Consider or Choose? The Role and Power of Consideration Sets

Yi-Chun Akchen
School of Management, University College London, London E14 5AB, United Kingdom, yi-chun.akchen@ucl.ac.uk

Dmitry Mitrofanov
Carroll School of Management, Boston College, dmitry.mitrofanov@bc.edu

Consideration sets play a crucial role in discrete choice modeling, where customers often form consideration sets in the first stage and then use a second-stage choice mechanism to select the product with the highest utility. While many recent studies aim to improve choice models by incorporating more sophisticated second-stage choice mechanisms, this paper takes a step back and goes into the opposite extreme. We simplify the second-stage choice mechanism to its most basic form and instead focus on modeling customer choice by emphasizing the *role* and *power* of the first-stage consideration set formation. To this end, we study a model, parameterized solely by a distribution over consideration sets with a bounded rationality interpretation. Intriguingly, we show that this model is characterized by the axiom of symmetric demand cannibalization, enabling complete statistical identification. The latter finding highlights the *critical role* of consideration sets in the identifiability of two-stage choice models. We also examine the model's implications for assortment planning, proving that the optimal assortment is revenue-ordered within each partition block created by consideration sets. Despite this compelling structure, we establish that the assortment problem under this model is NP-hard even to approximate, highlighting how consideration sets contribute to nontractability, even under the simplest uniform second-stage choice mechanism. Finally, using real-world data, we show that the model achieves prediction performance comparable to other advanced choice models. Given the simplicity of the model's second-stage phase, this result showcases the enormous *power* of first-stage consideration set formation in capturing customers' decision-making processes.

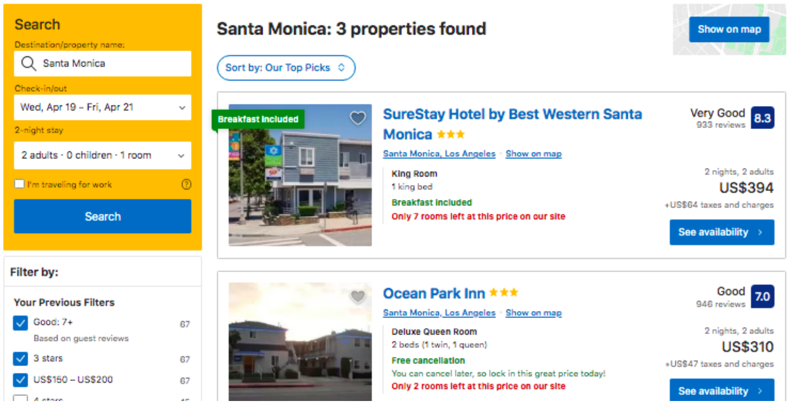
Key words: discrete choice, consideration sets, symmetric cannibalization, assortment optimization, bounded rationality, identification

History: First version: February 8, 2023; Second version: June 12, 2024; This version: February 19, 2025. Forthcoming in *Management Science*.

1. Introduction

The rise of the digital economy and technological advancements has fundamentally changed how we shop and make purchasing decisions. With an overwhelming variety of products and services available online, consumers, when making a choice, must navigate an enormous amount of information, including detailed product descriptions, customer reviews, and ratings. In an ideal scenario, fully rational individuals would make their purchase decisions by thoroughly assessing the features of each alternative, calculating its utility, and selecting the option with the highest utility, provided they have unlimited time and resources to perform such an evaluation. In reality, individuals have physical and cognitive limitations and thus only consider a subset of the available alternatives. Economists and psychologists term such smaller sets of alternatives as "consideration sets" (Wright and Barbour 1977) or "evoked sets" (Howard and Sheth 1969, Brisoux and Laroche 1981). The notion of consideration sets has been well-documented in the marketing literature when studying consumer behavior (Hauser 2014) and various heuristics have been proposed to model the formation of the consideration sets, such as screening rules based on product prices or other features (Pras and Summers 1975, Gilbride and Allenby 2004, Jedidi and Kohli 2005). The concept of consideration sets is further supported by the psychology literature, which questions consumers' ability to consistently evaluate every product within the offer set (Miller 1956, Hauser and Wernerfelt 1990, Iyengar and Lepper 2000).

In fact, the concept of consideration set formation goes beyond merely interpreting consumer behavior; it also has substantial practical implications, particularly in developing more comprehensive and accurate choice models. To this end, the seminal work by Hauser (1978) uses a goodness-of-fit statistic to demonstrate that consideration sets can explain nearly three-quarters of the variation in choice data, whereas a logit model, which is based solely on consumer preferences, explains only a quarter. This insight has contributed to the widespread adoption of *two-stage* choice models that integrate consideration sets into consumers' decision-making processes and many papers provide evidence of their superior prediction performance (Silk and Urban 1978, Hauser and Gaskin 1984, Gensch 1987). Within this two-stage framework, consumers initially form consideration sets, often using screening rules and simple heuristics (Hutchinson and Gigerenzer 2005). In the second stage, they apply a *choice mechanism* to select and purchase the product that maximizes their utility from the options in the consideration set (Ben-Akiva and Boccara 1995, Shocker et al. 1991). Specifically, we can illustrate two-stage choice models with an example of online hotel booking, where customers face an abundance of alternatives and information. On a travel website, they encounter details such as the hotel's location, services, amenities, quality, view, ratings, photos, and reviews, among other characteristics. Analyzing all this information for every hotel within a limited time is impractical. Instead, customers might use simple heuristics to narrow their options. Figure 1 shows a customer applying screening criteria – rating (7+), price ($150-$200 per night), and three-star quality – to reduce five hundred options to a consideration set of three alternatives. A customer then is expected to use a choice mechanism to evaluate these three considered alternatives and select the one offering the highest utility.



Destination/property name: Santa Monica
Check-in/out: Wed, Apr 19 – Fri, Apr 21
2-night stay
2 adults · 0 children · 1 room
□ I'm traveling for work

Filter by:
Your Previous Filters
☑ Good: 7+ 67
Based on guest reviews
☑ 3 stars 67
☑ US$150 – US$200 67

Santa Monica: 3 properties found

SureStay Hotel by Best Western Santa Monica ★★★
Santa Monica, Los Angeles · Show on map
Very Good 8.3
933 reviews
King Room
1 king bed
Breakfast Included
Only 7 rooms left at this price on our site
2 nights, 2 adults
US$394
+US$64 taxes and charges

Ocean Park Inn ★★★
Santa Monica, Los Angeles · Show on map
Good 7.0
946 reviews
Deluxe Queen Room
2 beds (1 twin, 1 queen)
Free cancellation
You can cancel later, so lock in this great price today!
Only 2 rooms left at this price on our site
2 nights, 2 adults
US$310
+US$47 taxes and charges

Figure 1 **An example of a two-stage decision-making process, where a customer uses screening rules to form a consideration set.**

Given the evidence showing the importance of capturing customers' consideration set formation, it is unsurprising that two-stage choice models incorporating this process have gained significant popularity over the past several decades. Many of these models leverage advanced choice mechanisms to further enhance their predictive accuracy and data-fitting capabilities (Aouad et al. 2021, Jagabathula et al. 2024). A common feature of these models is their foundation in nonparametric choice modeling principles, drawing inspiration from machine learning models and algorithms. Their primary characteristic is an advanced and refined second-stage choice mechanism which can be as sophisticated as the choice mechanisms proposed by Farias et al. (2013) and Chen and Mišić (2022). In this paper, we adopt an opposing approach: instead of exploring more complex second-stage choice mechanisms, we take a step back to examine the fundamental properties of decision-making processes driven solely by consideration sets, where the second-stage choice mechanism is kept as simple as possible. More specifically, we examine a class of nonparametric models defined solely by a distribution over consideration sets, which we refer to as the *consideration set model (CSM)*. Through the lens of this generic consideration-based model, we examine the role and power of consideration sets in choice model representation, demand cannibalization, choice prediction, and assortment optimization. We also discuss the implications for other choice models that incorporate consideration set structures. Specifically, we make the following contributions:

- *Considering rather than choosing.* We study a nonparametric choice model which is fully characterized by a distribution over the consideration sets. This distribution can be viewed as the probability mass of various customer types in the market, or it can indicate the stochasticity in how a customer forms a consideration set. After forming a consideration set, we assume that the customer does not rely on a particular choice mechanism to select the "best" product among those considered, but instead chooses any of them in a uniformly random way. Overall, this consideration set model represents a *simplified* version of the general class of two-stage choice models, as it does not incorporate a specific second-stage choice mechanism.

- *The role of consideration sets in choice model identifiability.* Surprisingly, we show that the consideration set model can be fully identified from the collection of choice probabilities. We further provide a closed-form expression on how one can compute the model parameters by means of the choice probabilities. In other words, the consideration set model can be uniquely reconstructed from its observed choice probabilities on assortments, making it one of the identifiable choice models with the highest degree of freedom in the existing literature. This finding on identifiability conveys an important insight: if a two-stage choice model cannot be uniquely identified – having several solutions that fit the data equally well even with unlimited choice data – it is not solely because of the first-stage consideration process. In other words, consideration sets themselves do not play a *critical role* in choice model nonidentifiability.
- *The role of the consideration sets in modeling demand cannibalization.* We also characterize the necessary and sufficient conditions for the choice probabilities to be consistent with our model. To this end, we demonstrate that with a mild condition imposed on the no-purchase probabilities, a choice model is a consideration set model if, and only if, the introduction of a product symmetrically reduces the market share of another product in each assortment. We refer to this condition as the *symmetric demand cannibalization axiom.* More specifically, this axiom is based on the assumption that for any two products i and j, the decrease in demand for product i when product j is introduced is the same as the decrease in demand for product j when product i is introduced. This axiomatic characterization highlights a fundamental limitation of any choice model based solely on consideration set formation, as consideration sets alone are only capable of capturing symmetric demand cannibalization. It also demonstrates that the choice mechanism, rather than the consideration set formation in the first stage, plays a significant role in representing demand cannibalization comprehensively.
- *The critical role of consideration sets in driving the intractability of the assortment problem.* We also explore the impact of consideration sets on downstream applications. In particular, we focus on assortment optimization problems. We first provide a precise characterization of the optimal assortment, showing that it is revenue-ordered within each block belonging to the partition of the product universe induced by the consideration sets. This structural property leads to a polynomial-time optimal algorithm when the number of customer types is fixed. However, even with this compelling optimal assortment structure in place, the problem of assortment optimization under the general consideration set model remains NP-hard to approximate within a factor of $O(n^{1/2-\epsilon})$ for any $\epsilon > 0$. This result implies that accounting for consideration sets in choice models inherently complicates downstream applications, even if the second-stage choice mechanism is significantly simplified.

- *Empirical study: the prediction power of consideration sets.* Finally, we conduct numerical experiments based on a real-world dataset (Bronnenberg et al. 2008) to show the prediction power of consideration sets when modeling consumer choice. We propose an estimation methodology to calibrate the consideration set model and empirically evaluate its predictive performance. Despite being a special case of the ranking-based and the mixed multinomial logit (mixed MNL) models, the consideration set model provides a predictive performance that is highly competitive with these more general models. This emphasizes that consideration sets *alone* can be responsible for the strong predictive performance of the two-stage consider-then-choose models in real-world prediction tasks.

1.1. Related Literature on Consideration Sets

The concept of consideration sets originated in the fields of marketing and psychology and has been studied extensively for several decades, with numerous studies exploring how consumers form and use consideration sets in their purchasing decisions. For a more detailed overview, we refer readers to the survey papers by Roberts and Lattin (1997) and Hauser (2014). Overall, it is widely accepted in the literature that consumers generally make decisions through a two-stage process (Swait and Ben-Akiva 1987). Specifically, consumers first consider a subset of products (i.e., form a consideration set) and then select one item from that set to purchase. In fact, abundant empirical evidence supports the concept of consideration sets. For example, Hauser (1978) adapts an information-theoretic viewpoint and shows that a model which incorporates the consideration set concept can explain up to 78% of the variation in the customer choice. Then, Hauser and Wernerfelt (1990) empirically investigate the consideration set size for various product categories and report that the mean consideration set size, which is the number of brands that a customer considers before making a choice, is relatively small: 3.9 for deodorant, 4.0 for coffee, and 3.3 for frozen dinner. To this end, our numerical analysis in Section 5.3, examining the size of consideration sets across various product categories using real-world grocery sales data, is consistent with the finding that consideration set sizes tend to be relatively small.

Over the past few decades, numerous theories and models have been developed to explain how consumers form their consideration sets. From a decision theory perspective, consumers are expected to continue searching for new products as long as the utility (or satisfaction) gained from finding a better option exceeds the cost of the search (Ratchford 1982, Roberts and Lattin 1991). Additionally, cognitive limitations may lead consumers to rely on simple heuristic rules to construct their consideration sets. These rules include elimination by aspects (Tversky 1972), conjunctive and disjunctive rules (Pras and Summers 1975, Brisoux and Laroche 1981, Laroche et al. 2003, Gilbride and Allenby 2004, Jedidi and Kohli 2005), and compensatory rules (Keeney et al. 1993, Hogarth and Karelaia 2005). In this paper, we take a more general and flexible perspective on the formation of consideration sets. Rather than focusing on specific rules or heuristics that consumers use to construct their consideration sets, we model the possibility that a consideration set sampled by a customer could be any subset of products. To this end, our paper is related to the work of Jagabathula et al. (2024), where the authors model customer choice using the consider-then-choose (CTC) model, which is parameterized by the joint distribution over consideration sets and rankings. Different from the CTC model, which uses a mixture of rankings as its second-stage choice mechanism, our model employs the simplest possible choice mechanism. Consequently, it can be viewed as a special case of the model proposed by Jagabathula et al. (2024). Furthermore, our model is also a special case of the ranking-based model (Block et al. 1959, Farias et al. 2013) and the mixed MNL model (Train 2009) as discussed in Section 2.3.

More recently, the concept of a consideration set has been gaining a lot of attention in the field of operations management. Feldman et al. (2019) propose an algorithm for determining the optimal assortment, incorporating the unique features of the ranking-based model and the assumption of small consideration sets. Similarly, Aouad et al. (2021) examine assortment optimization under the CTC choice model, where customers select the product with the highest rank from the intersection of their consideration set and the offered assortment. In Wang and Sahin (2018), the authors present a mathematical model that represents the trade-off between a product's expected utility and the search cost associated with it. Mitrofanov et al. (2024) investigate the assortment optimization problem under a two-stage choice model, where customers initially use non-parametric dominance relationships to narrow down their options and then make a selection from the shortlisted products using the multinomial logit (MNL) model. Chitla et al. (2023) use the consideration set concept in order to build the structural model and study the multihoming behavior of users in the ride-hailing industry. Interestingly, it was shown in the paper by Jagabathula et al. (2024) that the consideration set approach is particularly advantageous in the online platform setting or in the grocery retail setting where we might expect the noise in the sales transaction data because of the stockouts. To this end, there is a lot of evidence in the literature that online platforms might not have real-time information on the product availability in the grocery retail stores which could be the major cause for the stockouts (Knight and Mitrofanov 2024) despite the AI-enabled technology to alleviate this problem (Knight et al. 2023, Kim et al. 2024).

2. Model Description

In this section, we begin with an overview of choice modeling before introducing the *consideration set model*. Then, we connect this model to other choice models in the economics, marketing, and operations literature.

2.1. Preview of Choice Modeling and Motivation

We begin this subsection by introducing several notations. We consider a universe $N \equiv \{1, 2, \ldots, n\}$ of n products. Each assortment or offer set S is a subset of N, i.e., $S \subseteq N$. When a set of products S is offered, a customer or an agent chooses either a product in S or the "default" option 0. Depending on the context, the default option can be the no-purchase option or any product outside the universe N. Following the standard assumption in the literature, the default option is always assumed to be available to customers. Then, to simplify notation, we denote $N^+ = N \cup \{0\}$ and $S^+ = S \cup \{0\}$ for any $S \subseteq N$. Throughout the paper, $\mathbb{I}[A]$ or $\mathbb{I}_A$ denotes the indicator function that is equal to 1 if condition A is satisfied and 0, otherwise.

A choice model can be described by a mapping or function $\mathbb{P}$ that takes as input an assortment S and outputs a probability distribution over the elements of S^+. This probability distribution represents the likelihood of each product in the assortment being chosen by a consumer. Specifically, $\mathbb{P}_j(S)$ represents the probability that a customer selects product j from the offer set S, while $\mathbb{P}_0(S)$ represents the probability of choosing the default option. Note that when the offer set is empty, the default option is always chosen, i.e., $\mathbb{P}_0(\emptyset) = 1$. More formally, a choice model specifies the choice probabilities $\{\mathbb{P}_j(S) \colon j \in S^+, S \subseteq N\}$ that satisfy the standard probability laws: $\mathbb{P}_j(S) \geq 0$ for all $j \in S^+$ and $\sum_{j \in S^+} \mathbb{P}_j(S) = 1$, for all $S \subseteq N$. Note that discrete choice models are widely used to predict consumers' purchase decisions in operations, marketing, and economics research. For further details, we refer readers to Ben-Akiva et al. (1985) and Train (2009). While early research on choice modeling primarily focused on parametric models such as the multinomial logit (MNL) model, the mixed MNL model, and the nested logit model (Train 2009), recent years have seen a rapid surge in consumer choice data, driving the development of nonparametric choice models. These models aim to enhance the accuracy of consumer decision predictions and provide greater flexibility in fitting the data (Block et al. 1959, Farias et al. 2013, Aouad et al. 2021, Jagabathula et al. 2024, Chen and Mišić 2022).

More specifically, recent research in nonparametric choice modeling has primarily focused on developing more advanced and sophisticated second-stage choice mechanisms, with or without incorporating a consideration set formation stage, to better fit consumer choice data. In contrast, our paper takes the opposite approach by employing the simplest possible second-stage choice mechanism. This approach enables us to explore the role and power of first-stage consideration set formation and examine how incorporating consideration sets influences the characteristics and applicability of choice models. To this end, we introduce and analyze a model, referred to as the *consideration set model*, which is described in detail below.

2.2. Consideration Set Model (CSM)

In what follows, we introduce the consideration set model. As elaborated later in Section 2.3, this choice model is *not* based on entirely new principles but is instead nested within several well-established choice models. This structure enables us to extend our findings to other choice models that implicitly or explicitly incorporate consideration set structures. In essence, the consideration set model represents a probability distribution over sets of products, with each set $C \subseteq N$ corresponding to a consideration set. Specifically, let $\mathcal{C}$ be a collection of subsets of N and $\boldsymbol{\lambda}$ is a probability distribution over $\mathcal{C}$ such that $\lambda_C \geq 0$ for all $C \in \mathcal{C}$ and $\sum_{C \in \mathcal{C}} \lambda_C = 1$. Each $C \in \mathcal{C}$ denotes a *consideration set*, which is a subset of N. Furthermore, we assume that each consideration set C specifies a set of preference relations between elements in N^+ as follows.

Definition 1. Each consideration set $C \in \mathcal{C}$ of the consideration set model $(\mathcal{C}, \boldsymbol{\lambda})$ represents a set of preference relations between items in N^+ as follows: (a) $i \sim j$, for all $i, j \in C$; (b) $i \succ 0$, for all $i \in C$; and (c) $0 \succ j$, for all $j \neq C$.

In Definition 1, following the convention, we use $i \succ j$ to denote "i is preferred to j" and use $i \sim j$ to denote that i and j are equally preferred. Note that we do not need to specify preference relations between the items outside of the consideration set C, since those items are dominated by the default (i.e., no-purchase) option 0 which is always available. However, without loss of generality, one can still assume that $i \sim j$ for all $i, j \notin C$.

Following the standard interpretation in choice modeling, the consideration set model $(\mathcal{C}, \boldsymbol{\lambda})$ can be viewed either as an individual customer's stochastic decision rule or as a representation of customer segments in the market. In the former case, we assume that with probability λ_C, a customer samples a consideration set C according to distribution $\boldsymbol{\lambda}$ before making a final choice. In the latter case, each consideration set $C \in \mathcal{C}$ is associated with a customer type, and λ_C represents the proportion of customers of type C in the market. Before we show how to compute the choice probabilities under the consideration set model, we further provide an alternative parametrization of the model, which specifies a conditional distribution $\boldsymbol{\lambda}_{\cdot|S}$ given an offered assortment S.

Definition 2. A *conditional set distribution* $\boldsymbol{\lambda}_{\cdot|S}$ is defined with respect to a distribution $\boldsymbol{\lambda}$ over subsets in N, for every $S \subseteq N$, as follows:

$$\lambda_{C'|S} = \begin{cases} \sum_{C \subseteq \mathcal{C}} \lambda_C \cdot \mathbb{I}\left[C \cap S = C'\right], & \text{if } C' \subseteq S \text{ and } C' \neq \emptyset, \\ 0, & \text{if } C' \nsubseteq S \text{ and } C' \neq \emptyset, \end{cases} \tag{1}$$

along with $\lambda_{\emptyset|S} = 1 - \sum_{C' \subseteq S : C' \neq \emptyset} \lambda_{C'|S}$, where $\mathbb{I}[A]$ is an indicator function.

In fact, Equation (1) specifies the distribution over the sets within the offered assortment S. For example, assume that $N = \{1,2,3,4,5\}$ and an assortment $S = \{1,2,3\}$ is offered. Then the consideration set $C = \{2,3,4\}$ is equivalent to the consideration set $C' = C \cap S = \{2,3\}$. Put differently, if a customer who is considering buying products $C = \{2,3,4\}$ enters a store that only offers products $S = \{1,2,3\}$, then the customer would only consider buying product $C \cap S = \{2,3\}$, as product 4, despite being considered, is not offered. Intuitively, Definition 2 specifies that the probability that the customer samples a conditional set C' from the offer set S is the sum of the probabilities of sampling consideration sets C such that $C \cap S = C'$. With the conditional set distribution defined in Definition 2, there are two ways to describe a consumer's decision-making process when selecting from an assortment S. In the first approach, the customer is assumed to sample a consideration set $C \subseteq N$ according to the distribution $\boldsymbol{\lambda}$ and then make a final choice from the intersection $C \cap S$. Alternatively, the customer can be assumed to sample a conditional set $C' \subseteq S$ directly, based on the conditional set distribution $\boldsymbol{\lambda}_{\cdot|S}$, and then make a final choice from C'. Both approaches lead to identical choice probabilities, making them equivalent.

It is straightforward to verify that the conditional set distribution $\boldsymbol{\lambda}_{\cdot|S}$ satisfies the standard conditions such as $\lambda_{C'|S} \geq 0$ for all $C' \subseteq N$ and $\sum_{C' \subseteq N} \lambda_{C'|S} = 1$. In addition, $\boldsymbol{\lambda}_{\cdot|S}$ is consistent with the "monotonicity property" where for all $S_2 \subseteq S_1 \subseteq N$, we have that $\lambda_{C'|S_2} \geq \lambda_{C'|S_1}$ for all $C' \subseteq S_2$. More specifically, for any $C \subseteq N$, such that $C \cap S_1 = C'$ is satisfied for $C' \subseteq S_2$, we have $C \cap S_2 = C'$. Next, we formally define the probability of choosing an item j from the assortment $S \subseteq N$ under the consideration set model $(\mathcal{C}, \boldsymbol{\lambda})$ by means of the conditional set distribution $\boldsymbol{\lambda}_{\cdot|S}$.

Definition 3. Given a consideration set model $(\mathcal{C}, \boldsymbol{\lambda})$, the probability to choose product j from an assortment S is computed as follows:

$$\mathbb{P}_j^{(\mathcal{C},\boldsymbol{\lambda})}(S) = \sum_{C' \subseteq S : j \in C'} \frac{\lambda_{C'|S}}{|C'|}, \tag{2}$$

if $j \in S$ and 0, otherwise.

In other words, Equation (2) implies that when an assortment S is offered, a customer forms a consideration set C' with probability $\lambda_{C'|S}$ and, within this set, assigns equal preference to all products (as outlined in preference set (a) in Definition 1) while considering every product preferable to the default (no-purchase) option (as specified in preference set (b) in Definition 1). To this end, we assume that a *customer does not rely on a specific second-stage choice mechanism to determine the "best" product from C' and therefore would select any product from C' with equal likelihood* which leads to a factor $\lambda_{C'|S}/|C'|$ in the summation. Recall that we intentionally have this assumption to make the second-stage choice mechanism as simple as possible, allowing us to focus exclusively on analyzing the role and power of the consideration set formation layer in choice modeling. Alternatively, this assumption can be justified by the bounded rationality of customers who, constrained by cognitive and physical limitations, are unable to differentiate and rank all options within the consideration set (Simon 1955).

As mentioned above, we can also model the customer's decision-making process by first sampling C from $\boldsymbol{\lambda}$ and then choosing a product from $C \cap S$. In this case, we have the following expression to compute the probability of choosing item j from offer set S:

$$\mathbb{P}_j^{(\mathcal{C},\boldsymbol{\lambda})}(S) = \sum_{C \subseteq \mathcal{C}: j \in C} \frac{\lambda_C}{|C \cap S|}, \tag{3}$$

if $j \in S$, and 0, otherwise. Similar to Equation (2), the factor $|C \cap S|$ comes from the fact that a customer equally prefers to buy a product from $C \cap S$ after she/he forms the consideration set $C \subseteq N$ and thus chooses one from $C \cap S$ uniformly at random. According to Definition 2, the choice probabilities $\mathbb{P}_j^{(\mathcal{C},\boldsymbol{\lambda})}(S)$ in Equations (2) and (3) are equivalent (by rearranging the sums). We next present a simple example to illustrate the calculation of choice probabilities $\mathbb{P}_j^{(\mathcal{C},\boldsymbol{\lambda})}(S)$.

Example 1. Consider a universe of $n = 5$ products, i.e., $N = \{1,2,3,4,5\}$, and assume that customers make choices in accordance with a consideration set model $(\mathcal{C}, \boldsymbol{\lambda})$, where $\mathcal{C} = \{C_1, C_2, C_3\}$ such that $C_1 = \{1,3,5\}$, $C_2 = \{2,3,4,5\}$, and $C_3 = \{3,4,5\}$, alongside $(\lambda_{C_1}, \lambda_{C_2}, \lambda_{C_3}) = (0.1, 0.6, 0.3)$. One way to interpret this modeling setup is that an individual customer, before making a final choice, samples a consideration set from $\{C_1, C_2, C_3\}$ with probabilities 0.1, 0.6, and 0.3, respectively. Another interpretation implies that the customer population consists of three customer types C_1, C_2, and C_3, where the probability mass of customer types C_1,C_2, and C_3 is 10%, 60%, and 30%, respectively.

Next, assuming that assortment $S = \{1,2\}$ is offered, we can compute the conditional set distribution $\boldsymbol{\lambda}_{\cdot|S}$ using Equation (1) and obtain $\lambda_{\{1\}|S} = \lambda_{C_1} = 0.1$, $\lambda_{\{2\}|S} = \lambda_{C_2} = 0.6$, $\lambda_{\{1,2\}|S} = 0$, and $\lambda_{\emptyset|S} = 1 - 0.6 - 0.1 = 0.3$. Consequently, we can compute choice probabilities under the assortment S as follows: $\mathbb{P}_1^{(\mathcal{C},\lambda)}(S) = \lambda_{\{1\}|S}/|\{1\}| = 0.1$, $\mathbb{P}_2^{(\mathcal{C},\lambda)}(S) = \lambda_{\{2\}|S}/|\{2\}| = 0.6$, and $\mathbb{P}_0^{(\mathcal{C},\lambda)}(S) = 1 - 0.1 - 0.6 = 0.3$. Alternatively, we can directly factor the intersections $C \cap S$ into our computations as shown in Equation (3) which would result in the same choice probabilities: $\mathbb{P}_1^{(\mathcal{C},\lambda)}(S) = \lambda_{C_1}/|\{C_1 \cap S\}| = 0.1$, $\mathbb{P}_2^{(\mathcal{C},\lambda)}(S) = \lambda_{C_2}/|\{C_2 \cap S\}| = 0.6$, and $\mathbb{P}_0^{(\mathcal{C},\lambda)}(S) = 0.3$.

Similarly, when $S' = \{1,2,4,5\}$ is an offered assortment, one can show that $\lambda_{\{1,5\}|S'} = 0.1$, $\lambda_{\{2,4,5\}|S'} = 0.6$, and $\lambda_{\{4,5\}|S} = 0.3$, leading to choice probabilities $\mathbb{P}_1^{(\mathcal{C},\lambda)}(S') = \lambda_{\{1,5\}|S'}/2 = 0.05$, $\mathbb{P}_2^{(\mathcal{C},\lambda)}(S') = \lambda_{\{2,4,5\}|S'}/3 = 0.2$, $\mathbb{P}_4^{(\mathcal{C},\lambda)}(S') = \lambda_{\{2,4,5\}|S'}/3 + \lambda_{\{4,5\}|S'}/2 = 0.35$, and $\mathbb{P}_5^{(\mathcal{C},\lambda)}(S') = \lambda_{\{1,5\}|S'}/2 + \lambda_{\{2,4,5\}|S'}/3 + \lambda_{\{4,5\}|S'}/2 = 0.4$. By comparing the choice probabilities under the two different assortments S and S', we can clearly see that the consideration set model is not restricted by the independence of irrelevant alternatives (IIA) axiom (Luce 2012, Hausman and McFadden 1984) and thus not subsumed by the MNL model. Specifically, $\mathbb{P}_2^{(\mathcal{C},\lambda)}(S)/\mathbb{P}_1^{(\mathcal{C},\lambda)}(S) = 6 \neq 4 = \mathbb{P}_2^{(\mathcal{C},\lambda)}(S')/\mathbb{P}_1^{(\mathcal{C},\lambda)}(S')$.

Finally, we highlight that although we simplify the choice mechanism by assuming customers select items from the considered set of products uniformly at random, this does not imply that customers have equal probabilities of purchasing any two offered products that appear in the same consideration set. Moreover, our model can readily accommodate scenarios where a customer decides not to purchase any product from the offered assortment. Following the previous example, consider a model specification $(\mathcal{C}, \boldsymbol{\lambda})$ in which a representative customer is characterized by $\mathcal{C} = \{C_4, C_5, \emptyset\}$, with $C_4 = \{1\}$, $C_5 = \{1,2\}$, and $(\lambda_{C_4}, \lambda_{C_5}, \lambda_{\emptyset}) = (0.3, 0.4, 0.3)$. When an assortment $S = \{1,2\}$ is offered, a customer would choose product 1 with a probability of 0.5, product 2 with a probability of 0.2, and opt not to purchase any offered product with a probability of 0.3. It is important to note that, in this scenario, the customer does not show indifference between products 1 and 2 appearing in the same consideration set $C_5 = \{1,2\}$. In fact, the customer is more likely to choose product 1 due to the presence of a smaller, nested consideration set C_4 within C_5 and also has the option not to purchase any product from the assortment.

2.3. Connection to Other Choice Models

Upon closer examination, it becomes clear that the consideration set model, while not built on an entirely new foundation, is both sophisticated and highly flexible. This nonparametric model offers up to $2^n - 1$ degrees of freedom, precisely matching the number of parameters required to define the distribution $\boldsymbol{\lambda}$ over the 2^n subsets in N. Moreover, our model is a special case of several well-established choice models.

The consideration set model, in fact, can be viewed as a special case of the mixed MNL model. To recap, the mixed MNL model is characterized by k segments, with each segment $\ell \in \{1, 2, \dots, k\}$ accounting for a probability mass λ_ℓ of the market. Customers in segment ℓ make purchasing decisions based on an MNL model with parameters $(w_{\ell 1}, w_{\ell 2}, \dots, w_{\ell n})$. For a given assortment S, the choice probability under the mixed MNL model is calculated as $\mathbb{P}_j(S) = \sum_{\ell=1}^k \lambda_\ell \cdot \frac{w_{\ell j}}{1+\sum_{i \in S} w_{\ell i}}$. This formulation demonstrates that the consideration set model can be effectively represented within the structure of the mixed MNL model. Specifically, suppose $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$ and $\boldsymbol{\lambda} = (\lambda_{C_1}, \dots, \lambda_{C_k})$. We can then construct a mixed MNL model as follows. Let $\mathcal{M}$ be a sufficiently large constant. For each C_ℓ, $\ell = 1, 2, \dots, k$, we define a customer segment in the mixed MNL model with population weight $\lambda_\ell \equiv \lambda_{C_\ell}$, and set $w_{\ell i} = \mathcal{M}$ for $i \in C_\ell$ and $w_{\ell i} = 0$, otherwise. As $\mathcal{M}$ approaches infinity, it becomes evident that the corresponding choice probabilities in this mixed MNL model converge to those in Equation (3). Furthermore, the consideration set model can also be viewed as a special case of the ranking-based model. This relationship is formally stated in the following lemma.

