Demand Estimation under Uncertain Consideration Sets

Srikanth Jagabathula
Leonard N. Stern School of Business, New York University, New York, NY 10012, sjagabat@stern.nyu.edu

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

Gustavo Vulcano
School of Business, Universidad Torcuato di Tella, and CONICET, Buenos Aires, Argentina, gvulcano@utdt.edu

To estimate customer demand, choice models rely both on what the individuals do and do not purchase. A customer may not purchase a product because it was not offered, but also because it was not considered. To account for this behavior, existing literature has proposed the so-called consider-then-choose (CTC) models, which posit that customers sample a consideration set and then choose the most preferred product from the intersection of the offer set and the consideration set. CTC models have been studied quite extensively in the marketing literature. More recently, they have gained popularity within the Operations Management literature to make assortment and pricing decisions. Despite their richness, CTC models are difficult to estimate in practice because firms do not observe customers’ consideration sets. Therefore, the common assumption in operations has been that customers consider everything on offer, so the offer set is the same as the consideration set. This raises the following question: when firms only collect transaction data, do CTC models offer any predictive advantage over the classic choice models? More precisely, under what conditions do CTC models outperform (if ever) classic choice models in terms of prediction accuracy?

In this work, we study a general class of CTC models. We propose techniques to estimate these models efficiently from sales transaction data. We then compare their performance against the classic approach. We find that CTC models outperform standard choice models when there is noise in the offer set information and the noise is asymmetric across the training and test offer sets, but otherwise offer no particular predictive advantage over the classic approach. We also demonstrate the benefits of using CTC models in real-world retail settings. In particular, we show that CTC models calibrated on retail transaction data are better at long-term and warehouse level sales forecasts. We also evaluate their performance in the context of an online platform setting: a peer-to-peer car sharing company. In this context, offer sets are even difficult to define. We observe a remarkable performance of CTC models over standard choice models.

Key words: choice-based demand, consideration sets, consider-then-choose models, peer-to-peer platforms, retail operations, revenue management.

1. Introduction

Over the last two decades there has been growing interest in the operations management (OM) academic field and in the industry practice to incorporate sophisticated demand models, which provide high quality inputs for critical tasks such as inventory management, dynamic pricing and assortment planning. Examples of such models include the multinomial logit (MNL), the nested logit, the mixed logit and the latent class MNL model, which are traditional in the marketing and economics fields but novel in terms of their applicability in operational contexts. These models have been widely studied resulting in the development of specific estimation (e.g., Newman et al. (2014), Vulcano et al. (2012), Jagabathula et al. (2020b)) or assortment optimization algorithms (e.g., Talluri and Van Ryzin (2004), Feldman and Topaloglu (2015), Davis et al. (2014)). More recently, new demand models have been proposed (e.g., the Markov chain model by Blanchet et al. (2016)), and others like the rank list-based model (e.g., Farias et al. (2013)) and the exponential model (e.g., Alptekinoğlu and Semple (2016)) have been revisited, jointly with the expansion of the application of choice-based demand models to the operation of online platforms (e.g., Lee and Lee (2012)). Companies in different industry sectors, from airlines to retailers and more recently, online sharing economy platforms, have been testing and incorporating some of these models into their operational capabilities.

The usual data source to calibrate these demand models are records of past transactions, either sales transaction data in the case of retail operations and revenue management, or bookings from past interactions between peers in the case of online platforms. For each transaction, the understanding is that the client (or an agent in the platform) selected one option among a set of alternatives or offer set, where this offer set is usually defined as a full category assortment in the retail case, or the full set of available options within a radius in spatial choice models (e.g., car sharing). Given the transaction data, most choice models are trained assuming that the chosen option is preferred over all the other products on offer.

But customers may not consider everything on offer. For instance, in retailing a customer selecting from the coffee category may not evaluate the full assortment and might consider only a subset of products (e.g., decaf) in a choice instance. In online platforms, an agent may evaluate only compact cars available within a 0.2-mile radius. If we ignore these consideration sets, we make the incorrect inference that the chosen product is preferred over products not even considered, leading to model bias. To deal with this issue, the so-called consider-then-choose (CTC) models have been proposed in the literature. These models posit that customers sample a consideration set and then choose the most preferred product from the intersection of the offer set and the consideration set (e.g., Howard and Sheth (1969), Alba and Chattopadhyay (1985), Hauser and Wernerfelt (1990)).
CTC models have been studied quite extensively in the marketing literature (e.g., see Roberts and Lattin (1997)). They also gained recent popularity within the OM literature to make assortment and pricing decisions (e.g., see Feldman et al. (2019); Aouad et al. (2020)).

Despite their richness, CTC models are often difficult to fit in practice because customers’ consideration sets are not observed—we know the choice and the offer set but the consideration set itself could be any subset of the full category containing the chosen product. When consideration sets are not observed, CTC models may not be identified and it is unclear what, if any, predictive advantage they offer over simply assuming that customers consider everything on offer as classic models do. Existing literature is mostly silent on these issues. It offers broad empirical evidence that customers do indeed form consideration sets and studies specific instances of the CTC models. But it mostly takes it as given that firms should fit CTC models over classic choice models even when consideration sets are not observed.

In this paper, we systematically study a very general class of CTC models when we only observe customer choices and offer sets. Theoretically, we show that CTC models in general are not identified, but specific instances are indeed identified from transaction data alone. We also show that the general CTC model class is equivalent to the random utility maximization (RUM) class of models (which assumes that customers consider everything on offer), indicating that CTC models do not offer any additional modeling flexibility over classic choice models. To empirically evaluate the CTC models, we develop techniques to estimate them from transaction data. Our empirical analysis on both synthetic and real-world data shows that CTC models outperform classic choice models when offer sets are not observed perfectly, or they are “noisy,” and the noise is asymmetric across the training and test data. Here, the offer set noise associated with a product refers to a product being erroneously recorded as offered or stocked out when in fact it was not. We say noise is asymmetric if the degree of offer set noise associated with a product differs between training and test datasets (e.g., because factors affecting stock outs vary over time). On the other hand when the noise is symmetric, CTC models have a comparable performance, offering little to no predictive advantage over fitting classic models.

Noise in offer sets is quite common in practice. In retailing, for instance, offer set descriptions are often unreliable because of potential inventory inaccuracies (e.g., DeHoratius and Raman (2008)). For example, Kang and Gershwin (2005) analyze the accuracy of inventory records of a global retailer and find that only 51% of them match actual inventory on average, with the worst store experiencing more than 67% mismatch. In online sharing platforms, offer or availability sets are not even well defined. For example, in the context of a peer-to-peer car sharing platform, an offer set can be defined as all the available car listings on the platform or more reasonably, as all the
car listings within a radius of the customer’s location. Either definition is arbitrary and prone to errors.

