta, 0)$ by choosing an arbitrary $\delta > 0$. In this case, $F_l(x) = 0$ for every $x \quad z^l_{i_1,S_1} \quad \delta$.

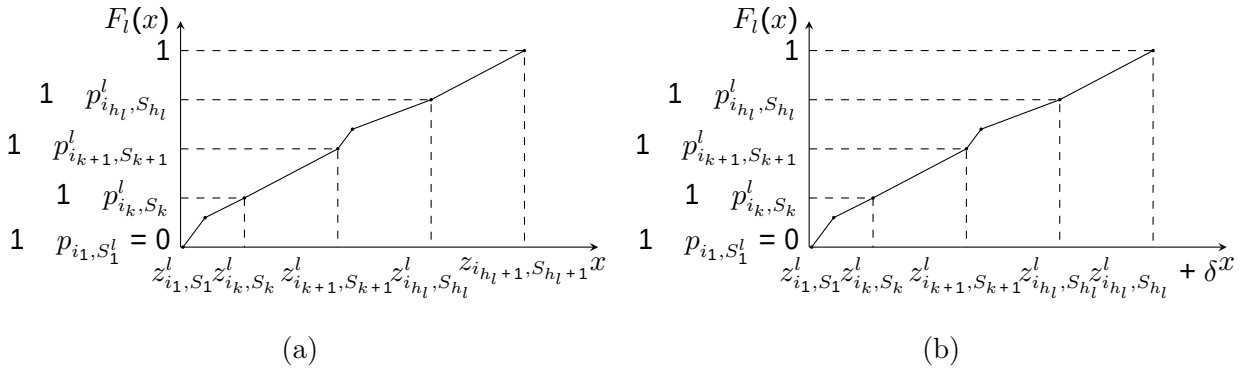


Figure EC.2 An illustration of the construction of the marginal distribution F_l when: (a) there is $p_{i,S} = 0$ for some $(i,S) \in \mathcal{I}_S$ with $g(i) = l$ (the case where $l_i < m_i$) and (b) $p_{i,S} > 0$ for all $(i,S) \in \mathcal{I}_S$ with $g(i) = l$ (the case where $h_l = m_l$).

The above construction gives marginal distribution functions $(F_l : l = 1, \ldots, K)$ which are absolutely continuous and strictly increasing within its support. We next show that the constructed marginal distributions yield the given choice probabilities $(p_{i,S} : i \in S)$, for any assortment $S \in \mathcal{S}$, when they are used in the optimality conditions (4). In other words, given $\boldsymbol{p}_{\mathcal{S}}$ and $\lambda : \mathcal{S} \to \mathbb{R}$, $\{v_i : i \in N\}$, we next verify that

$$\nu_i + F_i^{-1}(1 - p_{i,S}) - \lambda(S) + \lambda_{i,S} = 0, \quad \lambda_{i,S}\, p_{i,S} = 0, \quad \text{and} \quad \lambda_{i,S} \geq 0, \quad \forall (i,S) \in \mathcal{I}_{\mathcal{S}}.$$

For any $(i,S) \in \mathcal{S}$ with $g(i) = l$ and $p_{i,S} > 0$, we have from the construction of F_l that $F_l(z^l_{i,S}) = 1 - p_{i,S}$. Then for such $p_{i,S}$, we have the optimality condition $\nu_i + F_l^{-1}(1 - p_{i,S}) - \lambda(S) + \lambda_{i,S} = 0$ readily hold since $z^l_{i,S} = \lambda(S) - \nu_i$ and the optimality conditions also stipulate that $\lambda_{i,S} = 0$ when $p_{i,S} > 0$.
For any $(i,S) \in \mathcal{I}_{\mathcal{S}}$ with $g(i) = l$ such that $p_{i,S} = 0$, we have from Steps (a) and (c) of the above construction that $\lambda(S) - \nu_i = z^l_{i,S} - z_{i_{h_l+1},S_{h_l+1}} = F_l^{-1}(1) = F_l^{-1}(1 - p_{i,S})$. Then if we take $\lambda_{i,S} = \lambda(S) - \nu_i - z_{i_{h_l+1},S_{h_l+1}} = \lambda(S) - \nu_i - \lambda(S_{h_l+1}) + \nu_{i_{h_l+1}}$, we again readily have $\nu_i + F_i^{-1}(1 - p_{i,S}) - \lambda(S) + \lambda_{i,S} = 0$. This completes the verification that for any choice data $\boldsymbol{p}_{\mathcal{S}}$ satisfying (20), there exists marginal distributions $\{F_l : l = 1, \ldots, K\}$ which yield $\boldsymbol{p}_{\mathcal{S}}$ as the G-MDM choice probabilities.

Lastly, the condition in (20) is equivalent to testing if there exists $(\lambda_S : S \in \mathcal{S})$, $(\nu_i : i \in N)$ and $\epsilon > 0$ such that for all groups $l = 1, \ldots, K$, for all $(i,S), (j,T) \in \mathcal{I}_{\mathcal{S}}$ with $g(i) = g(j) = l$:

$$\lambda_S - \nu_i \leq \lambda_T - \nu_j + \epsilon \quad \text{if} \quad p_{i,S} < p_{j,T},$$
$$\lambda_S - \nu_i = \lambda_T - \nu_j \quad \text{if} \quad 0 < p_{i,S} = p_{j,T},$$

This is possible in polynomial time using a linear program by letting the above conditions be formulated constraints and maximizing ϵ. This linear program includes $|\mathcal{S}|$ variables for $(\lambda_S : S \in \mathcal{S})$, and n variables for $(\nu_i : i \in N)$, and 1 variable for ϵ, and $O(n^2|\mathcal{S}|^2)$ constraints.