Lemma 1. *The class of consideration set models is nested in the class of ranking-based models, and the reverse statement does not hold.*

We formally prove Lemma 1 in Section EC.1.1. Herein, we illustrate the idea of the proof with a straightforward example. Suppose the consideration set model consists of a single consideration set, $C_1 = \{1, 2\}$. This model can be represented as a ranking-based model comprising two rankings, $\sigma_1 = \{1 \succ 2 \succ 0\}$ and $\sigma_2 = \{2 \succ 1 \succ 0\}$, each assigned a weight of 0.5. In other words, to construct an equivalent ranking-based model, we interpret each consideration set C as an equal-weighted average of $|C|!$ rankings, where each ranking is a permutation of the elements in C followed by the no-purchase option 0. By averaging over all permutations of C, we ensure that customers assign equal preference to each product in C. It is worth noting that Lemma 1 also establishes that the consideration set model is a member of the random utility maximization (RUM) class (Thurstone 1927, McFadden et al. 1973). In the RUM framework, each alternative is associated with a random utility, and customers select the alternative with the highest utility once the randomness is revealed, effectively maximizing their utility. It is well-known that the RUM class is equivalent to the ranking-based model class, as the revealed utilities of products can be sorted to form a ranking over the products (Block et al. 1959, Farias et al. 2013). Finally, we note that our general consideration set model, where $\boldsymbol{\lambda}$ is a distribution over consideration sets, subsumes the more restrictive consideration set distribution parameterized by independent attention parameters (Manzini and Mariotti 2014). This directly follows from the parameterization of the distribution $\boldsymbol{\lambda}$ in the CSM model: $\lambda_C = \prod_{i \in C} \gamma_i \cdot \prod_{i \in N \setminus C}(1 - \gamma_i)$, where $\boldsymbol{\gamma}$ is the vector of attention parameters specifying the independent consideration set formation in the paper by Manzini and Mariotti (2014). We formally state this result in the following lemma.

Lemma 2. *The distribution over consideration sets proposed by Manzini and Mariotti (2014) is a special case of distribution $\boldsymbol{\lambda}$ of the CSM model.*

3. Identifiability and Axiomatic Characterization

In this section, we establish the identifiability of the consideration set model and present its axiomatic characterization. Additionally, we explore the implications of these findings for general two-stage choice models, highlighting how both the “consider” and “choose” steps influence the representational power of these models. For ease of notation, we let $N_j \equiv \{S \subseteq N : j \in S\}$ denote the collection of assortments that include product j. Also, throughout this section, we let *choice data* refer to the collection of the ground-truth choice probabilities $\{\mathbb{P}_j(S) : j \in S^+, S \subseteq N\}$. For now, we assume that the exact value of choice probabilities is provided to us, that is, we ignore potential finite sample issues. This is a reasonable assumption if the number of transactions under each assortment is large. We relax this assumption in Section EC.4.1 when estimating the consideration set model from real-world grocery retail data.

3.1. Identification of the Consideration Set Model

Recall that the consideration set model is fully characterized by the distribution $\boldsymbol{\lambda}$ over subsets. Therefore, the identification of the consideration set model reduces to the identification of the distribution $\boldsymbol{\lambda}$. For simplicity, we refer to $\boldsymbol{\lambda}$ as the consideration set model throughout this section. In what follows, we present a collection of results that provide different ways to obtain the distribution $\boldsymbol{\lambda}$ in closed form. The first result requires the specification of the choice probabilities for selecting an item j across the assortments in N_j to compute λ_C for each $C \in N_j$.

THEOREM 1. *Suppose that the collection of probabilities to choose product j, i.e., $\{\mathbb{P}_j(S): S \subseteq N_j\}$, is available and consistent with an underlying consideration set model. Then, we can reconstruct the model by uniquely computing λ_C, for every $C \in N_j$, as follows:*

$$\lambda_C = \sum_{X \in N_j} \mathbb{I}\Big[|C \cup X| \geq n-1\Big] \cdot n^{|C \cup X|-n+1}(-1)^{|C|+|X|-n+1}\mathbb{P}_j(X). \tag{4}$$

Proof sketch: For each $C \in N_j$, we first define three functions as follows: $\chi_C(X) = 1/|C \cap X|$, $\psi_C(X) = \mathbb{I}_{|C \cup X|=n-1} \cdot (-1)^{|C|+|X|-n+1}$, and $\varphi_C(X) = \mathbb{I}_{|C \cup X|=n} \cdot n \cdot (-1)^{|C|+|X|-n+1}$. The first function $\chi_C(X)$ is used to account for the size of the intersection $C \cap X$ between an assortment X and a consideration set C. The remaining two functions $\psi_C(X)$ and $\varphi_C(X)$ jointly form the orthonormal basis to $\chi_C(X)$. By varying the assortment X and observing the corresponding changes in the choice probability $\mathbb{P}_j(X)$, we can infer the probability weight λ_C of a consideration set C that includes product j. With these three functions in hand, we rewrite the choice probability, defined in Equation (3), by means of the function $\chi_C(X)$ as follows:

$$\mathbb{P}_j(X) = \sum_{C \in N_j} \lambda_C \chi_C(X). \tag{5}$$

We also notice that $\psi_C(X)$ and $\varphi_C(X)$ share the same factor $(-1)^{|C|+|X|-n+1}$ and, therefore, their summation can be simplified as follows:

$$\psi_C(X) + \varphi_C(X) = \mathbb{I}\Big[|C \cup X| \geq n-1\Big] \cdot n^{|C \cup X|-n+1}(-1)^{|C|+|X|-n+1}. \tag{6}$$

Next, we claim that for all $C, C' \in N_j$,

$$\sum_{X \in N_j} \chi_{C'}(X)(\psi_C(X) + \varphi_C(X)) = \mathbb{I}\left[C' = C\right]. \tag{7}$$

Invoking the claim, Equation (4) in the theorem follows immediately, since

$$\sum_{X \in N_j} \mathbb{P}_j(X) \cdot \mathbb{I}\Big[|C \cup X| \geq n-1\Big] \cdot n^{|C \cup X|-n+1}(-1)^{|C|+|X|-n+1}$$

$$
\begin{aligned}
&= \sum_{X \in N_j} \sum_{C' \in N_j} \lambda_{C'} \cdot \chi_{C'}(X) \cdot (\psi_C(X) + \varphi_C(X)) \qquad \Big[\text{by Equations (5) and (6)}\Big] \\
&= \sum_{C' \in N_j} \lambda_{C'} \cdot \mathbb{I}\left[C' = C\right] \qquad \Big[\text{by the claim stated above as Equation (7)}\Big] \\
&= \lambda_C .
\end{aligned}
$$

Therefore, in order to complete the proof of Equation (4), it is sufficient to prove the claim in Equation (7). Since the proof of the claim is quite involved, we relegate it to Section EC.2.1 in the e-companion.

In what follows next, we complete the proof of Theorem 1 by showing that the distribution $\boldsymbol{\lambda}$ is unique. First, note that Equation (5) relates probability distribution $\boldsymbol{\lambda}$ over the consideration sets to the choice probability $\mathbb{P}_j(X)$ through the system of linear equations which can be represented as $\boldsymbol{y} = A \cdot \boldsymbol{\lambda}$, where $\boldsymbol{y} = (\mathbb{P}_j(X))_{X:j\in X}$ and $\boldsymbol{\lambda} = (\lambda_C)_{C:C\in N_j}$ are two vectors of length 2^{n-1}. Then, Equation (4) provides another relationship between choice frequencies $\mathbb{P}_j(X)$ and the model parameters $\boldsymbol{\lambda}$ in a linear form as $\boldsymbol{\lambda} = B \cdot \boldsymbol{y}$. Given that A is a $2^{n-1} \times 2^{n-1}$ dimensional matrix, proving the uniqueness of $\boldsymbol{\lambda}$ distribution reduces to showing that $\det(A) \neq 0$. This is easy to see: as we have $\boldsymbol{\lambda} = B \cdot \boldsymbol{y} = B \cdot A \cdot \boldsymbol{\lambda}$ for any $\boldsymbol{\lambda}$, it implies that $I = B \cdot A$ and $1 = \det(I) = \det(B) \cdot \det(A)$. Consequently, $\det(A) \neq 0$. We refer the readers to Section EC.2.1 in the e-companion for the complete proof of the theorem. □

Thus, Theorem 1 allows us to compute the probability mass of each consideration set $C \subseteq N$ such that $C \neq \emptyset$. We can then find $\lambda_\emptyset$ directly by using the equation $\sum_{C\subseteq N} \lambda_C = 1$. It follows from Theorem 1 that the identification of λ_C relies only on the access to the probabilities of choosing an arbitrary item j in C under various assortments, i.e, j can be any item in C. As a result, for each C, there are at least $|C|$ ways to compute λ_C. We can also formulate the corollary below that specifies the necessary conditions for the data generation process to be consistent with the consideration set model.

Corollary 1. *Consider two products j and k. Suppose that the choice data for j and k, $\{\mathbb{P}_j(S)\colon S \subseteq N_j\}$ and $\{\mathbb{P}_k(S)\colon S \subseteq N_k\}$, are consistent with an underlying consideration set model. Then, for every consideration set $C \subseteq N$, the following equation is satisfied:*

$$\sum_{X\in N_j} \mathbb{I}_{|C\cup X| \geq n-1} \cdot n^{|C\cup X|-n+1} (-1)^{|C|+|X|-n+1} \mathbb{P}_j(X) = \sum_{X\in N_k} \mathbb{I}_{|C\cup X| \geq n-1} \cdot n^{|C\cup X|-n+1} (-1)^{|C|+|X|-n+1} \mathbb{P}_k(X).$$

Note that this corollary follows directly from the Theorem 1. Alternatively, we can recover the underlying parameters of the consideration set model from the collection of choice probabilities $\{\mathbb{P}_0(S)\colon S \subseteq N\}$, where we only need to have access to the probability to choose the default option under assortments $S \subseteq N$. In particular, we have the following result:

Lemma 3. *Suppose that the collection of probabilities to choose the default option, i.e., $\{\mathbb{P}_0(S) \colon S \subseteq N\}$, is consistent with an underlying consideration set model. Then, we have that*

$$\lambda_C = \sum_{X \subseteq C} (-1)^{|C|-|X|} \mathbb{P}_0(N \setminus X). \tag{8}$$

In fact, Lemma 3 follows from a particular form of the inclusion-exclusion principle stated in Graham (1995). For any finite set Z, if $f \colon 2^Z \to \mathbb{R}$ and $g \colon 2^Z \to \mathbb{R}$ are two real-valued set functions defined on the subsets of Z such that $g(X) = \sum_{Y \subseteq X} f(Y)$, then the inclusion-exclusion principle states that $f(Y) = \sum_{X \subseteq Y} (-1)^{|Y|-|X|} g(X)$. Our result then follows from setting $f(Y)$ to λ_C and defining $g(X) = \mathbb{P}_0(N \setminus X) = \sum_{C \subseteq X} \lambda_C$, where the second equality holds since a customer of type C would choose the outside option 0 from $N \backslash X$ if and only if $C \subseteq X$. Importantly, while recovering the distribution $\boldsymbol{\lambda}$ over consideration sets in the most general case would require observing the default choice probabilities for 2^n different assortments, practical evidence suggests that consideration sets typically have limited cardinality (Hauser and Wernerfelt 1990, Hauser 2014), which significantly reduces the number of required assortments. Specifically, if the size of consideration sets is bounded by a finite number m, then, by Lemma 3, identifying the distribution over consideration sets $\{C \subseteq N \mid |C| \leq m\}$ only requires computing the default choice probability $\mathbb{P}_0(N \backslash X)$ for sets X where $|X| \leq m$. This involves only $\sum_{i=0}^{m} \binom{n}{n-i} = O(n^m)$ different assortments.

Also, note that in real-world retail settings, it can be challenging to observe the probability that customers choose the default option. Consequently, Theorem 1 has higher practical importance in identifying parameters of the consideration set model than Lemma 3, although Equation (8) is less involved than Equation (4). Lemma 3 also connects to the classic decision theory findings of Block et al. (1959) and Falmagne (1978), as it recognizes that the sum in Equation (8) can be reformulated as a Block-Marshak polynomial. We will revisit these classical results in Section 3.3. Finally, after combining together Theorem 1 and Lemma 3, we can formulate another set of necessary conditions imposed on the choice data for the data generation process to be consistent with the consideration set model.

Corollary 2. *Suppose that the choice data $\{\mathbb{P}_j(S) \colon S \subseteq N_j\}$ and $\{\mathbb{P}_0(S) \colon S \subseteq N\}$ are consistent with an underlying consideration set model. Then, for every set $C \subseteq N$ such that $j \in C$, we must have*

$$\sum_{X \subseteq C} (-1)^{|C|-|X|} \mathbb{P}_0(N \setminus X) = \sum_{X \in N_j} \mathbb{I}\Big[|C \cup X| \geq n-1\Big] \cdot n^{|C \cup X|-n+1} (-1)^{|C|+|X|-n+1} \mathbb{P}_j(X).$$

3.2. Implications of Theorem 1

We first highlight that most nonparametric choice models are not identifiable from the choice data $\{\mathbb{P}_j(S) : j \in S^+, S \subseteq N\}$ alone. The examples include the ranking-based model (Farias et al. 2013) and the decision forest model (Chen and Mišić 2022), among others. In particular, Sher et al. (2011) show that the ranking-based model cannot be identified when $n \geq 4$. Intuitively, the nonidentifiability of the aforementioned nonparametric models can be explained by the fact that the parameter space grows much faster with n than the number of available equations that match predicted and actual choice probabilities $\{\mathbb{P}_j(S)\}_{j\in S^+, S\subseteq N}$ to identify model parameters. Specifically, these equations can be written in the form $\mathbb{P}_j^{\mathtt{MD}}(S) = \mathbb{P}_j(S)$, for $j \in S^+$ and $S \subseteq N$, where $\mathtt{MD}$ is a choice model and $\mathbb{P}_j^{\mathtt{MD}}(S)$ is the predicted probability to choose item j from assortment S under the choice model $\mathtt{MD}$.

For instance, a ranking-based model requires the estimation of $O(n!)$ parameters, which is the number of all possible rankings over N^+. Meanwhile, the number of available equations to identify model parameters is upper bounded by $O(n \cdot 2^n)$, where the factor 2^n is the total number of assortments $S \subseteq N$ and the factor $O(n)$ is the number of choices $j \in S^+$ under an assortment S. As $n! \gg n2^n$, the ranking-based model is not identifiable. In contrast, the number of parameters in the consideration set model is $2^n - 1$, which is the number of parameters to specify a distribution $\boldsymbol{\lambda} = (\lambda_C)_{C\in 2^N}$ over consideration sets. As $2^n < n \cdot 2^n$, it is not very surprising that the consideration set model does not suffer from the overparameterization faced by the ranking-based model. In fact, to the best of our knowledge, the consideration set model is one of the most flexible choice rules (i.e., it has the highest degrees of freedom) which are still identifiable from the choice data alone. Note that identifiability can benefit the downstream applications of the choice model. We empirically demonstrate the value of identifiability in Section EC.6 in the e-companion.

Given the complete identifiability of the first-stage consideration set formation, Theorem 1 suggests that if a two-stage choice model is non-identifiable, the first-stage consideration set formation itself is not responsible for that. Additionally, given that the consideration set model has $2^n - 1$ parameters and there are at most $O(n \cdot 2^n)$ equations available to identify a choice model, Theorem 1 indicates that any two-stage choice model characterized both by the general distribution over consideration sets as well as by the second-stage choice mechanism is at high risk of being non-identifiable. Specifically, any choice mechanism that models a purchase decision based on the first-stage consideration set could significantly increase the degree of freedom by at least $\Omega(n)$ factor, resulting in a choice model with $\Omega(n \cdot 2^n)$ parameters. For example, Jagabathula et al. (2024) demonstrates that the choice model, characterized by a joint distribution function over consideration sets and complete rankings, requires estimating $n! \cdot 2^n$ parameters and cannot be identified solely from sales transaction data.

Finally, we note that Theorem 1 is not only about identifiability but also provides a *closed-form* expression to compute all the parameters of the consideration set model. This further distinguishes the consideration set model from other parametric and non-parametric choice models. The MNL model is one of the very few models that have a similar property. Another example is a consider-then-choose (CTC) model studied by Jagabathula et al. (2024) where the authors show that the model is *partially identifiable* as the marginal distribution over consideration sets can be identified from the choice data. Jagabathula et al. (2024) also investigate a special case of the CTC model, named the GCC model, where all customers follow the same single ranking σ to make decisions after sampling a consideration set. They provide closed-form expressions to estimate the parameters of the GCC model. However, this model is quite restrictive and suffers from one-directional cannibalization which implies that all customers have to follow the same preference order (Jagabathula et al. 2024).

3.3. Axiomatic Characterization

In the previous section, we outlined various necessary conditions that choice data must satisfy to be consistent with the consideration set model (see Corollaries 1 and 2). While these necessary conditions are valuable for ensuring consistency with the consideration set model, Corollaries 1 and 2 are primarily algebraic in nature and lack significant intuitive interpretation. This encourages us to develop axioms based on customers' revealed preferences that could characterize the consideration set model and provide more intuition. To this end, we propose two axioms, called *default regularity* and *symmetric demand cannibalization.*

The axiom of default regularity relates to the classic decision theory developed by Block et al. (1959) and Falmagne (1978). Specifically, Block et al. (1959) define the Block-Marshak polynomial of the choice probabilities as follows:

$$H(i,S) = \sum_{X:S\subseteq X} (-1)^{|X|-|S|} \cdot \mathbb{P}_i(S), \qquad \forall S \subseteq N,\ i \in S^+.$$

Block et al. (1959) and Falmagne (1978) jointly show that the choice data belongs to the RUM class, i.e., the choice data is consistent with a ranking-based model, if and only if the inequality $H(i,S) \geq 0$ holds for *all* alternatives $i \in S^+$ and assortments $S \subseteq N$. Our first axiom, the *default regularity*, is specifically equivalent to the nonnegativeness of the Block-Marshak polynomial $H(0,S)$ of only the default (i.e., no-purchase) option 0. Thus, the constraints imposed on the choice probabilities by *default regularity* axiom is no more restrictive than the constraints imposed by the RUM class. In what follows we formally state the default regularity condition.

Definition 4. (Default Regularity) A collection of choice probabilities $\{\mathbb{P}_j(S) : j \in S^+, S \subseteq N\}$ satisfies the default regularity property if $H(0,S) \geq 0$ for all assortments $S \subseteq N$.

In other words, the axiom of default regularity imposes the RUM type of restriction *only* to the default choice probabilities in the choice data. Our second axiom, symmetric cannibalization, imposes restrictions on the choice data related to purchasing a specific product $j \in N$.

Definition 5. (Symmetric cannibalization) A collection of choice probabilities $\{\mathbb{P}_j(S) : j \in S^+, S \subseteq N\}$ satisfies the symmetric cannibalization property if for all assortments $S \subseteq N$ and $j, k \in S$ such that $j \neq k$, we have $\mathbb{P}_j(S \setminus \{k\}) - \mathbb{P}_j(S) = \mathbb{P}_k(S \setminus \{j\}) - \mathbb{P}_k(S)$.

The axiom of symmetric demand cannibalization relates to the concept of *demand cannibalization*, where the sales or market share of one product decreases due to the presence of a competing product. This axiom states that for any pair of products, j and k in S, the influence of product k on demand for product j is equal to the influence of product j on demand for product k across all product assortments $S \subseteq N$. This indicates a *symmetric* pattern in how products cannibalize each other's demand. In what follows below, we present our main theorem which characterizes the consideration set model through the axioms defined above.

Theorem 2. *The collection of choice probabilities $\{\mathbb{P}_j(S) : j \in S^+, S \subseteq N\}$ is consistent with a consideration set model with unique distribution $\boldsymbol{\lambda}$ if and only if it satisfies the axioms of default regularity and symmetric cannibalization.*

The proof is relegated to the e-companion (see Section EC.2.2). As it can be seen therein, establishing necessity is rather straightforward, but establishing sufficiency is more involved as it requires two auxiliary lemmas. We prove sufficiency by constructing the distribution $\boldsymbol{\lambda}$ and then the uniqueness of $\boldsymbol{\lambda}$ follows directly from the Theorem 1. From the proof, we can also notice that the non-negativity of the Block-Marshak polynomial $H(0, S)$ of the no-purchase option also ensures that the parameters of the consideration set model computed from the choice probabilities, as described by Lemma 3, are well-defined, i.e., $\lambda_C = \sum_{X \subseteq C} (-1)^{|C|-|X|} \mathbb{P}_0(N \backslash X) = \sum_{\bar{C} \subseteq \bar{X}} (-1)^{|\bar{X}|-|\bar{C}|} \mathbb{P}_0(\bar{X}) = H(0, \bar{C}) \geq 0$, where $\bar{X} = N \backslash X$ denotes the set complement for a set $X \subseteq N$. This is because the functional form in the axiom of default regularity resembles the formula used to calculate the probability mass of the consideration set distribution presented in Lemma 3.

Interestingly, Theorem 2 suggests that in a general two-stage choice model, it is the choice mechanism, rather than the consideration set formation, that is responsible for capturing the asymmetric demand cannibalization among products. The theorem reveals a fundamental limitation of the consideration set formation in the first stage. While the consideration set formation can explain the heterogeneity of customers' preferences (as illustrated in Example 1), it falls short of completely capturing demand substitution in the way that the ranking-based model or other general RUM models can. In other words, in order to capture complex inter-product substitution using a two-stage choice model more accurately, one should focus on developing the second-stage choice mechanism. Although the symmetric demand cannibalization can be considered as a limitation, in Section 5 we demonstrate that the consideration set model remains competitive in prediction performance compared to the ranking-based model when tested on real-world data. Thus, the effect of demand cannibalization in real-world settings may not be that asymmetric.

We also note that Theorem 2 can be used to verify if the choice generation process is indeed consistent with a consideration set model. It can also be observed that the default regularity axiom when combined with the symmetric cannibalization axiom ensures a well-known *regularity* property (or *weak rationality*), i.e, $\mathbb{P}_j(S_1) \geq \mathbb{P}_j(S_2)$ whenever $S_1 \subseteq S_2$. which plays an important role in the economics literature (Rieskamp et al. 2006). We state it formally as follows.

Lemma 4. *If a choice model $\mathbb{P}$ satisfies the axioms of symmetric cannibalization and default regularity, then for all $S_1, S_2 \subseteq N$ such that $S_1 \subseteq S_2$ we have $\mathbb{P}_j(S_1) \geq \mathbb{P}_j(S_2)$ for any $j \in S_1$.*

Thus, Theorem 2 and Lemma 4 jointly imply that the only restriction that the consideration set model imposes on top of the general RUM class of choice models, such as the ranking-based model, is the symmetric demand cannibalization property. Finally, we formulate a lemma that states that if the two axioms are satisfied, then the cannibalization effect of item k on item j diminishes when we enlarge the assortment.

Lemma 5. *If a choice model $\mathbb{P}$ satisfies the axioms of symmetric cannibalization and default regularity, then $\Delta_k \mathbb{P}_j(S_1) \geq \Delta_k \mathbb{P}_j(S_2)$ if $S_1 \subseteq S_2$, where $\Delta_k \mathbb{P}_j(S) = \mathbb{P}_j(S \setminus \{k\}) - \mathbb{P}_j(S)$.*

In other words, it follows from the lemma that if item k cannibalizes item j in assortment S_1, then item k also cannibalizes item j in assortment $S_2 \subseteq S_1$. Equivalently, if item k does not cannibalize item j in assortment S_1 then item k also does not cannibalize item j in assortment $S_2 \supseteq S_1$. We omit the proofs of Lemmas 4 and 5, as they follow straightforwardly.

We finish this section by adding two additional remarks. First, we note that the axiomatic characterization of choice models in the economics literature is usually established for *parametric* models. Luce's independence of irrelevant alternatives (IIA) axiom (Luce 2012, Hausman and McFadden 1984) is one of the most popular examples and is used to demonstrate the limitation of the MNL model both in the economics and operation fields. A more recent example is provided by Echenique and Saito (2019) where the authors extend the Luce model by proposing a set of axioms that relax the IIA property. Among the nonparametric models, the ranking-based model is one of a few models that are characterized by axioms. As we discussed above, the choice data under the ranking-based model can be characterized by the Block-Marshal polynomial $H(i, S)$ such that $H(i, S) \geq 0$ for all $i \in S^+$ and $S \subseteq N$; see Barberá and Pattanaik (1986) and McFadden and Richter (1990). Consequently, our paper contributes to the rare examples of axiomatic nonparametric choice models.

We also note that Theorem 2 indicates that neither the MNL nor the consideration set model subsumes each other. It is straightforward to check that the MNL model does not satisfy the symmetric cannibalization property unless its attraction parameters are the same for all items, meaning that all items have equal market shares. It is also obvious to see that the MNL model does not subsume the consideration set model, as the latter model is with a much higher degree of freedom. Symmetric cannibalization property also differentiates the consideration set model from the broad class of choice models with a single preference order (Manzini and Mariotti 2014, Jagabathula et al. 2024). In practice, as we will see in Section 5, the assumption of symmetric demand cannibalization does not seriously impair the predictive performance of the consideration set model, as it performs closely to the mixed MNL model and the ranking-based model.

4. Assortment Optimization

In this section, we investigate how accounting for consideration sets in choice modeling might influence the design and complexity of operations strategies in downstream applications. Specifically, we focus on the assortment optimization problem, which aims to identify the optimal set of products to offer to customers to maximize revenue. Throughout this section, let r_i denote the revenue associated with product $i \in N$. Without loss of generality, we assume the products are ordered such that $r_1 \geq r_2 \geq \ldots \geq r_n > 0$. In this context, the default option generally refers to either a customer leaving without making a purchase (i.e., the no-purchase option) or choosing a product outside the designated set (i.e., the outside option), both of which result in zero revenue. For ease of notation, we let $(\mathcal{C}, \boldsymbol{\lambda})$ denote a consideration set model where $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ and $\lambda_{C_j} \equiv \lambda_j > 0$ for $j \in K \equiv \{1, 2, \ldots, k\}$. Additionally, for a customer type associated with a consideration set C, we denote the expected revenue under assortment S as $\text{Rev}_C(S)$:

$$\text{Rev}_C(S) = \begin{cases} \sum_{i \in C \cap S} r_i / |C \cap S|, & \text{if } |C \cap S| \geq 1, \\ 0, & \text{if } |C \cap S| = 0. \end{cases} \tag{9}$$

Hence, $\sum_{j \in K} \lambda_j \cdot \text{Rev}_{C_j}(S)$ is the expected revenue from all the customer segments in the population. We can thus state the assortment optimization problem under the consideration set model as follows:

$$\max_{S \subseteq N} \left\{ \sum_{j \in K} \lambda_j \cdot \text{Rev}_{C_j}(S) \right\}. \tag{10}$$

In addition, without loss of generality, we assume that $\cup_{j \in K} C_j = N$. Specifically, if a product $i \notin \cup_{j \in K} C_j$, then $i \notin C_j$ for any $j \in K$, meaning no customer in the market considers purchasing it. As a result, the product has zero demand across all assortments and does not affect the choice probabilities of other products, making it irrelevant to the assortment decision. Therefore, it can be excluded from the product universe N.

Finally, note that assortment optimization is an important application of choice modeling, widely used in practical tasks such as menu design and product recommendation. For a comprehensive overview, we refer readers to the monograph by Kok et al. (2008). In this section, we first establish foundational results regarding the optimal solution to Problem (10), followed by an analysis of its computational complexity.

4.1. Analysis of Optimal Assortments

In this subsection, we provide an exact characterization of the optimal assortment structure. We begin with some definitions. First, we let $b \in \{0,1\}$ be a binary variable. Next, we define $\eta_b(C)$ as an operator over a set $C \subseteq N$, such that $\eta_1(C) = C$ and $\eta_0(C) = \bar{C} = N \backslash C$. In other words, when $b = 1$, the operator $\eta_b(\cdot)$ functions as the identity mapping, whereas for $b = 0$, it returns the complement of the set. Then, given a binary vector $\mathbf{b} \in \{0,1\}^k$ of length k, we define a *block* $I_{\mathbf{b}}$ in the following way:

$$I_{\mathbf{b}} = \bigcap_{j=1}^{k} \eta_{b_j}(C_j). \tag{11}$$

For example, if $\mathbf{b} = (0,1,0)$, then $I_{\mathbf{b}} = \bar{C}_1 \cap C_2 \cap \bar{C}_3$. We also let $\mathcal{B} = \{0,1\}^k$ and define $\mathcal{I}$ as the collection of all blocks such that $\mathcal{I} = \{I_{\mathbf{b}} \mid \mathbf{b} \in \mathcal{B}\}$. Consequently, as demonstrated in the proof of the upcoming theorem, the non-empty sets in $\mathcal{I} = \{I_{\mathbf{b}} \mid \mathbf{b} \in \mathcal{B}\}$ form a partition of N.

Second, we define a set of products S_1 as *revenue-ordered* within its superset S_2 if there exists a threshold $r \in \mathbb{R}$ such that $S_1 = \{i \in S_2 \mid r_i > r\}$. In other words, a revenue-ordered set S_1 consists only of the highest-revenue products in S_2. Notably, under this definition, the empty set is also considered revenue-ordered, as it can be obtained by setting the threshold r arbitrarily high. In fact, the notion of a revenue-ordered structure is well-established in the assortment optimization literature. In particular, the seminar work by Talluri and Van Ryzin (2004) shows that the optimal assortment S^* under the MNL model is revenue-ordered within the product universe N. With these definitions in place, we can now present the main theorem that characterizes the optimal assortment for Problem (10).

Theorem 3. *If S^* is an optimal assortment for Problem* (10)*, then for every block $I_{\mathbf{b}} \in \mathcal{I}$, the subset $S^*_{\mathbf{b}} \equiv S^* \cap I_{\mathbf{b}}$ must be revenue-ordered within $I_{\mathbf{b}}$.*

While the formal proof of this theorem is relegated to Section EC.3.1 in the e-companion, herein we discuss several intuitive insights derived from it. First, note that the products within a block $I_{\mathbf{b}}$ are always considered together by customers – if one product in the block is considered by a customer, all other products in that block are considered by the same customer as well. In addition, due to the uniform choice mechanism in the second stage of the consideration set model, each product within the block generates the same demand if offered. Therefore, once deciding to include $n_{\mathbf{b}}$ products from block $I_{\mathbf{b}}$ into the assortment, it is optimal to select the $n_{\mathbf{b}}$ most expensive products. This strategy ensures that the demand is concentrated on the highest-revenue products in each block, maximizing overall revenue.

Overall, these insights demonstrate that optimal assortments under the consideration set model (CSM) exhibit a clear and well-defined structure – a characteristic often absent when solving assortment problems under other choice models, such as the ranking-based model or the mixed MNL model. Moreover, leveraging the intuition outlined earlier, Theorem 3 enables a polynomial-time algorithm to solve the assortment problem (10), provided the number of consideration sets, k, is bounded by a constant. We formally state this result below.