Furthermore, the so-called “test” offer set inputs used for predictions tend to be noisier (i.e., more different from the actual ones) than the training offer sets. Choice models must be given an offer set to make predictions but the firm faces a high degree of uncertainty about future offer sets. For example, a retailer is uncertain about next week’s offer set because existing products may stock out or new products might be introduced or replenished. Future offer sets are even more uncertain for online sharing platforms because availability is determined in real time not by the platform but by individual providers in the market, whose decisions are difficult to predict.

We thus argue that the conditions needed for CTC models to outperform classic choice models are often met in practice. To support our arguments, we present two real-world case studies, one in the context of retailing and the other one in the context of online sharing economy platforms, where the CTC models significantly outperform classic choice models. Our findings suggest that firms must use CTC models, particularly in the context of spatial choice in online sharing economy platforms (e.g., online peer-to-peer car sharing platform where cars are located at different spots).

1.1. Summary of results

Unlike existing literature which studies specific instances of CTC models, we study the following general class. The population preferences are characterized by a joint distribution over rank lists of the full set of products and over subsets of the full assortment (i.e., over the consideration sets). This joint distribution is common across all customers. In each choice instance, a customer samples a consideration set and a preference list, and purchases the most preferred product in the choice set, which results from the intersection between the sampled consideration set and the exhibited offer set.

We make the following remarks about our model. First, we note that we use a product-based as opposed to a feature-based consideration set definition. In a product-based consideration set definition, the model directly specifies the probability of consideration for each product. By contrast, a feature-based definition assumes that customers screen on features, considering only those products whose attributes are within pre-specified acceptable ranges (e.g., see Jagabathula and Rusmevichientong (2017)). Product-based consideration sets are more general and tractable to analyze. They can also readily subsume feature-dependence, as we show in this paper. Second, a key distinguishing aspect of our model is that we allow the distribution over consideration sets to be general, in contrast to the bulk of the existing literature which has generally restricted it to belong to specific classes. Third, while the joint distribution can be general in principle, for estimation, we approximate it with an appropriately defined mixture distribution (see Section 4).
The main goal of our proposal is to better understand if the CTC modeling framework has any operational value in practice. Specifically, we make the following contributions:

- **Identification conditions.** We first show that the CTC model class is equivalent to the RUM class, in that the set of choice probabilities induced by the CTC model class is the same as the those induced by the RUM class (see Proposition 2). This result shows that the CTC model class is no richer than the RUM class but is more flexible in terms of its practical usage as it can accommodate consideration set formation. We also show that CTC models are not identified from sales transaction data alone—not just the joint distribution, but also marginal distributions over consideration sets and preference lists (see Proposition 3). To ensure identification, we need to impose restrictions. We show that if we restrict the no-purchase option to be the least preferred product in all the preference lists, then the marginal distribution over consideration sets is identified (see Corollary 1). If we further restrict all the customers to choose using the same preference list, then the ranking is also identified (see Proposition 6). We also establish other results which investigate cases when consideration sets are “small” (see Proposition 5 and Corollary 1) and conditions to verify whether the observed data are consistent with a specific instance of the CTC model class.

- **Methodology to estimate the parameters of the CTC models.** For estimation, we approximate the general CTC model class with a mixture of what we call independent consideration set (ICS) models (see Section 4.1). An ICS model is specified by a single ranking and an independent distribution over consideration sets in which a customer generates a consideration set by including each product independently with a certain probability. We propose an expectation-maximization (EM) algorithm to estimate the mixture. For each mixture component, we propose an outer-approximation and cutting-plane method to ensure convergence to the maximum likelihood estimate.

- **Numerical experiments on synthetic data:** Because the consideration set formation is explicitly modeled in consider-then-choose type of frameworks, it is highly likely that its predictive performance is robust to the noise in the definition of the offer sets in comparison with competitive benchmarks (e.g., MNL and rank-based models). We verify this conjecture by explicitly adding noise, erroneously including or excluding products from the offer sets. We find that the CTC models outperform classic models when the noise is asymmetric between the training and test offer sets (e.g., the sets of items that are exposed to noise in training and test do not intersect). Their performance on the other hand is comparable when noise is symmetric. This result shows that when using choice models in practice, it is not sufficient to choose the model that provides the best predictive performance on the historical data set, but it is important to understand how the prediction task might differ from the training task in the data generation process (e.g., product availability).
Empirical analysis: better demand predictions for the retail industry and online platforms.

To support our findings on synthetic data, we compare choice models under several real world scenarios in retailing and online platforms where we are likely to face significant noise in the offer set definitions. On real-world grocery transaction data, we find that the relative performance of CTC models over the benchmarks improves as the level of noise in the test offer sets increases, for instance, when making long-term predictions. On a data set obtained from an online car-sharing platform, we show that CTC models outperform classic models because of the significant uncertainty in future availability sets. We also show that CTC models offer the flexibility to use non-linear models, such as decision trees and random forests, in modeling the consideration set distribution. These models are more interpretable and offer better prediction accuracy (up to 53.7% improvement in the root mean squared error metric).

The remainder of this paper is organized as follows. Section 2 positions our work within the existing literature. Section 3 defines our model and the theoretical results. We describe our data model and estimation algorithm in Section 4. The evaluation of the model starts in Section 5 with synthetic data experiments, followed by experiments on real data in Sections 6 (retailing) and 7 (online car sharing platform). Finally, our concluding remarks are discussed in Section 8.

2. Literature review

CTC models are built upon the key concept of consideration sets. This notion has recently been gaining attention in the OM literature, but is well studied in the marketing and psychology fields, dating back to the papers by Campbell (1969), Howard and Sheth (1969) and Wright and Barbour (1977).