EC.4.2. Proof of Corollary 3

Proof. When $K = 1$, $G = N$, we have $g(i) = g(j)$ for all i, j and the result follows.

EC.4.3. Proof of Theorem 6

Proof. To prove Theorem 6, we first present a lemma that can construct special cases of choice probabilities obtained from the MNL model.

Lemma EC.5. *For any fixed n, let $S, T \subseteq N$ with $|S|, |T| \geq 2$. Let $i \in S$ and $j \in T$. Then there exists a set of positive integers $x_1, ..., x_n$ such that*

$$\frac{x_i}{\sum_{k \in S} x_k} < \frac{x_j}{\sum_{k \in T} x_k}, \tag{EC.13}$$

as long as $i \neq j$ or $S \neq T$.

Proof. We prove Lemma EC.5 holds for a set of positive integers such that $x_1 = 2$ and $x_{n+1} = x_n \sum_{i=1}^{n} x_i$ when $n \geq 2$ by induction. Base Case: when $n = 3$, the possible subsets are $\{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}$. We have $x_1 = 2$, $x_2 = 4$, and $x_3 = 24$. By substituting x_1, x_2, x_3 to (EC.13) for all i, j, S, T such that $i \in S, j \in T$, and $i \neq j$ or $S \neq T$, Lemma EC.5 holds. Induction Step: Assume that when $n = k$, Lemma EC.5 holds for the set of positive integers $x_1, ..., x_k$ that satisfies $x_1 = 2$ and $x_{k+1} = x_k \sum_{i=1}^{k} x_i$ when $k \geq 2$. We next prove that the statement is true when $n = k + 1$. In the rest of the proof, we prove the statement by contradiction.

(1) If $x_{k+1} = x_i = x_j$ in (EC.13), we have

$$\frac{x_{k+1}}{\sum_{k \in S \setminus \{k+1\}} x_k + x_{k+1}} = \frac{x_{k+1}}{\sum_{k \in T \setminus \{k+1\}} x_k + x_{k+1}}.$$

We also have $\frac{x_1}{\sum_{k \in S \setminus \{k+1\}} x_k} = \frac{x_1}{\sum_{k \in T \setminus \{k+1\}} x_k}$ from the induction assumption, which implies that

$$\sum_{k \in S \setminus \{k+1\}} x_k = \sum_{k \in T \setminus \{k+1\}} x_k.$$

So, a contradiction exists.

(2) If $x_{k+1} \neq x_i$, and $x_{k+1} \neq x_j$, and $k + 1 \in S \setminus T$, we have

$$\frac{x_i}{\sum_{k \in S \setminus \{k+1\}} x_k + x_{k+1}} = \frac{x_j}{\sum_{k \in T \setminus \{k+1\}} x_k + x_{k+1}}.$$

We also have $\frac{x_i}{\sum_{k \in S \setminus \{k+1\}} x_k} \neq \frac{x_j}{\sum_{k \in T \setminus \{k+1\}} x_k}$ from the induction assumption, which implies

$$\frac{x_i}{\sum_{k \in S \setminus \{k+1\}} x_k + x_{k+1}} \neq \frac{x_j}{\sum_{k \in T \setminus \{k+1\}} x_k + x_{k+1}}.$$

So, a contradiction exists.

(3) If $x_i = x_{k+1}$ and $x_j \neq x_{k+1}$, we have $\{k+1\} \in S$, and $\{k+1\} \notin T$,and $\frac{x_{k+1}}{\sum_{k\in S\cap\{k+1\}} x_k + x_{k+1}} = \frac{x_j}{\sum_{k\in T} x_k}$. This implies

$$x_{k+1}\sum_{k\in T} x_k = x_j\Big(\sum_{k\in S\cap\{k+1\}} x_k + x_{k+1}\Big).$$

By sorting terms, we have

$$x_{k+1} = \frac{x_j \sum_{k\in S\cap\{k+1\}} x_k}{\sum_{k\in T} x_k \quad x_j} < x_j \sum_{k\in S\cap\{k+1\}} x_k < x_k \sum_{i=1}^{k} x_k.$$

Then, a contradiction exists.

(4) If $x_i = x_{k+1}$, and $x_j \neq x_{k+1}$, and $k+1 \in T$, we have $\frac{x_{k+1}}{\sum_{k\in S\cap\{k+1\}} x_k + x_{k+1}} = \frac{x_j}{\sum_{k\in T\cap\{k+1\}} x_k + x_{k+1}}$. This implies

$$\frac{x_{k+1} \quad x_j}{\sum_{k\in S\cap\{k+1\}} x_k + x_{k+1} \quad x_j} = \frac{x_j \quad x_j}{\sum_{k\in T\cap\{k+1\}} x_k + x_{k+1} \quad x_j} = 0,$$

if both denominators are not equal to zero. Since $x_{k+1} = x_k \sum_{i=1}^{k} x_k$, we have

$$x_{k+1} \quad x_j > x_k \sum_{i=1}^{k} x_k \quad x_j x_k = x_k \sum_{i=1, i\neq j}^{k} x_k > 0,$$

which implies

$$\frac{x_{k+1} \quad x_j}{\sum_{k\in S\cap\{k+1\}} x_k + x_{k+1} \quad x_j} > 0.$$