Proposition 1. *There exists a polynomial-time optimal algorithm for the assortment problem* (10) *if the number of consideration sets in the CSM model, k, is bounded by a constant.*

Proof: We prove this proposition by invoking Theorem 3. First, we express each block $I_{\mathbf{b}}$ as $\{i_1^{\mathbf{b}}, i_2^{\mathbf{b}}, \ldots, i_{|I_{\mathbf{b}}|}^{\mathbf{b}}\}$, where $i_j^{\mathbf{b}}$ is the jth most expensive product in block $I_{\mathbf{b}}$. Then, by invoking Theorem 3, the optimal assortment S^* must belong to the following collection of assortments:

$$\mathcal{S}_{\text{OPT}} = \left\{ S = \bigcup_{\mathbf{b}\in\mathcal{B}} \{i_1^{\mathbf{b}}, i_2^{\mathbf{b}}, \ldots, i_{n_{\mathbf{b}}}^{\mathbf{b}}\} \;\middle|\; n_{\mathbf{b}} \in \{0,1,\ldots,|I_{\mathbf{b}}|\},\ \forall \mathbf{b}\in\mathcal{B} \right\}.$$

Therefore, we can find the optimal solution by enumerating all assortments in $\mathcal{S}_{\text{OPT}}$ and calculating each assortment's expected revenue. Given that $|\mathcal{B}|= 2^k$, the number of assortments in $\mathcal{S}_{\text{OPT}}$ is bounded as follows:

$$|\mathcal{S}_{\text{OPT}}| \le \prod_{\mathbf{b}\in\mathcal{B}} (|I_{\mathbf{b}}|+1) \le \prod_{\mathbf{b}\in\mathcal{B}} (n+1) \le (n+1)^{2^k}.$$

Furthermore, computing the expected revenue for each assortment requires a runtime of $O(nk)$. Consequently, the total runtime of the algorithm is $O(nk)\cdot(n+1)^{2^k} = O(k\cdot n^{2^k+1})$, which remains polynomial in n if k is upper bounded by a constant. □

In what follows, we discuss several implications of Proposition 1. First, this result can be contrasted with the complexity of assortment optimization under the mixed MNL model. To begin with, recall that the CSM model is a special case of the mixed MNL model, where each customer type follows an MNL model that assigns "infinite" weights to considered products (see Section 2.3 for details). In the mixed MNL setting, it is well-established that the assortment optimization problem is generally intractable. Specifically, no polynomial-time optimal algorithm exists, even for the case of just two customer segments, i.e., $k = 2$ (Rusmevichientong et al. 2014). To address this, Désir et al. (2022) proposed a fully polynomial-time approximation scheme (FPTAS) that provides a $(1-\epsilon)$-optimal solution for the mixed MNL model when the number of customer segments k is bounded by a constant. In contrast, Proposition 1 demonstrates that for the CSM model – a special case of the mixed MNL model – the assortment optimization problem becomes more tractable. Specifically, under the same assumption of a constant number of customer types as in Désir et al. (2022), the optimal solution under the CSM model can be found using a polynomial-time optimal algorithm. This highlights that the CSM model, compared to the general mixed MNL model, results in a more computationally tractable assortment optimization problem.

In the interest of practical implementation, we also demonstrate how the assortment problem (12) can be formulated and solved as a mixed-integer linear program (MILP) in a simpler and more efficient manner. Note that the objective function in the assortment problem (10) takes the form of a linear-fractional sum, enabling the application of standard linearization techniques (Charnes and Cooper 1962, Şen et al. 2018) to reformulate it as follows:

$$\begin{aligned}
\underset{\mathbf{x}\in\{0,1\}^n,\mathbf{h}\geq\mathbf{0},\mathbf{u}\geq\mathbf{0}}{\text{maximize}} \quad & \sum_{j\in K}\sum_{i\in C_j}\lambda_j\cdot r_i\cdot h_{ij} && && \text{(12a)}\\
\text{subject to} \quad & h_{ij}\leq x_i, && \forall i\in C_j,\ j\in K, && \text{(12b)}\\
& h_{ij}\leq u_j, && \forall i\in C_j,\ j\in K, && \text{(12c)}\\
& u_j+x_i\leq h_{ij}+1, && \forall i\in C_j,\ j\in K, && \text{(12d)}\\
& \sum_{i\in C_j} h_{ij}\leq 1, && \forall j\in K, && \text{(12e)}
\end{aligned}$$

where $\mathbf{x}$ is a binary decision vector such that $x_i=1$ if and only if product i is included in the assortment. Importantly, the four sets of constraints in the optimization problem jointly ensure that $h_{ij}=x_i/\sum_{i\in C_j}x_i$ when $\sum_{i\in C_j}x_i\neq 0$. Also, note that the aforementioned constraints naturally ensure that $h_{ij}=0$ if $\sum_{i\in C_j}x_i=0$. In the interest of space, we evaluate the scalability and effectiveness of the MILP (12) in Section EC.3.4 of the e-companion.

4.2. Hardness and Inapproximability

We further demonstrate that relaxing the assumption of a bounded number of customer segments makes the assortment problem (10) computationally hard, even if the size of each consideration set is restricted. This result is formally stated as follows.

Proposition 2. *The assortment problem* (10) *is NP-hard even if the size of each consideration set is upper bounded by a constant.*

We prove this hardness result by constructing a reduction from the vertex cover problem, a well-known NP-hard problem (Garey and Johnson 1979). The proof, provided in Section EC.3.2 of the e-companion, shows that the assortment problem remains NP-hard even under the restricted condition that each consideration set contains at most two products (i.e., $|C| \leq 2$ for all $C \in \mathcal{C}$). Interestingly, Propositions 1 and 2 jointly imply that it is the variety of customer types, represented by distinct consideration sets, rather than the size of the consideration sets themselves, that fundamentally drives the computational complexity of the assortment problem under the CSM model.

Furthermore, in the general case where neither the size of the consideration sets nor the number of customer types is bounded by a constant, the assortment problem (10) can be shown to be *NP-hard even to approximate*. We formally state this result as follows[1].

Theorem 4. *The assortment optimization problem* (10) *is NP-hard to approximate within factor $O(n^{\frac{1}{2}-\epsilon})$ for any fixed $\epsilon > 0$.*

We prove this theorem by constructing a reduction that transforms any instance of the maximum independent set problem on an n-vertex graph, known to be NP-hard to approximate within an $O(n^{1-\epsilon})$ factor (Håstad 1999), into an instance of the assortment problem (10) of n products and n consideration sets. We refer the readers to Section EC.3.3 of the e-companion for details. Note that Aouad et al. (2018) use a reduction from the maximum independent set problem to show that the assortment optimization problem under the ranking-based model is NP-hard to approximate within factor $O(n^{1-\epsilon})$. While our construction of the problem instances resembles that of Aouad et al. (2018), our retrieval procedure to construct an independent set from an assortment solution is quite different and involved, resulting in the $O(n^{\frac{1}{2}-\epsilon})$ factor (see the proof of Claim EC.2 in Section EC.3.3 of the e-companion).

In addition, it is worth noting that Theorem 4 establishes a lower bound of $O(\sqrt{n})$ for the inapproximability factor of the assortment problem (10). An upper bound of $O(n)$ – matching the inapproximability for the ranking-based model – can be easily obtained by approximating Problem (10) with an assortment of all products (i.e., when $S = N$). While the exact inapproximability factor of Problem (10) remains an open question, it is striking that the assortment optimization problem under the CSM model already exhibits an inapproximability factor of at least $O(\sqrt{n})$.

From an operational perspective, Theorem 4 highlights the challenges of incorporating consideration sets into choice models. Even with the simplest possible second-stage mechanism, the CSM model leads to an assortment optimization problem that is hard to approximate. This underscores the crucial role of consideration sets in the tractability of assortment optimization problems: *any choice model that explicitly accounts for consideration set formation is likely to be computationally intractable* unless additional structural assumptions are imposed to simplify the consideration set formation process.

[1] We sincerely thank Danny Segev for helping us develop this theorem.

5. Empirical Study: Prediction Performance

In this section, we compare the predictive performance of the consideration set model against state-of-the-art benchmark models using the IRI Academic dataset (Bronnenberg et al. 2008).

5.1. Data Preprocessing, Performance Metrics

The IRI Academic Dataset consists of consumer packaged goods (CPG) purchase transaction data over a chain of grocery stores in two large Behavior Scan markets in the USA. In this dataset, each item is represented by its universal product code (UPC) and we aggregate all the items with the same vendor code (comprising digits 3 through 7 in a 13-digit-long UPC code) into a unique "product". Next, to alleviate data sparsity, we first include in our analyses only the products with a relatively high market share (i.e., products with at least 1% market share) and then aggregate the remaining products into the outside option. In order to streamline our case study, we focus on the top fifteen product categories out of thirty-one that have the highest number of unique products (see Table 1) and consider the first four weeks in the year 2007 for our analyses.

We represent our sales transactions with the set of the tuples $\{(S_t, i_t)\}_{t\in\mathcal{T}}$, where i_t is the purchased product, S_t is the offered assortment and $\mathcal{T}$ denotes the collection of all transactions. For every purchase instance in the dataset which is characterized by a tuple (S_t, i_t), we have the week and the store ID of the purchase which allows us to approximately construct the offer set S_t by taking the union of all the products that were purchased within the same category as i_t, during the same week, and at the same store.

Next, we split our sales transaction data into the training set, which consists of the first two weeks of our data and is used for the calibration of the choice models, and the test/hold-out set, which consists of the last two weeks of our data and is used to compute the prediction performance scores defined below. In what follows, we use the mean absolute percentage error (MAPE) to measure the predictive performance of the choice models:

$$\text{MAPE} = \frac{1}{\sum_{S\in\mathcal{S}} \tau_o(S)} \cdot \sum_{S\in\mathcal{S}} \tau_o(S) \sum_{i\in S^+} \left| \frac{\hat{p}_{i,S} - \bar{p}_{i,S}}{\bar{p}_{i,S}} \right|, \tag{13}$$

where $\bar{p}_{i,S}$ is the empirical choice frequency computed directly from the sales transaction data, i.e., $\bar{p}_{i,S} = \tau_o(S,i)/(\sum_{i\in N^+} \tau_o(S,i))$ and $\tau_o(S,i)$ is the number of times alternative $i \in S^+$ was chosen under assortment S in the test dataset, $\hat{p}_{i,S}$ is the predictive probability of choosing item $i \in S^+$ from the offer set S by a specific choice model, $\mathcal{S}$ is the set of unique assortments in the test dataset, and $\tau_o(S) = \sum_{i\in S^+} \tau_o(S,i)$ is the number of observed transactions under assortment S. In the interest of space, we report the predictive outcome based on out-of-sample KL-divergence in Section EC.5.1 of the e-companion. For both metrics, a lower score indicates better performance.

Product Category	# products	# assortments	# transactions	# types	set size
Beer	19	721	759,968	39	1.36
Coffee	17	603	749,867	38	1.86
Deodorant	13	181	539,761	108	2.40
Frozen Dinners	18	330	1,963,025	40	1.29
Frozen Pizza	12	138	584,406	54	2.21
Household Cleaners	21	883	562,615	41	1.30
Hotdogs	15	533	202,842	51	1.82
Margarine/Butter	11	27	282,649	36	1.93
Milk	18	347	476,899	80	2.84
Mustard/Ketchup	16	644	266,291	38	1.87
Salty Snacks	14	152	1,476,847	86	1.84
Shampoo	15	423	574,711	45	1.81
Soup	17	315	1,816,879	44	1.37
Spaghetti/Italian Sauce	12	97	552,033	45	1.84
Tooth Brush	15	699	392,079	37	2.07

Table 1 Descriptive statistics of the IRI dataset after preprocessing. The first four columns show the category name, the number of products, the number of unique observed assortments, and the number of transactions. The last two columns report the number of customer types ($|\mathcal{C}|$) and the average consideration set size in the estimated CSM model.

5.2. Consideration Set Model Estimation Results

We begin by discussing the insights gained from calibrating the CSM model. For brevity, a detailed description of the CSM calibration method is provided in the e-companion. Specifically, Section EC.4 introduces an estimation approach based on the maximum likelihood estimation (MLE) framework. The core idea is to reformulate the MLE problem as a large-scale concave maximization problem, where the objective is the log-likelihood function and the linear constraints map the distribution of the consideration sets, $\boldsymbol{\lambda}$, to the choice probabilities. To solve this problem optimally, we employ the column generation technique. This approach has also been used in previous studies, including van Ryzin and Vulcano (2014) for estimating ranking-based models and Chen and Mišić (2022) for estimating decision forest models from sales data. Additional details can be found in Section EC.4 of the e-companion.

After estimating the consideration set model, we can emphasize several key observations. First, the fifth column of Table 1 reports the number of unique customer types (i.e., $|\mathcal{C}|$) in the estimated model. This column reveals that the number of consideration sets is moderate, ranging from 36 to 108 across all product categories, which suggests sparsity in the number of customer types. Second, the last column of Table 1 presents the weighted average size of the consideration sets in the estimated model $(\mathcal{C}, \boldsymbol{\lambda})$, with each weight corresponding to the probability λ_C of a consideration set $C \in \mathcal{C}$. From this column, we observe that the typical customer considers a relatively small number of products, with consideration set sizes ranging from 1.3 to 2.8 across all categories, despite some categories featuring as many as 18–21 products. This finding aligns with prior empirical research in behavioral economics and marketing, which consistently shows that consumers tend to consider only a limited number of alternatives before making their final choice (Hauser and Wernerfelt 1990, Hauser 2014).

5.3. Brand Choice Prediction Results

In this subsection, we compare the predictive performance of the CSM model against six benchmark models. The first two benchmarks are the *independent demand model* and the *MNL model*. The independent model, though widely used in practice, does not capture substitution effects. The MNL model, widely used in both academic research and practical applications, is calibrated using a maximum likelihood estimation (MLE) approach in a straightforward way. The third benchmark, the *mixed MNL model*, is estimated using the expectation-maximization (EM) algorithm (Train 2009) with $K = 10$ latent classes. The fourth benchmark is the *ranking-based model*, which is prominent in the operations management literature (Farias et al. 2013, van Ryzin and Vulcano 2014). Like the mixed MNL model, the ranking-based model is estimated via the MLE framework (van Ryzin and Vulcano 2014, 2017). Notably, both the mixed MNL and the ranking-based models subsume the CSM model studied in this paper as they are equivalent to the RUM class, and thus they are highly competitive benchmarks. The fifth model is the *Markov chain model* studied by Blanchet et al. (2016), which captures demand substitution by Markov chains. We estimate this model by the EM algorithm (Şimşek and Topaloglu 2018). While the Markov chain model also belongs to the RUM class, Berbeglia et al. (2022) empirically demonstrate that this model has superior predictive performance relative to the mixed MNL and ranking-based models across several datasets. The last benchmark model is the *decision forest model* proposed by Chen and Mišić (2022), which can also be estimated by an MLE approach. The decision forest model is outside of the RUM class as it can subsume any discrete choice model. To alleviate the computational effort required for cross-validation, we estimate the decision forest model using trees with a depth of three.

Table 2 summarizes the out-of-sample predictive performance of the consideration set model and the benchmark models, evaluated using the MAPE score on test data. The models are denoted as follows: ID (independent demand), MNL (MNL model), MMNL (mixed MNL), RBM (ranking-based model), MC (Markov chain), DF (decision forest), and CSM (consideration set model). As expected, the independent demand and MNL models exhibit significantly worse predictive performance compared to the CSM. Although the MNL model is not subsumed by the CSM (see Section 3.3), the latter consistently outperforms it in prediction accuracy.

The consideration set model also achieves comparable predictive performance to the mixed MNL and ranking-based models, both of which represent the general RUM class. Furthermore, the CSM remains competitive with the Markov chain model, which has been shown to have an edge over other RUM-based models (Berbeglia et al. 2022). Of all the benchmarks, the decision forest model achieves the best predictive accuracy on real-world transaction data, as measured by the MAPE score, and also performs well in terms of KL-divergence (see Section EC.5.1). This result is expected, given that the decision forest model lies outside the RUM class and offers the greatest flexibility in capturing complex customer preferences. While the nonparametric nature and high flexibility of the decision forest model make it powerful for fine-grained predictions, they come at the cost of significantly increased computational complexity. Specifically, its large degree of freedom makes downstream operational tasks, such as assortment optimization, much more challenging compared to the CSM model (Akchen and Mišić 2021).

The key takeaway from this study is the effectiveness of the CSM model in accurately predicting customer choices, even with the simplest uniformly random second-stage choice mechanism. This underscores the pivotal role of first-stage consideration set formation within the two-stage choice framework and highlights its substantial influence on choice modeling.

In Table 2, we also present two variations of the consideration set model. The first, CSM2, includes only consideration sets with at most two products, i.e., $|C| \leq 2$ for all C in $\mathcal{C}$. Interestingly, CSM2 only slightly underperforms the general CSM in predictive accuracy. This result is consistent with earlier findings on small consideration set sizes (see Table 1) and aligns with empirical studies in the literature (Hauser and Wernerfelt 1990, Hauser 2014). The second variant, the CSM model blended with rankings (CSMR), is a mixture of the CSM model and the ranking-based model, enhancing the latter's predictive performance by accounting for ties between products (see Lemma 1). While CSMR remains within the RUM class, it outperforms the standard ranking-based model, which cannot explicitly handle product ties, and performs comparably to the Markov chain model. These findings are consistent with the study by Désir et al. (2021), which demonstrates that integrating a smoothed mixture of rankings can substantially enhance predictive performance.

5.4. Additional Analyses

In the interest of space, we relegate additional analyses and experiments to the e-companion. In Section EC.5.1, we compare the CSM model with benchmark models using an additional performance metric, KL-divergence. Our findings confirm that the insights from Table 2 remain consistent when evaluated with alternative metrics, highlighting the robustness of our results. In Section EC.5.2, we examine the computational efficiency of the CSM model compared to the ranking-based model in the estimation process. Specifically, we analyze how the in-sample log-likelihood of both models evolves over a finite runtime. The results show that the estimation algorithm for the CSM model converges significantly faster to a near-optimal solution, requiring much less time than the ranking-based model.

Category	ID	MNL	MMNL	RBM	MC	DF	CSM	CSM2	CSMR
Beer	2.26	2.03	1.88	1.96	1.87	1.63	1.90	1.87	1.85
Coffee	3.61	3.23	2.80	2.94	2.76	2.61	2.96	2.92	2.83
Deodorant	1.02	0.84	0.80	0.80	0.81	0.76	0.79	0.80	0.79
Frozen Dinners	1.90	1.65	1.49	1.47	1.46	1.24	1.49	1.46	1.44
Frozen Pizza	2.32	1.89	1.54	1.48	1.54	1.17	1.54	1.60	1.49
Household Cleaners	1.74	1.52	1.45	1.49	1.41	1.37	1.46	1.42	1.41
Hotdogs	3.08	2.81	2.54	2.56	2.53	2.52	2.57	2.61	2.51
Margarine/Butter	1.63	1.44	1.21	1.21	1.19	0.81	1.22	1.22	1.21
Milk	3.97	3.09	2.50	2.60	2.54	2.44	2.54	2.65	2.54
Mustard/Ketchup	2.58	1.93	1.74	1.84	1.73	1.86	1.79	1.76	1.72
Salty Snacks	2.17	1.51	1.22	1.18	1.25	1.10	1.25	1.31	1.28
Shampoo	1.39	1.05	0.96	1.07	0.98	0.87	0.97	0.95	0.96
Soup	2.14	1.85	1.74	1.78	1.70	1.34	1.71	1.71	1.69
Spaghetti/Italian Sauce	2.29	1.72	1.46	1.43	1.47	0.98	1.47	1.47	1.42
Tooth Brush	1.79	1.40	1.24	1.33	1.24	1.28	1.29	1.28	1.24
Average	2.26	1.86	1.64	1.67	1.63	1.46	1.66	1.68	1.63

Table 2 Out-of-sample prediction performance results measured by MAPE (in unit of 10^{-1}).

In Section EC.5.3, we explore the role of the symmetric cannibalization property introduced in Section 3.3 in the predictive performance of the CSM model relative to the mixed MNL model. As noted earlier, symmetric cannibalization is a defining feature of the CSM model that enhances its tractability compared to other models in the RUM class. However, this property may also limit its ability to fully capture customer purchasing behavior. Our analysis identifies a correlation between the mixed MNL model's predictive performance over the CSM model and the degree of demand cannibalization asymmetry, suggesting that deviations from symmetric cannibalization contribute to the CSM model's occasional underperformance.

Finally, in Section EC.6, we demonstrate the operational value of model identifiability in choice modeling. Using assortment planning as a revenue management application, we show that non-identifiable choice models can lead to significant variability in the optimal assortments they produce, resulting in reduced average revenue performance. To illustrate this, we compare the CSM model, which is identifiable, with the ranking-based model, which is non-identifiable, using the IRI dataset. The results demonstrate the benefits of choice model identifiability in achieving stable and reliable operational outcomes.

6. Concluding Remarks

In this paper, we explore a class of consideration-based choice models that are fully defined by the distribution over consideration sets (i.e., the consideration set model) and examine the fundamental role and power of consideration sets in discrete choice modeling. We first prove that the consideration set model is identifiable from choice data in closed form and results in symmetric demand cannibalization. Then, we demonstrate the operational significance of consideration sets in choice modeling through the emphasis on assortment planning. To this end, we show that the optimal assortment is blockwise revenue-ordered under the consideration set model, leading to a polynomial-time optimal algorithm for the assortment problem if the number of consideration sets in the model is bounded by a constant. However, in general, the assortment optimization problem under the consideration set model is computationally hard even to approximate, although the model has the simplest possible second-stage choice mechanism. Finally, we empirically highlight the competitive predictive performance of the consideration set model despite its symmetric demand cannibalization property. To conclude, this paper examined the role and power of accounting for consideration sets in choice-based demand modeling, with the hope of motivating further research on consideration-set-based choice models and their applications in operations management.

Acknowledgments

We sincerely thank the department editor, the associate editor, and the three anonymous referees for their thoughtful comments that helped significantly improve this work. We are also grateful to Jacob Feldman and Danny Segev for their valuable feedback on the results in the assortment optimization section, which has greatly strengthened our analysis in that part of the study.

This work has benefited from generous research support provided by the UCL School of Management at University College London and the Carroll School of Management at Boston College.

References

Akçakuş, İrem, Velibor V Mišić. 2021. Exact logit-based product design. *arXiv preprint arXiv:2106.15084* .

Akchen, Yi-Chun, Velibor V Mišić. 2021. Assortment optimization under the decision forest model. *arXiv preprint arXiv:2103.14067* .

Aouad, Ali, Vivek Farias, Retsef Levi. 2021. Assortment optimization under consider-then-choose choice models. *Management Science* **67**(6) 3368–3386.

Aouad, Ali, Vivek Farias, Retsef Levi, Danny Segev. 2018. The approximability of assortment optimization under ranking preferences. *Operations Research* **66**(6) 1661–1669.

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

Ben-Akiva, Moshe, Bruno Boccara. 1995. Discrete choice models with latent choice sets. *International journal of Research in Marketing* **12**(1) 9–24.

Ben-Akiva, Moshe E, Steven R Lerman, Steven R Lerman, et al. 1985. *Discrete choice analysis: theory and application to travel demand*, vol. 9. MIT press.

Berbeglia, Gerardo, Agustín Garassino, Gustavo Vulcano. 2022. A comparative empirical study of discrete choice models in retail operations. *Management Science* **68**(6) 4005–4023.

Bertsimas, Dimitris, Velibor V Mišić. 2019. Exact first-choice product line optimization. *Operations Research* **67**(3) 651–670.

Bertsimas, Dimitris, John N Tsitsiklis. 1997. *Introduction to linear optimization*. Athena Scientific.

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

Block, Henry David, Jacob Marschak, et al. 1959. Random orderings and stochastic theories of response. Tech. rep., Cowles Foundation for Research in Economics, Yale University.

Brisoux, Jacques E, Michel Laroche. 1981. Evoked set formation and composition: An empirical investigation under a routinized response behavior situation. *ACR North American Advances* .

Bronnenberg, Bart J, Michael W Kruger, Carl F Mela. 2008. Database paper - the IRI marketing data set. *Marketing science* **27**(4) 745–748.

Charnes, Abraham, William W Cooper. 1962. Programming with linear fractional functionals. *Naval Research logistics quarterly* **9**(3-4) 181–186.

Chen, Yi-Chun, Velibor V Mišić. 2022. Decision forest: A nonparametric approach to modeling irrational choice. *Management Science* .

Chitla, Sandeep, Srikanth Jagabathula, Dmitry Mitrofanov, Maxime Cohen. 2023. Customers' multi-homing in ride-hailing: Empirical evidence from uber and lyft. *NYU Stern School of Business* .

Désir, Antoine, Vineet Goyal, Srikanth Jagabathula, Danny Segev. 2021. Mallows-smoothed distribution over rankings approach for modeling choice. *Operations Research* **69**(4) 1206–1227.

Désir, Antoine, Vineet Goyal, Jiawei Zhang. 2022. Capacitated assortment optimization: Hardness and approximation. *Operations Research* **70**(2) 893–904.

Desrosiers, Jacques, Marco E Lübbecke. 2005. A primer in column generation. *Column generation*. Springer, 1–32.

Echenique, Federico, Kota Saito. 2019. General luce model. *Economic Theory* **68**(4) 811–826.

Evgeniou, Theodoros, Constantinos Boussios, Giorgos Zacharia. 2005. Generalized robust conjoint estimation. *Marketing Science* **24**(3) 415–429.

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

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

Feldman, Jacob, Alice Paul, Huseyin Topaloglu. 2019. Assortment optimization with small consideration sets. *Operations Research* **67**(5) 1283–1299.

Feldman, Jacob, Dennis J Zhang, Xiaofei Liu, Nannan Zhang. 2022. Customer choice models vs. machine learning: Finding optimal product displays on alibaba. *Operations Research* **70**(1) 309–328.

Garey, Michael R, David S Johnson. 1979. Computers and intractability: A guide to the theory of NP-completeness .

Gensch, Dennis H. 1987. A two-stage disaggregate attribute choice model. *Marketing Science* **6**(3) 223–239.

Gilbride, Timothy J, Greg M Allenby. 2004. A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science* **23**(3) 391–406.

Graham, Ronald L. 1995. *Handbook of combinatorics*. Elsevier.

Gurobi Optimization, LLC. 2024. *Gurobi Optimizer Reference Manual*. URL `https://www.gurobi.com`.

Håstad, Johan. 1999. Clique is hard to approximate within $n^{1-\varepsilon}$.

Hauser, John R. 1978. Testing the accuracy, usefulness, and significance of probabilistic choice models: An information-theoretic approach. *Operations Research* **26**(3) 406–421.

Hauser, John R. 2014. Consideration-set heuristics. *Journal of Business Research* **67**(8) 1688–1699.

Hauser, John R, Steven P Gaskin. 1984. Application of the "defender" consumer model. *Marketing science* **3**(4) 327–351.

Hauser, John R, Birger Wernerfelt. 1990. An evaluation cost model of consideration sets. *Journal of consumer research* **16**(4) 393–408.

Hausman, Jerry, Daniel McFadden. 1984. Specification tests for the multinomial logit model. *Econometrica: Journal of the econometric society* 1219–1240.

Hogarth, Robin M, Natalia Karelaia. 2005. Simple models for multiattribute choice with many alternatives: When it does and does not pay to face trade-offs with binary attributes. *Management Science* **51**(12) 1860–1872.

Howard, John A, Jagdish N Sheth. 1969. The theory of buyer behavior. *New York* **63** 145.

Hutchinson, John MC, Gerd Gigerenzer. 2005. Simple heuristics and rules of thumb: Where psychologists and behavioural biologists might meet. *Behavioural processes* **69**(2) 97–124.

Iyengar, Sheena S, Mark R Lepper. 2000. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology* **79**(6) 995.

Jagabathula, Srikanth, Dmitry Mitrofanov, Gustavo Vulcano. 2022. Personalized retail promotions through a directed acyclic graph–based representation of customer preferences. *Operations Research* **70**(2) 641–665.

Jagabathula, Srikanth, Dmitry Mitrofanov, Gustavo Vulcano. 2024. Demand estimation under uncertain consideration sets. *Operations Research* **72**(1) 19–42.

Jedidi, Kamel, Rajeev Kohli. 2005. Probabilistic subset-conjunctive models for heterogeneous consumers. *Journal of Marketing Research* **42**(4) 483–494.

Keeney, Ralph L, Howard Raiffa, Richard F Meyer. 1993. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press.

Khot, Subhash. 2010. Inapproximability of NP-complete problems, discrete fourier analysis, and geometry. *Proceedings of the International Congress of Mathematicians*. World Scientific, 2676–2697.

Kim, Youngsoo, Benjamin Knight, Dmitry Mitrofanov, Yuqian Xu. 2024. Algorithm-enabled decision support and worker learning: a large-scale field experiment. *Available at SSRN 4976809* .

Knight, Benjamin, Dmitry Mitrofanov. 2024. Disclosing low product availability: An online platform's strategy for mitigating stockout risk. *Accepted at Management Science* .

Knight, Benjamin, Dmitry Mitrofanov, Serguei Netessine. 2023. Ai-enabled technology and gig workforce: The role of experience, skill level, and task complexity. *Working Paper* .

Kok, A Gurhan, Marshall L Fisher, Ramnath Vaidyanathan. 2008. Assortment planning: Review of literature and industry practice. *Retail supply chain management* **122**(1) 99–153.

Laroche, Michel, Chankon Kim, Takayoshi Matsui. 2003. Which decision heuristics are used in consideration set formation? *Journal of Consumer Marketing* .

Luce, R Duncan. 2012. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Manzini, Paola, Marco Mariotti. 2014. Stochastic choice and consideration sets. *Econometrica* **82**(3) 1153–1176.

McFadden, Daniel, Marcel K Richter. 1990. Stochastic rationality and revealed stochastic preference. *Preferences, Uncertainty, and Optimality, Essays in Honor of Leo Hurwicz.* 161–186.

McFadden, Daniel, et al. 1973. Conditional logit analysis of qualitative choice behavior .

McLachlan, GJ, T Krishnan. 2007. *The EM algorithm and extensions*. John Wiley & Sons.

Miller, George A. 1956. The magic number seven plus or minus two: Some limits on our capacity for processing information. *Psychological review* **63** 91–97.

Mišić, Velibor V. 2020. Optimization of tree ensembles. *Operations Research* **68**(5) 1605–1624.

Mitrofanov, Dmitry, Huseyin Topaloglu, Yuheng Wang. 2024. Choice modeling, assortment optimization, and estimation when customers are non-rational: Multinomial logit model with non-parametric dominance. *Available at SSRN: https://ssrn.com/abstract=4958971* .

Pras, Bernard, John Summers. 1975. A comparison of linear and nonlinear evaluation process models. *Journal of Marketing Research* **12**(3) 276–281.

Ratchford, Brian T. 1982. Cost-benefit models for explaining consumer choice and information seeking behavior. *Management Science* **28**(2) 197–212.

Rieskamp, Jörg, Jerome R Busemeyer, Barbara A Mellers. 2006. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature* **44**(3) 631–661.

Roberts, John H, James M Lattin. 1991. Development and testing of a model of consideration set composition. *Journal of Marketing Research* **28**(4) 429–440.

Roberts, John H, James M Lattin. 1997. Consideration: Review of research and prospects for future insights. *Journal of Marketing Research* **34**(3) 406–410.

Rusmevichientong, Paat, David Shmoys, Chaoxu Tong, Huseyin Topaloglu. 2014. Assortment optimization under the multinomial logit model with random choice parameters. *Production and Operations Management* **23**(11) 2023–2039.

Şen, Alper, Alper Atamtürk, Philip Kaminsky. 2018. A conic integer optimization approach to the constrained assortment problem under the mixed multinomial logit model. *Operations Research* **66**(4) 994–1003.

Sher, Itai, Jeremy T Fox, Patrick Bajari, et al. 2011. Partial identification of heterogeneity in preference orderings over discrete choices. Tech. rep., National Bureau of Economic Research.

Shocker, Allan D, Moshe Ben-Akiva, Bruno Boccara, Prakash Nedungadi. 1991. Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing letters* **2** 181–197.

Silk, Alvin J, Glen L Urban. 1978. Pre-test-market evaluation of new packaged goods: A model and measurement methodology. *Journal of marketing Research* **15**(2) 171–191.

Simon, Herbert A. 1955. A behavioral model of rational choice. *The quarterly journal of economics* 99–118.

Şimşek, A Serdar, Huseyin Topaloglu. 2018. An expectation-maximization algorithm to estimate the parameters of the markov chain choice model. *Operations Research* **66**(3) 748–760.

Srinivasan, Venkataraman, Allan D Shocker. 1973. Linear programming techniques for multidimensional analysis of preferences. *Psychometrika* **38**(3) 337–369.

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

Swait, Joffre, Moshe Ben-Akiva. 1987. Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B: Methodological* **21**(2) 91–102.

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

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

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

Tversky, Amos. 1972. Elimination by aspects: A theory of choice. *Psychological review* **79**(4) 281.

van Ryzin, G., G. Vulcano. 2014. A market discovery algorithm to estimate a general class of nonparametric choice models. *Management Science* **61**(2) 281–300.

van Ryzin, Garrett, Gustavo Vulcano. 2017. An expectation-maximization method to estimate a rank-based choice model of demand. *Operations Research* **65**(2) 396–407.

Wang, Ruxian, Ozge Sahin. 2018. The impact of consumer search cost on assortment planning and pricing. *Management Science* **64**(8) 3649–3666.

Wright, Peter, Fredrick Barbour. 1977. *Phased decision strategies: Sequels to an initial screening*. Graduate School of Business, Stanford University.