It has long been recognized that consumers usually make choices in a two-stage process (Swait and Ben-Akiva (1987); Lynch (1991); Roberts and Lattin (1997)). First, they identify a small subset of products for further evaluation, the so-called consideration set, and then purchase the most preferred product from this subset. It is hard to observe whether a product which is not purchased has been included or not in a consumer’s consideration set, as it might even depend on a number of factors not necessarily related to the consumer’s preferences. Nevertheless, there is ample empirical evidence in the literature of the consider-then-choose behavior of customers. In his seminal paper, Hauser (1978) shows that a model based on the consideration set concept accounts for as much as 78% of the explainable uncertainty in purchase transaction data. Hauser and Wernerfelt (1990) empirically observe that customers consider on average only 3 brands of deodorants, 4 brands of shampoos, 4 brands of laundry detergents, and 4 brands of coffee. In a follow-up paper, Hauser (2014) reports that the average size of the consideration set of consumer packaged goods in US is a tenth of the total number of brands in the product category.
The notion of consideration sets might arise from the limited information gathering ability of consumers because they incur a search cost to learn detailed information about the products (Ratchford (1982)). The underlying justification is that consumers keep searching for products until the marginal expected gain from the search is less than the marginal search cost. Another argument to build consideration sets is related to cognitive heuristics, which is popular in the marketing and psychology literature while being of great importance for managerial decisions in advertising, product development, and strategic planning, e.g., conjunctive, disjunctive, compensatory, and elimination by aspects (Tversky (1972); Montgomery and Svenson (1976); Hauser (2014); Hogarth and Karelaia (2005)).

In the OM-related literature, the prevailing assumption aligned with the classical discrete choice literature has been that the consideration set is equivalent to the offer set. It has only been recently that more sophisticated consider-then-choose models of demand have been incorporated. Aouad et al. (2020) study the problem of assortment optimization under several variants of a choice model defined by two elements: a collection of consideration sets and a collection of customer types represented by rank lists. Different constraints in the definition of these collections lead to different special versions of the model (e.g., limiting the number of features that consumers use to filter a subset of alternatives). The authors develop a dynamic programming framework to study the computational aspects of assortment optimization under variants of these consider-then-choose premises. They show that for many empirically vetted assumptions on how customers consider and choose, their resulting dynamic program is efficient.

Feldman and Topaloglu (2018) consider the assortment optimization problem under the MNL model when consideration sets for different customer types are nested, whereas Feldman et al. (2019) focus on the assortment optimization problem when customers choose in accordance with the rank list model of demand but under small consideration sets. Wang and Sahin (2018) present a consider-then-choose model where the consideration set is formed by balancing the incremental expected utility of a product and the related search cost. The subsequent choice behavior within a consideration set is governed by the MNL model. Given the hardness of the assortment optimization problem, they propose as an approximation a solution that may exclude some high-attractiveness products from the offer set. Jagabathula and Rusmevichientong (2017) propose a model where first customers consider the set of products with prices less than the threshold and choose the most preferred product from the set considered. They develop a tractable nonparametric expectation maximization (EM) algorithm to fit the model to transaction data and design an efficient algorithm to determine the profit-maximizing combination of offer set and price. Jagabathula and Vulcano (2018) propose a framework to estimate individual consumer preferences under some heuristic rules used by consumers to form their consideration sets (e.g., they consider only products under...
promotion jointly with the ones purchased in the previous store visit). In Jagabathula et al. (2020a), the authors use the notion of consideration sets in order to construct personalized directed acyclic graphs (DAGs), which represent customers’ strong preferences, in a data-driven fashion. In spite of a recent popularity of CTC models in the OM field, most of the existing papers in OM mainly focus on various optimization problems under CTC choice rule ignoring the question of when CTC type of models can stand out and outperform classical benchmarks. In contrast, in this paper, we want to better understand the application area of CTC models and develop the methodology to calibrate these models from sales transactions data.

Our proposal here follows a different perspective and builds upon the modeling approach introduced by Manzini and Mariotti (2014). These authors study a choice model where the consideration set formation is stochastic and defined by the realization of the attention parameter of every alternative. This attention parameter is equivalent to our propensity parameter when the offer set is fully observable. After forming a consideration set, a consumer purchases the product that maximizes a preference relation within considered products. One of their main results states that this random choice rule is the only one for which the impact of removing an alternative on the choice probability of any other alternative, is asymmetric and menu-independent. The potential of the Manzini-Mariotti model in operational contexts was first evaluated by Gallego and Li (2017), who verified in a case study in the airline industry that its ability to fit booking data outperforms both the MNL and mixtures of MNLs in most of the markets evaluated. They also show that the related assortment optimization problem runs in polynomial time even with capacity constraints. In our paper, we study a more general class of CTC models and focus on understanding their operational value in practice.

3. Model specification and technical results

In this section we formally introduce consider-then-choose (CTC) models, followed by the presentation of related identification conditions.

3.1. Model description

We consider a universe $N$ of products $\{a_1, a_2, \ldots, a_n\}$, in addition to the ‘no-purchase’ or ‘outside’ option $a_0$. We denote $N^+ = N \cup \{a_0\}$.

Customers arrive to the store sequentially. In each choice instance, a customer is presented with a subset $S \subseteq N$ of products and chooses either one of the products in $S$ or the outside option $a_0$. We let $P_j(S)$ denote the probability that a customer chooses product $a_j \in S$ and $P_0(S)$ the probability that the customer chooses the outside option. Our goal is to model this choice process through a probabilistic model that specifies all the choice probabilities $\{P_j(S) : a_j \in S^+, S \subseteq N\}$, where we
use $S^+$ to denote the set $S \cup \{a_0\}$. We assume that the choice probabilities satisfy the standard probability laws: $\mathbb{P}_j(S) \geq 0$ for all $a_j \in S^+$ and $\sum_{a_j \in S^+} \mathbb{P}_j(S) = 1$ for all $S \subseteq N$.

To explicitly account for the fact that customers may not consider all the offered products before making a choice, we assume that their choice behavior follows a two-stage consider-then-choose (CTC) model. In the first stage, the customer forms a consideration set $C \subseteq N$ and in the second stage, selects either a product from the choice set $S \cap C$ or the outside option $a_0$. In this model, for a product to be purchased, it must be both offered and considered. The seller restricts customers’ choices by deciding which products to offer. But the customer further restricts her choices to just the ones in her consideration set. Two usual reasons to justify this approach are: (i) the customer has strong unobserved preferences (which prevent her from ever buying certain products), or (ii) the cognitive overload that prevents her from evaluating all the products on offer before choosing.

To describe customer preferences, let $\mathcal{S}_n$ denote the set of all full rankings or permutations of products in $N^+$ with cardinality $(n+1)!$, where we are also accounting for the no purchase option. The preference ordering or ranking of the products in $N^+$ is described by a bijective ranking function $\sigma: N^+ \to \{0, 1, \ldots, n\}$ specifying a preference rank $\sigma(a_j)$ for each product $a_j$. The preference ordering $\sigma$ induces an antireflexive, antisymmetric, and transitive preference relation $\succ_\sigma$, defined as $a \succ_\sigma b$ if and only if $\sigma(a) < \sigma(b)$. Customer preferences are described by a joint probability distribution $\nu: \mathcal{S}_n \times 2^N \to [0, 1]$, where $2^N$ denotes the collection of all subsets of the set $N$. In each choice instance, a customer samples a ranking $\sigma$ and consideration set $C \subseteq N$ with probability $\nu(\sigma, C)$ and chooses the product $\arg \min \{\sigma(a_j): a_j \in (S \cap C)^+\}$, where $(S \cap C)^+ = (S \cap C) \cup \{a_0\}$.