We have $j \in T$, which implies the denominator $\sum_{k\in T\cap\{k+1\}} x_k + x_{k+1} \quad x_j > 0$. So,

$$\frac{x_{k+1} \quad x_j}{\sum_{k\in S\cap\{k+1\}} x_k + x_{k+1} \quad x_j} > 0 \text{ if } \frac{x_j \quad x_j}{\sum_{k\in T\cap\{k+1\}} x_k + x_{k+1} \quad x_j} = 0$$

when $x_{k+1} = x_k \sum_{i=1}^{k} x_k$. Thus, a contradiction exists.

This completes the proof.

Recall that the probability of choosing product i in assortment S under an MNL model is $p_{i,S} = \frac{e^{\nu_i}}{\sum_{j\in S} e^{\nu_j}}$. Then, there exists $\nu_i = \ln x_i$ for all $i \in N$, such that the instance in Lemma EC.5 is an MNL instance.

To show a G-MDM has a positive measure, it suffices to show that G-MDM with $K = 1$, which is APU, has a positive measure. Equipped with Lemma EC.5, we prove Theorem 6 as follows. Let $\boldsymbol{p}_S$ be an instance following the manners in Lemma EC.5. Since MNL is a special case of APU, $\boldsymbol{p}_S$ is APU-representable. We next show that any instance $\boldsymbol{p}'_S$ that lies in the ball centered at $\boldsymbol{p}_S$ with the radius $\epsilon > 0$ is an APU instance. Let $0 < \epsilon < \min\{p_{i,S} \quad p_{j,T} | \forall (i,S),(j,T) \in I_S\}$. We perturb $\boldsymbol{p}_S$ to be $\boldsymbol{p}'_S$ by letting $p'_{i,S} = p_{i,S} + \epsilon$ and $p'_{j,S} = p_{j,S} \quad \epsilon$ by arbitrarily choosing $(i,S),(j,S) \in I_S$, and keeping other entries of $\boldsymbol{p}'_S$ the same as $\boldsymbol{p}_S$. Since $\boldsymbol{p}_S$ is APU-representable, we have

$$\lambda_S \quad \nu_i > \lambda_T \quad \nu_j \quad \text{if} \quad p_{i,S} < p_{j,T} \quad \forall (i,S),(j,T) \in I_S.$$

We have

$$\lambda_S \quad \nu_i > \lambda_T \quad \nu_j \quad \text{if} \quad p_{i,S} + \epsilon < p_{i,T} \quad \forall (i,S),(j,T) \in I_S,$$
$$\lambda_T \quad \nu_i > \lambda_S \quad \nu_j \quad \text{if} \quad p_{i,T} < p_{j,S} \quad \epsilon \quad \forall (i,T),(j,S) \in I_S,$$

since $\epsilon < \min\{p_{i,S} \quad p_{j,T}\}, \forall (i,S),(j,T) \in I_S$. From the construction of $\boldsymbol{p}^0_S$, equivalently, we have

$$\lambda_S \quad \nu_i > \lambda_T \quad \nu_j \quad \text{if} \quad p^0_{i,S} < p^0_{j,T} \quad \forall (i,S),(j,T) \in I_S,$$
$$\lambda_T \quad \nu_i > \lambda_S \quad \nu_j \quad \text{if} \quad p^0_{i,T} < p^0_{j,S} \quad \forall (i,T),(j,S) \in I_S.$$

Thus, $\boldsymbol{p}^0_S$ is APU-representable.

EC.4.4. Proof of Proposition 6

Proof. Recall the notation $S^0 = S \cup \{A\}$, and the collection of all G-MDM choice probability vectors $\boldsymbol{x}_A = (x_{i,A} : i \in A)$ for the new assortment A which are consistent with the observed choice data $\boldsymbol{p}_S$ be given by U_G. The main idea of the proof is that the second, third, and fourth sets of constraints of (21) model the closure of the G-MDM feasible region of I_{S^0} similar as the limit of MDM in Proposition 5. The first set of constraints of (21) fix $\boldsymbol{x}_S = \boldsymbol{p}_S$, which ensures that $(\boldsymbol{x}_A, \boldsymbol{p}_S) \in P_G(S^0)$ Next, we prove the constraints of (21) indeed models the closure of U_G by adjusting the proof of Proposition 5.

The optimal value of this optimization formulation is not affected by the presence of the following additional constraints: $0 \quad \lambda_S \quad \nu_i \quad 1$ for all $(i,S) \in I_{S^0}$. Indeed, this is because the signs of the differences $\{\lambda_S \quad \nu_i \quad \lambda_T + \nu_j : (i,S),(j,T) \in I_{S^0}\}$ are not affected by these additional constraints. Taking ϵ to be smaller than $1/(2n|S|)$, for example, ensures that there is a feasible assignment for $(\lambda_S \quad \nu_i : (i,S) \in I_{S^0})$ within the interval $[0,1]$ even if all these variables take distinct values.