Electronic Companion

EC.1. Supplementary Proofs for Section 2

EC.1.1. Proof of the Lemma 1

First, we argue that the consideration set model is nested in the class of the ranking-based models. Let $(\mathcal{C}, \boldsymbol{\lambda})$ be a consideration set model. We will construct the ranking-based model in which, each $C \in \mathcal{C}$ is associated with $|C|!$ rankings that have the same probability weight. Overall, the support of the constructed ranking-based model consists of $\sum_{C \in \mathcal{C}} |C|!$ rankings. In what follows, we describe the construction of the corresponding ranking-based model in detail.

For each consideration set C, let $\mathcal{P}_C$ denote the collection of all permutations of elements in C. Then, for each element of this permutation $\rho = (i_1, i_2, \ldots, i_{|C|}) \in \mathcal{P}_C$, we construct a ranking $\sigma(\rho)$ which can be represented as follows:

$$\sigma(\rho) = \{i_1 \succ i_2 \succ \ldots i_{|C|} \succ 0\}.$$

Next, we assign a probability weight of $\lambda_C/|C|!$ to each ranking $\sigma(\rho)$ for all $\rho \in \mathcal{P}_C$. Note that the rank positions of products less preferred than the outside option 0 are not important, as these products will not be purchased anyway, since the outside option is assumed to be always available.

Finally, we show that the consideration set model $(\mathcal{C}, \boldsymbol{\lambda})$ and the constructed ranking-based model result in the same choice probabilities. To this end, let us compute the probability of purchasing item j from assortment S according to the ranking-based model:

$$\begin{aligned}
\mathbb{P}_j(S) &= \sum_{C \subseteq \mathcal{C}} \sum_{\sigma \in \mathcal{P}_C} \frac{\lambda_C}{|C|!} \cdot \mathbb{I}\left[j \text{ is the most preferred product in } C \cap S \text{ according to } \sigma\right] \\
&= \sum_{C \subseteq \mathcal{C}} \frac{\lambda_C}{k!} \cdot \mathbb{I}[j \in C \cap S] \cdot (k-p-1)! \cdot \binom{k}{k-p} \cdot p! \quad \left[\text{where } k = |C| \text{ and } k - p = |C \cap S|\right] \\
&= \sum_{C \subseteq \mathcal{C}} \frac{\lambda_C}{k!} \cdot \mathbb{I}[j \in C \cap S] \cdot (k-p-1)! \cdot \frac{k!}{p!\,(k-p)!} \cdot p! \\
&= \sum_{C \subseteq \mathcal{C}} \frac{\lambda_C}{k-p} \cdot \mathbb{I}[j \in C \cap S] \\
&= \sum_{C \subseteq \mathcal{C}} \frac{\lambda_C}{|C \cap S|} \cdot \mathbb{I}[j \in C \cap S],
\end{aligned}$$

where the last expression is used to compute the choice probability under the consideration set model (see Equation (3)).

To complete the proof, we now provide an example that shows that there exists a ranking-based model that cannot be represented by any consideration set model. Let N denote the universe of two items plus the default option 0, i.e., $N = \{1, 2\}$. Consider two rankings $\sigma_1 = \{1 \succ 2 \succ 0\}$ and

$\sigma_2 = \{2 \succ 1 \succ 0\}$, along with a probability distribution $\boldsymbol{\mu}$ such that $\mu_{\sigma_1} + \mu_{\sigma_2} = 1$ and $\mu_{\sigma_1} \neq \mu_{\sigma_2}$. Under this ranking-based model specified by $\boldsymbol{\mu}$, we have that

$$\mu_{\sigma_1} = \mathbb{P}_2(\{2\}) - \mathbb{P}_2(\{1,2\}) \neq \mathbb{P}_1(\{1,2\}) - \mathbb{P}_1(\{1\}) = \mu_{\sigma_2},$$

which violates the symmetric cannibalization property (Definition 5) that all consideration set models have to satisfy (Theorem 2). □

EC.2. Supplementary Proofs for Section 3

EC.2.1. Proof of Theorem 1

We follow the proof sketch in Section 3.1. For every $C \in N_j$ we define boolean functions $\chi_C : N_j \to \mathbb{R}$, $\psi_C : N_j \to \mathbb{R}$, and $\varphi_C : N_j \to \mathbb{R}$ by

$$\begin{aligned}
\chi_C(X) &= \frac{1}{|C \cap X|},\\
\psi_C(X) &= \mathbb{I}\Big[|C \cup X| = n-1\Big] \cdot (-1)^{|C|+|X|-n+1},\\
\varphi_C(X) &= \mathbb{I}\Big[|C \cup X| = n\Big] \cdot n \cdot (-1)^{|C|+|X|-n+1},
\end{aligned}$$

where $\mathbb{I}[A]$ is an indicator function that is equal to 1, if condition A is satisfied, and 0 otherwise. Note that for all $X \in N_j$ we have that

$$\mathbb{P}_j(X) = \sum_{C \in N_j} \lambda_C \cdot \chi_C(X), \tag{EC.1}$$

$$\lambda_C = \sum_{X \in N_j} \mathbb{P}_j(X) \cdot (\psi_C(X) + \varphi_C(X)), \tag{EC.2}$$

Then, for all $C_1, C_2 \in N_j$ we claim that

$$\sum_{X \in N_j} \chi_{C_1}(X) \cdot (\psi_{C_2}(X) + \varphi_{C_2}(X)) = \begin{cases} 1, \text{ if } C_1 = C_2,\\ 0, \text{ otherwise.}\end{cases}$$

Consequently, it follows from the claim that

$$\begin{aligned}
&\sum_{X \in N_j} \mathbb{I}\Big[|C \cup X| \geq n-1\Big] \cdot n^{|C \cup X|-n+1} \cdot (-1)^{|C|+|X|-n+1} \cdot \mathbb{P}_j(X)\\
&= \sum_{X \in N_j} \mathbb{P}_j(X) \cdot (\psi_C(X) + \varphi_C(X)) = \sum_{X \in N_j} \sum_{C_1 \in N_j} \lambda_{C_1} \cdot \chi_{C_1}(X) \cdot (\psi_C(X) + \varphi_C(X))\\
&= \sum_{C_1 \in N_j} \lambda_{C_1} \cdot \sum_{X \in N_j} \chi_{C_1}(X) \cdot (\psi_C(X) + \varphi_C(X)) = \lambda_C, \ \big[\text{by the above claim}\big].
\end{aligned}$$

Then, to complete the proof of the theorem, it is sufficient to prove the claim and show the uniqueness of the solution. In the following proof, we slightly abuse the notation for the calligraphic $\mathcal{C}$, using $\mathcal{C}^n_m = \frac{n!}{m!(n-m)!}$ to denote the binomial coefficient. This choice is made to enhance readability, as the standard notation $\binom{n}{m}$ can become cumbersome when multiple parentheses appear in the same equation. We use either '$-$' or '$\backslash$' to denote the set subtraction. We prove the claim by considering four different cases.

I) First, consider the case $C_1 = C_2 = C$:

In what follows below, we first claim that $\sum_{X\in N_j} \chi_C(X)\cdot\psi_C(X) = \frac{(|C|-n)}{|C|}$ and our second claim is that $\sum_{X\in N_j} \chi_C(X)\cdot\varphi_C(X) = \frac{n}{|C|}$, and thus, it follows from those two claims that $\sum_{X\in N_j} \chi_{C_1}(X)\cdot(\psi_{C_2}(X)+\varphi_{C_2}(X)) = 1$. We first prove the first claim in the following way:

$$\sum_{X\in N_j} \chi_C(X)\cdot\psi_C(X)$$

$$= \sum_{X\in N_j} \mathbb{I}\Big[|C\cup X| = n-1\Big]\cdot(-1)^{|C|+|X|-n+1}\cdot\frac{1}{|C\cap X|}$$

$$= \sum_{X_1\subseteq N-C}\ \sum_{X_2\subseteq C: j\in X_2} \mathbb{I}\Big[|C\cup X_1| = n-1\Big]\cdot(-1)^{|C|+|X_1|+|X_2|-n+1}\cdot\frac{1}{|X_2|}$$

[We write X as a disjoint union of X_1 and X_2, where $X_2 = X\cap C$ and $X_1 = X - C$. Note that $j\in X_2$ since $j\in C$ and $j\in X$.]

$$= \sum_{X_1\subseteq N-C} \mathbb{I}\Big[|C\cup X_1| = n-1\Big]\cdot\sum_{X_2\subseteq C: j\in X_2}(-1)^{|C|+|X_1|+|X_2|-n+1}\cdot\frac{1}{|X_2|}$$

$$= \sum_{X_1\subseteq N-C} \mathbb{I}\Big[|C\cup X_1| = n-1\Big]\cdot\sum_{X_2\subseteq C: j\in X_2}(-1)^{|X_2|}\cdot\frac{1}{|X_2|}$$

$$= (n-|C|)\cdot\sum_{k=1}^{|C|}\frac{(-1)^k\cdot\mathcal{C}_{k-1}^{|C|-1}}{k}$$

[j must be in X_2. Also, X_1 can be any set that consists of all $N\backslash C$ except one element]

$$= (n-|C|)\cdot\sum_{k=1}^{|C|}(-1)^k\cdot\frac{(|C|-1)!}{k\cdot(k-1)!\,(|C|-k)!} = \frac{(n-|C|)}{|C|}\cdot\sum_{k=1}^{|C|}(-1)^k\cdot\frac{|C|!}{k!\,(|C|-k)!}$$

$$= \frac{(n-|C|)}{|C|}\sum_{k=1}^{|C|}(-1)^k\cdot\mathcal{C}_k^{|C|} = \frac{(n-|C|)}{|C|}\left(\sum_{k=0}^{|C|}(-1)^k\cdot\mathcal{C}_k^{|C|} - 1\right) = \frac{(|C|-n)}{|C|}.$$

Then, we prove the second claim in the following way:

$$\sum_{X\in N_j} \chi_C(X)\cdot\varphi_C(X)$$

$$= \sum_{X\in N_j} \mathbb{I}\Big[|C\cup X| = n\Big]\cdot n\cdot(-1)^{|C|+|X|-n+1}\cdot\frac{1}{|C\cap X|}$$

$$= \sum_{X_1\subseteq N-C}\ \sum_{X_2\subseteq C: j\in X_2} \mathbb{I}\Big[|C\cup X_1| = n\Big]\cdot n\cdot(-1)^{|C|+|X_1|+|X_2|-n+1}\cdot\frac{1}{|X_2|}$$

[We write X as the disjoint union of X_1 and X_2, where $X_2 = X\cap C$ and $X_1 = X - C$. Note that $j\in X_2$ since $j\in C$ and $j\in X$.]

$$= \sum_{X_1\subseteq N-C} \mathbb{I}\Big[|C\cup X_1| = n\Big]\cdot n\cdot\sum_{X_2\subseteq C: j\in X_2}(-1)^{|C|+|X_1|+|X_2|-n+1}\cdot\frac{1}{|X_2|}$$

$$= (-n)\cdot\sum_{X_2\subseteq C: j\in X_2}(-1)^{|X_2|}\cdot\frac{1}{|X_2|}$$

[Since X_1 can only be $N-C$. Thus, $|X_1| = n-|C|$.]

$$= (-n) \cdot \sum_{k=1}^{|C|} \frac{(-1)^k \cdot \mathcal{C}_{k-1}^{|C|-1}}{k} \qquad \big[\text{Since } j \text{ must be in } X_2.\big]$$

$$= (-n) \cdot \sum_{k=1}^{|C|} (-1)^k \cdot \frac{(|C|-1)!}{k \cdot (k-1)!\,(|C|-k)!} = -\frac{n}{|C|} \sum_{k=1}^{|C|} (-1)^k \cdot \mathcal{C}_k^{|C|} = \frac{n}{|C|}.$$

II) Second, consider the case $C_1 \neq C_2$, $C_1 \cap C_2 = \{j\}$:

In what follows below we first claim that $\sum_{X \in N_j} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = 0$ and our second claim is that $\sum_{X \in N_j} \chi_{C_1}(X) \cdot \varphi_{C_2}(X) = 0$, and thus, it follows from those claims that $\sum_{X \in N_j} \chi_{C_1}(X) \cdot (\psi_{C_2}(X) + \varphi_{C_2}(X)) = 0$. We prove the first claim in the following way:

$$\sum_{X \in N_j} \chi_{C_1}(X) \cdot \psi_{C_2}(X)$$

$$= \sum_{X \in N_j} \mathbb{I}\Big[|C_1 \cup X| = n-1\Big] \cdot (-1)^{|C_1|+|X|-n+1} \cdot \frac{1}{|C_2 \cap X|}$$

$$= \sum_{X_1 \subseteq N - C_1} \sum_{X_2 \subseteq C_1 : j \in X_2} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot (-1)^{|C_1|+|X_1|+|X_2|-n+1} \cdot \frac{1}{|C_2 \cap X_1| + |C_2 \cap X_2|}$$

$\big[$We write X as disjoint union of X_1 and X_2, where $X_2 = X \cap C_1$ and $X_1 = X - C_1$. Note that $j \in X_2$ since $j \in C_1$ and $j \in X$.$\big]$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot \sum_{X_2 \subseteq C_1 : j \in X_2} (-1)^{|C_1|+|X_1|+|X_2|-n+1} \cdot \frac{1}{|C_2 \cap X_1| + |C_2 \cap X_2|}$$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot \frac{1}{|X_1 \cap C_2| + 1} \cdot \sum_{X_2 \subseteq C_1 : j \in X_2} (-1)^{|C_1|+|X_1|+|X_2|-n+1} \qquad \big[\text{since } C_1 \cap C_2 = \{j\}, X_2 \subseteq C_1\ \big]$$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot \frac{1}{|X_1 \cap C_2| + 1} \cdot \sum_{X_2 \subseteq C_1 : j \in X_2} (-1)^{|X_2|} \qquad \big[\text{Since } |C_1| + |X_1| = n-1\big]$$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot \frac{1}{|X_1 \cap C_2| + 1} \cdot (-1) \cdot \sum_{k=0}^{|C_1|-1} (-1)^k \cdot \mathcal{C}_k^{|C_1|-1} = 0.$$

Then, we prove the second claim in the following way:

$$\sum_{X \in N_j} \chi_{C_1}(X) \cdot \varphi_{C_2}(X)$$

$$= \sum_{X \in N_j} \mathbb{I}\Big[|C_1 \cup X| = n\Big] \cdot n \cdot (-1)^{|C_1|+|X|-n+1} \cdot \frac{1}{|C_2 \cap X|}$$

$$= \sum_{X_1 \subseteq N - C_1} \sum_{X_2 \subseteq C_1 : j \in X_2} \mathbb{I}\Big[|C_1 \cup X_1| = n\Big] \cdot n \cdot (-1)^{|C_1|+|X_1|+|X_2|-n+1} \cdot \frac{1}{|C_2 \cap X_1| + |C_2 \cap X_2|}$$

$\big[$We write X as the disjoint union of $X_1 = X - C_1$ and $X_2 = X \cap C_1$. Note that X_2 includes j.$\big]$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n\Big] \cdot n \cdot \sum_{X_2 \subseteq C_1 : j \in X_2} (-1)^{|C_1|+|X_1|+|X_2|-n+1} \cdot \frac{1}{|C_2 \cap X_1| + |C_2 \cap X_2|}$$

$$= \frac{-n}{|C_2 \cap X_1| + 1} \cdot \sum_{X_2 \subseteq C_1 : j \in X_2} (-1)^{|X_2|} \quad \big[\text{since } C_1 \cap C_2 = \{j\}, X_2 \subseteq C_1\big]$$

$$= \frac{-n}{|C_2 \cap X_1| + 1} \cdot (-1) \cdot \sum_{k=0}^{|C_1|-1} (-1)^k \cdot \mathcal{C}_k^{|C_1|-1} = 0.$$

III) Third, consider the case $C_1 \neq C_2$, $|C_1 \cap C_2| > 1$ and $C_1 \not\subseteq C_2$:

In what follows below we first claim that $\sum_{X \in N_j} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = 0$ and our second claim is that $\sum_{X \in N_j} \chi_{C_1}(X) \cdot \varphi_{C_2}(X) = 0$. It follows from those two claims that $\sum_{X \in N_j} \chi_{C_1}(X) \cdot (\psi_{C_2}(X) + \varphi_{C_2}(X)) = 0$. We prove the first claim in the following way:

$$\sum_{X \in N_j} \chi_{C_1}(X) \cdot \psi_{C_2}(X)$$

$$= \sum_{X \in N_j} \mathbb{I}\Big[|C_1 \cup X| = n-1\Big] \cdot (-1)^{|C_1|+|X|-n+1} \cdot \frac{1}{|C_2 \cap X|}$$

$$= \sum_{X_1 \subseteq N - C_1} \sum_{X_2 \subseteq C_1 : j \in X_2} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot (-1)^{|C_1|+|X_1|+|X_2|-n+1} \cdot \frac{1}{|C_2 \cap X_1| + |C_2 \cap X_2|}$$

$\big[$We write X as the disjoint union of X_1 and X_2, where $X_2 = X \cap C_1$ and $X_1 = X - C_1$. Note that $j \in X_2$.$\big]$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot \sum_{X_2 \subseteq C_1 : j \in X_2} (-1)^{|C_1|+|X_1|+|X_2|-n+1} \cdot \frac{1}{|C_2 \cap X_1| + |C_2 \cap X_2|}$$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot \sum_{X_2 \subseteq C_1 : j \in X_2} (-1)^{|X_2|} \cdot \frac{1}{|C_2 \cap X_1| + |C_2 \cap X_2|}$$

$\big[$where $|C_2 \cap X_2|$ varies from 1 to $|C_2 \cap C_1|\big]$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot \sum_{Y \subseteq C_2 \cap C_1 : j \in Y} \sum_{Z \subseteq C_1 - C_2} (-1)^{|Y|+|Z|} \cdot \frac{1}{|C_2 \cap X_1| + |C_2 \cap Y|}$$

$\big[$where $X_2 = Y \cup Z$ such that Y is any subset of $C_2 \cap C_1$ that includes j and Z is any subset of $C_1 - C_2\big]$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot \sum_{Y \subseteq C_2 \cap C_1 : j \in Y} \frac{1}{|C_2 \cap X_1| + |C_2 \cap Y|} (-1)^{|Y|} \cdot \sum_{Z \subseteq C_1 - C_2} (-1)^{|Z|}$$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot \sum_{Y \subseteq C_2 \cap C_1 : j \in Y} \frac{1}{|C_2 \cap X_1| + |C_2 \cap Y|} (-1)^{|Y|} \cdot \sum_{k=0}^{|C_1 - C_2|} (-1)^k \cdot \mathcal{C}_k^{|C_1 - C_2|}$$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n-1\Big] \cdot \sum_{Y \subseteq C_2 \cap C_1 : j \in Y} \frac{1}{|C_2 \cap X_1| + |C_2 \cap Y|} (-1)^{|Y|} \cdot 0 = 0.$$

Then, we prove the second claim in the following way:

$$\sum_{X \in N_j} \chi_{C_1}(X) \cdot \varphi_{C_2}(X)$$

$$= \sum_{X \in N_j} \mathbb{I}\Big[|C_1 \cup X| = n\Big] \cdot n \cdot (-1)^{|C_1|+|X|-n+1} \cdot \frac{1}{|C_2 \cap X|}$$

$$= \sum_{X_1 \subseteq N - C_1} \sum_{X_2 \subseteq C_1 : j \in X_2} \mathbb{I}\Big[|C_1 \cup X_1| = n\Big] \cdot n \cdot (-1)^{|C_1|+|X_1|+|X_2|-n+1} \cdot \frac{1}{|C_2 \cap X_1| + |C_2 \cap X_2|}$$

$\big[$We write X as disjoint union of $X_1 = X - C_1$ and $X_2 = X \cap C_1$. Note that $j \in X_2$.$\big]$

$$= \sum_{X_1 \subseteq N - C_1} \mathbb{I}\Big[|C_1 \cup X_1| = n\Big] \cdot n \cdot \sum_{X_2 \subseteq C_1 : j \in X_2} (-1)^{|C_1|+|X_1|+|X_2|-n+1} \cdot \frac{1}{|C_2 \cap X_1| + |C_2 \cap X_2|}$$

$$=(-n)\cdot\sum_{X_2\subseteq C_1:j\in X_2}(-1)^{|X_2|}\cdot\frac{1}{|C_2\cap X_1|+|C_2\cap X_2|}$$

$\big[$where $|C_2\cap X_2|$ varies from 1 to $|C_2\cap C_1|\big]$

$$=(-n)\cdot\sum_{Y\subseteq C_2\cap C_1:j\in Y}\;\sum_{Z\subseteq C_1-C_2}(-1)^{|Y|+|Z|}\cdot\frac{1}{|C_2\cap X_1|+|C_2\cap Y|}$$

$\big[$where $X_2=Y\cup Z$ such that Y is any subset of $C_2\cap C_1$ that includes j and Z is any subset of $C_1-C_2\}\big]$

$$=(-n)\cdot\sum_{Y\subseteq C_2\cap C_1:j\in Y}\frac{1}{|C_2\cap X_1|+|C_2\cap Y|}(-1)^{|Y|}\cdot\sum_{Z\subseteq C_1-C_2}(-1)^{|Z|}$$

$$=(-n)\cdot\sum_{Y\subseteq C_2\cap C_1:j\in Y}\frac{1}{|C_2\cap X_1|+|C_2\cap Y|}(-1)^{|Y|}\cdot\sum_{k=0}^{|C_1-C_2|}(-1)^k\cdot\mathcal{C}_k^{|C_1-C_2|}$$

$$=(-n)\cdot\sum_{Y\subseteq C_2\cap C_1:j\in Y}\frac{1}{|C_2\cap X_1|+|C_2\cap Y|}(-1)^{|Y|}\cdot 0=0.$$

IV) Fourth, consider the case $C_1\neq C_2$, $|C_1\cap C_2|>1$ and $C_1\subseteq C_2$:

In what follows below we first simplify summations $\sum_{X\in N_j}\chi_{C_1}(X)\cdot\psi_{C_2}(X)$ and $\sum_{X\in N_j}\chi_{C_1}(X)\cdot\varphi_{C_2}(X)$, and then, show that $\sum_{X\in N_j}\chi_{C_1}(X)\cdot(\psi_{C_2}(X)+\varphi_{C_2}(X))=0$. First, we have that

$$\sum_{X\in N_j}\chi_{C_1}(X)\cdot\psi_{C_2}(X)$$

$$=\sum_{X\in N_j}\mathbb{I}\Big[|C_1\cup X|=n-1\Big]\cdot(-1)^{|C_1|+|X|-n+1}\cdot\frac{1}{|C_2\cap X|}$$

$$=\sum_{X_1\subseteq N-C_1}\;\sum_{X_2\subseteq C_1:j\in X_2}\mathbb{I}\Big[|C_1\cup X_1|=n-1\Big]\cdot(-1)^{|C_1|+|X_1|+|X_2|-n+1}\cdot\frac{1}{|C_2\cap X_1|+|C_2\cap X_2|}$$

$\big[$We write X as the disjoint union of $X_1=X-C_1$ and $X_2=X\cap C_1$. Note that $j\in X_2\big]$

$$=\sum_{X_1\subseteq N-C_1}\;\sum_{X_2\subseteq C_1:j\in X_2}\mathbb{I}\Big[|C_1\cup X_1|=n-1\Big]\cdot(-1)^{|X_2|}\cdot\frac{1}{|C_2\cap X_1|+|C_2\cap X_2|}$$

$$=\sum_{X_1\subseteq N-C_1}\;\sum_{X_2\subseteq C_1:j\in X_2}\mathbb{I}\Big[|C_1\cup X_1|=n-1\Big]\cdot(-1)^{|X_2|}\cdot\frac{1}{|C_2\cap X_1|+|X_2|}\qquad\big[\text{Since } C_2\cap X_2=C_2\cap X\cap C_1=X\cap C_1=X_2\big]$$

$$=\sum_{X_1\subseteq N-C_1}\;\sum_{m=1}^{|C_1|}\mathbb{I}\Big[|C_1\cup X_1|=n-1\Big]\cdot\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{|C_1|-1}}{m+|X_1\cap C_2|}\qquad\big[\text{since } |X_2| \text{ varies from 1 to } |C_1|\ \big]$$

$$=(|C_2|-|C_1|)\cdot\sum_{m=1}^{|C_1|}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{|C_1|-1}}{m+|C_2|-|C_1|-1}+(n-|C_2|)\cdot\sum_{m=1}^{|C_1|}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{|C_1|-1}}{m+|C_2|-|C_1|}.$$

$\big[$In the sum over X_1, for $|C_2|-|C_1|$ times, $C_1\cup X_1$ misses an element in C_2, then $|X_1\cap C_2|=|C_2|-|C_1|-1$;

for $n-|C_2|$ times, $C_1\cup X_1$ misses an element in $N-C_2$, then $|X_1\cap C_2|=|C_2|-|C_1|$.$\big]$

Also, we have that

$$\sum_{X\in N_j}\chi_{C_1}(X)\cdot\varphi_{C_2}(X)$$

$$
\begin{aligned}
&= \sum_{X\in N_j} \mathbb{I}\Big[|C_1\cup X| = n\Big]\cdot n\cdot(-1)^{|C_1|+|X|-n+1}\cdot\frac{1}{|C_2\cap X|}\\
&= \sum_{X_1\subseteq N-C_1}\ \sum_{X_2\subseteq C_1: j\in X_2} \mathbb{I}\Big[|C_1\cup X_1| = n\Big]\cdot n\cdot(-1)^{|C_1|+|X_1|+|X_2|-n+1}\cdot\frac{1}{|C_2\cap X_1|+|C_2\cap X_2|}
\end{aligned}
$$

[We write X as the disjoint union of $X_1 = X - C_1$ and $X_2 = X\cap C_1$. Note that $j\in X_2$]

$$
\begin{aligned}
&= \sum_{X_1\subseteq N-C_1} \mathbb{I}\Big[|C_1\cup X_1| = n\Big]\cdot n\cdot \sum_{X_2\subseteq C_1: j\in X_2}(-1)^{|C_1|+|X_1|+|X_2|-n+1}\cdot\frac{1}{|C_2\cap X_1|+|C_2\cap X_2|}\\
&= (-n)\cdot\mathbb{I}\Big[X_1 = N - C_1\Big]\cdot\sum_{X_2\subseteq C_1: j\in X_2}(-1)^{|X_2|}\cdot\frac{1}{|C_2\cap X_1|+|C_2\cap X_2|}\\
&= (-n)\cdot\mathbb{I}\Big[X_1 = N - C_1\Big]\cdot\sum_{X_2\subseteq C_1: j\in X_2}(-1)^{|X_2|}\cdot\frac{1}{|C_2\cap X_1|+|X_2|}\\
&= (-n)\cdot\sum_{X_2}(-1)^{|X_2|}\cdot\frac{1}{|C_2|-|C_1|+|X_2|}\\
&= (-n)\cdot\sum_{m=1}^{|C_1|}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{|C_1|-1}}{m+|C_2|-|C_1|} \qquad \big[\text{where } |X_2| \text{ varies from 1 to } |C_1|\big].
\end{aligned}
$$

After simplifying $\sum_{X\in N_j}\chi_{C_1}(X)\cdot\psi_{C_2}(X)$ and $\sum_{X\in N_j}\chi_{C_1}(X)\cdot\varphi_{C_2}(X)$ as above, we have

$$
\begin{aligned}
&\sum_{X\in N_j}\chi_{C_1}(X)\cdot(\psi_{C_2}(X)+\varphi_{C_2}(X))\\
&= (|C_2|-|C_1|)\cdot\sum_{m=1}^{|C_1|}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{|C_1|-1}}{m+|C_2|-|C_1|-1} - |C_2|\cdot\sum_{m=1}^{|C_1|}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{|C_1|-1}}{m+|C_2|-|C_1|}\\
&= p\cdot\sum_{m=1}^{i}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p-1} - (p+i)\cdot\sum_{m=1}^{i}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p}, \qquad \big[\text{where we denote } p\equiv|C_2|-|C_1|,\ \ i\equiv|C_1|\ \big]\\
&= -\sum_{m=1}^{i}\frac{(-1)^m\cdot(m-1)\cdot\mathcal{C}_{m-1}^{i-1}}{m+p-1} + \sum_{m=1}^{i}(-1)^m\cdot\mathcal{C}_{m-1}^{i-1} - (p+i)\cdot\sum_{m=1}^{i}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p}
\end{aligned}
$$

[letting the p at the beginning expand as $p = (m+p-1)-(m-1)$]

$$
\begin{aligned}
&= -\sum_{m=1}^{i}\frac{(-1)^m\cdot(m-1)\cdot\mathcal{C}_{m-1}^{i-1}}{m+p-1} - (p+i)\cdot\sum_{m=1}^{i}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p}\\
&= -\sum_{m=1}^{i}\frac{(-1)^m\cdot(m-1)\cdot\mathcal{C}_{m-1}^{i-1}}{m+p-1} - \sum_{m=1}^{i}(-1)^m\cdot\mathcal{C}_{m-1}^{i-1} + \sum_{m=1}^{i}\frac{(-1)^m\cdot m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p} - i\cdot\sum_{m=1}^{i}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p}
\end{aligned}
$$

[expanding the last term by breaking the factor $(p+i)$]

$$
\begin{aligned}
&= -\sum_{m=1}^{i}\frac{(-1)^m\cdot(m-1)\cdot\mathcal{C}_{m-1}^{i-1}}{m+p-1} + \sum_{m=1}^{i}\frac{(-1)^m\cdot m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p} - i\cdot\sum_{m=1}^{i}\frac{(-1)^m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p}\\
&= -\sum_{m=1}^{i}\frac{(-1)^m\cdot(m-1)\cdot\mathcal{C}_{m-1}^{i-1}}{m+p-1} + \sum_{m=1}^{i}\frac{(-1)^m\cdot m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p} - \sum_{m=1}^{i}\frac{(-1)^m\cdot m\cdot\mathcal{C}_{m}^{i}}{m+p}\\
&= \sum_{m=0}^{i-1}\frac{(-1)^m\cdot m\cdot\mathcal{C}_{m}^{i-1}}{m+p} + \sum_{m=1}^{i}\frac{(-1)^m\cdot m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p} - \sum_{m=1}^{i}\frac{(-1)^m\cdot m\cdot\mathcal{C}_{m}^{i}}{m+p}
\end{aligned}
$$

$$
\begin{aligned}
&=\sum_{m=0}^{i-1}\frac{(-1)^m\cdot m\cdot\mathcal{C}_m^{i-1}}{m+p}+\sum_{m=1}^{i}\frac{(-1)^m\cdot m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p}-\sum_{m=1}^{i}\frac{(-1)^m\cdot m\cdot(\mathcal{C}_{m-1}^{i-1}+\mathcal{C}_m^{i-1})}{m+p}\\
&=\sum_{m=0}^{i-1}\frac{(-1)^m\cdot m\cdot\mathcal{C}_m^{i-1}}{m+p}+\sum_{m=1}^{i}\frac{(-1)^m\cdot m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p}-\sum_{m=1}^{i}\frac{(-1)^m\cdot m\cdot\mathcal{C}_{m-1}^{i-1}}{m+p}-\sum_{m=1}^{i-1}\frac{(-1)^m\cdot m\cdot\mathcal{C}_m^{i-1}}{m+p}\\
&=\sum_{m=0}^{i-1}\frac{(-1)^m\cdot m\cdot\mathcal{C}_m^{i-1}}{m+p}-\sum_{m=1}^{i-1}\frac{(-1)^m\cdot m\cdot\mathcal{C}_m^{i-1}}{m+p}\\
&=\sum_{m=0}^{i-1}\frac{(-1)^m\cdot m\cdot\mathcal{C}_m^{i-1}}{m+p}-\sum_{m=0}^{i-1}\frac{(-1)^m\cdot m\cdot\mathcal{C}_m^{i-1}}{m+p}=0.
\end{aligned}
$$

In order to complete the proof, we show the uniqueness of the probability distribution function λ in our setting. First, note that Equation (EC.1) relates probability distribution λ over consideration sets to the choice frequencies $\mathbb{P}_j(X)$ through the system of linear equations:

$$
\mathbb{P}_j(X)=\sum_{C\in N_j}\lambda_C\cdot\chi_C(X),\ \forall\ X\subseteq N\Longleftrightarrow\boldsymbol{y}=A\cdot\boldsymbol{\lambda}, \tag{EC.3}
$$

where $\boldsymbol{y}=(y_X)_{X\subseteq N_j}$ denotes the $|2^{N-1}|\times 1$ vector of choice fractions and $\boldsymbol{\lambda}=(\lambda_C)_{C\in N_j}$ denotes the $|2^{N-1}|\times 1$ vector that represents the probability distribution function over consideration sets. A is the $|2^{N-1}|\times|2^{N-1}|$ matrix such that A's entry corresponding to the row X and column C is equal to $\chi_C(X)$. As a result, the relation between the choice probabilities and the underlying model can be represented in a compact form as $\boldsymbol{y}=A\cdot\boldsymbol{\lambda}$. Then the proof of the uniqueness of λ reduces to showing that $\det(A)\neq 0$. It follows from Equation (EC.2) that

$$
\lambda_C=\sum_{X\in N_j}\mathbb{P}_j(X)\cdot(\psi_C(X)+\varphi_C(X)),\ \forall C\in N_j\Longleftrightarrow\boldsymbol{\lambda}=B\cdot\boldsymbol{y},
$$

which provides another relationship between choice frequencies $\mathbb{P}_j(X)$ and the model parameters λ in a linear form as $\boldsymbol{\lambda}=B\cdot\boldsymbol{y}$, where B is the $|2^{N-1}|\times|2^{N-1}|$ matrix such that B's entry corresponding to the row C and column X is equal to $\psi_C(X)+\varphi_C(X)$. Therefore, we have the following set of equalities:

$$
\begin{aligned}
\boldsymbol{\lambda}&=B\cdot\boldsymbol{y}=B\cdot A\cdot\lambda,\quad\big[\text{by Equation (EC.3)}\big]\\
&\Longrightarrow\ I=B\cdot A\Longrightarrow\ \det(I)=\det(B)\cdot\det(A)\\
&\Longrightarrow 1=\det(B)\cdot\det(A)\ \Longrightarrow\ \det(A)\neq 0.
\end{aligned}
$$

□

EC.2.2. Proof of Theorem 2

As in the proof of Theorem 1 in Section EC.2.1, we slightly abuse the notation for the calligraphic $\mathcal{C}$, using $\mathcal{C}_m^n=\frac{n!}{m!(n-m)!}$ to denote the binomial coefficient. This choice is made to enhance readability, as the standard notation $\binom{n}{m}$ can become cumbersome when multiple parentheses appear in the same equation.