Figure 1 illustrates the choice process for a particular store visit given a joint distribution function over consideration sets and full rankings, and an offer set $S$. The choice probability $\mathbb{P}_j(S)$ under this model is given by

$$\mathbb{P}_j(S) = \sum_{\sigma \in \mathcal{S}_n} \sum_{C \subseteq N} \nu(\sigma, C) \cdot \mathbf{1}[a_j \in (S \cap C)^+] \cdot \mathbf{1}[\sigma(a_j) \succ_\sigma \sigma(a_k) \forall a_k \in (S \cap C)^+, a_k \neq a_j], \quad (1)$$

where $\mathbf{1}[A]$ is the standard indicator function taking the value 1 if condition $A$ is satisfied and the value 0 otherwise. We further assume that the empty condition $A = \emptyset$ is always satisfied. We say that choice data $(\mathbb{P}_j(S): a_j \in S^+, S \in \mathcal{S})$ for some collection of subsets $S \subseteq 2^N$ is consistent with an underlying CTC model if there exists a distribution $\nu(\cdot, \cdot)$ that satisfies Equation (1) for all $a_j \in S$ and $S \in \mathcal{S}$.

The CTC model as stated is not amenable to estimation, and therefore, we parameterize it as a mixture of what we call independent consideration set (ICS) models; as shown below, this parameterization is without loss of generality. A single-class ICS model is defined by a single preference ordering $\succ$ and a product form consideration set distribution. In particular, each product $a_j$ is...
Choice process. There are $n = 5$ items in the product universe plus the no purchase option 0. A customer samples a consideration set $C = \{1, 2, 3\}$ and a ranking $\sigma$ according to some pre-specified joint distribution over consideration sets and rankings. The offer set is $S = \{1, 2, 5\}$, leading to $(S \cap C)^+ = \{1, 2, 0\}$. The available product in $C$ that ranks highest in $\sigma$ is 1. Therefore, item 1 is the product that a customer would buy in this particular store visit under CTC model.

associated with an inclusion probability (or propensity) parameter $\theta_j \in [0, 1]$, which denotes the probability that a customer includes product $a_j$ in her consideration set. Customers make product inclusion decisions independently of each other, and therefore, the probability $\lambda(C)$ of sampling consideration set $C$ is equal to $\prod_{a_j \in C} \theta_j \prod_{a_j \notin C}(1 - \theta_j)$. The ICS model was studied in Manzini and Mariotti (2014), and it can be shown that the probability of choosing product $a_j$ from offer set $S$ is

$$P_j(S) = \theta_j \prod_{a_i \in S: a_i \succ a_j} (1 - \theta_i),$$

which corresponds to the probability of the event that product $a_j$ is considered but none of the products preferred over $a_j$ are.

The ICS model is a special case of the CTC model class in which customers have homogeneous preferences and product consideration decisions are made independently of other products. Finite mixtures of these models provide more modeling flexibility, and in fact, subsume the entire class of CTC models, as shown in the proposition below, the proof of which is given in Appendix A2.

**Proposition 1.** Every CTC model with the joint probability distribution $\nu(\cdot, \cdot)$ over consideration sets and full rankings can be represented as a finite mixture of $K$ ICS models with preference orders $\sigma_h$, attention parameters $\theta_h$, and mixture weights $\nu(\sigma_h, C_h)$ for $h \in \{1, ..., K\}$ and for some value of $K$.

Our next goal is to study the problem of identifying customer consideration sets from sales transaction data alone. Such identification is not possible in general because if all we know is that a customer chose product $a$ from offer set $S$, then the customer may have considered any subset of products containing product $a$, all the way from the singleton set $\{a\}$ to the entire offer set $S$; we would not have any basis to select one subset over the other. Turns out however that when we have sufficiently many observations and the model is sufficiently restricted, identification is indeed possible, as we show next.
3.2. Identification conditions for the CTC model class

In this section, we derive various conditions under which the CTC models are identified. But first, it is natural to wonder how the CTC class is related to the classic random-utility maximization (RUM) class of models (Block and Marschak, 1960). The RUM class is the most studied choice model class in the literature and includes popular models, such as the MNL, nested logit (NL), and mixture of MNLs (MMNL) models. At the core, it assumes that customers sample utility values for products from some underlying joint distribution and choose the product with the highest utility. It is equivalently described by a distribution over product preference lists (Block and Marschak (1960); Strauss (1979); Farias et al. (2013); van Ryzin and Vulcano (2014, 2017)), so customers choose the most preferred product according to the sampled ranking list. Formally, we establish the following result.

**Proposition 2.** The collection of choice probabilities \( \{ P_j(S) : S \subseteq N, a_j \in N \} \) is consistent with an underlying RUM model if and only if it is also consistent with an underlying CTC model.

The proof of the proposition is rather straightforward and is provided in Appendix A2 for the sake of completeness. The above result shows that the CTC model is comparable to the RUM class in terms of modeling power. Yet, it does offer the ability to capture consideration sets explicitly and to parameterize the consideration set formation in order to incorporate product and customer level features. It seems particularly appropriate for some specific business contexts, as we will illustrate in subsequent sections.

We next focus on model identification. To this end, our goal is to investigate if we can uniquely infer model parameters from the collection of choice probabilities \( \{ P_j(S) : S \subseteq N, a_j \in N \} \). Unsurprisingly, CTC models are not identified from sales transaction data in general.

**Proposition 3.** Neither the marginal distribution over the rankings nor the marginal distribution over the considerations sets of a CTC model can be identified from sales transactions data alone.

The proof of the proposition is provided in Appendix A2. Our result shows that neither the distribution over rankings nor the distribution over consideration sets is identified which is similar to the result that RUM models are not identified from sales transaction data for \( n \geq 4 \) (Sher et al., 2011).

To ensure identification, we need to impose restrictions. Our next result shows that if the no-purchase option is the least preferred product in all the rankings, then the marginal distribution over consideration sets is identified.