Let F denote the feasible values for the variables $(\lambda_S : S \in S^0), (\nu_i : i \in N), (x_{i,S} : (i,S) \in I^0_S)$ satisfying the constraints as the closure of $P_G(S^0)$. Equipped with this feasible region F, we have the following deductions for $(\lambda_S : S \in S^0), (\nu_i : i \in N), (x_{i,S} : (i,S) \in I^0_S)$ in F: For any $(i,S),(j,T) \in I_{S^0}$ with $g(i) = g(j)$:

(i) we have $\lambda_S \quad \nu_i < \lambda_T \quad \nu_j$ if and only if $\delta_{i,j,S,T} = 1$ and $\delta_{j,i,T,S} = 0$, due to the second set of constraints of (21); in this case, we have from the third set of constraints of (21) that $0 \quad x_{i,T} \quad x_{i,S} \quad 1$;

(ii) likewise, we have $\lambda_S \quad \nu_i > \lambda_T \quad \nu_j$ if and only if $\delta_{i,j,S,T} = 0$ and $\delta_{j,i,T,S} = 1$, due to the second set of constraints; in this case, we have from the third set of constraints of (21) that $0 \quad x_{i,S} \quad x_{i,T} \quad 1$.

(iii) finally, $\lambda_S \quad \nu_i = \lambda_T \quad \mu_j$ if and only if $\delta_{i,j,S,T} = 0$ and $\delta_{j,i,T,S} = 0$; here we have from the fourth set of constriants of (21) that $x_{i,S} = x_{i,T}$.

Thus the binary variables $\{\delta_{i,j,S,T} : (i,S),(j,T) \in S^0$ with $g(i) = g(j)\}$ suitably model the constraints of $P_G(S^0)$. By adding the first set of constraints $\boldsymbol{x}_S = \boldsymbol{p}_S$. We have $\boldsymbol{x}_A \in U_G$. Therefore (21) computes the worst-case expected revenue $\underline{r}(A)$.

EC.4.5. Proof of Proposition 7

Proof. The proof of Proposition 7 follows directly from the proof of Proposition 6 by replacing $I_{S'}$ to I_S.

EC.5. An Algorithm for Evaluating the Limit of MDM when $|S|$ is Small

Algorithm 1: An algorithm solves the limit of MDM polynomial in n

Input: Observed choice probabilities $\boldsymbol{p}_S$, collection $\mathcal{S}$, product universe N.

Output: MDM choice probabilities $\boldsymbol{x}_S$, optimal loss f , optimal ranking of assortments τ .

$\boldsymbol{T}$ {all rankings of $(S : S \in \mathcal{S})$};

f $+\infty$ keeps tracking of the optimal value of Problem (15);

$\boldsymbol{x}_S$ $\boldsymbol{0}$ keeps tracking of the optimal solution;

for $\tau \in \boldsymbol{T}$ do

  Solve

$$\begin{aligned} \min_{x_S} \quad & \text{loss}(\boldsymbol{x}_S, \boldsymbol{p}_S) \\ \text{s.t.} \quad & x_{i,S} \quad x_{i,T} \quad 0, \quad \text{if } \tau(T) < \tau(S) \quad \forall (i,S),(i,T) \in I_S, \\ & \sum_{i \in S} x_{i,S} = 1, \quad \forall S \in \mathcal{S}, \\ & x_{i,S} \quad 0, \quad \forall (i,S) \in I_S. \end{aligned} \tag{Limit-LP}$$

  f the output optimal objective value of (Limit-LP);

  $\boldsymbol{x}_S$ the output optimal solution of (Limit-LP);

  if $f < f$ then

    $\boldsymbol{x}_S$ $\boldsymbol{x}_S$;

    f f ;

    τ τ;

  end

end

In Algorithm 1, for a fixed $\boldsymbol{\lambda}$, we just need to solve a convex optimization problem with $O(n|S|)$ continuous variables and $O(n|S|^2)$ linear constraints to compute the limit loss. There are $m!$ possible $\boldsymbol{\lambda}$. Thus, Algorithm 1 is polynomial in the alternative size n.

EC.6. Illustrative Examples

EC.6.1. An example to show the non-convexity of MDM feasible region

Example EC.3. $\boldsymbol{x}_S$ is MDM-representable because $x_{1,A} < x_{1,B}$,$x_{2,A} < x_{2,C}$ and $x_{3,B} < x_{3,C}$ implies $\lambda_A > \lambda_B$, $\lambda_A > \lambda_C$ and $\lambda_B > \lambda_C$. The values $\lambda_A = 12, \lambda_B = 10, \lambda_C = 8$ satify this. $\boldsymbol{y}_S$ is MDM-representable because $y_{1,A} > y_{1,B}$, $y_{2,A} > y_{2,C}$ and $y_{3,B} > y_{3,C}$ implies $\lambda_A < \lambda_B$, $\lambda_A < \lambda_C$ and $\lambda_B < \lambda_C$. The values $\lambda_A = 8, \lambda_B = 10, \lambda_C = 12$ satify this. $\boldsymbol{w} = 0.4\boldsymbol{x}_S + 0.6\boldsymbol{y}_S$ is a convex combination of $\boldsymbol{x}_S$

and $\boldsymbol{y}_S$ but it can not be represented by MDM because $w_{1,A} > w_{1,B}$, $w_{1,A} < w_{2,C}$ and $w_{3,B} > w_{3,C}$ which implies $\lambda_A < \lambda_B$, $\lambda_A > \lambda_C$ and $\lambda_B < \lambda_C$, i.e., $\lambda_B < \lambda_C < \lambda_A < \lambda_B$. This means $\boldsymbol{w}$ can not be represented by MDM.