The Auxiliary Lemmas

We first present two lemmas with proofs which are used below to prove Theorem 2. First, we establish the following combinatorial identity.

Lemma EC.1. *For all n and m such that $m \leq n$ the combinatorial identity below holds*

$$\sum_{k=0}^{m-1} (-1)^k \cdot \mathcal{C}_k^{m-1} \cdot \left(\frac{n-m}{n-m+k} - \frac{n}{n-m+1+k} \right) = 0.$$

Proof: We prove it by induction on m:

Base case: $m = 1$.

$$\sum_{k=0}^{m-1} (-1)^k \cdot \mathcal{C}_k^{m-1} \cdot \left(\frac{n-m}{n-m+k} - \frac{n}{n-m+1+k} \right) = (1-1) = 0$$

Induction hypothesis: $m = p$.

$$\begin{aligned}
0 &= \sum_{k=0}^{p-1} (-1)^k \cdot \mathcal{C}_k^{p-1} \cdot \left(\frac{n-p}{n-p+k} - \frac{n}{n-p+1+k} \right) \\
&= (p-1)! \cdot \left[\sum_{k=0}^{p-1} (-1)^k \cdot \frac{(n-p)}{(p-1-k)! \cdot k! \cdot (n-p+k)} - \sum_{k=0}^{p-1} (-1)^k \cdot \frac{n}{(p-1-k)! \cdot k! \cdot (n-p+1+k)} \right] \\
&= (p-1)!\,(n-p) \left[\sum_{k=0}^{p-1} (-1)^k \cdot \frac{1}{(p-1-k)! \cdot k! \cdot (n-p+k)} - \frac{n}{(n-p)} \cdot \sum_{k=0}^{p-1} (-1)^k \cdot \frac{1}{(p-1-k)! \cdot k! \cdot (n-p+1+k)} \right].
\end{aligned}$$

Thus, it implies

$$0 = \sum_{k=0}^{p-1} (-1)^k \cdot \frac{1}{(p-1-k)! \cdot k! \cdot (n-p+k)} - \frac{n}{(n-p)} \cdot \sum_{k=0}^{p-1} (-1)^k \cdot \frac{1}{(p-1-k)! \cdot k! \cdot (n-p+1+k)}.$$

Induction step: $m = p + 1$.

$$\begin{aligned}
&\sum_{k=0}^{p} (-1)^k \cdot \mathcal{C}_k^{p} \cdot \left(\frac{n-p-1}{n-p-1+k} - \frac{n}{n-p+k} \right) \\
&= p! \cdot \left[\sum_{k=0}^{p} (-1)^k \cdot \frac{(n-p-1)}{(p-k)! \cdot k! \cdot (n-p-1+k)} - \sum_{k=0}^{p} (-1)^k \cdot \frac{n}{(p-k)! \cdot k! \cdot (n-p+k)} \right] \\
&= p! \cdot \left[\sum_{k=0}^{p} (-1)^k \cdot \frac{(n-p-1)}{(p-k)! \cdot k! \cdot (n-p-1+k)} - \frac{n}{n-p} \cdot \sum_{k=0}^{p} (-1)^k \cdot \frac{n-p}{(p-k)! \cdot k! \cdot (n-p+k)} \right] \\
&= p! \cdot \left[\sum_{k=0}^{p} (-1)^k \cdot \frac{(n-p-1+k-k)}{(p-k)! \cdot k! \cdot (n-p-1+k)} - \frac{n}{n-p} \cdot \sum_{k=0}^{p} (-1)^k \cdot \frac{n-p+k-k}{(p-k)! \cdot k! \cdot (n-p+k)} \right]
\end{aligned}$$

$$= p! \cdot \left[\sum_{k=0}^{p} (-1)^k \cdot \frac{(-k)}{(p-k)! \cdot k! \cdot (n-p-1+k)} - \frac{n}{n-p} \cdot \sum_{k=0}^{p} (-1)^k \cdot \frac{-k}{(p-k)! \cdot k! \cdot (n-p+k)} \right]$$

$\left[\text{since } \sum_{k=0}^{p} (-1)^k \cdot \frac{1}{(p-k)! \cdot k!} = \frac{1}{p!} \cdot \sum_{k=0}^{p} (-1)^k \cdot \mathcal{C}_k^p = 0\right]$

$$= p! \cdot \left[\sum_{k=1}^{p} (-1)^k \cdot \frac{(-k)}{(p-k)! \cdot k! \cdot (n-p-1+k)} - \frac{n}{n-p} \cdot \sum_{k=1}^{p} (-1)^k \cdot \frac{-k}{(p-k)! \cdot k! \cdot (n-p+k)} \right]$$

$$= p! \cdot \left[\sum_{k=1}^{p} (-1)^k \cdot \frac{-1}{(p-k)! \cdot (k-1)! \cdot (n-p-1+k)} - \frac{n}{n-p} \cdot \sum_{k=1}^{p} (-1)^k \cdot \frac{-1}{(p-k)! \cdot (k-1)! \cdot (n-p+k)} \right]$$

$$= p! \cdot \left[\sum_{k=0}^{p-1} (-1)^k \cdot \frac{1}{(p-k-1)! \cdot k! \cdot (n-p+k)} - \frac{n}{n-p} \cdot \sum_{k=1}^{p} (-1)^k \cdot \frac{1}{(p-k-1)! \cdot k! \cdot (n-p+k+1)} \right]$$

$$= 0 \qquad [\text{by the induction hypothesis}].$$

□

Then, we show that if the symmetric cannibalization axiom holds then choice data satisfy the equality below.

Lemma EC.2. *If for all $S \subseteq N$ and $i, j \in S$ we have that*

$$\mathbb{P}_j(S) - \mathbb{P}_j(S \setminus \{i\}) = \mathbb{P}_i(S) - \mathbb{P}_i(S \setminus \{j\})$$

then for all $S \subseteq N$ and $j \in S$ we have that

$$\mathbb{P}_j(S) = \sum_{Y \subseteq S} \mathbb{P}_0(Y) \cdot \left(\mathbb{I}[j \notin Y] \cdot \sum_{k=0}^{|Y|} (-1)^k \frac{\mathcal{C}_k^{|Y|}}{|S| - |Y| + k} - \mathbb{I}[j \in Y] \cdot \sum_{k=0}^{|Y|-1} (-1)^k \frac{\mathcal{C}_k^{|Y|-1}}{|S| - |Y| + k + 1} \right).$$

Proof: We prove it by induction on $|S|$:

<u>*Base case:*</u> $|S| = 1$.

$$\mathbb{P}_j(\{j\}) = 1 - \mathbb{P}_0(\{j\})$$

<u>*Induction hypothesis:*</u> the equation holds to compute $\mathbb{P}_j(S \setminus \{i\})$ for all $i \in N \setminus \{j\}$.

$$\mathbb{P}_j(S \setminus \{i\}) = \sum_{Y \subseteq S \setminus \{i\}} \mathbb{P}_0(Y) \cdot \left(\mathbb{I}[j \notin Y] \cdot \sum_{k=0}^{|Y|} (-1)^k \frac{\mathcal{C}_k^{|Y|}}{|S| - 1 - |Y| + k} - \mathbb{I}[j \in Y] \cdot \sum_{k=0}^{|Y|-1} (-1)^k \frac{\mathcal{C}_k^{|Y|-1}}{|S| - |Y| + k} \right).$$

<u>*Induction step:*</u> the equation holds to compute $\mathbb{P}_j(S)$.

$$\begin{aligned}
&\mathbb{P}_j(S) = \mathbb{P}_i(S) + \mathbb{P}_j(S \setminus \{i\}) - \mathbb{P}_i(S \setminus \{j\}) \\
\Rightarrow &\sum_{i \in S \setminus \{j\}} \mathbb{P}_j(S) = \sum_{i \in S \setminus \{j\}} \mathbb{P}_i(S) + \sum_{i \in S \setminus \{j\}} \mathbb{P}_j(S \setminus \{i\}) - \sum_{i \in S \setminus \{j\}} \mathbb{P}_i(S \setminus \{j\}) \\
\Rightarrow &\sum_{i \in S \setminus \{j\}} \mathbb{P}_j(S) = 1 - \mathbb{P}_0(S) - \mathbb{P}_j(S) + \sum_{i \in S \setminus \{j\}} \mathbb{P}_j(S \setminus \{i\}) - \sum_{i \in S \setminus \{j\}} \mathbb{P}_i(S \setminus \{j\})
\end{aligned}$$

$$\begin{aligned}
&\Rightarrow \sum_{i\in S\setminus\{j\}} \mathbb{P}_j(S) = 1-\mathbb{P}_0(S)-\mathbb{P}_j(S)+\left(\sum_{i\in S\setminus\{j\}}\mathbb{P}_j(S\setminus\{i\})\right)-(1-\mathbb{P}_0(S\setminus\{j\}))\\
&\Rightarrow |S|\cdot\mathbb{P}_j(S)=\mathbb{P}_0(S\setminus\{j\})-\mathbb{P}_0(S)+\sum_{i\in S\setminus\{j\}}\mathbb{P}_j(S\setminus\{i\})\\
&\Rightarrow \sum_{i\in S\setminus\{j\}}\mathbb{P}_j(S\setminus\{i\})=\mathbb{P}_j(S)\cdot|S|+\mathbb{P}_0(S)-\mathbb{P}_0(S\setminus\{j\}).
\end{aligned}$$

Therefore, it is sufficient to prove that

$$\begin{aligned}
\sum_{i\in S\setminus\{j\}}\mathbb{P}_j(S\setminus\{i\})=\sum_{Y\subseteq S}\mathbb{P}_0(Y)\cdot\Bigg(&\mathbb{I}[j\notin Y]\cdot\sum_{k=0}^{|Y|}(-1)^k\frac{\mathcal{C}_k^{|Y|}}{|S|-|Y|+k}\\
&-\mathbb{I}[j\in Y]\cdot\sum_{k=0}^{|Y|-1}(-1)^k\frac{\mathcal{C}_k^{|Y|-1}}{|S|-|Y|+k+1}\Bigg)\cdot|S|+\mathbb{P}_0(S)-\mathbb{P}_0(S\setminus\{j\}).
\end{aligned}$$

This way we complete the proof:

$$\begin{aligned}
&\sum_{i\in S\setminus\{j\}}\mathbb{P}_j(S\setminus\{i\})\\
&=\sum_{i\in S\setminus\{j\}}\sum_{Y\subseteq S\setminus\{i\}}\mathbb{P}_0(Y)\cdot\left(\mathbb{I}[j\notin Y]\cdot\sum_{k=0}^{|Y|}(-1)^k\frac{\mathcal{C}_k^{|Y|}}{|S|-1-|Y|+k}-\mathbb{I}[j\in Y]\cdot\sum_{k=0}^{|Y|-1}(-1)^k\frac{\mathcal{C}_k^{|Y|-1}}{|S|-|Y|+k}\right)\\
&\quad\big[\text{by induction hypothesis}\big]\\
&=\sum_{Y\subseteq S}\mathbb{P}_0(Y)\cdot\Bigg(\mathbb{I}[j\notin Y]\cdot\sum_{k=0}^{|Y|}(-1)^k\frac{\mathcal{C}_k^{|Y|}}{|S|-1-|Y|+k}\cdot(|S|-|Y|-1)\\
&\qquad\qquad-\mathbb{I}[j\in Y]\cdot\sum_{k=0}^{|Y|-1}(-1)^k\frac{\mathcal{C}_k^{|Y|-1}}{|S|-|Y|+k}\cdot(|S|-|Y|)\Bigg)+\mathbb{P}_0(S)-\mathbb{P}_0(S\setminus\{j\})
\end{aligned}$$

[(1) in the summation $\sum_{i\in S\setminus\{j\}}\sum_{Y\subseteq S\setminus\{i\}}$, Y can be any subset of S but S and $S\setminus\{j\}$;

(2) in the summation $\sum_{i\in S\setminus\{j\}}\sum_{Y\subseteq S\setminus\{i\}}$, if $j\in Y$, then Y is assigned to a specific subset of S for $|S|-|Y|$ times;

(3) in the summation $\sum_{i\in S\setminus\{j\}}\sum_{Y\subseteq S\setminus\{i\}}$, if $j\notin Y$, then Y is assigned to a specific subset of S for $|S|-|Y|-1$ times.]

$$\begin{aligned}
&=\sum_{Y\subseteq S}\mathbb{P}_0(Y)\cdot\Bigg(\mathbb{I}[j\notin Y]\cdot\sum_{k=0}^{|Y|}(-1)^k\frac{\mathcal{C}_k^{|Y|}}{|S|-1-|Y|+k}\cdot(|S|-|Y|-1)\\
&\qquad\qquad-\mathbb{I}[j\in Y]\cdot\sum_{k=0}^{|Y|-1}(-1)^k\frac{\mathcal{C}_k^{|Y|-1}}{|S|-|Y|+k+1}\cdot|S|\Bigg)+\mathbb{P}_0(S)-\mathbb{P}_0(S\setminus\{j\})\\
&\quad\big[\text{by invoking Lemma EC.1, where } m=|Y| \text{ and } n=|S|\big]\\
&=\sum_{Y\subseteq S}\mathbb{P}_0(Y)\cdot\Bigg(\mathbb{I}[j\notin Y]\cdot\sum_{k=0}^{|Y|}(-1)^k\frac{\mathcal{C}_k^{|Y|}}{|S|-|Y|+k}\cdot|S|\\
&\qquad\qquad-\mathbb{I}[j\in Y]\cdot\sum_{k=0}^{|Y|-1}(-1)^k\frac{\mathcal{C}_k^{|Y|-1}}{|S|-|Y|+k+1}\cdot|S|\Bigg)+\mathbb{P}_0(S)-\mathbb{P}_0(S\setminus\{j\})\\
&\quad\big[\text{by invoking Lemma EC.1, where } m=|Y|+1 \text{ and } n=|S|\big]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{Y \subseteq S} \mathbb{P}_0(Y) \cdot \Bigg(\mathbb{I}[j \notin Y] \cdot \sum_{k=0}^{|Y|} (-1)^k \frac{\mathcal{C}_k^{|Y|}}{|S|-|Y|+k} \\
&\qquad - \mathbb{I}[j \in Y] \cdot \sum_{k=0}^{|Y|-1} (-1)^k \frac{\mathcal{C}_k^{|Y|-1}}{|S|-|Y|+k+1} \Bigg) \cdot |S| + \mathbb{P}_0(S) - \mathbb{P}_0(S \setminus \{j\})
\end{aligned}
$$

□

Proof of Theorem 2

We first prove the "⇐" direction, i.e., sufficiency. To simplify the exposition, we also let $\bar{X} := N \setminus X$ and $X^+ := X \cup \{0\}$. Let $\langle S \rangle$ denote the power set of S, i.e., $\langle S \rangle = 2^S$, and let $A \uplus B$ denote $\{a \cup b : a \in A, b \in B\}$ for any sets A, B. we claim that a choice model that satisfies the symmetric cannibalization and default regularity is a stochastic set model with a probability distribution function λ over preselected sets.

$$
\begin{aligned}
&\mathbb{P}_j(S) \\
&= \sum_{Y \subseteq S} \mathbb{P}_0(Y) \cdot \Bigg(\mathbb{I}[j \notin Y] \cdot \sum_{k=0}^{|Y|} (-1)^k \frac{\mathcal{C}_k^{|Y|}}{|S|-|Y|+k} - \mathbb{I}[j \in Y] \cdot \sum_{k=0}^{|Y|-1} (-1)^k \frac{\mathcal{C}_k^{|Y|-1}}{|S|-|Y|+k+1} \Bigg) \\
&\quad \big[\text{by invoking Lemma EC.2}\big] \\
&= \sum_{X_1 \subseteq S} \mathbb{P}_0(S \setminus X_1) \cdot \Bigg(\mathbb{I}[j \in X_1] \cdot \sum_{k=0}^{|S|-|X_1|} (-1)^k \frac{\mathcal{C}_k^{|S|-|X_1|}}{|X_1|+k} - \mathbb{I}[j \notin X_1] \cdot \sum_{k=0}^{|S|-|X_1|-1} (-1)^k \frac{\mathcal{C}_k^{|S|-|X_1|-1}}{|X_1|+k+1} \Bigg) \\
&\quad \big[\text{where } X_1 = S \setminus Y\big] \\
&= \sum_{X_1 \subseteq S} \mathbb{P}_0(S \setminus X_1) \cdot (-1)^{-|X_1|} \cdot \Bigg(\mathbb{I}[j \in X_1] \cdot \sum_{k=0}^{|S|-|X_1|} (-1)^{k+|X_1|} \frac{\mathcal{C}_k^{|S|-|X_1|}}{|X_1|+k} \\
&\qquad + \mathbb{I}[j \notin X_1] \cdot \sum_{k=0}^{|S|-|X_1|-1} (-1)^{k+|X_1|+1} \frac{\mathcal{C}_k^{|S|-|X_1|-1}}{|X_1|+k+1} \Bigg) \\
&= \sum_{X_1 \subseteq S} \mathbb{P}_0(S \setminus X_1) \cdot (-1)^{-|X_1|} \cdot \Bigg(\mathbb{I}[j \in X_1] \cdot \sum_{k=|X_1|}^{|S|} (-1)^k \frac{\mathcal{C}_{k-|X_1|}^{|S|-|X_1|}}{k} \\
&\qquad + \mathbb{I}[j \notin X_1] \cdot \sum_{k=|X_1|+1}^{|S|} (-1)^k \frac{\mathcal{C}_{k-|X_1|-1}^{|S|-|X_1|-1}}{k} \Bigg) \\
&= \sum_{X_1 \subseteq S} \mathbb{P}_0(S \setminus X_1) \cdot (-1)^{-|X_1|} \cdot \Bigg(\mathbb{I}[j \in X_1] \cdot \sum_{\substack{C_1 \supseteq X_1: \\ C_1 \subseteq S}} \frac{(-1)^{|C_1|}}{|C_1|} \\
&\qquad + \mathbb{I}[j \notin X_1] \cdot \sum_{\substack{C_1 \supseteq \{X_1 \cup \{j\}\}: \\ C_1 \subseteq S}} \frac{(-1)^{|C_1|}}{|C_1|} \Bigg) \\
&= \sum_{X_1 \subseteq S} \mathbb{P}_0(S \setminus X_1) \cdot (-1)^{-|X_1|} \cdot \sum_{\substack{C_1 \supseteq X_1: \\ C_1 \subseteq S, \\ j \in C_1}} \frac{(-1)^{|C_1|}}{|C_1|}
\end{aligned}
$$

$$= \sum_{X_1 \subseteq S} \sum_{\substack{C_1 \in \langle S \setminus X_1 \rangle \uplus X_1: \\ j \in C_1}} \frac{(-1)^{|C_1|-|X_1|}}{|C_1|} \cdot \mathbb{P}_0(S \setminus X_1)$$

$$= \sum_{X_1 \subseteq S} \sum_{\substack{C_1 \in \langle S \setminus X_1 \rangle \uplus X_1: \\ j \in C_1}} \frac{(-1)^{|C_1|-|X_1|}}{|C_1|} \cdot \mathbb{P}_0(S \setminus X_1) \cdot \sum_{X_2 \subseteq N \setminus S} \sum_{C_2 \in \langle \{N \setminus S\} \setminus X_2 \rangle} (-1)^{|C_2|}$$

$\big[$since the summation over X_2 and C_2 is equal to 1 if $X_2 = N \setminus S$ and 0, otherwise$\big]$

$$= \sum_{X_1 \subseteq S} \sum_{\substack{C_1 \in \langle S \setminus X_1 \rangle \uplus X_1: \\ j \in C_1}} \frac{(-1)^{|C_1|-|X_1|}}{|C_1|} \cdot \mathbb{P}_0(S \setminus X_1) \cdot \sum_{X_2 \subseteq N \setminus S} (-1)^{-|X_2|} \cdot \sum_{C_2 \in \langle \{N \setminus S\} \setminus X_2 \rangle \uplus X_2} (-1)^{|C_2|}$$

$$= \sum_{X_1 \subseteq S} \sum_{\substack{C_1 \in \langle S \setminus X_1 \rangle \uplus X_1: \\ j \in C_1}} \sum_{X_2 \subseteq N \setminus S} \sum_{C_2 \in \langle \{N \setminus S\} \setminus X_2 \rangle \uplus X_2} \frac{(-1)^{|C_1|+|C_2|-|X_1|-|X_2|}}{|C_1|} \cdot \mathbb{P}_0(S \setminus X_1)$$

$$= \sum_{X_1 \subseteq S} \sum_{X_2 \subseteq N \setminus S} \sum_{\substack{C_1 \in \langle S \setminus X_1 \rangle \uplus X_1: \\ j \in C_1}} \sum_{C_2 \in \langle \{N \setminus S\} \setminus X_2 \rangle \uplus X_2} \frac{(-1)^{|C|-|X|}}{|C_1|} \cdot \mathbb{P}_0(S \setminus X_1)$$

$\big[$where $X = X_1 \cup X_2$ and $C = C_1 \cup C_2$.$\big]$

$$= \sum_{X_1 \subseteq S} \sum_{X_2 \subseteq N \setminus S} \sum_{\substack{C_1 \in \langle S \setminus X_1 \rangle \uplus X_1: \\ j \in C_1}} \sum_{C_2 \in \langle \{N \setminus S\} \setminus X_2 \rangle \uplus X_2} \frac{(-1)^{|C|-|X|}}{|C_1|} \cdot \mathbb{P}_0(N \setminus X)$$

$\big[$since the summation is nonzero only if $X_2 = N \setminus S$$\big]$

$$= \sum_{\substack{C \subseteq N: \\ j \in C}} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \frac{1}{|C \cap S|} \cdot \mathbb{P}_0(N \setminus X)$$

$$= \sum_{\substack{C \subseteq N: \\ j \in C}} \frac{1}{|C \cap S|} \cdot \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X)$$

$$= \sum_{C \subseteq N} \lambda_C \cdot \mathbf{I}[j \in S] \cdot \mathbf{I}[j \in C] \cdot \frac{1}{|C \cap S|}, \text{ where } \lambda_C = \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X),$$

which is exactly the equation to compute the probability of purchasing $j \in S$ under the offer set $S \subseteq N$ under the consideration set model. We note that this model is well-defined since we have that

$$\lambda_C = \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X) = \sum_{\bar{C} \subseteq \bar{X}} (-1)^{|\bar{X}|-|\bar{C}|} \cdot \mathbb{P}_0(\bar{X}) = H(0, \bar{C}) \geq 0,$$

where the non-negativeness is provided by the default regularity. Moreover, it follows from Theorem 1 that $\boldsymbol{\lambda}$ is defined uniquely.

Finally, to prove the necessity of the theorem (the "⇒" direction), it suffices to show that $\mathbb{P}_j(S \setminus \{k\}) - \mathbb{P}_j(S)$ is invariant to the exchange of the indexes j and k, which is shown below.

$$\mathbb{P}_j(S \setminus \{k\}) - \mathbb{P}_j(S)$$
$$= \sum_{\substack{C \subseteq N: \\ j \in C}} \frac{\lambda_C}{|C \cap \{S \setminus \{k\}\}|} - \sum_{\substack{C \subseteq N: \\ j \in C}} \frac{\lambda_C}{|C \cap S|}$$

$$
\begin{aligned}
&= \sum_{\substack{C\subseteq N:\\ j\in C\\ k\in C}} \frac{\lambda_C}{|C\cap\{S\setminus\{k\}\}|} + \sum_{\substack{C\subseteq N:\\ j\in C\\ k\notin C}} \frac{\lambda_C}{|C\cap\{S\setminus\{k\}\}|} - \sum_{\substack{C\subseteq N:\\ j\in C}} \frac{\lambda_C)}{|C\cap S|}\\
&= \sum_{\substack{C\subseteq N:\\ j\in C\\ k\in C}} \frac{\lambda_C}{|C\cap\{S\setminus\{k\}\}|} + \sum_{\substack{C\subseteq N:\\ j\in C\\ k\notin C}} \frac{\lambda_C}{|C\cap\{S\setminus\{k\}\}|} - \sum_{\substack{C\subseteq N:\\ j\in C\\ k\in C}} \frac{\lambda_C}{|C\cap S|} - \sum_{\substack{C\subseteq N:\\ j\in C\\ k\notin C}} \frac{\lambda_C}{|C\cap S|}\\
&= \sum_{\substack{C\subseteq N:\\ j\in C\\ k\in C}} \frac{\lambda_C}{|C\cap\{S\setminus\{k\}\}|} + \sum_{\substack{C\subseteq N:\\ j\in C\\ k\notin C}} \frac{\lambda_C}{|C\cap S|} - \sum_{\substack{C\subseteq N:\\ j\in C\\ k\in C}} \frac{\lambda_C}{|C\cap S|} - \sum_{\substack{C\subseteq N:\\ j\in C\\ k\notin C}} \frac{\lambda_C}{|C\cap S|}\\
&= \sum_{\substack{C\subseteq N:\\ j\in C\\ k\in C}} \frac{\lambda_C}{|C\cap\{S\setminus\{k\}\}|} - \sum_{\substack{C\subseteq N:\\ j\in C\\ k\in C}} \frac{\lambda_C}{|C\cap S|} = \sum_{\substack{C\subseteq N:\\ j\in C\\ k\in C}} \frac{\lambda_C}{|C\cap S|-1} - \sum_{\substack{C\subseteq N:\\ j\in C\\ k\in C}} \frac{\lambda_C}{|C\cap S|}.
\end{aligned}
$$

EC.3. Supplementary Proofs and Results for Section 4

EC.3.1. Proof of Theorem 3

First, it is worth observing that each product i belongs to exactly one block $I_{\mathbf{b}}\in\mathcal{I}$. This follows directly from the construction of the blocks $I_{\mathbf{B}}\in\mathcal{I}$. Specifically, a product i belongs to a block $I_{\mathbf{b}}$ if and only if the following two conditions are satisfied: (a) product i is included in all consideration sets C_j where $b_j=1$; and (b) product i is not included in any consideration sets C_j where $b_j=0$. If either of these two conditions is violated for a specific $I_{\mathbf{b}}\in\mathcal{I}$, it is clear that product i cannot belong to that block, as determined by Equation (11). Consequently, each product i uniquely belongs to the block $I_{\mathbf{b}}$ where $b_j=\mathbb{I}[i\in C_j]$ for all $j\in K$. It also implies that the non-empty sets in $\mathcal{I}=\{I_{\mathbf{b}}\mid\mathbf{b}\in\mathcal{B}\}$ form a partition of N. Thus, in what follows below, we can rewrite the revenue function $\mathrm{Rev}_{C_j}(S)$ defined in Equation (9). Specifically, when $|S\cap C_j|\geq 1$, we have the following:

$$
\mathrm{Rev}_{C_j}(S)=\frac{\sum_{i\in S\cap C_j} r_i}{|S\cap C_j|}=\frac{\sum_{\mathbf{b}\in\mathcal{B}}\sum_{i\in S\cap C_j\cap I_{\mathbf{b}}} r_i}{\sum_{\mathbf{b}\in\mathcal{B}}|S\cap C_j\cap I_{\mathbf{b}}|}=\frac{\sum_{\mathbf{b}\in\mathcal{B}:b_j=1}\sum_{i\in S\cap I_{\mathbf{b}}} r_i}{\sum_{\mathbf{b}\in\mathcal{B}:b_j=1}|S\cap I_{\mathbf{b}}|},\tag{EC.4}
$$

where the second equality holds because $\mathcal{I}=\{I_{\mathbf{b}}\mid\mathbf{b}\in\mathcal{B}\}$ partitions N, and the third equality follows from the fact that, for any binary vector $\mathbf{b}\in\mathcal{B}$ and any $j\in K$, we have the following equality:

$$
S\cap C_j\cap I_{\mathbf{b}}=\begin{cases} S\cap I_{\mathbf{b}}, & \text{if } b_j=1,\\ \emptyset, & \text{if } b_j=0.\end{cases}
$$

In what follows next, we fix an arbitrary assortment $S\subseteq N$. We claim that if $S\cap I_{\mathbf{b}}$ is not revenue-ordered in any block $I_{\mathbf{b}}\in\mathcal{I}$, then S is not an optimal assortment. We prove this claim by construction. To this end, we assume that $S\cap I_{\mathbf{b}^\dagger}$ is not revenue-ordered in block $I_{\mathbf{b}^\dagger}$ for a binary vector $\mathbf{b}^\dagger$. Next, we construct a new assortment S' such that: (a) $S'\cap I_{\mathbf{b}}=S\cap I_{\mathbf{b}}$ for all $\mathbf{b}\neq\mathbf{b}^\dagger$; and (b) $S'\cap I_{\mathbf{b}^\dagger}$ consists of the most expensive $|S\cap I_{\mathbf{b}^\dagger}|$ products in $I_{\mathbf{b}^\dagger}$. In other words, under the partition $\mathcal{I}$, the two assortments S and S' coincide except in the block $I_{\mathbf{b}^\dagger}$. Note that given $S \cap I_{\mathbf{b}^\dagger}$ is not revenue-ordered, we know that $S \cap I_{\mathbf{b}^\dagger}$ is not the empty set. Furthermore, it can be seen that $\sum_{i \in S \cap I_{\mathbf{b}}} r_i = \sum_{i \in S' \cap I_{\mathbf{b}}} r_i$ for any $\mathbf{b} \neq \mathbf{b}^\dagger$ and $\sum_{i \in S \cap I_{\mathbf{b}^\dagger}} r_i < \sum_{i \in S' \cap I_{\mathbf{b}^\dagger}} r_i$, while $|S \cap I_{\mathbf{b}}| = |S' \cap I_{\mathbf{b}}|$ for all $\mathbf{b} \in \mathcal{B}$. Those observations imply that S' achieves a higher revenue than S'. To formalize this claim, we consider the revenue function $\text{Rev}_{C_j}(S)$ of each customer type j as specified in Equation (EC.4). We analyze the following two cases:

- If $b_j^\dagger = 0$, we have $\text{Rev}_{C_j}(S) = \text{Rev}_{C_j}(S')$. This is because the products in block $I_{\mathbf{b}^\dagger}$ are not considered by customer type j at all if $b_j^\dagger = 0$.
- If $b_j^\dagger = 1$, then we have the following expression:

$$\begin{aligned}\text{Rev}_{C_j}(S) =& \frac{\sum_{i \in S \cap I_{\mathbf{b}^\dagger}} r_i + \sum_{\mathbf{b} \in \mathcal{B}: \mathbf{b} \neq \mathbf{b}^\dagger, b_j = 1} \sum_{i \in S \cap I_{\mathbf{b}}} r_i}{\sum_{\mathbf{b} \in \mathcal{B}: b_j = 1} |S \cap I_{\mathbf{b}}|} \\ <& \frac{\sum_{i \in S' \cap I_{\mathbf{b}^\dagger}} r_i + \sum_{\mathbf{b} \in \mathcal{B}: \mathbf{b} \neq \mathbf{b}^\dagger, b_j = 1} \sum_{i \in S' \cap I_{\mathbf{b}}} r_i}{\sum_{\mathbf{b} \in \mathcal{B}: b_j = 1} |S' \cap I_{\mathbf{b}}|} = \text{Rev}_{C_j}(S').\end{aligned}$$

As the non-emptiness of $S \cap I_{\mathbf{b}^\dagger}$ implies the non-emptiness of $I_{\mathbf{b}^\dagger}$, we know that there must exist at least one customer type $j^* \in K$ such that $b_{j^*}^\dagger = 1$. Specifically, if $b_j^\dagger = 0$ for all $j \in K$, the non-emptiness of $I_{\mathbf{b}^\dagger}$ implies that there would exist a product $i \in I_{\mathbf{b}^\dagger} = \cap_{j \in K} \eta_{b_j^\dagger}(C_j) = \cap_{j \in K} \eta_0(C_j) = \cap_{j \in K} \bar{C}_j$, violating the assumption of $\cup_{j \in K} C_j = N$ at the beginning of Section 4. Due to the existence of this customer type j^*, we have $\text{Rev}(S) = \sum_{j \in K} \lambda_j \cdot \text{Rev}_{C_j}(S) < \sum_{j \in K} \lambda_j \cdot \text{Rev}_{C_j}(S') = \text{Rev}(S')$. Thus, S is not optimal. □

EC.3.2. Proof of Proposition 2

We prove the statement by constructing a reduction from the vertex cover problem, which is known to be an NP-hard problem (Garey and Johnson 1979), to the assortment optimization problem under the stochastic set model. First, we recall how the latter problem is defined. To start with, let G denote a graph with a collection of nodes $V = \{1, ..., n\}$ and edges E, i.e., $G = (V, E)$. Next, we say that a subset $U \subseteq V$ is a vertex cover if for every edge $e = (i, j) \in E$ either $i \in U$ or $j \in U$. Then, the vertex cover problem addresses the following question: "Given a graph $G = (V, E)$ and an integer k, is there a vertex cover of size at most k?"