**Proposition 4.** Suppose that a collection of choice probabilities \( \{ P_0(S) : S \subseteq N \} \) are consistent with an underlying CTC model \( \nu(\cdot, \cdot) \) where the no-purchase option is the least preferred product in

Electronic copy available at: https://ssrn.com/abstract=3410019
all the rankings in the support. Then, the marginal distribution \( \lambda(\cdot) \) over consideration sets, defined as \( \lambda(C) = \sum_{\sigma \in \mathcal{A}} \nu(\sigma, C) \), is fully identified. Specifically, we have

\[
\lambda(C) = \sum_{X \subseteq C} (-1)^{|C| - |X|} \mathbb{P}_0(N \setminus X).
\]

The proof of the proposition follows immediately from a particular form of the inclusion-exclusion principle stated in Graham (1995). For any finite set \( Z \), if \( f: 2^Z \rightarrow \mathbb{R} \) and \( g: 2^Z \rightarrow \mathbb{R} \) are two real-valued set functions defined on the subsets of \( Z \) such that \( g(X) = \sum_{Y \subseteq X} f(Y) \) and \( X \subseteq Z \), then the inclusion-exclusion principle states that \( Y \subseteq Z \) and \( f(Y) = \sum_{X \subseteq Y} (-1)^{|Y| - |X|} g(X) \). In our setting, because the no-purchase option is the least preferred product in all the rankings, a customer chooses the no-purchase option only if the offer set and the sampled consideration set are disjoint; that is, \( \mathbb{P}_0(S) = \sum_{C \subseteq N \setminus S} \lambda(C) \).\(^1\) Our result then follows from replacing \( f(Y) \) with \( \lambda(C) \) and defining \( g(X) = \mathbb{P}_0(N \setminus X) = \sum_{C \subseteq X} \lambda(C) \). For completeness, we provide an alternative proof of this result from first principles in Appendix A2.

Empirical evidence in the marketing literature suggests that the size of the consideration sets for most customers in different categories is relatively small, e.g., Hoyer (1984) concludes that the median number of laundry detergents that a consumer considers before making a purchase is one. When the size of consideration sets is bounded above by \( k \), with \( k < n \), it follows from Proposition 4 that to recover \( \lambda \), we need choice probabilities under offer sets of size \( n - k \) or larger.

**Corollary 1.** Consider a CTC model in which customers sample consideration sets of size at most \( k \) for some \( 1 \leq k \leq n \); that is, \( \lambda(C) = 0 \) whenever \( |C| > k \). Furthermore, suppose that the no-purchase option is the least preferred product in all the preference lists in the support. Then, the distribution \( \lambda \) over consideration sets can be identified using choice probabilities under offer sets of size \( n - k \) or larger, i.e., from the collection \( \{\mathbb{P}_0(S) : |S| \geq n - k\} \).

When the consideration sets are small, Corollary 1 argues that it is sufficient to collect choice probabilities for the no-purchase alternative from large offer sets. In many applications, however, firms cannot offer very large offer sets to its customers because of space constraint either in a physical store or on the relevant locations (e.g., top slots) of a website. The next proposition shows that when consideration sets are small, firms can identify \( \lambda \) by offering only small offer sets:

**Proposition 5.** Consider a CTC model in which customers sample consideration sets of size at most \( k \) for some \( 1 \leq k \leq n \) and the no-purchase option is the least preferred product in all the

\(^1\) Note that we can obtain the system of \( 2^n \) linear equations to estimate probabilities to sample \( 2^n \) consideration sets by writing down the equation \( \mathbb{P}_0(S) = \sum_{C \subseteq N \setminus S} \lambda(C) \) for all possible offer sets \( S \).
rankings in the support. Let \( \{ P_0(S) : S \subseteq N, |S| \leq k \} \) be a collection of choice probabilities that are consistent with such a CTC model. Then, we have

\[
\lambda(C) = \sum_{X \subseteq N} \sum_{Y \supseteq X \cup C} (-1)^{|Y| - |X \Delta C|} \cdot I[|X \cup C| \leq k < |Y|] \cdot P_0(X),
\]

where \( X \Delta C \) denotes the symmetric difference \((X \setminus C) \cup (C \setminus X)\).

The proof of the proposition is involved. It requires establishing several combinatorial identities. We present it in Appendix A2. The proposition shows that when the consideration sets are of size at most \( k \), then the consideration set distribution can be recovered using choice probabilities of offer sets of size at most \( k \).

### 3.3. The General Consideration Set (GCS) Model

The results in the previous section focus on the recovery of the marginal distribution over consideration sets. To ensure complete identification, we need to restrict our CTC model further. We consider the class of models in which customers are homogeneous in their preference orderings, that is, the model is described by a single preference list, but are heterogeneous in their consideration sets. We call this model the general consideration set (GCS) model. Similar to the CTC class, we parameterize the GCS models as a mixture of ICS models sharing the same ranking. It follows from the proof of the Proposition 3 that such a restriction is necessary for identification.

More precisely, we assume that the GCS model is defined by a single preference ordering \( \succ \) and a distribution \( \lambda(\cdot) \) over consideration sets. A customer samples a consideration set according to \( \lambda \) and then chooses the most preferred product according to \( \succ \) from the set \((S \cap C)^+\). When there is a single preference ordering, the assumption that the no-purchase option is ranked last is without loss of generality because otherwise, the customer will never purchase a product \( a_j \) that is less preferred than \( a_0 \), which is always considered. In this case, product \( a_j \) could simply be eliminated from the product universe \( N \).

With the above restriction, we show that the choice rule is also identified. Specifically, we have the following result:

**Proposition 6.** Suppose that the collection of choice probabilities \( \{ P_j(S) : a_j \in S, |S| \leq 2 \} \) are consistent with an underlying GCS model. Then, for all \( 1 \leq i, j \leq n \) and \( i \neq j \), we have that \( \sigma(a_j) < \sigma(a_i) \) if \( P_i(\{a_i\}) > P_i(\{a_i, a_j\}) \).

The argument follows by contradiction. Suppose that there exists a set \( \{a_i, a_j\} \) for which \( P_i(\{a_i\}) > P_i(\{a_i, a_j\}) \), but for which \( \sigma(a_j) > \sigma(a_i) \). Then it must hold from the GCS model definition that \( P_i(\{a_i\}) \leq P_i(\{a_i, a_j\}) \), which is a contradiction. Unsurprisingly, the GCS model is not as rich as the CTC model class. In particular, it is a special case of the RUM choice rule.
Proposition 7. The GCS choice model is a special case of the RUM choice rule, that is, \( \text{GCS} \subseteq \text{RUM} \), but \( \text{GCS} \neq \text{RUM} \).