$\boldsymbol{x}_S$

Alternative	A={1,2}	B={1,3}	C={2,3}
1	0.3	0.9	
2	0.7		0.8
3		0.1	0.2

$\boldsymbol{y}_S$

Alternative	A={1,2}	B={1,3}	C={2,3}
1	0.75	0.1	
2	0.25		0.2
3		0.9	0.8

$\boldsymbol{w} = 0.4\boldsymbol{x}_S + 0.6\boldsymbol{y}_S$

Alternative	A={1,2}	B={1,3}	C={2,3}
1	0.57	0.42	
2	0.43		0.44
3		0.58	0.56

EC.6.2. An example to show the grouping effect

If we have two groupings of the products denoted by $\mathcal{G}_1$ and $\mathcal{G}_2$ where each group in $\mathcal{G}_1$ is a subset of a group in $\mathcal{G}_2$, then the set of choice probabilities that G-MDM captured with $\mathcal{G}_2$ is a subset of that captured with $\mathcal{G}_1$. We provide an example as follows.

Example EC.4. Consider the choice probabilities in Table EC.6 below with $n = 4$ products over two assortments S and T. This can be represented by G-MDM only when the number of groups $K \quad 2$.

Table EC.6 Choice probabilities that cannot be generated with a single group for $n = 4, m = 2$.

Alternative	$S = \{1,2,3\}$	$T = \{1,2,4\}$
1	$p_{1,S} = 0.28$	$p_{1,T} = 0.25$
2	$p_{2,S} = 0.40$	$p_{2,T} = 0.20$
3	$p_{3,S} = 0.32$	-
4	-	$p_{4,T} = 0.55$

If all the alternatives are in the same group, using Corollary 3, we have $\lambda(S) \quad \nu_2 < \lambda(S) \quad \nu_1 < \lambda(T) \quad \nu_1 < \lambda(T) \quad \nu_2$ since $p_{2,S} > p_{1,S} > p_{1,T} > p_{2,T}$, which implies $\nu_2 < \nu_1$ and $\nu_2 > \nu_1$. This means the choice probabilities in Table EC.6 cannot be represented by G-MDM with $K = 1$. Now consider $K = 2$ and $\mathcal{G} = \{\{1,3\},\{2,4\}\}$. Then we have $\lambda(S) \quad \nu_3 < \lambda(S) \quad \nu_1 < \lambda(T) \quad \nu_1$ since $p_{3,S} > p_{1,S} > p_{1,T}$ and $\lambda(T) \quad \nu_4 < \lambda(S) \quad \nu_2 < \lambda(T) \quad \nu_2$ since $p_{4,T} > p_{2,S} > p_{2,T}$. The conditions are satisfied for $\lambda(S) = 1, \lambda(T) = 3, \nu_1 = 1, \nu_2 = 1.5, \nu_3 = 3, \nu_4 = 4$, since we get $\quad 2 < 0 < 2$ and $\quad 1 < \quad 0.5 < 1.5$. This means the choice probabilities in Table EC.6 can be represented by G-MDM with $K = 2$.

EC.6.3. An example to show the nonconvexity of G-MDM feasible region

Example EC.5. We focus on a single group with $K = 1$. Both the choice probabilities $\boldsymbol{p}_S$ and $\boldsymbol{q}_S$ are G-MDM-representable but the convex combination of $\boldsymbol{p}_S$ and $\boldsymbol{q}_S$ can not be represented by G-MDM with $K = 1$. One can check $\boldsymbol{p}_S$ can be represented by G-MDM where $p_{1,A} < p_{2,A} < p_{1,C} < p_{1,B} < p_{2,B} < p_{3,A} < p_{3,C}$ implies $\lambda_A \quad \nu_1 > \lambda_A \quad \nu_2 > \lambda_C \quad \nu_1 > \lambda_B \quad \nu_1 > \lambda_B \quad \nu_2 > \lambda_A \quad \nu_3 > \lambda_C \quad \nu_3$.

The values $\lambda_A = 12, \lambda_B = 8, \lambda_C = 10$ and $\nu_1 = 3, \nu_2 = 4, \nu_3 = 10$ satisfy this. Similarly, $\boldsymbol{q}_S$ can be represented by G-MDM where $q_{3,A} < q_{3,C} < q_{2,A} < q_{2,B} < q_{1,A} < q_{1,B} < q_{1,C}$ implies $\lambda_A \quad \nu_3 > \lambda_C \quad \nu_3 > \lambda_A \quad \nu_2 > \lambda_B \quad \nu_2 > \lambda_A \quad \nu_1 > \lambda_B \quad \nu_1 > \lambda_C \quad \nu_1$. The values $\lambda_A = 11, \lambda_B = 10, \lambda_C = 8$ and $\nu_1 = 10, \nu_2 = 4, \nu_3 = 0$ satisfy this. However $\boldsymbol{r}_S$ can not be represented by G-MDM since we have $r_{1,A} > r_{2,A}$ which implies $\nu_1 > \nu_2$ and $r_{1,B} = r_{2,B} > 0$ which implies $\nu_1 = \nu_2$, both of which cannot be simultaneously satisfied.