In what follows, we describe the reduction Φ that maps any instance of the vertex cover (VC) problem, i.e., $I_{\text{VC}} = (G, k)$, to the instance of the assortment optimization (AO) problem (10), i.e., I_{AO}, where $\Phi(I_{\text{VC}})$ is defined as follows:

- For each vertex $j \in V$, we introduce a product j with the price of 1 dollar each.
- We introduce an additional product $n + 1$ to the instance I_{AO} with the price of 3 dollars. Therefore, instance I_{AO} consists of $n + 1$ products in total.

- For each edge $(i,j) \in E$, there is a corresponding customer type which is characterized by a preselected set $C^E_{(i,j)}$ such that $C^E_{(i,j)} = \{i,j\}$. Let us label these customers as "edge customer types". We denote the collection of "edge customer types" by $\mathcal{C}^E$ such that $\mathcal{C}^E = \{C^E_{(i,j)} \mid (i,j) \in E\}$. We assign the same weight of $1/(|E|+|V|/3)$ to every customer in $\mathcal{C}^E$.
- For each vertex $j \in V$, there is a corresponding customer type which is characterized by a preselected set C^V_j such that $C^V_j = \{j, n+1\}$. Let us label these customers as "vertex customer types". We denote the collection of "vertex customer types" as $\mathcal{C}^V$ such that $\mathcal{C}^V = \{C^V_i \mid i \in V\}$. We assign the weight of $1/(3|E|+|V|)$ to every customer in $\mathcal{C}^V$. As a result, we have $|E|+|V|$ customer types in total, i.e., the collection $\mathcal{C}$ of customer types is $\mathcal{C} = \mathcal{C}^E \cup \mathcal{C}^V$. Note that the resulting distribution over subsets λ_C constructed above satisfies the property that the sum of all the weights of customers in $\mathcal{C}$ is equal to 1.

Let us first denote L as the normalizing constant, which is equal to $1/(|E|+|V|/3)$ that we will use to simplify the exposition. Then, following the aforementioned hardness result of the vertex cover problem, to establish the hardness of the assortment optimization problem under the consideration set problem it is sufficient to prove that Φ satisfies two properties:

- **Claim 1:** For a vertex cover of the size at most k in the instance I_{VC}, there is an optimal assortment in instance I_{AO} that has expected revenue of at least $(|E|+|V|-k/3) \cdot L$.
- **Claim 2:** Reciprocally, given the optimal assortment in instance I_{AO} that has expected revenue of at least $(|E|+|V|-k/3) \cdot L$, there exists an instance I_{VC} with a vertex cover of the size at most k.

Proof of Claim 1: Let us assume that $U \subseteq V$ is a vertex cover of the graph G and its cardinality is less than k, i.e., $|U| \leq k$. Then we state that the assortment $S_U = \{i \mid i \in U\} \cup \{n+1\}$ results into the expected revenue of at least $(|E|+|V|-k/3) \cdot L$. In what follows below we prove this statement which concludes the proof of Claim 1.

$$\left[\text{Expected revenue under assortment } S_U\right] = \sum_{C \in \mathcal{C}} \lambda_C \cdot \text{Rev}_C(S_U)$$

$$= \sum_{C \in \mathcal{C}^E} \lambda_C \cdot \text{Rev}_C(S_U) + \sum_{C \in \mathcal{C}^V} \lambda_C \cdot \text{Rev}_C(S_U)$$

$$= \sum_{C \in \mathcal{C}^E} \lambda_C \cdot 1 + \sum_{C \in \mathcal{C}^V} \lambda_C \cdot \text{Rev}_C(S_U) \quad \text{[since for every edge } (i,j) \in E \text{ either } i \text{ or } j \text{ is covered]}$$

$$\text{[and the price is the same for both products } i \text{ and } j \text{ and it is equal to 1]}$$

$$= \sum_{C \in \mathcal{C}^E} \lambda_C + \sum_{j \in U} \lambda_{C^V_j} \cdot \text{Rev}_{(C^V_j)}(S_U) + \sum_{j \in V \setminus U} \lambda_{C^V_j} \cdot \text{Rev}_{(C^V_j)}(S_U)$$

$$= \sum_{C \in \mathcal{C}^E} \lambda_C + \sum_{j \in U} \lambda_{C^V_j} \cdot 1/2 \cdot (1+3) + \sum_{j \in V \setminus U} \lambda_{C^V_j} \cdot 3$$

$$= \sum_{C \in \mathcal{C}^E} L + \sum_{j \in U} L/3 \cdot 1/2 \cdot (1+3) + \sum_{j \in V \setminus U} L/3 \cdot 3$$

$= L\cdot(|E|+2/3\cdot|U|+(|V|-|U|)) = L\cdot(|E|+|V|-|U|/3) \geq (|E|+|V|-k/3)\cdot L.$

Proof of Claim 2: Let S_+ denote an optimal assortment in instance I_{AO} such that its expected revenue is greater than or equal to $|E|+|V|-k/3$. It is easy to see that $n+1$ is part of the optimal assortment S_+ because adding product $n+1$ to any assortment would increase the expected revenue. Thus, we assume that $n+1\in S_+$ and $S_+ = S\cup\{n+1\}$ without loss of generality.

First of all, it is clear that all customer types in $\mathcal{C}^V$ have a positive contribution to the expected revenue under the optimal assortment S_+ because $n+1\in S_+$, i.e., the number of customer types in $\mathcal{C}^V$ that have a positive contribution to the revenue is equal to n. Moreover, we state that all the customer types in $\mathcal{C}^E$ have a positive contribution to the expected revenue under the optimal assortment S_+ and we prove that statement at the very end. For now, if we assume that the latter statement holds, then for every customer type $C^E_{(i,j)}\in\mathcal{C}^E$ to have a positive contribution to the revenue it should be the case that either $i\in S$ or $j\in S$ and thus we have that U_S, such that $U_S=\{\text{vertex } j \mid j\in S\}$, is a vertex cover. Note that in this case, the number of customer types in the set $\mathcal{C}^E$ that have a positive contribution to the revenue under the optimal assortment is equal to $|E|$. Next, we can obtain the following set of equations and inequalities.

$$\begin{aligned}
&\left[\text{Expected revenue under assortment } S_+\right] = \sum_{C\in\mathcal{C}}\lambda_C\cdot\text{Rev}_C(S_+)\\
&= \sum_{C\in\mathcal{C}^E}\lambda_C\cdot\text{Rev}_C(S_+) + \sum_{C\in\mathcal{C}^V}\lambda_C\cdot\text{Rev}_C(S_+)\\
&= |E|\cdot L + \sum_{C\in\mathcal{C}^V}\lambda_C\cdot\text{Rev}_C(S_+)\\
&= |E|\cdot L + \sum_{j\in S}\lambda_{C^V_j}\cdot\text{Rev}_{(C^V_j)}(S_+) + \sum_{j\in V\setminus S}\lambda_{C^V_j}\cdot\text{Rev}_{(C^V_j)}(S_+)\\
&= |E|\cdot L + |S|/2\cdot L/3 + 3|S|/2\cdot L/3 + \sum_{j\in V\setminus S}\lambda_{C^V_j}\cdot\text{Rev}_{(C^V_j)}(S_+)\\
&= |E|\cdot L + 2|S|\cdot L/3 + (|V|-|S|)\cdot L\\
&= (|E|+|V|-|S|/3)\cdot L\\
&\geq (|E|+|V|-k/3)\cdot L \;\; [\textbf{by the assumption in the Claim 2}]\\
&\Rightarrow |U_S| = |S| \leq k.
\end{aligned}$$

Then, to conclude the proof of Claim 2 it is sufficient to prove the aforementioned statement that all the customer types in $\mathcal{C}^E$ have a positive contribution to the expected revenue under the optimal assortment S_+. We prove that statement by contradiction. Suppose that there exists a customer type $C^E_{(i',j')}\in\mathcal{C}^E$ which does not contribute to the expected revenue, i.e., $i'\notin S_+$ and $j'\notin S_+$. Then, adding i' to S_+ will have a combination of two effects: (1) it will make customer type $C^E_{(i',j')}$ positively contribute to the expected revenue and increase the expected revenue by

L; and (2) contribution of the customer of type $C_{i'}^V$ to the expected revenue will decrease by $(3-2)\cdot L/3 = L/3$. As a result, adding item i' to the assortment S_+ has a net positive effect on the expected revenue (i.e., it increases the revenue by $2L/3$) which contradicts the fact that S_+ is an optimal assortment. □

EC.3.3. Proof of Theorem 4

In what follows, we first describe a reduction Φ from any instance $\mathcal{I}$ of the Maximal Independent Set (Max-IS) to an instance $\Phi(\mathcal{I})$ of the assortment problem (10), which consists of n products and n consideration sets. We then utilize the inapproximability result of Max-IS (Håstad 1999), which states that the Max-IS problem is NP-hard to approximate within factor $O(n^{1-\epsilon})$, to show that the assortment problem (10) is also NP-hard to approximate. In this proof, we use $[n]$ to denote $\{1,2,\dots,n\}$.

Let a Max-IS instance $\mathcal{I}$ be defined on a graph $G=(V,E)$, where $V=\{v_1,\dots,v_n\}$ is a set of vertices and E is a set of edges. For each vertex $v_i\in V$, we use $N^-(i)$ to denote the indices of v_i's neighbors whose indices are smaller than i, i.e.,

$$N^-(i)=\{j\in[n]\mid (v_i,v_j)\in E \text{ and } j<i\}. \tag{EC.5}$$

Now we describe the mapping Φ. For each vertex $v_i\in V$, we introduce a product i with price $r_i=n^{2i}/\alpha$, where $\alpha=1/(\sum_{i=1}^n n^{-2i})$. For each vertex $v_i\in V$, we also construct a consideration set $C_i=\{i\}\cup N^-(i)$, which has weight $\lambda_i=\alpha/n^{2i}$. Next, we state the following two claims:

Claim EC.1. *For any independent set $U\subseteq V$ in the instance $\mathcal{I}$, there exists a corresponding assortment S_U in the instance $\Phi(\mathcal{I})$ such that $Rev(S_U)\geq|U|$.*

Claim EC.2. *Given any assortment S in the instance $\Phi(\mathcal{I})$ such that $Rev(S)=\Omega(n^{\frac{1}{2}+\frac{\epsilon}{2}})$, there exists a corresponding independent set U_S of size $\Omega(n^{\frac{\epsilon}{2}})$ in the instance $\mathcal{I}$.*

Before proving those two claims, we make the following two statements. First, the construction of the mapping Φ in our proof follows the one in Aouad et al. (2018), except that we do not need to specify the preference over products in each C_i. Claim 1 also appears in a similar form in the proof of inapproximability of the assortment optimization problem under the ranking-based model in Aouad et al. (2018). However, Claim 2 is significantly different from the counterpart in Aouad et al. (2018), as it requires a very different argument to construct the independent set U_S, leading to the narrowed $O(\sqrt{n})$ gap instead of an $O(n)$ gap.

Second, the two claims jointly lead to the inapproximability of the assortment problem. Our argument proceeds as follows. Based on the analyses provided by Håstad (1999), we know that there does not exist a polynomial time algorithm to find an independent set of size at least $n^{\frac{\epsilon}{2}}$ given that the n-size vertex graph instance $\mathcal{I}$ has an independent set of size $n^{1-\frac{\epsilon}{2}}$ (Khot 2010). Now, let us focus on such a Max-IS instance $\mathcal{I}$. Suppose by contradiction that there exists a polynomial-time approximation algorithm $\mathcal{A}$ which solves the assortment problem within factor $O(n^{\frac{1}{2}-\epsilon})$. We can then use this algorithm $\mathcal{A}$ to obtain an assortment S' in instance $\Phi(\mathcal{I})$ with expected revenue $\Omega(n^{\frac{1}{2}+\frac{\epsilon}{2}})$, since

$$\text{Rev}(S') \geq \frac{R(S^*)}{c_1 \cdot n^{\frac{1}{2}-\epsilon}} \geq \frac{n^{1-\frac{\epsilon}{2}}}{c_1 \cdot n^{\frac{1}{2}-\epsilon}} = (1/c_1) \cdot n^{\frac{1}{2}+\frac{\epsilon}{2}},$$

where $c_1 > 0$ is an absolute constant. Here, the first inequality follows from the definition of $\mathcal{A}$, and the second inequality follows from Claim 1. In particular, since we assume that $\mathcal{I}$ has an independent set of size $n^{1/2-\epsilon}$, it follows from Claim 1 that there exists an assortment resulting in revenue that is greater than $n^{1/2-\epsilon}$. This implies that $R(S^*) \geq n^{1/2-\epsilon}$. Given that, for S' such that $\text{Rev}(S') = \Omega(n^{\frac{1}{2}+\frac{\epsilon}{2}})$, it follows from Claim 2 that we can construct an independent set $U_{S'}$ that is of the size $\Omega(n^{\frac{\epsilon}{2}})$. Therefore, it implies that we are able to use a polynomial-time algorithm to construct a $\Omega(n^{\frac{\epsilon}{2}})$-size independent set which leads to a contradiction to the inapproximability result of the Max-IS problem.

In what follows below we prove the aforementioned two claims which completes the proof of the theorem.

Proof of Claim 1: Let U be the independent set in the claim and let us define $S_U \equiv \{i \mid v_i \in U\}$. Since $N^-(i) \cap U = \emptyset$ if $v_i \in U$, we know that $C_i \cap S_U = \{i\}$. Therefore,

$$\text{Rev}\,(S_U) = \sum_{i=1}^{n} \lambda_i \cdot \text{Rev}_{C_i}\,(S_U) \geq \sum_{i \in S_U} \lambda_i \cdot r_i = |U|.$$

Proof of Claim 2: Let S be the assortment in the claim. We define two collections of the consideration sets,

$$G = \left\{ i \in [n] \;\middle|\; \text{Rev}_{C_i}(S) \geq \frac{r_i}{\sqrt{n}} = \frac{n^{2i}}{\alpha\sqrt{n}} \right\} \quad \text{and} \quad B = \left\{ i \in [n] \;\middle|\; \text{Rev}_{C_i}(S) < \frac{r_i}{\sqrt{n}} = \frac{n^{2i}}{\alpha\sqrt{n}} \right\},$$

which we denote as the "Good" and "Bad" collections of consideration sets, respectively. We make the following three statements about the "Good" collection of consideration sets G.

First, we argue that $G \subseteq S$. Let i, by contradiction, be a product such that $i \in G$ but $i \notin S$. Then, we have that

$$\text{Rev}_{C_i}(S) \leq \frac{n^{2(i-1)}}{\alpha} = \frac{n^{2i}}{\alpha \cdot n^2} < \frac{n^{2i}}{\alpha \cdot \sqrt{n}}, \tag{EC.6}$$

which contradicts the definition of the "Good" collection of consideration sets G. In the aforementioned chain of equalities and inequalities (EC.6), the first inequality follows because of the fact that all other products in $C_i \cap S$ have a price of at most $n^{2(i-1)}/\alpha$.

Second, we argue that $|G| = \Omega(n^{\frac{1}{2}+\frac{\epsilon}{2}})$. To this end, the assumption imposed on the expected revenue of S in Claim 2 leads to the following set of equalities and inequalities:

$$c_2 \cdot n^{\frac{1}{2}+\frac{\epsilon}{2}} \le \mathrm{Rev}(S) = \sum_{i\in G} \lambda_i \cdot \mathrm{Rev}_{C_i}(S) + \sum_{i\in B} \lambda_i \cdot \mathrm{Rev}_{C_i}(S) < \sum_{i\in G} 1 + \sum_{i\in B} \frac{1}{\sqrt{n}} \le |G| + \sqrt{n}, \tag{EC.7}$$

where $c_2 > 0$ is an absolute constant. In the aforementioned chain of equalities and inequalities, the second inequality follows because $\mathrm{Rev}_{C_i}(S) \le r_i \le n^{2i}/\alpha$ and $\lambda_i = \alpha/n^{2i}$. Overall, inequality (EC.7) implies that $|G| = \Omega(n^{\frac{1}{2}+\frac{\epsilon}{2}})$. In other words, asymptotically, only consideration sets belonging to the "Good" collection of consideration sets G are contributing to the revenue $\mathrm{Rev}(S)$.

Third, we argue that for each $i \in G$, we have that $|C_i \cap S| \le 2\sqrt{n}$. We prove this statement by contradiction. Assume by contradiction that $|C_i \cap S| \ge 2\sqrt{n} + 1$, then we have that

$$\mathrm{Rev}_{C_i}(S) = \frac{\sum_{j\in C_i\cap S} r_j}{|C_i \cap S|} \le \frac{\sum_{j\in C_i\cap S} r_j}{2\sqrt{n}+1} < \frac{2 \cdot n^{2i}/\alpha}{2\sqrt{n}} = \frac{n^{2i}}{\alpha\sqrt{n}},$$

which contradicts the fact that $i \in G$. In the aforementioned expression, the second inequality follows because of the following chain of equalities and inequalities:

$$\sum_{j\in C_i\cap S} r_j < \frac{n^{2i}}{\alpha} + \frac{n^{2i-2}}{\alpha} + \frac{n^{2i-4}}{\alpha} + \ldots = \frac{n^{2i}}{\alpha} \cdot \left(1 + \frac{1}{n^2} + \frac{1}{n^4} + \ldots\right) < \frac{n^{2i}}{\alpha} \cdot 2,$$

which is a valid expression as long as $n > 1$.

Next, taking into account all three statements related to the "Good" collection of the consideration sets G, we construct the independent set U_S from S as follows. We begin with an empty set, i.e., $U_S = \emptyset$. Then, starting from the largest index i_1 in G, we add i_1 to U_S and delete the elements in C_{i_1} from G, i.e., $G \leftarrow G \backslash C_{i_1}$. We then proceed to the next largest index i_2 in the updated G, add i_2 to U_S, and remove the elements in C_{i_2} from G. This process is repeated until G becomes empty. The resulting U_S forms an independent set because the neighbors of each vertex are removed from G once the vertex is added to U_S. We also observe that $|U_S| = \Omega(n^{\frac{\epsilon}{2}})$, because each time a vertex is added, we remove at most $|C_i \cap G| \le |C_i \cap S| \le 2\sqrt{n}$ elements from G. Thus, we at least can add

$$\frac{|G|}{2\sqrt{n}} = \frac{\Omega(n^{\frac{1}{2}+\frac{\epsilon}{2}})}{2\sqrt{n}} = \Omega\left(n^{\frac{\epsilon}{2}}\right)$$

items to the set U_S which shows that $|U_S| = \Omega\left(n^{\frac{\epsilon}{2}}\right)$. □

EC.3.4. Scalability of the Mixed-Integer Linear Program (12)

To numerically test the scalability of the mixed-integer linear program (12) when solving the assortment problem (10), we first generate random problem instances as follows. We vary the number of products n and the number of consideration sets $k = |\mathcal{C}|$ while fixing a constant s which is the size of consideration sets. Specifically, for each $C \in \mathcal{C}$, we sample the consideration set C from the set $N = \{1, 2, \ldots, n\}$ restricting its size by s, uniformly at random. We repeat this sampling for all k consideration sets in $\mathcal{C}$. We further let $\lambda_C = 1/k$ and set $s = 5$, motivated by the empirical evidence from the literature (Hauser and Wernerfelt 1990, Hauser 2014) and from Table 1 where we can see that customers usually consider only a few products. We let $n \in \{250, 500, 750, 1000\}$ and $k \in \{n, 3n, 5n\}$, where the scale factor between n and k is consistent with our empirical estimation outcomes in Table 1. We set the time limit to 20 minutes when solving the mixed-integer linear program (MILP) using Gurobi. If Gurobi fails to find the optimal solution within the time limit, we report the bound of the optimality gap returned by the solver. Thus, for each pair (n, k), we randomly generate ten instances and calculate the average runtime (in minutes) and the optimality gap. The scalability results are reported in Table EC.1.

n	k	Time (min)	Gap (%)
250	250	0.00	0.00
250	750	5.99	0.00
250	1250	17.71	0.68
500	500	0.08	0.00
500	1500	20.03	0.57
500	2500	20.05	1.28
750	750	0.30	0.01
750	2250	20.07	0.87
750	3750	20.13	1.30
1000	1000	0.53	0.01
1000	3000	20.14	1.16
1000	5000	20.23	1.59

Table EC.1 Scalability of the MILP approach when solving the assortment problem.

It follows from this table that within a twenty-minute time limit, even for large instances like $(n, k) = (1000, 5000)$, the MILP achieves a solution with an optimality gap of no more than 2%. This underscores our claim that MILP (12) is efficient and can be useful for practical applications. Additionally, the performance of the MILP could be further enhanced by adopting the mixed-integer conic reformulation suggested by Şen et al. (2018).

EC.4. Model Estimation

EC.4.1. Estimation Methodology

We propose a maximum likelihood estimation (MLE) procedure based on the expectation maximization (EM) algorithm and the column generation algorithm method. Overall, our estimation methodology follows van Ryzin and Vulcano (2014, 2017), who estimate the ranking-based model using the EM algorithm and column generation method. A similar framework has also been adapted to estimate the decision forest model (Chen and Mišić 2022).

We assume we have the sales transactions denoted by $\{(S_t, i_t)\}_{t=1,\ldots,T}$, which consist of purchase records over T periods. A purchase record at time t is characterized by a tuple (S_t, i_t), where $S_t \subseteq N$ denotes the subset of products on offer in period t and $i_t \in S_t \cup \{0\}$ denotes the product purchased in period t. In addition, we let $\mathcal{S}$ denote the set of unique assortments observed in the data and let m denote the cardinality of this set, i.e., $m = |\mathcal{S}|$. We term each assortment in $\mathcal{S}$ as a *historical assortment*, i.e., an assortment that was offered in the past. For each historical assortment $S \in \mathcal{S}$ and $i \in S^+ \equiv S \cup \{0\}$, we let $\tau(S, i)$ denote the number of times the tuple (S, i) appears in the sales transactions $\{(S_t, i_t)\}_{t=1,\ldots,T}$.

EC.4.1.1. Maximum Likelihood Estimation (MLE). To estimate the model from data $\{(S_t, i_t)\}_{t=1,\ldots,T}$, we first write down the following optimization problem with respect to a fixed collection $\bar{\mathcal{C}}$ of consideration sets:

$$P^{\mathrm{MLE}}\left(\bar{\mathcal{C}}\right): \quad \underset{\boldsymbol{\lambda} \geq 0, \mathbf{v}}{\text{maximize}} \quad \mathcal{L}(\mathbf{v}) \tag{EC.8a}$$

$$\text{such that} \quad \sum_{C \in \bar{\mathcal{C}}} (\mathbf{A}_S)_{i,C}\, \lambda_C = v_{i,S} \qquad \forall S \in \mathcal{S},\ i \in N^+, \tag{EC.8b}$$

$$\sum_{C \in \bar{\mathcal{C}}} \lambda_C = 1, \tag{EC.8c}$$

where λ_C is an element of the distribution $\boldsymbol{\lambda} \in \mathbb{R}^{|\bar{\mathcal{C}}|}$ over the sets in $\bar{\mathcal{C}}$, $v_{i,S}$ is the aggregate likelihood that customers would pick the alternative i from the offer set S, and the objective is the log-likelihood function $\mathcal{L}(\mathbf{v}) \equiv \sum_{S \in \mathcal{S}} \sum_{i \in S^+} \tau(S, i) \cdot \log(v_{i,S})$, which is a concave function. The matrix $\mathbf{A}_S$ is of size $(n+1) \times |\bar{\mathcal{C}}|$ and it maps from the model $\left(\bar{\mathcal{C}}, \boldsymbol{\lambda}\right)$ to the its choice probability $v_{i,S}$ of item i under historical assortment S. The element $(\mathbf{A}_S)_{i,C}$ is equal to $1/|S \cap C|$ if $i \in S \cap C$ and 0, otherwise.

First, it is worth emphasizing that when $\bar{\mathcal{C}}$ includes all subsets of N (i.e., $\bar{\mathcal{C}}$ is the power set of N) then optimization problem (EC.8) solves the MLE problem under the consideration set model. When $\bar{\mathcal{C}}$ does not include all the subsets in N, we refer the problem $P^{\mathrm{MLE}}\left(\bar{\mathcal{C}}\right)$ specified above as a *restricted* optimization problem. Second, we note that the maximization problem (EC.8) is concave and thus we can use a convex programming solver to find the optimal solution. Alternatively, we will exploit an expectation-maximization (EM) algorithm to obtain the optimal solution to the restricted problem $P^{\mathrm{MLE}}\left(\bar{\mathcal{C}}\right)$. We find that this EM algorithm is more efficient than existing general convex programming solvers. For now, we assume $\bar{\mathcal{C}}$ is fixed and we will come back to discuss how to enlarge it during the estimation procedure.

EC.4.1.2. Solving MLE with the EM Algorithm. The EM algorithm is a method for finding the maximum likelihood estimates of parameters in statistical models where the data are incomplete or there are unobserved latent variables. The algorithm proceeds in two steps: the expectation step (E-step), where the expectation of the complete-data log-likelihood is calculated given the current parameter estimates, and the maximization step (M-step), where the parameters are updated to maximize the expected complete-data log-likelihood. In the context of our MLE problem $P^{\text{MLE}}\left(\bar{\mathcal{C}}\right)$ specified above, the unobserved latent variable is the probability mass λ_C of each customer type $C \in \bar{\mathcal{C}}$, since customer types are not directly observed from the sales transaction data. We start our EM algorithm with arbitrary initial parameter estimate $\boldsymbol{\lambda}^{\text{est}}$. Then, we repeatedly apply the "E" and "M" steps, which are described below, until convergence.

E-step: If customer types associated with each transaction are known to us then we would be able to represent the complete-data log-likelihood function as $\mathcal{L}^{\text{complete}} = \sum_{C \in \bar{\mathcal{C}}} \tau(C) \cdot \log \lambda_C + \text{constant}$, where $\tau(C)$ is the number of transactions made by customers of the type C. However, in the context of our problem, the customer types are not observed in the data and we thus replace $\tau(C)$ from the aforementioned complete-data log-likelihood function by its conditional expectation $\mathbf{E}\left[\tau(C) \mid \boldsymbol{\lambda}^{\text{est}}\right]$ which allows us to obtain $\mathbf{E}\left[\mathcal{L}^{\text{complete}} \mid \boldsymbol{\lambda}^{\text{est}}\right]$. To obtain the conditional expectation, we first apply the Bayes' rule

$$\mathbb{P}\left(C \mid S, i, \boldsymbol{\lambda}^{\text{est}}\right) = \frac{\mathbb{P}\left(i \mid C, S, \boldsymbol{\lambda}^{\text{est}}\right) \cdot \mathbb{P}\left(C \mid S, \boldsymbol{\lambda}^{\text{est}}\right)}{\sum_{C \in \bar{\mathcal{C}}} \mathbb{P}\left(i \mid C, S, \boldsymbol{\lambda}^{\text{est}}\right) \cdot \mathbb{P}\left(C \mid S, \boldsymbol{\lambda}^{\text{est}}\right)} = \frac{\mathbb{P}\left(i \mid C, S\right) \cdot \lambda_C^{\text{est}}}{\sum_{C \in \bar{\mathcal{C}}} \mathbb{P}\left(i \mid C, S\right) \cdot \lambda_C^{\text{est}}}, \tag{EC.9}$$

where $\mathbb{P}\left(C \mid S, i, \boldsymbol{\lambda}^{\text{est}}\right)$ is the probability that an item i from the offer set S is purchased by a customer of type C when $\boldsymbol{\lambda}^{\text{est}}$ represents the probability mass of different customer types and $\mathbb{P}\left(i \mid C, S\right)$ is the probability to choose item i from the offer set S by a customer of type C. Consequently, the aforementioned Equation (EC.9) allows us to compute $\hat{\tau}(C)$, which is the expected number of transactions made by customers of the type C by $\hat{\tau}(C) = \mathbf{E}\left[\tau(C) \mid \boldsymbol{\lambda}^{\text{est}}\right] = \sum_{S \in \mathcal{S}} \sum_{i \in S^+} \tau(S, i) \cdot \mathbb{P}\left(C \mid S, i, \boldsymbol{\lambda}^{\text{est}}\right)$. We thus compute the expectation of the complete-data log-likelihood function as $\mathbf{E}\left[\mathcal{L}^{\text{complete}} \mid \boldsymbol{\lambda}^{\text{est}}\right] = \sum_C \hat{\tau}(C) \cdot \log \lambda_C^{\text{est}}$.

M-step: By maximizing the expected complete data log-likelihood function $\mathbf{E}\left[\mathcal{L}^{\text{complete}} \mid \boldsymbol{\lambda}^{\text{est}}\right]$, we obtain the optimal solution $\lambda_C^* = \hat{\tau}(C) / \sum_C \hat{\tau}(C)$ which is used to update $\boldsymbol{\lambda}^{\text{est}}$.

Finally, we note that the EM algorithm has been widely used in the operations management (OM) field to estimate various choice models when customer types are not directly observed in the data. Examples include the mixed MNL model (Train 2009), the ranking-based model (van Ryzin and Vulcano 2017), the Markov chain choice model (Şimşek and Topaloglu 2018), the consider-then-choose (CTC) model (Jagabathula et al. 2024), and DAG-based choice models (Jagabathula et al. 2022).

EC.4.1.3. Consideration Set Discovery Algorithm. In theory, as mentioned above, to solve the MLE problem one can exploit the aforementioned EM algorithm and apply it to the optimization problem $P^{\text{MLE}}\left(\bar{\mathcal{C}}\right)$ when the collection of customer segments is equal to the power set of N, i.e., $\bar{\mathcal{C}} = 2^N$. However, this approach becomes highly intractable as the number of products n increases. In this case, the number of decision variables would grow exponentially with n.