The proof of the proposition is provided in Appendix A2. It exhibits an example of a choice model that belongs to the RUM class but not to the GCS class.

In all the results above, we assumed that the collection of observed choice probabilities is consistent with an underlying CTC model. To verify that it is indeed the case, we establish here a set of necessary and sufficient conditions that the observed choice probabilities must satisfy to ensure consistency with an underlying GCS model.

Proposition 8. The collection of choice probabilities \( \{P_j(S): a_j \in S^+, S \subseteq N\} \) is consistent with a GCS model with unique parameters \( \sigma \) and consideration set distribution \( \lambda \) such that \( \lambda(C) > 0 \) for all \( |C| \leq 3 \) and \( \lambda(C) \geq 0 \) whenever \( |C| > 3 \) if and only if it satisfies the following conditions:

Condition 1. For all offer sets \( S \subseteq N \) and \( a_1, a_2 \in S \) such that \( a_1 \neq a_2 \) if \( P_1(S \setminus \{a_2\}) \neq P_1(S) \), then it must hold that \( P_2(S \setminus \{a_1\}) = P_2(S) \).

Condition 2. For all offer sets \( S, S' \subseteq N \) and \( a_1, a_2 \in S \cap S' \) such that \( a_1 \neq a_2 \) if \( P_1(S \setminus \{a_2\}) \) \( \geq \) \( P_1(S) \), then it must hold that \( P_1(S' \setminus \{a_2\}) \geq P_1(S') \).

Condition 3. For all offer sets \( S \subseteq N \), we have that \( \sum_{X \subseteq S} (-1)^{|S|-|X|} P_0(N \setminus X) \geq 0 \) with a strict inequality when \( |S| \leq 3 \).

Proposition 8 is similar to the set of conditions established in (Manzini and Mariotti, 2014, Theorem 1) for the case when the consideration set distribution \( \lambda \) has the product form due to the independence of the attention (or propensity) parameters. Our result extends their result to a general consideration set distribution \( \lambda \). Condition 1 is similar to the I-Asymmetry assumption in Manzini and Mariotti (2014), which states that either product \( a_2 \) influences the sales of product \( a_1 \) or vice versa, but not both (note that the influence may either be an increase or decrease). In other words, influence is one-directional and two products cannot influence the sales of each other. Condition 2 states that if product \( a_2 \) cannibalizes the sales of product \( a_1 \) in one offer set, then it must continue to do that in all the offer sets. That is, the direction of influence is consistent across all the offer sets. Condition 3 is a technical restriction to ensure the existence of a valid probability distribution function \( \lambda \) over the consideration sets. The strict inequality in Condition 3 is needed to ensure that the preference list over products in \( N \) satisfies the transitivity requirement. The proof of Proposition 8 is presented in Appendix A2. As it can be observed therein, establishing necessity is straightforward, but establishing sufficiency requires significant amount of work.
4. Data model and estimation methodology

In this section, we propose techniques to estimate a mixture of ICS models from observed sales transaction data. We first present an algorithm to fit a single-class ICS model to the data, which we then extend using the expectation-maximization (EM) framework to fit a finite mixture of ICS models.

We use the standard maximum likelihood estimation (MLE) technique and formulate the MLE problem as a mixed integer nonlinear program (MINLP). We then reduce the MINLP problem to solving a sequence of mixed integer linear programs (MILPs) and propose two algorithms (outer-approximation and cutting plane; see Appendix A3) in order to implement the solution process. We also show a way to incorporate product features into the consideration set model, using three widely used methods in machine learning—logistic-, decision tree-, and random forest-regressions.

Throughout the section we assume access to sales data consisting of purchase transactions over $T$ periods. Every purchasing instance is represented by a tuple $(a_j, S_t)$ for $t \in \{1, ..., T\}$, where $S_t$ denotes the subset of products offered in period $t$ and $a_j$ denotes the product purchased.

4.1. Independent consideration set (ICS) model

To formulate the likelihood function under this model, we define binary linear ordering variables $\delta_{kj}$, $\forall j, k$, $k \neq j$, where $\delta_{kj} = 1$ if product $a_k$ is preferred over product $a_j$ in the preference list $\succ$ (or, equivalently, $\sigma$), and $\delta_{kj} = 0$ otherwise. Note that $\delta$ is an alternative parameterization of the preference order $\succ$. Then, the log-likelihood function under the single class ICS model can be shown to be

$$L(\theta, \delta) = \sum_{t=1}^{T} \left[ \log \theta_{jt} + \sum_{a_k \in S_t: k \neq j} \delta_{kj} \log(1 - \theta_k) \right],$$

and the maximum likelihood estimation (MLE) problem can be formulated as follows:

$$\max_{\theta, \delta} L(\theta, \delta) \quad (4)$$

s.t.: $\delta_{jk} + \delta_{kj} = 1$, $\forall j, k$, $j < k$; 

$\delta_{jk} + \delta_{kp} + \delta_{pj} \leq 2$, $\forall j, k, p$, $j \neq k \neq p$; 

$\delta_{jk} \in \{0, 1\}$, $\forall j, k$; 

$0 \leq \theta_j \leq 1$, $\forall j$, 

where constraints (5) and (6) ensure that $\delta$ indeed represents a total order. In particular, the set of constraints (5) ensures that either $a_j$ is preferred over $a_k$ or vice versa, and the set of constraints (6) imposes the total ordering among any three products.
4.1.1. Estimation methodology. To be able to solve the above problem, we reformulate it as follows. First, we introduce a new variable $\tau$ defined as $\tau_{kj} = \delta_{kj} \theta_k$, $\forall j,k$ and rewrite the likelihood function in the following way

$$L(\theta, \tau) = \sum_{t=1}^{T} \left[ \log \theta_{jt} + \sum_{a_k \in S_t: k \neq j_t} \log(1 - \tau_{kj_t}) \right].$$