$\boldsymbol{p}_S$

Alternative	A={1,2,3}	B={1,2}	C={1,3}
1	0.1	0.4	0.25
2	0.2	0.6	-
3	0.7	-	0.75

$\boldsymbol{q}_S$

Alternative	A={1,2,3}	B={1,2}	C={1,3}
1	0.6	0.65	0.8
2	0.3	0.35	-
3	0.1	-	0.2

$\boldsymbol{r}_S = 0.6\boldsymbol{p}_S + 0.4\boldsymbol{q}_S$

Alternative	A={1,2,3}	B={1,2}	C={1,3}
1	0.3	0.5	0.47
2	0.24	0.5	-
3	0.46	-	0.53

EC.7. Additional Useful Details on the Experiments Presented in the Paper

In this section, we give details on the implementation of the experiments. We used a MacBook Pro Laptop with a 2 GHz 4 core Intel Core i5 processor for all experiments.

EC.7.1. Implementation Details of Experiment 1

Data Generation of Experiment 1

(1) Collection information.

(a) A size of alternatives $n = 1000$ where the collection size $|S| = m$ varies from 100 to 1000 in steps of 100.

(b) In a run, the collection with a smaller size is nested by the one with a larger size.

(c) In each instance, each alternative is chosen into an assortment with the same probability $p = 0.005$. This ensures that the average size of the assortments in the data is about 5. The distinct assortments with at least size 2 are randomly generated with no repeat. Across different instances, the generation of assortments in the same collection size is independent.

(2) Observed choice probabilities.

(a) Given the colletion information, $\boldsymbol{p}_S$ is generated as a MNL instance where the deterministic utilities in MNL follow a standard normal distribution.

(b) The choice probabilities $\boldsymbol{p}_S$ are perturbed by Gaussian noise with mean 0 and standard deviation 0.01. After the perturbation, regularization of all choice probabilities and normalization of the choice probabilities in an assortment are applied. Let α denote the proportion of entries of $\boldsymbol{p}_S$ being perturbed. We test the instances with $\alpha = \{0.25, 0.5, 0.75, 1\}$ respectively. Let $\tilde{\boldsymbol{p}}_S$ denote the perturbed choice probabilities. We have $\tilde{\boldsymbol{p}}_S = \boldsymbol{p}_S(\mathbf{1} + \boldsymbol{\epsilon}\boldsymbol{\delta})$, where $\mathbf{1}$ is a all ones matrix, $\boldsymbol{\epsilon}$ is the Gaussian noise and $\boldsymbol{\delta}$ is a matrix of binary variables to indicate whether the entries are chosen to be modified. We have $\mathbb{P}(\delta_{i,S} = 1) = \alpha, \forall (i, S) \in \mathcal{I}_S$.

Checking the representability of MDM For a collection of observed choice probabilities $\boldsymbol{p}_S$, according to Theorem 1, we check the representability of MDM with the following linear program:

$$\begin{aligned} \max_{\epsilon} \quad & \epsilon \\ \text{s.t.} \quad & \lambda_S \quad \lambda_T \quad \epsilon \quad 0, \text{ if } p_{i,S} < p_{i,T} \quad \forall (i,S),(i,T) \in \mathcal{I}_S, \\ & \lambda_S \quad \lambda_T = 0, \text{ if } p_{i,S} = p_{i,T} > 0 \quad \forall (i,S),(i,T) \in \mathcal{I}_S. \end{aligned} \tag{EC.14}$$

If the optimal value of (EC.14) is strictly positive, then $\boldsymbol{p}_S$ can be represented by MDM. Otherwise, $\boldsymbol{p}_S$ cannot be represented by MDM.

Checking the representability of MNL For a collection of observed choice probabilities $\boldsymbol{p}_S$, we check the representability of MNL with the following linear program:

$$\begin{aligned} \max \quad & \sum_{i \in N} \nu_i \\ \text{s.t.} \quad & p_{i,S} \sum_{j \in \nu_j} \quad \nu_i = 0, \quad \forall (i,S) \in \mathcal{I}_S, \\ & \nu_i \quad 0, \quad \forall i \in N. \end{aligned} \tag{EC.15}$$

If the optimal value of (EC.15) is strictly positive, then $\boldsymbol{p}_S$ can be represented by MNL. Otherwise, $\boldsymbol{p}_S$ cannot be represented by MNL.

Checking the representability of RUM For a collection of observed choice probabilities $\boldsymbol{p}_S$, we check the representability of RUM with the following linear program:

$$\begin{aligned} \max \quad & 0 \\ \text{s.t.} \quad & \sum_{\sigma \in \quad _n} \lambda(\sigma) \mathrm{I}[\sigma, i, S] \quad p_{i,S} = 0, \quad \forall (i,S) \in \mathcal{I}_S, \\ & \sum_{\sigma \in \quad _n} \lambda(\sigma) = 1, \quad \lambda(\sigma) \quad 0, \quad \forall \sigma \in \quad _n. \end{aligned} \tag{EC.16}$$

If (EC.16) is feasible, then $\boldsymbol{p}_S$ can be represented by RUM. Otherwise, $\boldsymbol{p}_S$ cannot be represented by RUM.

In Experiment 1, for each α, the representability of a model is tested over 1000 instances of the same collection size. We report the proportion of the representable instances of the tested model and an average of computational time over these 1000 instances for each collection size.