Therefore, we propose an alternative way to solve the MLE problem which is based on the *column generation* (CG) procedure (Bertsimas and Tsitsiklis 1997), a widely used procedure to solve large-scale optimization problems with linear constraints. In accordance with the CG framework, instead of solving the full-scale optimization problem $P^{\mathrm{MLE}}(2^N)$ directly, we will repeatedly execute the following two steps: (i) find the optimal solution $(\mathbf{v}, \boldsymbol{\lambda})$ to the restricted problem $P^{\mathrm{MLE}}(\bar{\mathcal{C}})$ by the EM algorithm described in Section EC.4.1.2; (ii) solve a subproblem (which will be described later) to find a new column and expand the restricted problem $P^{\mathrm{MLE}}(\bar{\mathcal{C}})$ by concatenating the column. Notice that each column in the complete problem $P^{\mathrm{MLE}}(2^N)$ corresponds to a consideration set. Therefore, when we introduce a new column C^* to the restricted problem $P^{\mathrm{MLE}}(\bar{\mathcal{C}})$, we equivalently augment the *support* of the distribution $\boldsymbol{\lambda}$ from $\bar{\mathcal{C}}$ to $\bar{\mathcal{C}} \cup \{C^*\}$. We will repeat these two steps until we reach optimality or meet a stopping criterion such as a runtime limit. In the following, we provide the details on how we augment the support of the distribution $\boldsymbol{\lambda}$ by solving a subproblem, which will complete the description of our MLE framework for calibrating the consideration set model.

Let $(\mathbf{v}, \boldsymbol{\lambda})$ be the optimal primal solution of the optimization problem $P^{\mathrm{MLE}}\left(\bar{\mathcal{C}}\right)$, which is defined for a fixed collection $\bar{\mathcal{C}}$ of consideration sets. Let $(\boldsymbol{\alpha}, \beta)$ denote the dual solution of the optimization problem $P^{\mathrm{MLE}}\left(\bar{\mathcal{C}}\right)$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_S)_{S\in\mathcal{S}}$ corresponds to the first set of constraints, with each $\boldsymbol{\alpha}_S$ being a $(n+1)$-dimensional vector, and β corresponds to the unit-sum constraint (see Section EC.4.2 for further notes). We then solve the following CG subproblem:

$$\max_{C\subseteq N}\left[\sum_{S\in\mathcal{S}}\sum_{i\in N^+}\alpha_{S,i}\cdot(\mathbf{A}_S)_{i,C}+\beta\right]=\max_{C\subseteq N}\left[\sum_{S\in\mathcal{S}}\left(\alpha_{S,0}\cdot\mathbb{I}\left[C\cap S=\emptyset\right]+\sum_{i\in S}\frac{\alpha_{S,i}\cdot\mathbb{I}\left[i\in C\right]}{|C\cap S|}\right)+\beta\right], \tag{EC.10}$$

where $(\mathbf{A}_S)_{i,C}$ is defined in Section EC.4.1.1 and $\alpha_{S,i}$ is the element of $\boldsymbol{\alpha}_S$ for $i\in S^+$. In Problem (EC.10), we assume "$0/0 = 0$" to simply the notation. Let C^* be the optimal solution to the CG subproblem (EC.10). We add C^* to the set $\bar{\mathcal{C}}$, which is the support of the $\boldsymbol{\lambda}$, if and only if the optimal objective value of the subproblem is positive. This indicates that adding C^* to the set $\bar{\mathcal{C}}$ improves the objective value of $P^{\mathrm{MLE}}\left(\bar{\mathcal{C}}\right)$. Consequently, if the optimal objective value of the subproblem (EC.10) is not positive, we terminate our algorithm, as it indicates that the current set $\bar{\mathcal{C}}$, along with the distribution $\boldsymbol{\lambda}$, already represents the consideration set model with maximum likelihood.

Finally, it remains to solve the subproblem (EC.10). We exploit a mixed-integer linear optimization (MILP) formulation similar to the assortment problem (12). Here, in addition to the variables

u_S and $q_{S,i}$ introduced for the linearization, we also introduce binary variable z_S to present the indicator $\mathbb{I}[C \cap S]$. Along with binary variable x_i that represent $\mathbb{I}[i \in C]$, we have the following MILP that solves the CG subproblem (EC.10)

$$P^{\text{CG-sub}}(\boldsymbol{\alpha}): \quad \underset{\mathbf{x},\mathbf{z},\mathbf{u},\mathbf{q}}{\text{maximize}} \quad \sum_{S\in\mathcal{S}} \left[\alpha_{S,0} \cdot z_s + \sum_{i\in S} \alpha_{S,i} \cdot q_{S,i} \right], \tag{EC.11a}$$

$$\text{such that} \quad 0 \le q_{S,i} \le x_i, \qquad \forall S \in \mathcal{S}, i \in S, \tag{EC.11b}$$

$$0 \le q_{S,i} \le u_S, \qquad \forall S \in \mathcal{S}, i \in S, \tag{EC.11c}$$

$$u_S + x_i \le q_{S,i} + 1, \qquad \forall S \in \mathcal{S}, i \in S, \tag{EC.11d}$$

$$z_S + \sum_{i\in S} q_{S,i} = 1, \qquad \forall S \in \mathcal{S}, \tag{EC.11e}$$

$$x_i \in \{0,1\}, \ z_S \in \{0,1\}, \ u_S \in [1/n, 1], \quad \forall i \in N, S \in \mathcal{S}. \tag{EC.11f}$$

Overall, the proposed estimation procedure consists of solving a finite sequence of MILPs. The size of each MILP scales in $O(n+m)$ in the number of variables and in $O(nm)$ in the number of constraints, where n is the number of products and m is the number of historical assortments. Thus, not surprisingly, estimating a consideration set model can be more tractable than estimating a ranking-based model van Ryzin and Vulcano (2014), as the corresponding CG subproblem of the latter model scales in $O(n^2+m)$ in the number of variables and in $O(n^3+nm)$ in the number of constraints. We relegate additional discussion to Section EC.4.2.

Additionally, we note that the consideration sets we estimate can be any subset of N, meaning any set of products. To enhance interpretability and enable potential applications in new product development and pricing, we also demonstrate how product features (such as price) can be incorporated into our model calibration. In Section EC.4.3, we introduce a model estimation framework that integrates contextual information through conjunctive models, disjunctive models, and compensatory models – well-known heuristic rules for constructing consideration sets. For an overview, see Section 1.1.

EC.4.2. Additional Discussion

In this section, we discuss the estimation procedure described in Section EC.4.1.

Obtaining the optimal dual variables $\boldsymbol{\alpha}$ and β. We can obtain the optimal dual solution once we know the optimal primal solution. Notice that our MLE problem (EC.8) is of the same form as Problem (7) in van Ryzin and Vulcano (2014), where the difference is only in the entries in the constraint matrix $\mathbf{A}$. In particular, each column of our constraint matrix $\mathbf{A} = (\mathbf{A}_S)_{S\in\mathcal{S}}$ encodes the decision of a consideration set under historical assortments $S \in \mathcal{S}$. In contrast, each column of the constraint matrix $\mathbf{A}$ in van Ryzin and Vulcano (2014) encodes the decision of a ranking. Since our MLE problem takes the same form as van Ryzin and Vulcano (2014), the dual variables $\boldsymbol{\alpha}$ and

β satisfy the same KKT conditions except that the constraint matrix has different coefficients. In the end, given the optimal primal variable, the optimal dual variable can be obtained following Equations (8) and (9) in van Ryzin and Vulcano (2014).

Optimality of the EM algorithm. Note that given a collection of consideration sets $\bar{\mathcal{C}}$, we use the EM algorithm to find the optimal distribution $\boldsymbol{\lambda}$ over sets in $\bar{\mathcal{C}}$ that maximizes the log-likelihood objective, i.e., to solve $P^{\text{MLE}}(\bar{\mathcal{C}})$. Recall that $P^{\text{MLE}}(\bar{\mathcal{C}})$ is a concave maximization problem with a strictly concave objective. Therefore, any locally optimal solution of $P^{\text{MLE}}(\bar{\mathcal{C}})$ implies the globally optimal solution of $P^{\text{MLE}}(\bar{\mathcal{C}})$. As the EM algorithm guarantees that each M-step improves the objective value of $P^{\text{MLE}}(\bar{\mathcal{C}})$ and thus finds the local maximum (McLachlan and Krishnan 2007), we know that the EM approach of solving the optimization problem (EC.8) guarantees global optimality.

Only solving a finite number of times of the subproblem/MILP. This is because at each iteration, we introduce a new variable λ_C that corresponds to a consideration set C to the full program $P^{\text{MLE}}(\bar{\mathcal{C}})$ for $\bar{\mathcal{C}} = 2^N$. As there are only a finite number of consideration sets (theoretically at most 2^n even though we have a much smaller set of consideration sets in practice according to numerous literature in marketing and Table 1), the estimation procedure thus only solves a finite number of MILPs.

EC.4.3. Feature-based Model Estimation

In Section EC.4.1, we demonstrated how the consideration set model $(\mathcal{C}, \boldsymbol{\lambda})$ can be calibrated from sales transaction data in its most general form. Specifically, each consideration set $C \in \mathcal{C}$ in the model can be any subset of the product universe N, where each product is characterized by a unique feature vector. This approach eliminates the need to explicitly incorporate product features into our estimation framework. It also aligns with existing literature in operations research on choice modeling estimation. Notably, the estimation of nonparametric choice models, including ranking-based models (Farias et al. 2013, van Ryzin and Vulcano 2014), the Markov chain model (Blanchet et al. 2016, Şimşek and Topaloglu 2018), the consider-then-choose model (Jagabathula et al. 2024), and the decision forest model (Chen and Mišić 2022), follows a similar convention.

In contrast, as discussed in Section 1.1, customers may form their consideration sets based on heuristic models that explicitly incorporate product features. In this section, we integrate several well-documented feature-based heuristic rules from the marketing science literature into our estimation framework. Specifically, we assume that each product $i \in N$ is represented by a feature vector $\boldsymbol{\chi}_i \in \mathbb{R}^d$ in a d-dimensional vector space. We define $[a] \equiv 1, 2, \ldots, a$, where a is a positive integer.

EC.4.3.1. Conjunctive Rule. The conjunctive rule heuristics assumes that a product is included in the consideration set if it satisfies a sequence of screening criteria. Each criterion is defined by a feature $p \in [d]$ and a threshold value t_p. A product i satisfies the screening criterion for feature p if and only if $\chi_{i,p} \leq t_p$. Mathematically, the conjunctive model is parameterized by $(E, \mathbf{t}_E)$, where $E \subseteq [d]$ is a set of features, and $\mathbf{t}_E = (t_e)_{e \in E}$ is the vector of threshold values associated with the features in E. A conjunctive model $(E, \mathbf{t}_E)$ leads to a consideration set C in the following way:

$$C = \bigcap_{e \in E} \{ i \in N \mid \chi_{i,e} \leq t_e \} . \tag{EC.12}$$

Note that without loss of generality, we only need to consider the screening rules with the smaller-or-equal-to sign (i.e., $\leq$) since one can always introduce an additional feature e' such that $\chi_{i,e'} = -\chi_{i,e}$ for $i \in N$. Conjunctive models are among the most popular non-compensatory models in the marketing science literature and have been supported by abundant empirical studies (Pras and Summers 1975, Brisoux and Laroche 1981, Laroche et al. 2003, Gilbride and Allenby 2004, Jedidi and Kohli 2005).

Let $\mathcal{C}_{\text{conjunctive}}$ be the collection of all consideration sets that can be specified by the conjunctive models. Since our goal is to calibrate a consideration set model by modeling the consideration set formation by means of the conjunctive rule, we solve the following MLE problem:

$$\text{maximize} \left[P^{\text{MLE}} \left(\bar{\mathcal{C}} \right) \mid \bar{\mathcal{C}} \subseteq \mathcal{C}_{\text{conjunctive}} \right] , \tag{EC.13}$$

where P^{MLE} is defined as in Problem (EC.8) and $\bar{\mathcal{C}}$ is the collection of customer segments. As discussed in Section EC.4.1, if one can enumerate all sets in $\mathcal{C}_{\text{conjunctive}}$, we can simply let $\bar{\mathcal{C}} = \mathcal{C}_{\text{conjunctive}}$ and apply the EM algorithm to obtain the probability mass of every consideration set from the set $\bar{\mathcal{C}}$. However, given that, in theory, the number of customer segments can be exponential in the number of items in the product universe, we also propose solving Problem (EC.13) by the column generation approach. To this end, instead of solving the subproblem (EC.10) where a candidate consideration set can be any subset in N, we ensure that the candidate consideration set belongs to the collection $\mathcal{C}_{\text{conjunctive}}$, i.e., we solve the following optimization problem:

$$\max_{C \subseteq N} \left[\sum_{S \in \mathcal{S}} \sum_{i \in N+} \alpha_{S,i} \cdot (\mathbf{A}_S)_{i,C} + \beta \,\middle|\, C \in \mathcal{C}_{\text{conjunctive}} \right] . \tag{EC.14}$$

We introduce a MILP formulation to solve Problem (EC.14). In order to avoid using the Big-M parameter in the formulation, which usually results in a weaker system of constraints, we use relative values instead of the exact values of product features. Let L_p be the number of unique values in $\{\chi_{i,p}\}_{i \in N}$, i.e., the number of unique values of the feature p observed in the sales data (Mišić 2020, Akçakuş and Mišić 2021). Next, we let $v_{p,\ell}$ denote the ℓth lowest value in L_p such that

$v_{p,1} < v_{p,2} < \cdots < v_{p,L_p-1} < v_{p,L_p}$. Then, we let $v_{p,0} = -\infty$ and $\tau_{p,i}$ be the value of $\ell \in \{1,\ldots,L_p\}$ such that $\chi_{i,p} = v_{p,\ell}$, i.e., $\tau_{p,i}$ indicates the position of product i in the ranking of all unique values of feature p.

After specifying $(\tau_{p,\ell})_{p\in[d],\ell\in[L_p]}$, in what follows below, we propose an MILP formulation to solve Problem (EC.14). We further assume that the conjunctive model consists of at most R screening rules, i.e., $|E|\leq R$ in Equation (EC.12). The value of R can be either set to be d for the most general conjunctive model or to be a value smaller than d, as it is widely assumed that consumers have cognitive and physical limitations and cannot take into account an unlimited set of features when making purchasing decisions. In the latter case, one can view R as a fixed parameter of the model and determine its value by cross-validation.

As above, $\mathbf{x}, \mathbf{z}, \mathbf{u}$, and $\mathbf{q}$ are our decision variables that have the same interpretation as in Problem (EC.11). In addition, we introduce a new binary decision variable $\lambda_{r,p,\ell}$ which is equal to 1 if the rth screening rule is $\chi_{i,p} \leq v_{p,\ell}$ and 0, otherwise. Then, we provide our MILP formulation as follows:

$$\begin{aligned}
P^{\text{CG-sub}}_{\text{conjunctive}}(\boldsymbol{\alpha}): \quad & \underset{\mathbf{x},\mathbf{z},\mathbf{u},\mathbf{q},\boldsymbol{\lambda}}{\text{maximize}} && \sum_{S\in\mathcal{S}} \left[\alpha_{S,0}\cdot z_s + \sum_{i\in S} \alpha_{S,i}\cdot q_{S,i} \right] && && \text{(EC.15a)}\\
& \text{such that} && \text{Constraint (EC.11b) - (EC.11f)} && && \text{(EC.15b)}\\
& && \sum_{p=1}^{d}\sum_{\ell=1}^{L_p} \lambda_{r,p,\ell} = 1, && \forall r\in[R], && \text{(EC.15c)}\\
& && x_i \leq \sum_{p=1}^{d}\sum_{\ell:\ell\geq\tau_{p,i}} \lambda_{r,p,\ell}, && \forall i\in N, r\in[R], && \text{(EC.15d)}\\
& && \sum_{r=1}^{R}\sum_{p=1}^{d}\sum_{\ell:\ell\geq\tau_{p,i}} \lambda_{r,p,\ell} \leq x_i + d - 1, && \forall i\in N, && \text{(EC.15e)}\\
& && \lambda_{r,p,\ell}\in\{0,1\}, && \forall r\in[R], p\in[d], \ell\in L_p, && \text{(EC.15f)}
\end{aligned}$$

where constraint (EC.15b) ensures that variables $\mathbf{x}, \mathbf{z}, \mathbf{u}$, and $\mathbf{q}$ satisfy the same constraints as in Problem (EC.11). Constraint (EC.15c) ensures that each screening rule is associated with exactly one feature and one threshold value. Constraints (EC.15d) and (EC.15e) ensure that the consideration sets are formed by a conjunctive rule. In particular, Constraint (EC.15d) states that product i is in the consideration set "only if" it satisfies all screening rules. Constraint (EC.15e) completes the "if" direction of the statement. Notice that we allow the screening rules to be repeated. Therefore, one can apply up to $|R|$ screening rules for the conjunctive model by solving Problem (EC.15). While we inevitably have to introduce new binary decision variables $\lambda_{r,p,\ell}$ to characterize the conjunctive model, we can further decrease the number of binary variables by relaxing the binary restriction imposed on the consideration set decisions $\mathbf{x}$, since they remain integral after relaxation.

EC.4.3.2. Disjunctive Rule. Similarly to the conjunctive rule, the disjunctive rule heuristic is also defined by a set of screening criteria. Differently, the disjunctive rule implies that a product belongs to a customer's consideration set if at least one of the screening criteria is satisfied. Mathematically, a consideration set C can be represented by a disjunctive model if there exists a collection of features E and threshold values such that

$$C = \bigcup_{e \in E} \{i \in N \mid \chi_{i,e} \leq t_e\}. \tag{EC.16}$$

While the disjunctive model is usually benchmarked as an alternative to the conjunctive model, it receives much less attention in the literature. For more details, we refer the reader to the empirical studies conducted by Pras and Summers (1975), Gilbride and Allenby (2004), Jedidi and Kohli (2005). To calibrate the consideration set model under this heuristic, we follow the aforementioned procedure described in Section EC.4.3.1 with the only difference that we solve the column generation subproblem in the following way:

$$\max_{C \subseteq N} \left[\sum_{S \in \mathcal{S}} \sum_{i \in N^+} \alpha_{S,i} \cdot (\mathbf{A}_S)_{i,C} + \beta \;\middle|\; C \in \mathcal{C}_{\text{disjunctive}} \right],$$

where $\mathcal{C}_{\text{disjunctive}}$ is the collection of all consideration sets that can be represented by a disjunctive rule. One can solve this subproblem by formulating it as the following MILP, which is similar to Problem (EC.15).

$$\begin{aligned}
P^{\text{CG-sub}}_{\text{disjunctive}}(\boldsymbol{\alpha}): \quad & \underset{\mathbf{x},\mathbf{z},\mathbf{u},\mathbf{q},\boldsymbol{\lambda}}{\text{maximize}} && \sum_{S \in \mathcal{S}} \left[\alpha_{S,0} \cdot z_s + \sum_{i \in S} \alpha_{S,i} \cdot q_{S,i} \right] && && \text{(EC.17a)} \\
& \text{such that} && \text{Constraint (EC.11b) - (EC.11f)} && && \text{(EC.17b)} \\
& && \sum_{p=1}^{d} \sum_{\ell=1}^{L_p} \lambda_{r,p,\ell} = 1, && \forall r \in [R], && \text{(EC.17c)} \\
& && x_i \leq \sum_{r=1}^{R} \sum_{p=1}^{d} \sum_{\ell : \ell \geq \tau_{p,i}} \lambda_{r,p,\ell}, && \forall i \in N, && \text{(EC.17d)} \\
& && \sum_{p=1}^{d} \sum_{\ell : \ell \geq \tau_{p,i}} \lambda_{r,p,\ell} \leq x_i, && \forall i \in N, r \in [R], && \text{(EC.17e)} \\
& && \lambda_{r,p,\ell} \in \{0,1\}, && \forall r \in [R], p \in [d], \ell \in L_p. && \text{(EC.17f)}
\end{aligned}$$

The difference between Problem (EC.17) and Problem (EC.15) is in how the consideration set decision $\mathbf{x}$ and the screening rule decision $\boldsymbol{\lambda}$ factor in the system of the constraints in the optimization problem. In the former problem, constraint (EC.17d) states that a product is considered "only if" there exists a screening criteria which is satisfied, and Constraint (EC.17e) completes the argument by adding the "if" direction. Similarly to Problem (EC.15), one can relax the binary restriction imposed on $\mathbf{x}$ while solving Problem (EC.17) as $\mathbf{x}$ remains integral after this relaxation.

EC.4.3.3. Compensatory Rule. Both conjunctive and disjunctive rules belong to the class of non-compensatory decision processes, i.e., they assume that customers do not consider the trade-off between product features (e.g., not willing to trade higher price for higher quality). On the other hand, it is common in conjoint analysis (Srinivasan and Shocker 1973, Evgeniou et al. 2005) and in discrete-choice modeling (Ben-Akiva et al. 1985, Feldman et al. 2022) to assume that product utility is a composite of the contributions from each product feature. In particular, one can assume that the utility function is linear in product features. In this section, we follow this common approach and assume that a compensatory model is a part-worth vector $\boldsymbol{\pi} \in \mathbb{R}^{d+1}$. A product is in the consideration set if and only if its part-worth utility is nonnegative, i.e.,

$$C = \{i \in N \mid \pi_0 + \sum_{p=1}^{d} \pi_{i,p} \geq 0\}.$$

To estimate the consideration set model under this compensatory heuristics, we follow the aforementioned procedure described in Section EC.4.3.1 with the only difference that we solve the column generation subproblem in the following way:

$$\max_{C \subseteq N} \left[\sum_{S \in \mathcal{S}} \sum_{i \in N^+} \alpha_{S,i} \cdot (\mathbf{A}_S)_{i,C} + \beta \,\middle|\, C \in \mathcal{C}_{\text{linear}} \right],$$

where $\mathcal{C}_{\text{linear}}$ is the collection of all consideration sets that can be represented by a linear compensatory rule. This subproblem can be formulated as the following MILP:

$$P^{\text{CG-sub}}_{\text{linear}}(\boldsymbol{\alpha}): \quad \underset{\mathbf{x},\mathbf{z},\mathbf{u},\mathbf{q},\boldsymbol{\pi}}{\text{maximize}} \quad \sum_{S \in \mathcal{S}} \left[\alpha_{S,0} \cdot z_s + \sum_{i \in S} \alpha_{S,i} \cdot q_{S,i} \right] \tag{EC.18a}$$

$$\text{such that} \quad \text{Constraint (EC.11b) - (EC.11f)} \tag{EC.18b}$$

$$\sum_{p=1}^{d} \chi_{i,p} \pi_p + \pi_0 \leq x_i, \qquad \forall i \in N, \tag{EC.18c}$$

$$x_i - 1 + \epsilon \leq \sum_{p=1}^{d} \chi_{i,p} \pi_p + \pi_0, \qquad \forall i \in N, \tag{EC.18d}$$

where $\epsilon > 0$ is a sufficiently small constant. Constraints (EC.18c) and (EC.18d) ensure that a product is included in the consideration set if and only if the part-worth utility is nonnegative. Note that the subproblem $P^{\text{CG-sub}}_{\text{linear}}(\boldsymbol{\alpha})$ is invariant if one multiplies the vector $\boldsymbol{\pi}$ by a positive number. Therefore, we do not need to introduce the big-M constant in constraints (EC.18c) and (EC.18d).

EC.5. Additional Results on the Predictive Performance and Model Calibration

EC.5.1. Prediction Performance Measured by KL Divergence

We formally define the KL-divergence metric as follows:

$$\text{KL} = -\frac{1}{\sum_{S \in \mathcal{S}} \tau_o(S)} \cdot \sum_{S \in \mathcal{S}} \tau_o(S) \sum_{i \in S^+} \bar{p}_{i,S} \log \left(\frac{\hat{p}_{i,S}}{\bar{p}_{i,S}} \right), \tag{EC.19}$$

where parameters $\tau_o(S), \bar{p}_{i,S}$, and $\hat{p}_{i,S}$ are defined as in Section 5. Table EC.2 reports the predictive performance of each choice model introduced in Section 5, measured by the out-of-sample KL divergence. We observe qualitatively similar results in Table EC.2 as those presented in Table 2. In both tables, the decision forest model has the best predictive performance while the MNL and the independent demand models provide the worst performance. Although it is a special case, the consideration set model demonstrates predictive performance comparable to the ranking-based and mixed MNL models, with all three achieving similar KL divergence scores of approximately 3.30×10^{-2}. Furthermore, when combined with rankings, the consideration set model achieves performance comparable to the Markov chain model, with KL divergence scores of 3.15×10^{-2} for the CSMR (consideration set model with rankings) and 3.16×10^{-2} for the Markov chain model.

Category	ID	MNL	MMNL	RBM	MC	DF	CSM	CSM2	CSMR
Beer	5.61	4.45	3.87	4.03	3.75	2.98	3.97	3.86	3.73
Coffee	12.58	9.49	7.68	7.71	7.29	6.54	8.20	8.25	7.42
Deodorant	1.82	1.00	0.94	0.99	0.94	0.89	0.93	0.97	0.93
Frozen Dinners	3.54	2.94	2.51	2.41	2.36	1.89	2.46	2.44	2.40
Frozen Pizza	6.26	3.84	2.82	2.63	2.85	1.82	2.84	3.04	2.56
Household Cleaners	3.44	2.67	2.37	2.46	2.21	2.23	2.38	2.35	2.26
Hotdogs	8.61	6.51	5.39	5.44	5.28	5.56	5.50	5.89	5.29
Margarine/Butter	3.47	2.31	1.80	1.82	1.84	1.21	1.80	1.82	1.80
Milk	16.77	8.85	6.19	6.37	6.28	5.94	6.40	6.52	6.22
Mustard/Ketchup	7.20	3.87	3.27	3.48	3.11	3.65	3.33	3.31	3.15
Salty Snacks	4.22	2.28	1.56	1.45	1.67	1.37	1.64	1.79	1.55
Shampoo	2.78	1.52	1.27	1.53	1.25	1.09	1.32	1.35	1.28
Soup	5.03	3.70	3.19	3.13	2.99	2.10	3.04	3.15	2.95
Spaghetti/Sauce	6.98	4.19	3.35	3.25	3.34	1.54	3.42	3.51	3.29
Tooth Brush	5.08	2.69	2.29	2.76	2.25	1.73	2.54	2.62	2.42
Average	6.23	4.02	3.25	3.29	3.16	2.77	3.32	3.39	3.15

Table EC.2 Out-of-sample prediction performance results measured by KL-divergence (in unit of 10^{-2}).

EC.5.2. Faster Convergence in CSM Model Calibration

In this subsection, we aim to compare the consideration set model with the ranking-based model, as the two share structural similarities. Both models are nonparametric: the consideration set model is characterized by a distribution over subsets, while the ranking-based model is defined by a distribution over rankings. As discussed earlier in Section 5, neither model consistently outperforms the other in terms of predictive performance, as demonstrated in our case study in Sections 5 and EC.5.1.

Then, it is worth emphasizing that scalability is a common challenge when calibrating nonparametric choice models. In particular, solving the likelihood maximization problem under the ranking-based model using the column generation method can be computationally demanding, as its subproblem must be solved via an integer program (van Ryzin and Vulcano 2014). However, compared to the ranking-based model, the consideration set model is significantly more tractable. As discussed in Section EC.4.1, the column generation subproblem of the consideration set model can be formulated as an integer program with $O(n+m)$ binary variables and $O(n^2+nm)$ constraints, where n is the number of products and $m=|\mathcal{S}|$ is the number of the historical assortments in the dataset. This is computationally much simpler than the subproblem for the ranking-based model, which requires solving an integer program with $O(n^2+m)$ binary variables and $O(n^3+nm)$ constraints.

Table EC.3 provides numerical evidence of the computational tractability of the consideration set model. The second to fourth columns in the table present the number of products (n), assortments (m), and transactions ($|\mathcal{T}|$) for each product category following data preprocessing. The fifth and sixth columns present the estimation runtimes for the consideration set model (T_{CSM}) and the ranking-based model (T_{RBM}), respectively. A runtime limit of 30 minutes was imposed, and if the estimation procedure failed to terminate within this limit, the runtime is recorded as 1800.00 in Table EC.3.

Across the five product categories where the estimation procedures for both models fully converged to the optimal solution, the consideration set model demonstrated an average estimation speed that was 380% faster than that of the ranking-based model. In one category (*Milk*), the consideration set model successfully converged before the time limit, while the ranking-based model failed to do so. For the remaining categories, the estimation procedures for both models did not fully converge within the runtime limit due to the large number of assortments (m). However, we expect the consideration set model would still terminate earlier. In these cases, both models exhibit the tailing-off effect of the column generation method (Desrosiers and Lübbecke 2005), where convergence slows significantly as the algorithm approaches a sufficiently small optimality gap, with each iteration yielding only marginal improvements to the objective value.

We further highlight the computational efficiency advantage of using the consideration set model in the estimation process. In Figure EC.1, we illustrate the changes in the log-likelihood value while estimating both the consideration set and ranking-based models for the *Salty Snacks* category. Similar qualitative results are observed when considering other product categories. Specifically, we show how the in-sample log-likelihood for both models improves over time as the estimation algorithms run. From Figure EC.1, we observe that the consideration set model's estimation algorithm converges to a nearly optimal solution in a very short time (around 70 seconds), while the ranking-based model takes approximately four minutes to achieve the same log-likelihood value. However, as shown in the figure, the estimation algorithm of the ranking-based model eventually reaches a

Product Category	n	m	$\lvert\mathcal{T}\rvert$	T_{CSM}	T_{RBM}
Beer	19	721	759,968	1800.0	1800.0
Coffee	17	603	749,867	1800.0	1800.0
Deodorant	13	181	539,761	365.7	1355.4
Frozen Dinners	18	330	1,963,025	1800.0	1800.0
Frozen Pizza	12	138	584,406	111.6	439.7
Household Cleaners	21	883	562,615	1800.0	1800.0
Hotdogs	15	533	202,842	1800.0	1800.0
Margarine/Butter	11	27	282,649	10.4	20.5
Milk	18	347	476,899	1231.2	1800.0
Mustard/Ketchup	16	644	266,291	1800.0	1800.0
Salty Snacks	14	152	1,476,847	268.2	1786.4
Shampoo	15	423	574,711	1800.0	1800.0
Soup	17	315	1,816,879	1800.0	1800.0
Spaghetti/Italian Sauce	12	97	552,033	64.9	509.4
Tooth Brush	15	699	392,079	1800.0	1800.0

Table EC.3 Runtime comparison (in seconds) for estimating the consideration set model (CSM) and the ranking-based model (RBM) across fifteen product categories in the IRI dataset. The runtime is capped at 30 minutes (1,800 seconds).

higher log-likelihood value than that of the consideration set model, although the improvement is marginal.

Notably, for better illustration, we extend the trajectory of the consideration set model beyond the termination point as a horizontal line. In contrast, the estimation process for the ranking-based model fully terminates at $T_{\text{RBM}} = 1786.$ seconds. As the improvement in log-likelihood per data point falls below 0.0005 after $T = 600$ seconds for the ranking-based model, this marginal change is not noticeable at the scale of Figure EC.1. Therefore, for clarity, we plot the trajectory only up to $T = 600$ seconds. This further illustrates the tailing-off effect of the column generation method, where a substantial portion – approximately two-thirds – of the runtime is spent narrowing the final 0.1% of the optimality gap.

EC.5.3. Symmetry of the Demand Cannibalization

As discussed in Section 3.3, the symmetric cannibalization property is a key characteristic of the consideration set model. On one hand, this property differentiates the model from other members of the RUM class (e.g., the mixed MNL model) and enhances its tractability compared to the ranking-based model. On the other hand, it may limit the model's ability to fully capture customers' purchase behavior.