Note that with this change of variable the log-likelihood function becomes jointly concave in $\theta$ and $\tau$. We can then formulate the MLE problem in terms of the variables $(\delta, \tau, \theta)$:

$$\max_{\theta, \tau, \delta} L(\theta, \tau) \quad (9)$$

subject to:

$$\tau_{kj} \leq \theta_k, \quad \forall j,k, \quad (10)$$

$$\tau_{kj} \leq \delta_{kj}, \quad \forall j,k, \quad (11)$$

$$\tau_{kj} \geq \theta_k + \delta_{kj} - 1, \quad \forall j,k, \quad (12)$$

$$\tau_{kj} \geq 0, \quad \forall j,k, \quad (13)$$

$(\delta, \theta)$ satisfy $(5)-(8)$,

where linear constraints (10)-(13) ensure that $\tau_{kj} = \delta_{kj} \theta_k$, $\forall j,k$, given that $\delta$ is a binary variable, and constraints (5) and (6) ensure again a total order on $\delta$. We reformulate the MLE problem to have a linear objective function:

$$\max_{\theta, \tau, \delta, \mu} \mu \quad (14)$$

subject to $(\delta, \tau, \theta)$ satisfy $(5)-(8)$ and $(10)-(13)$,

$$\mu \leq L(\theta, \tau). \quad (15)$$

Note that if we know the ranking $\sigma$, then optimization problem (14) reduces to solving a globally concave maximization problem with a unique, closed form solution given by

$$\theta_j = \frac{\sum_{t=1}^{T} \mathbb{I}[a_{jt} = a_j]}{\sum_{t=1}^{T} \mathbb{I}[a_{jt} = a_j] + \sum_{t=1}^{T} \mathbb{I}[a_j \in S_t, a_j > a_{jt}]} \quad (16)$$

Next we show how to apply the outer-approximation method of Duran and Grossmann (1986) to solve the optimization problem (14)-(15). The proposed algorithm effectively exploits its structure, where we have linearity of the constraints involving the binary variables $\delta$, and convexity of the non-linear constraint (15) which only depends on continuous variables $\theta$ and $\tau$. In order to linearize the optimization problem, we use the outer-approximation of a convex set by the intersection of the collection of its supporting half-spaces. The broad idea of this algorithm is to approximate the convex constraint in the MINLP (i.e., constraint (15)) with a set of linear constraints. As a
result, solving the MINLP reduces to solving a sequence of MILPs, where at each iteration we add only one linear constraint to the MILP formulated in the previous iteration. Next, we provide the details of how to apply the outer-approximation algorithm to our MLE problem.

Let $C[x, y]$ denote a constraint which is a linear approximation of the constraint (15) at a point $(x, y)$, i.e.,

$$ C[x, y] := \left\{ \mu \leq \mathcal{L}(x, y) + \sum_{t=1}^{T} \frac{1}{x_{jt}} \cdot (\theta_{jt} - x_{jt}) + \sum_{t=1}^{T} \sum_{a_k \in S_t, k \neq j_t} 1 x_{jt} \cdot (\theta_{jt} - x_{jt}) + \sum_{t=1}^{T} \sum_{a_k \in S_t, k \neq j_t} 1 y_{kt} - 1 \cdot (\tau_{kt} - y_{kt}) \right\}. $$

Then we define the following optimization problem

$$ \max_{\theta, \tau, \theta, \mu} \mu \quad \text{(17)} $$

subject to $(\delta, \tau, \theta)$ satisfy (5) – (8) and (10) – (13),

$$ C[x, y] \forall (x, y) \in A \quad \text{(18)} $$

$$ \mu_L \leq \mu \leq \mu_U, $$

where we have replaced constraint (15) with the finite collection of linear constraints (18) at points in a set $A$ that is incrementally built. It follows from the convexity of constraint (15) that every point that satisfies constraint (15) also satisfies the collection of constraints (18) for every finite set $A$. We thus obtain an outer approximation. We also add bounds $\mu_L$ and $\mu_U$ to the log-likelihood function, which will be chosen in each iteration to tighten the interval containing the solution. This outer approximation defines the optimization subproblem as an MILP. Because of the potentially many continuous points required for outer-approximation, we solve a sequence of MILPs to build up increasingly tight relaxations of the original MINLP. The algorithm to calibrate the ICS model is provided below.

**ICS model calibration algorithm**

**Input** Given sales transaction data, do:

**Step 1** Sort products in decreasing number of sales and let $\sigma^{(0)}$ (i.e., $\delta^{(0)}$) denote the corresponding ranking. Compute $\theta^{(0)}$ using equation (16) given $\sigma^{(0)}$.

**Step 2** Obtain all possible rankings $\sigma^{(1)}, \sigma^{(2)}, ..., \sigma^{(m)}$, by swapping positions of any pair or two pairs of items in $\sigma^{(0)}$. Compute $\theta^{(i)}$ using equation (16) given $\sigma^{(i)}$ for all $i \in \{1, 2, ..., m\}$.

**Step 3** Set $\tau_{kj}^{(i)} := \delta_{k}^{(i)} \theta_{k}^{(i)}$, $\forall j, k$ and for all $i \in \{0, 1, 2, ..., m\}$. Recall that $\delta^{(i)}$ is an alternative parameterization of the ranking $\sigma^{(i)}$. Set $\mu_L := \max_{0 \leq i \leq m} \mathcal{L}(\theta^{(i)}, \tau^{(i)})$ and $\mu_U := \infty$. Set $i := m$.

**Step 4** While $|\mu_U - \mu_L| > \varepsilon$ and running time did not exceed the limit, do:

Set $i := i + 1$.

Solve optimization problem (17) with set $A = \{(\theta^{(k)}, \tau^{(k)})\}_{k=0}^{i-1}$ and obtain
solution \( \delta^{(i)}, \mu^{(i)} \). Note that in each iteration, we only add one constraint to the optimization problem solved in the previous iteration.

Update \( \theta^{(i)} \) using equation (16) given \( \delta^{(i)} \). Set \( \tau_{kj} := \delta_{kj}^{(i)} \theta_{k}^{(i)} \), \( \forall j, k \).

Set \( \mu_{L} := \max \{ \mathcal{L}(\theta^{(i)}, \tau^{(i)}), \mu_{L} \} \) and \( \mu_{U} = \mu^{(i)} \).

Endwhile

**Step 5** Find solution \( \theta^{*} = \theta^{(i^{*})} \) and \( \delta^{*} = \delta^{(i^{*})} \) where \( i^{*} := \arg\max_{0 \leq k \leq i} \mathcal{L}(\theta^{(k)}, \tau^{(k)}) \).

**Step 6** Stop.

Overall, the proposed algorithm consists of solving a finite sequence of MILPs. The size of each MILP scales in \( n \), quadratically in the number of variables and cubically in the number of constraints. It follows from existing results in Duran and Grossmann (1986) that this algorithm converges to the global optimum in the long run.

Empirically, we analyze the performance of the proposed algorithm to estimate the ICS model on the IRI Academic dataset (to be described later in Section 6). We limited its running time to 3 hours, and the precision was set to \( \varepsilon = 1e-6 \). It follows from Figure A4 in Appendix A3 that the optimality gap of the outer approximation algorithm (1) to calibrate the ICS model is 3.3% on average over 20 product categories, determined by the time limit, which indicates that this algorithm provides quite a reasonable performance in our setting.