EC.7.2. Implementation Details of Experiment 2

Data Generation of Experiment 2. Product size $n = 7$ with a collection size m taking values in $\{2, 3, 5, 10, 15, 20\}$ was setup in the experiment comparing the representational power and the computational time of MDM and RUM. We stop at the product size 7 because of RUM is intractable even for a small value of n. All other setups are the same as Experiment 1 except the generation of random assortments. Since the size of the available alternatives is small, we test on the cases where collections are randomly chosen from assortments with size 2 or 3.

The implementation details of Experiment 2 are the same as Experiment 1.

EC.7.3. Implementation Details of Experiment 4

Data Generation of Experiment 4 All steps of generating instances are the same of Experiment 2 but we now use uniformly distributed choice data instead of using the underlying MNL choice data with gaussian noise.

Limit computations for MDM, RUM and MNL In Experiment 3 of Section 7, we set the loss function to be $\min_{x_S} \sum_{S \subseteq \mathcal{S}} T_S \sum_{i \in S} p_{i,S} | x_{i,S} \quad p_{i,S} |$ and report the average loss and standard error of the limit of each model over 1000 instances. We use Gurobi solver for the limit computation of MDM and RUM and CVXPY solver for the Maximum Likelihood Estimation of MNL. For each instance $\boldsymbol{p}_S$, the limit of MDM can be computed by solving (16). The limit of RUM can be computed by solving a convex optimization problem with the constraints in Problem (EC.16) and the 1-norm objective function. For the computation of the limit of MNL, we first compute the Maximum Likelihood Estimator (MLE) of the parameter in the MNL model, where the MLE can be obtained by solving the following loglikelihood function

$$\begin{aligned} \text{argmax} \quad ll(\boldsymbol{\nu} | \boldsymbol{p}_S) &= \sum_{S \subseteq \mathcal{S}} T_S \sum_{i \in S} p_{i,S} \log\left(\frac{\exp v_i}{\sum_{j \in S} \exp v_j}\right) \\ &= \sum_{S \subseteq \mathcal{S}} T_S \left(\sum_{i \in S} p_{i,S} v_i \quad \log \sum_{j \in S} \exp v_j\right). \end{aligned}$$

Then, we compute the loss between the choice probability collection and the probability collection with the estimated MLE of MNL.

EC.7.4. Implementation Details of Experiment 7

For a given instance $\boldsymbol{p}_S$, checking the representability of MDM, MNL can be done by (EC.14) and (EC.15). WeWe check of the representability of the regular model by checking if $\boldsymbol{p}_S$ satisfies the following inequities:

$$p_{i,S} \quad p_{i,T} \text{ if } T \quad S, \quad \forall (i,S), (i,T) \in I_S.$$

EC.8. Numerical Experiments with G-MDM

Experiment 8 is devoted to exploring how imparting domain knowledge into the model via grouping reduces the length of prediction intervals (thereby pointing to lesser data requirements). Experiment 9 validates the effectiveness of the grouping identification procedure.

EC.8.0.1. The grouping effect of G-MDM The results of Experiment 8, reported in Figure EC.3, show how grouping alternatives leads to significantly narrower prediction intervals for choice probabilities in unseen assortments. Varying m = the number of assortments for which choice data are available, we utilize the mixed integer linear programs in Proposition 2 and 6 to construct and report prediction interval lengths averaged over revenues of unseen assortments. As expected, both

models lead to tighter intervals when more choice data is made available for training. However, the predicted intervals of the model using grouping information are 86% - 91% narrower than the model assuming no groups when m ranging from 10 to 15.

The effectiveness of the grouping procedure The results of Experiment 9 demonstrate the effectiveness of the group identification procedure. In this experiment, we generate 50 random underlying G-MDMs instances with a product size of 7 and a collection size ranging among {20, 60, 80, 100} and develop a procedure to identify group information by utilizing the representable conditions of G-MDM. The procedure begins by assessing the dissimilarity of the stochastic noise distributions of any pair of alternatives. Given an instance $\boldsymbol{p}_S$, we define $\boldsymbol{D} = (D(i,j) : i,j \in N)$ as a matrix with $D(i,j) = D(j,i)$ where $i \neq j$, and $D(i,i) = 0$, $\forall i \in N$. For any two alternatives i and j where $i \neq j$, we can explain $D(i,j)$ as the violation of representable conditions if alternative i and j are grouped together. Taking any MDM-representable choice data $\boldsymbol{p}_S$ as the input, $D(i,j)$ is computed as follows. Let $C_1 = \{(i,j,S,T) : (i,S),(j,T) \in I, p_{i,S} < p_{j,T}\}$ and $C_2 = \{(i,j,S,T) : (i,S),(j,T) \in I, 0 < p_{i,S} = p_{j,T}\}$. Let I_{C_1} and I_{C_2} denote the set of indices of the tuples in C_1 and C_2 respectively. We solve the following linear program:

$$
\begin{aligned}
\min_{y,\, \lambda, \nu, z,\, \epsilon} \quad & \sum_{k \in I_{C_1}} y_k + \sum_{l \in I_{C_2}} z_l \\
\text{s.t.} \quad & \lambda_S - \nu_i - \lambda_T + \nu_j - y_k \leq 0, && \forall k \in I_{C_1}, \\
& y_k \geq 0, \quad y_k \geq p_{i,S} - p_{j,T} - \epsilon, && \forall k \in I_{C_1}, \\
& \lambda_S - \nu_i - \lambda_T + \nu_j + \eta_l = 0, && \forall l \in I_{C_2}, \\
& \eta_l \geq -z_l, \quad \eta_l \leq z_l, && \forall l \in I_{C_2}, \\
& \lambda_S - \lambda_T \geq \epsilon \text{ if } p_{h,S} < p_{h,T}, && \forall (h,S),(h,T) \in I_S, \\
& \lambda_S - \lambda_T = 0 \text{ if } 0 < p_{h,S} = p_{h,T}, && \forall (h,S),(h,T) \in I_S,
\end{aligned}
\tag{EC.17}
$$

with $0 < \epsilon < 1/(2|S|)$. Using $\boldsymbol{D}$ as the input distance matrix and setting different numbers of groups K, K-Means clustering is performed to find the groups that minimize the aggregated distance. The elbow method is then used to determine the optimal number of groups under $\boldsymbol{p}_S$. We evaluate the clustering accuracy using this method over 50 instances for different collection sizes and report the average accuracy and the standard deviation in Figure EC.4. As more choice data is available for training, the clustering accuracy improves and the error decreases. Moreover, the identification procedure is able to achieve high accuracy for all collection sizes with underlying G-MDMs.

EC.9. Additional Useful Details on Experiments Relating to G-MDM

EC.9.1. Implementation Details of Experiment 8

The instances that are feasible for G-MDM with $K = 1$ in Experiment 2 with perturbation parameter $\alpha = 0.5$ are used for the prediction experiments. For each instance, a tested unseen assortment of

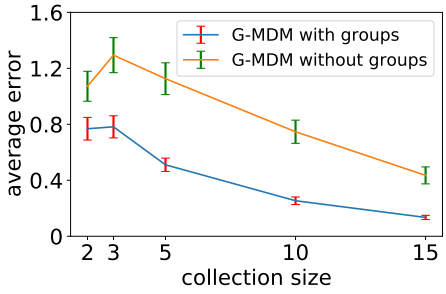


Figure EC.3 The grouping effect of G-MDM in Experiment 8

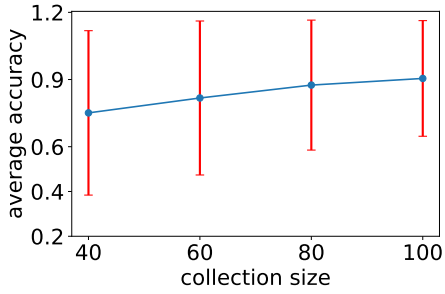


Figure EC.4 The grouping accuracy with MDM instances in Experiment 9

size 2 or 3 is randomly generated. For each instance, we utilize the MILPs in Proposition 2 and 6 to compute both predicted revenue intervals for the alternatives in the unseen assortment under G-MDM with 1 group and G-MDM without grouping assumption (K=n). We report the average length of the predicted intervals and standard error of each scenario over the testing instances.

EC.9.2. Implementation Details of Experiment 9

Data generation with underlying MDM choice data

Step 1: Uniformly generate $\nu_i, i \in N$.

Step 2: Randomly generate a collection of assortments $(S : S \in S, S \subseteq N, |S| \geq 2)$ with a given collection size.

Step 3: Consider two different ground-truth distribution $F_1(\cdot)$ and $F_2(\cdot)$ for G_1 and G_2 respectively.

Step 4: For each assortment, S with $S \in S$, compute the value of dual multiplier λ_S and the choice probabilities via bisection search with the following equations are satisfied:

$$p_{i,S} = 1 - F_{g(i)}(\lambda_S - \nu_i), \text{ and } \sum_{i \in S} p_{i,S} = 1.$$

The size of the collections of assortments varies from 40 to 100 in steps of 20. For each assortment collection size, we generate 1000 instances. For each instance, define $\boldsymbol{D} = (D(i,j) : i,j \in N)$ to measure the dissimilarity of the marginals of alternatives. Take the $\boldsymbol{p}_S$ as the input and compute the distance between any pair of alternatives (i,j) with $i,j \in N, i \neq j$ via solving Problem (EC.17). Set the number of clusters as 2 and perform K-Means clustering with $\boldsymbol{D}$. We measure the clustering accuracy by V-Measure with the true clusters and the clustering results. We report the average accuracy under each collection size.

The correctness of (EC.17) is provided as follows. $D(i,j)$ takes the optimal value of Problem (EC.17). $D(i,j)$ can be interpreted as the distance of the observed choice probabilities of alternative i and j to a G-MDM with i and j are grouped together and other alternatives are singletons. To see this, y_k's are the variables taking the negative parts of y_k's. Since the objective minimizes y_k's, y_k's

are the negative parts of $\lambda_S \quad \nu_i \quad \lambda_T + \nu_j$ when $p_{i,S} < p_{j,T}$, meaning the violation of the inequality conditions of (20) if alternative i and j are grouped together. if alternative i and j are grouped together. z_l's are the variables taking the absolute values of η_l's. Since the objective minimizes z_l's, z_l's are the absolute values of $\lambda_S \quad \nu_i \quad \lambda_T + \nu_j$ when $p_{i,S} = p_{j,T} > 0$, meaning the violation of the equality conditions of (20) if alternative i and j are grouped together. The last two constraints ensure that the variables $\boldsymbol{\lambda}$ satisfies the MDM-represetable conditions 5.