In this section, we empirically examine whether the symmetric cannibalization property adversely affects the predictive performance of the consideration set model when compared to the state-of-

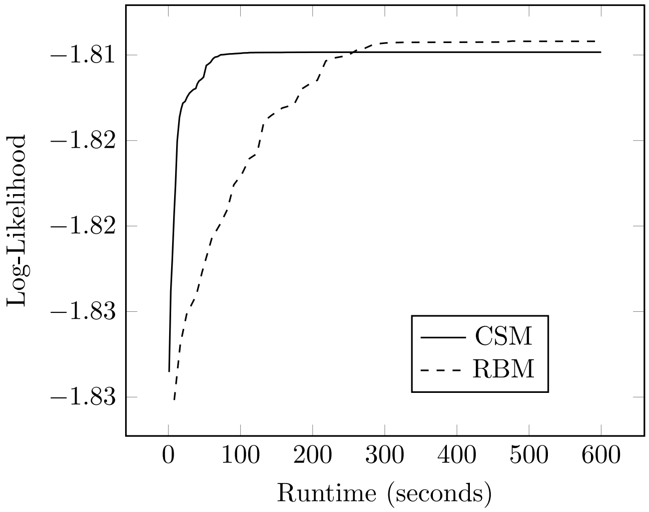


Figure EC.1 Change in the in-sample log-likelihood value over time during the MLE algorithm. The y-axis represents the log-likelihood, while the x-axis shows the runtime (in seconds) for the consideration set model (solid line) and the ranking-based model (dashed line) for the Salty Snacks category.

the-art mixed MNL model. First, recall that in Section 3.3, we defined a choice model as consistent with the symmetric cannibalization property if and only if the following equation holds:

$$[\mathbb{P}_j(S\setminus\{a_k\})-\mathbb{P}_j(S)]-[\mathbb{P}_k(S\setminus\{a_j\})-\mathbb{P}_k(S)]=0, \tag{EC.20}$$

for all assortments $S\subseteq N$ such $|S|\geq 2$ and pairs of products $j,k\in S$. To quantify the extent to which the symmetric cannibalization property is violated, we propose a *cannibalization asymmetry index*, defined as:

$$\frac{1}{|\{S:|S|\geq 2\}|}\sum_{S:|S|\geq 2}\frac{1}{\binom{|S|}{2}}\sum_{j\neq k}\left|\frac{[\mathbb{P}_j(S\setminus\{a_k\})-\mathbb{P}_j(S)]-[\mathbb{P}_k(S\setminus\{a_j\})-\mathbb{P}_k(S)]}{\mathbb{P}_j(S)+\mathbb{P}_k(S)}\right|. \tag{EC.21}$$

Intuitively, this index measures the degree to which the symmetric cannibalization property, as expressed in Equation (EC.20), is violated across all assortments $S\subseteq N$ such $|S|\geq 2$ and all pairs of products $j,k\in S$. The denominator, $\mathbb{P}_j(S)+\mathbb{P}_k(S)$, is included to ensure the comparability of violations across assortments of different sizes. For larger assortments, $\mathbb{P}_j(S)$ and $\mathbb{P}_k(S)$ tend to be smaller than the corresponding purchase probabilities in smaller assortments. This adjustment normalizes the index and mitigates the impact of assortment size on the measurement of violations. A higher cannibalization asymmetry index indicates a greater degree of violation of the symmetric cannibalization property in the choice data. Notably, if the choice data are fully consistent with the consideration set model, the cannibalization asymmetry index will be zero.

We further note that the cannibalization asymmetry index defined in Equation (EC.21) requires access to choice probabilities for nearly all assortments in the product universe. However, as shown

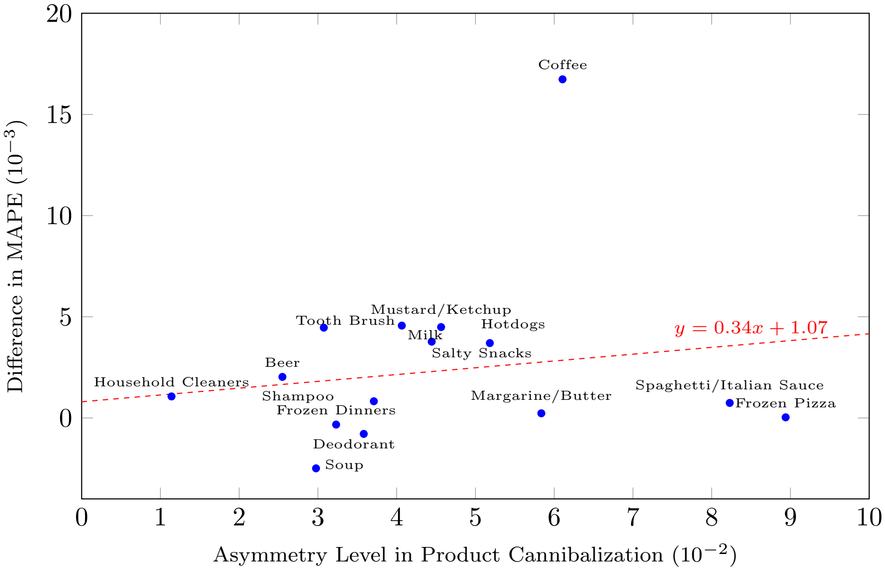


Figure EC.2 Scatter plot of the improvement of mixed MNL model over consideration set model in the prediction task against the cannibalization asymmetry index, for the fifteen product categories. The higher the improvement of the mixed MNL model over the consideration set model the higher the difference in their MAPE score.

in Table 1, our sales transaction data contain only a limited number of unique assortments for each product category, making it impossible to compute the index directly from the data. To address this limitation, we assume that the choice data generation process follows the mixed MNL model estimated from data (see Section 5). By calibrating the mixed MNL model, we can estimate the choice probabilities and compute the cannibalization asymmetry index for any collection of assortments. Note that the mixed MNL model is not constrained by the symmetric cannibalization property. Second, even when using the mixed MNL model as the ground truth for choice probabilities, calculating the cannibalization asymmetry index remains computationally challenging for product categories with a large number of items. To overcome this, we employ Monte Carlo simulations to approximate the index. Specifically, for each instance, we randomly sample an assortment S such that $|S| \geq 2$, then randomly select two indices j and k from S. We perform this process for a total of 10,000 instances and compute the index as the average over these simulations.

Figure EC.2 presents a scatterplot showing the improvement of the mixed MNL model over the consideration set model, measured by the difference in MAPE scores, against the cannibalization asymmetry index across fifteen product categories. The plot reveals significant variation in the predictive performance improvement of the mixed MNL model over the consideration set model across different product categories. To better understand this variation, we include a linear regression trendline, represented by a red dashed line. The trend suggests a correlation between the improvement in the predictive performance of the mixed MNL model and the cannibalization asymmetry index. This finding implies that violations of the symmetric cannibalization assumption may contribute to the underperformance of the consideration set model in prediction tasks. However, the results also highlight exceptions. For example, the consideration set model performs comparably to the mixed MNL model in categories such as *Spaghetti/Italian Sauce* and *Frozen Pizza*, despite their high asymmetry indices. Moreover, it outperforms the mixed MNL model in categories like *Soup* and *Deodorant*, which have intermediate asymmetry indices. What Figure EC.2 conveys is actually a positive message: when demand cannibalization is not highly asymmetric, the consideration set model remains competitive in its predictive power and is practical for real-world use.

EC.6. Empirical Analysis of Revenue Performance

In this section, we evaluate the performance of the consideration set model in a revenue management task using the IRI Dataset (Bronnenberg et al. 2008). Specifically, we focus on assortment planning as a classical revenue management application, as discussed in Section 4. Additionally, we explore the connection between end-to-end performance and the identifiability property established in Section 3. We demonstrate how the identifiability of a choice model limits the number of maximum-likelihood estimates that fit the data equally well, resulting in more stable operational decisions. To this end, we compare the performance of the consideration set model with the ranking-based model, which is non-identifiable.

EC.6.1. Uncertainty Set and Identifiability

Following the notation from Section EC.4.1.1 and given the data $\{(S_t, i_t)\}_{t\in\mathcal{T}}$, we assume that we have the estimated parameters of the consideration set model using the MLE framework (see Section EC.4.1). We let $\mathbf{v}^*_{\mathrm{CSM},S}$ be the choice probability vector over a historical assortment $S \in \mathcal{S} = \{S_1, \dots, S_m\}$ under this consideration set model. Herein, historical assortments are those appearing in the training dataset. The vector $\mathbf{v}^*_{\mathrm{CSM},S}$ has length $n+1$, where its i-th component represents the probability of purchasing product $i \in N^+$ from assortment S^+. If $i \notin S$, then the i-th component is zero. Next, we define $\mathbf{v}^*_{\mathrm{CSM}}$ as the vertical concatenation of $\mathbf{v}^*_{\mathrm{CSM},S_1}$, ..., and $\mathbf{v}^*_{\mathrm{CSM},S_m}$, resulting in a vector of length $m(n+1)$. This vector can also be expressed as follows:

$$\mathbf{v}^*_{\mathrm{CSM}} = \arg\max_{\mathbf{v}} \left\{ \mathcal{L}(\mathbf{v}) \mid \exists\, \boldsymbol{\lambda}_{\mathrm{CSM}} \geq \mathbf{0} :\ \mathbf{1}^T \boldsymbol{\lambda}_{\mathrm{CSM}} = 1,\ \mathbf{A}_S^{\mathrm{CSM}} \boldsymbol{\lambda}_{\mathrm{CSM}} = \mathbf{v}_S,\ \forall S \in \mathcal{S} \right\},$$

where, with a slight abuse of notation, $\mathbf{A}_S^{\mathrm{CSM}}$ is a $(n+1) \times 2^n$ matrix, which is the same as $\mathbf{A}_S$ in Problem (EC.8) for $\bar{\mathcal{C}} = 2^N$. Note that the matrix $\mathbf{A}_S^{\mathrm{CSM}}$ maps the distribution $\boldsymbol{\lambda}_{\mathrm{CSM}} \in \mathbb{R}_+^{2^n}$, representing the consideration set model, to the choice probability vector $\mathbf{v}_S$ for each historical assortment $S \in \mathcal{S}$. The objective $\mathcal{L}$ in the given expression corresponds to the log-likelihood function described in Section EC.4. We further define matrix $\mathbf{A}^{\mathrm{CSM}}$ as the vertical concatenation of $\mathbf{A}_{S_1}^{\mathrm{CSM}}$, $\mathbf{A}_{S_2}^{\mathrm{CSM}}$, ..., $\mathbf{A}_{S_m}^{\mathrm{CSM}}$, resulting in $\mathbf{A}^{\mathrm{CSM}}$ being an $m(n+1) \times 2^n$ matrix.

Similarly, we assume that we have the estimated parameters of the ranking-based model using the MLE framework (van Ryzin and Vulcano 2014, 2017). To this end, we let $\mathbf{v}^*_{\text{RBM},S}$ denote the choice probability vector over a historical assortment $S \in \mathcal{S}$ under this ranking-based model. Then, the vector $\mathbf{v}^*_{\text{RBM},S}$, of length $n+1$, has its i-th component indicating the probability of purchasing product $i \in N^+$ from the assortment S, or zero if $i \notin S$. Next, we define $\mathbf{v}^*_{\text{RBM}}$ as the vertical concatenation of $\mathbf{v}^*_{\text{RBM},S_1}, \ldots, \mathbf{v}^*_{\text{RBM},S_m}$, resulting in a vector of length $m(n+1)$, which can be equivalently expressed as follows:

$$\mathbf{v}^*_{\text{RBM}} = \arg\max_{\mathbf{v}} \left\{ \mathcal{L}(\mathbf{v}) \mid \exists \boldsymbol{\lambda}_{\text{RBM}} \geq \mathbf{0} : \ \mathbf{1}^T \boldsymbol{\lambda}_{\text{RBM}} = 1, \ \ \mathbf{A}^{\text{RBM}}_S \boldsymbol{\lambda}_{\text{RBM}} = \mathbf{v}_S, \ \ \forall S \in \mathcal{S} \right\},$$

where $\mathbf{A}^{\text{RBM}}_S$ is a $(n+1) \times (n+1)!$ matrix. Note that each column of $\mathbf{A}^{\text{RBM}}_S$ corresponds to a ranking-based purchasing decision under assortment S. Particularly, the matrix element $(\mathbf{A}^{\text{RBM}}_S)_{i,\sigma} = \mathbb{I}\left[i = \arg\min_{j \in S^+} \sigma(j)\right]$ indicates whether product $i \in N^+$ is the most preferred one under ranking σ in assortment S. Then, $\mathbf{A}^{\text{RBM}}$ is formed by vertically concatenating $\mathbf{A}^{\text{RBM}}_{S_1}, \ldots, \mathbf{A}^{\text{RBM}}_{S_m}$. Notably, both vectors $\mathbf{v}^*_{\text{CSM}}$ and $\mathbf{v}^*_{\text{RBM}}$ are uniquely determined due to the strict concavity of the log-likelihood function $\mathcal{L}$, ensuring well-defined MLE solutions for both models.

Next, we define the uncertainty sets as follows:

$$\mathcal{U}_{\text{CSM}} \equiv \left\{ \boldsymbol{\lambda}_{\text{CSM}} \geq \mathbf{0} \,|\, \mathbf{A}^{\text{CSM}} \boldsymbol{\lambda}_{\text{CSM}} = \mathbf{v}^*_{\text{CSM}}, \ \mathbf{1}^T \boldsymbol{\lambda}_{\text{CSM}} = 1 \right\},$$
$$\mathcal{U}_{\text{RBM}} \equiv \left\{ \boldsymbol{\lambda}_{\text{RBM}} \geq \mathbf{0} \,|\, \mathbf{A}^{\text{RBM}} \boldsymbol{\lambda}_{\text{RBM}} = \mathbf{v}^*_{\text{RBM}}, \ \mathbf{1}^T \boldsymbol{\lambda}_{\text{RBM}} = 1 \right\}.$$

These sets represent all MLE solutions for their respective models. Importantly, although $\mathbf{v}^*_{\text{CSM}}$ is unique, $\mathbf{A}^{\text{CSM}}$ may have a rank lower than 2^n, resulting in multiple distributions $\boldsymbol{\lambda}_{\text{CSM}}$ satisfying $\mathbf{A}^{\text{CSM}} \boldsymbol{\lambda}_{\text{CSM}} = \mathbf{v}^*_{\text{CSM}}$. Similarly, $\mathcal{U}_{\text{RBM}}$ can contain multiple solutions due to the lower rank of $\mathbf{A}^{\text{RBM}}$.

Although the dimensions of $\boldsymbol{\lambda}_{\text{CSM}}$ and $\boldsymbol{\lambda}_{\text{RBM}}$ differ, making it difficult to directly compare the sizes of the two uncertainty sets $\mathcal{U}_{\text{CSM}}$ and $\mathcal{U}_{\text{RBM}}$, Theorem 1 indicates that $\mathcal{U}_{\text{CSM}}$ is effectively "smaller" due to the identifiability of the consideration set model. As m increases, $\mathcal{U}_{\text{CSM}}$ converges to a single solution. In contrast, the ranking-based model remains non-identifiable for $n \geq 4$ (Sher et al. 2011), and $\mathcal{U}_{\text{RBM}}$ generally contains multiple solutions. This is reflected in the fact that the rank of the constraint matrix $\mathbf{A}^{\text{RBM}}$ satisfies $\text{rank}(\mathbf{A}^{\text{RBM}}) < (n+1)!$ whenever $n \geq 4$, making the system $\mathbf{A}^{\text{RBM}} \boldsymbol{\lambda}_{\text{RBM}} = \mathbf{v}^*_{\text{RBM}}$ under-determined.

EC.6.2. Experiment Setup

Notably, the multiplicity in the uncertainty set (either $\mathcal{U}_{\text{CSM}}$ or $\mathcal{U}_{\text{RBM}}$) can introduce significant variability when the corresponding choice model is used to make assortment decisions. Although all models within an uncertainty set achieve the same in-sample likelihood value, the optimal assortments identified by these models can differ substantially. To illustrate this phenomenon, we conduct a numerical experiment using the IRI Dataset. Below, we provide a high-level overview of the experiment, followed by a detailed explanation of each step.

Overview. The experiment consists of the following three steps:

Step 1: Model Sampling: We sample (with replacement) a set of consideration set models, denoted as $\boldsymbol{\lambda}^{(1)}_{\text{CSM}}, \boldsymbol{\lambda}^{(2)}_{\text{CSM}}, \dots, \boldsymbol{\lambda}^{(\xi)}_{\text{CSM}}$, from the uncertainty set $\mathcal{U}_{\text{CSM}}$. Similarly, we sample ranking-based models, $\boldsymbol{\lambda}^{(1)}_{\text{RBM}}, \boldsymbol{\lambda}^{(2)}_{\text{RBM}}, \dots, \boldsymbol{\lambda}^{(\xi)}_{\text{RBM}}$, from the uncertainty set $\mathcal{U}_{\text{RBM}}$.

Step 2: Optimal Assortment Selection: For each $i \in [\xi]$, we compute the optimal assortment $S^{(i)}_{\text{CSM}}$ under the consideration set model $\boldsymbol{\lambda}^{(i)}_{\text{CSM}}$ and the optimal assortment $S^{(i)}_{\text{RBM}}$ under the ranking-based model $\boldsymbol{\lambda}^{(i)}_{\text{RBM}}$.

Step 3: Revenue Evaluation: We evaluate the expected revenue associated with the assortments $S^{(1)}_{\text{CSM}}, \dots, S^{(\xi)}_{\text{CSM}}$ and $S^{(1)}_{\text{RBM}}, \dots, S^{(\xi)}_{\text{RBM}}$.

Recall that each consideration set model $\boldsymbol{\lambda}^{(i)}_{\text{CSM}}$ is guaranteed to maximize the in-sample likelihood within the class of consideration set models. Consequently, if the MLE is used to fit a consideration set model to data, then $S^{(i)}_{\text{CSM}}$ is a potential downstream assortment decision. Similarly, $S^{(i)}_{\text{RBM}}$ represents a potential assortment derived from a ranking-based model estimated using the MLE approach.

We compare the revenue performance of the assortments $\{S^{(i)}_{\text{CSM}}\}_{i\in[\xi]}$ and $\{S^{(i)}_{\text{RBM}}\}_{i\in[\xi]}$. We will show that assortments derived from ranking-based models exhibit greater variability in expected revenue, with some assortments achieving substantially lower revenue compared to those derived from consideration set models.

Before presenting the results, in what follows below, we provide a detailed description of the three steps outlined above:

Details on Step 1: Note that the uncertainty set $\mathcal{U}_{\text{CSM}}$ is a polyhedron, with each vertex representing a consideration set model. We will sample vertices of $\mathcal{U}_{\text{CSM}}$ using the following procedure, where $\mathcal{S}$ denotes the collection of historical assortments in the dataset:

Sample a model from the uncertainty set $\mathcal{U}_{CSM}$

1. Select an assortment $\tilde{S} \notin \mathcal{S}$ uniformly at random.
2. Sample a simplex vector $\tilde{\mathbf{v}} \in \Delta^{|\tilde{S}|+1}$ uniformly.
3. Solve the linear program:

$$\underset{\boldsymbol{\lambda}}{\text{maximize}} \quad \left\{ \sum_{i\in\tilde{S}^+} \sum_{C\subseteq\mathcal{C}} \tilde{v}_i \left(\mathbf{A}^{\text{CSM}}_{\tilde{S}}\right)_{i,C} \lambda_C \;\middle|\; \boldsymbol{\lambda} \in \mathcal{U}_{\text{CSM}} \right\}. \tag{EC.22}$$

4. Return the optimal solution of the linear program, denoted as $\tilde{\boldsymbol{\lambda}}_{\text{CSM}}$.

Notably, the assortment $\tilde{S}$ and its associated choice vector $\tilde{\mathbf{v}}$ represent a random perturbation of the maximum likelihood solutions. When the new assortment $\tilde{S} \notin \mathcal{S}$ and its choice vector $\tilde{\mathbf{v}}$ are introduced, the returned model $\tilde{\boldsymbol{\lambda}}_{\text{CSM}} \in \mathcal{U}_{\text{CSM}}$ is the one that best aligns with this new information $(\tilde{S}, \tilde{\mathbf{v}})$ while maintaining the in-sample likelihood.

Next, by replacing $\mathbf{A}^{\text{CSM}}$ and $\mathcal{U}_{\text{CSM}}$ in the optimization problem (EC.22) with $\mathbf{A}^{\text{RBM}}$ and $\mathcal{U}_{\text{RBM}}$, respectively, we can similarly sample a ranking-based model $\tilde{\boldsymbol{\lambda}}_{\text{RBM}}$ from the uncertainty set $\mathcal{U}_{\text{RBM}}$.

Finally, we note that the optimization problem (EC.22) is actually a large-scale linear program. When applied to the uncertainty set $\mathcal{U}_{\text{CSM}}$, the problem involves $O(2^n)$ variables; whereas for $\mathcal{U}_{\text{RBM}}$, the number of variables grows at a rate of $O(n!)$. Given this scale, we use the column generation method to solve the optimization problem (EC.22) efficiently, as directly formulating and solving the problem using commercial solvers like Gurobi (Gurobi Optimization 2024) is computationally impractical. In fact, the column generation approach used for this application is analogous to the method detailed in Section EC.4.1.3, and we omit further details for brevity.

Details on Step 2: Finding the optimal assortments is rather straightforward. For each consideration set model $\boldsymbol{\lambda}_{\text{CSM}}^{(i)}$, we solve the mixed-integer linear program (12) to obtain its optimal assortment. Similarly, for each ranking-based model $\boldsymbol{\lambda}_{\text{RBM}}^{(i)}$, we find the optimal assortment via a mixed-integer programming approach (Feldman et al. 2019, Bertsimas and Mišić 2019).

Details on Step 3: As mentioned in Section EC.5.3, the IRI dataset comprises real-world transaction data but obviously does not fully reveal a ground truth choice model. Consequently, we cannot directly evaluate the revenue performance of $S_{\text{CSM}}^{(i)}$ and $S_{\text{RBM}}^{(i)}$ for each $i \in [\xi]$. To address this limitation, we assume that the mixed MNL model estimated in Section 5.3 is our ground truth choice model. Then, we let $\mathbb{P}^{\text{MM}}(j \mid S)$ denote the predicted probability of choosing product j from assortment S under the estimated mixed MNL model. Consequently, the expected revenue under an assortment S is computed as $\text{Rev}^{\text{MM}}(S) = \sum_{j \in S} r_j \cdot \mathbb{P}^{\text{MM}}(j \mid S)$, where r_j is the revenue of product j. We obtain r_j from the IRI dataset by averaging the selling price of product j across all transactions. Using this definition, we obtain the revenues $\text{Rev}^{\text{MM}}(S_{\text{CSM}}^{(i)})$ and $\text{Rev}^{\text{MM}}(S_{\text{RBM}}^{(i)})$ for assortments $S_{\text{CSM}}^{(i)}$ and $S_{\text{RBM}}^{(i)}$, respectively. We further denote the expected revenue of the optimal assortment S_{MM}^{*} under the ground-truth mixed MNL model as $\text{Rev}^{*} \equiv \text{Rev}^{\text{MM}}(S_{\text{MM}}^{*})$. As a robustness check, we will later use the estimated decision forest model as the ground truth to evaluate the expected revenue of each assortment.

EC.6.3. Results: Revenue Performance

To begin with, we first follow the three steps outlined above to generate assortments $S_{\text{CSM}}^{(i)}$ and $S_{\text{RBM}}^{(i)}$ along with their corresponding revenues $\text{Rev}^{\text{MM}}(S_{\text{CSM}}^{(i)})$ and $\text{Rev}^{\text{MM}}(S_{\text{RBM}}^{(i)})$ for $i \in [\xi]$, where we set $\xi = 100$. Then, to reduce the variance when comparing the performance of these assortments, we use the same random perturbation $(\tilde{S}, \tilde{v})$ to generate $S_{\text{CSM}}^{(i)}$ and $S_{\text{RBM}}^{(i)}$ for each i. This experiment is repeated across five product categories from the IRI Dataset: *Deodorant*, *Frozen Pizza*, *Margarine*

Butter, *Salty Snacks*, and *Spaghetti/ Italian Sauce*. We chose these five categories because they have the smallest numbers of products after preprocessing (see Table 1). Note that solving the large-scale linear program (EC.22) under the ranking-based models via the column generation method is still computationally intensive, particularly since it is repeated $\xi = 100$ times. Therefore, we had to restrict our analysis to those five categories with smaller numbers of products.

We present the results of the experiment in Table EC.4. Each pair of rows in the table provides the descriptive statistics of the collections of revenues, $\mathcal{R}_{\text{CSM}} \equiv \{\text{Rev}^{\text{MM}}(S^{(i)}_{\text{CSM}}) \mid i \in [\xi]\}$ and $\mathcal{R}_{\text{RBM}} \equiv \{\text{Rev}^{\text{MM}}(S^{(i)}_{\text{RBM}}) \mid i \in [\xi]\}$, under a specific product category. The statistics include the minimum (third column), maximum (fourth column), average (fifth column), and standard deviation (last column) for each revenue set. For example, the second row shows that for the *Deodorant* category, the minimum revenue of the ξ assortments generated by the ranking-based model (i.e., the minimum of $\mathcal{R}_{\text{RBM}}$) is 3.058, the maximum is 3.126, the average is 3.085, and the standard deviation is 0.006.

Several key observations arise across these five product categories. First, the ranking-based model produces assortments with significantly higher variability in revenue compared to the consideration set model. For all categories, the minimum revenue of $\mathcal{R}_{\text{RBM}}$ is lower than that of $\mathcal{R}_{\text{CSM}}$, and the maximum of $\mathcal{R}_{\text{RBM}}$ is higher. This is further evidenced by the standard deviation of $\mathcal{R}_{\text{RBM}}$, which is consistently much larger than that of $\mathcal{R}_{\text{CSM}}$. This increased variability is expected, as the ranking-based model subsumes the consideration set model, making it more flexible. However, this flexibility can have mixed consequences. For instance, in the *Spaghetti/Italian Sauce* category, the ranking-based model achieves a similar minimum revenue as the consideration set model but attains a higher maximum, resulting in better overall performance. In contrast, for categories such as *Frozen Pizza*, *Margarine Butter*, and *Salty Snacks*, this flexibility proves detrimental. In these cases, the minimum revenue of $\mathcal{R}_{\text{RBM}}$ is substantially lower than that of $\mathcal{R}_{\text{CSM}}$, leading to a noticeably lower average revenue. For example, in the *Frozen Pizza* category, the minimum revenue of $\mathcal{R}_{\text{RBM}}$ is less than one-third of the minimum of $\mathcal{R}_{\text{CSM}}$, causing its average revenue to be 25% lower than that of $\mathcal{R}_{\text{CSM}}$. Overall, in four out of the five categories, the consideration set model outperforms the ranking-based model in average revenue, leading to more profitable assortments. On average, across all five categories, the mean of $\mathcal{R}_{\text{CSM}}$ is 9.1% higher than that of $\mathcal{R}_{\text{RBM}}$.

Finally, we note that except for the *Deodorant* category, the optimal assortments obtained by the consideration set model or the ranking-based model exhibit a noticeable gap when compared to the maximum revenue Rev^*. It is important to clarify that this gap should not be interpreted as the standard optimality gap in the mathematical optimization literature, as each assortment is solved *optimally* with respect to its corresponding choice model. Rather, this gap highlights the inherent challenges of data-driven assortment optimization in an end-to-end (i.e., data-to-decision) setting. Since the number m of historical assortments is typically much smaller than 2^n, the total

Category	Assortments	Min	Max	Avg.	Std.
Deodorant (Rev* = 3.137)	$S^{(i)}_{\text{CSM}}$	3.117	3.117	3.117	0.000
	$S^{(i)}_{\text{RBM}}$	3.058	3.126	3.085	0.006
Frozen Pizza (Rev* = 3.714)	$S^{(i)}_{\text{CSM}}$	3.555	3.555	3.555	0.000
	$S^{(i)}_{\text{RBM}}$	1.187	3.676	2.842	0.526
Margarine Butter (Rev* = 2.907)	$S^{(i)}_{\text{CSM}}$	2.308	2.375	2.354	0.031
	$S^{(i)}_{\text{RBM}}$	1.689	2.669	2.102	0.099
Salty Snacks (Rev* = 2.690)	$S^{(i)}_{\text{CSM}}$	2.109	2.162	2.114	0.014
	$S^{(i)}_{\text{RBM}}$	1.036	2.208	1.871	0.400
Spaghetti/ Italian Sauce (Rev* = 2.907)	$S^{(i)}_{\text{CSM}}$	2.139	2.242	2.152	0.035
	$S^{(i)}_{\text{RBM}}$	2.131	2.691	2.272	0.170

Table EC.4 Descriptive statistics (minimum, maximum, average, standard deviation) of the revenues for assortments $\{S^{(i)}_{\textsf{CSM}} \mid i \in [\xi]\}$ and $\{S^{(i)}_{\textsf{RBM}} \mid i \in [\xi]\}$, across five product categories in the IRI dataset, evaluated using the estimated mixed MNL model.

number of possible assortments, the available data are often underspecified, which exacerbates the under-determined nature of non-parametric choice models like the ranking-based model. For instance, in the *Salty Snacks* category, a ranking-based model that maximizes the likelihood of the data produces an assortment $S^{(i)}_{\text{RBM}}$ with revenue of 1.036 – only 40% of Rev*.

In summary, the consideration set model, as an identifiable and therefore more constrained class of choice models compared to the non-identifiable ranking-based model, exhibits lower variance in revenue performance. When adopting an estimate-then-optimize approach – estimating a choice model using maximum likelihood estimation and then solving the assortment optimization problem – the average performance of assortments generated by the consideration set model is 9% higher than those from the ranking-based model, as measured using the mixed MNL model ground truth. These findings highlight that in data-to-decision strategies, increased model flexibility may come at the cost of greater variation in decision quality, underscoring the trade-offs inherent in using more flexible and potentially non-identifiable models. In fact, to mitigate the risks of an overly flexible ranking-based model, one approach is to adopt a robust assortment optimization method (Sturt 2021), though this comes with significant computational challenges.

EC.6.4. Robustness Check

As a robustness check, we also evaluate the performance of assortments $S^{(i)}_{\text{CSM}}$ and $S^{(i)}_{\text{RBM}}$ ($i \in [\xi]$) using the decision forest model estimated in Section 5.3 as the ground truth choice model. Specifically, we let $\mathbb{P}^{\text{DF}}(j \mid S)$ be the predicted probability of choosing product j from the assortment S under the estimated decision forest model. Consequently, the revenue function is defined as $\text{Rev}^{\text{DF}}(S) = \sum_{j \in S} r_j \cdot \mathbb{P}^{\text{DF}}(j \mid S)$. Table EC.5 presents the descriptive statistics for $\mathcal{R}'_{\text{CSM}} \equiv$

Category	Assortments	Min	Max	Avg.	Std.
Deodorant (Rev* = 3.128)	$S_{\text{CSM}}^{(i)}$	2.939	2.939	2.939	0.000
	$S_{\text{RBM}}^{(i)}$	2.180	3.106	2.946	0.266
Frozen Pizza (Rev* = 3.649)	$S_{\text{CSM}}^{(i)}$	2.902	2.902	2.902	0.000
	$S_{\text{RBM}}^{(i)}$	0.294	2.930	1.621	0.569
Margarine Butter (Rev* = 2.173)	$S_{\text{CSM}}^{(i)}$	1.660	1.730	1.682	0.032
	$S_{\text{RBM}}^{(i)}$	0.194	2.139	1.751	0.468
Salty Snacks (Rev* = 2.230)	$S_{\text{CSM}}^{(i)}$	1.882	2.069	2.043	0.065
	$S_{\text{RBM}}^{(i)}$	0.115	2.094	1.455	0.773
Spaghetti/ Italian Sauce (Rev* = 2.180)	$S_{\text{CSM}}^{(i)}$	1.324	1.970	1.886	0.218
	$S_{\text{RBM}}^{(i)}$	0.098	2.023	1.244	0.663

Table EC.5 Descriptive statistics (minimum, maximum, average, standard deviation) of the revenues for assortments $\{S_{\text{CSM}}^{(i)} \mid i \in [\xi]\}$ and $\{S_{\text{RBM}}^{(i)} \mid i \in [\xi]\}$, across five product categories in the IRI dataset, evaluated using the estimated decision forest model.

$\{\text{Rev}^{\text{DF}}(S_{\text{CSM}}^{(i)}) \mid i \in [\xi]\}$ and $\mathcal{R}'_{\text{RBM}} \equiv \{\text{Rev}^{\text{DF}}(S_{\text{RBM}}^{(i)}) \mid i \in [\xi]\}$, which are the collections of revenues for assortments obtained under the consideration set model and the ranking-based model, respectively, for each product category. Similarly to the previous section, we further denote the expected revenue of the optimal assortment S_{DF}^* under the ground-truth decision forest model as $\text{Rev}^* \equiv \text{Rev}^{\text{DF}}(S_{\text{DF}}^*)$.

As expected, the results in Table EC.5, using the decision forest model as the ground truth, align closely with those in Table EC.4, which use the mixed MNL model as the ground truth. Consistent with our previous findings, the ranking-based model continues to generate assortments with significantly greater revenue variability compared to the consideration set model. In fact, across all categories, the minimum revenue of $\mathcal{R}'_{\text{RBM}}$ is lower than that of $\mathcal{R}'_{\text{CSM}}$, while its maximum is higher. Notably, the minimum of $\mathcal{R}'_{\text{RBM}}$ is substantially lower than the minimum of $\mathcal{R}_{\text{RBM}}$ from Table EC.4. As a result, when averaged across all five categories, the mean of $\mathcal{R}'_{\text{CSM}}$ is 33.4% higher than that of $\mathcal{R}'_{\text{RBM}}$, representing a significant increase from the 9.1% reported in Table EC.4.