**4.1.2. Feature-based modeling.** To capture the impact of product features on consideration set formation, we use the following three ways to make propensity parameter \( \theta_j \) a function of the product features, where \( x_{jk} \) is the observed \( k \)th feature of product \( a_j \):

- **Logistic-based ICS model (L-ICS).** We assume that customers have linear-in-parameters utility \( U_j \) from considering product \( a_j \in \mathcal{N} \), given by

\[
U_j = \beta_{j0} + \sum_k \beta_k x_{jk} + \varepsilon_j,
\]

where \( \varepsilon_j \) is a random variable distributed as a standard logistics, i.e., \( \varepsilon_j \sim \text{Logistic}(1) \). Therefore, product \( a_j \) is considered by an individual if and only if the utility from paying attention on it is non-negative, i.e.,

\[
a_j \in C \text{ iff } U_j = \beta_{j0} + \sum_k \beta_k x_{jk} + \varepsilon_j \geq 0.
\]

Then the propensity of product \( a_j \) is given by

\[
\theta_j = \Pr[a_j \in C] = \frac{\exp(\beta_{j0} + \sum_k \beta_k x_{jk})}{1 + \exp(\beta_{j0} + \sum_k \beta_k x_{jk})}.
\]
• **Decision tree-based ICS model (DT-ICS).** Here it is assumed that individuals decide which items to consider based on a tree with leaves \( m \in \{1, 2, ..., M\} \) to which we can associate a mean probability \( w_m \) of whether the item is going to be considered or not (see Murphy (2012)). Then we can write the probability to consider the item \( a_j \) in the following way:

\[
\theta_j = \Pr[a_j \in C] = \sum_{m=1}^{M} w_m I[x_j \in R_m] = \sum_{m=1}^{M} w_m \phi(x_j, v_m),
\]

where \( R_m \) is the \( m \)th region, i.e., the \( m \)th leaf; \( v_m \) encodes the choice of features to split on and their threshold values, on the path from the root to the \( m \)th leaf; and \( \phi(x_j, v_m) \) is equal to 1 if \( x_j \) belongs to the \( m \)th leaf, and equal to 0 otherwise.

• **Random forest-based ICS model (RF-ICS).** In this case, we assume that individuals first randomly sample a tree and then decide which items to consider based on the sampled tree (see Murphy (2012)). Note that random forest avoids the overfitting problem of decision trees by adding more trees instead of building one big tree. We can write the probability of considering the item \( a_j \) as follows:

\[
\theta_j = \Pr[a_j \in C] = \frac{1}{K} \sum_{k=1}^{K} f_k(x_j),
\]

where \( f_k(x_j) \) is the probability of considering item \( a_j \) according to the \( k \)th decision tree.

**Estimation methodology.** In a similar spirit to the preliminaries in Section 4.1, we can formulate the maximum likelihood estimation problem for the single class, logistic-based ICS (L-ICS) model with product features in such a way so that we can apply the outer-approximation algorithm in Appendix A3.1 in order to calibrate it.

On the other hand, the calibration of DT-ICS and RF-ICS models is more challenging. To this end we need to estimate both the ranking \( \sigma \) and decision tree (or random forest). Intuitively, both decision tree and random forest map product features into the binary outcome variable of whether the product is going to be considered or not, in non-linear way. Note that if the ranking \( \sigma \) is known, then the log-likelihood optimization is equivalent to calibrating a classification decision tree or random forest with the splitting criteria based on the entropy function. Then, given a decision tree or a random forest, the log-likelihood optimization problem reduces to solving a MILP to find \( \sigma \). We could then heuristically iterate between these two steps of finding \( \sigma \) and \( \theta \) until finding a fixed-point or a time limit is reached. As a practical matter in Section 7, we assume that the ranking \( \sigma \) is known a priori (see Section A5.2 for the details).

### 4.2. CTC model calibration: EM algorithm

To make estimation tractable, we represent the general distribution \( \nu(\cdot, \cdot) \) using a finite mixture of ICS models and then apply the EM algorithm. First, let us briefly discuss the main principles of
the EM procedure. We begin with arbitrary initial parameter estimates \( \hat{x}^{(0)} \). Then, we compute the conditional expected value of the log-likelihood function \( E[\log L(x) | \hat{x}^{(0)}] \) (the “E”, expectation, step). Next, the resulting expected log-likelihood function is maximized to compute new estimates \( \hat{x}^{(1)} \) (the “M”, maximization, step), and we repeat the algorithm until convergence or until a time limit is reached, to get a sequence of estimates \( \{ \hat{x}^{(q)}, q = 1, 2, \ldots \} \). We next briefly outline the E-step and M-step of every iteration and how we start the algorithm in the context of our estimation problem. Further specific details are relegated to the Appendix A3.4.

**Initialization:** we randomly allocate sales transaction to one of the \( K \) classes which allows us to compute initial weights \( \gamma^{(0)}_h \) of every class \( h \), and estimate initial parameters of our model: \( \sigma^{(0)}_h \) and \( \theta^{(q)}_h \) for all \( h \in \{1, \ldots, K\} \). To this end, we use the outer approximation algorithm for the ICS model.

**E-step:** we compute probability of every transaction at time \( t \) to belong to segment \( h \) based on the parameter estimates at the previous iteration and the purchasing transactions data.

**M-step:** first, we update the weights \( \gamma^{(q)}_h \) of every segment \( h \in \{1, 2, \ldots, K\} \) and then optimize the conditional expected value of the log-likelihood function, obtained in the previous step, by using the outer approximation algorithm for the ICS model. This way we obtain \( \theta^{(q)}_h \) and \( \sigma^{(q)}_h \) for all \( h \in \{1, \ldots, K\} \).

A key observation is that for sufficiently large \( K \) we can calibrate the CTC model by estimating the following three components: (1) segment probabilities \( \gamma_h \), (2) propensity parameters for each segment \( \theta_h \), and (3) the ranking \( \sigma_h \). Therefore, this parameterization of the CTC model is a natural approach when we have sparsity in customer segments. Even though it is well acknowledged that the convergence of EM algorithms is not guaranteed a priori, it was verified consistently in all our experiments.

In a similar way, we can also calibrate the GCS model with the EM algorithm (see Appendix A3.3 for details). Note that it is very straightforward to build upon Proposition 1 and show that GCS model can also be represented as a mixture of ICS models.

### 5. Study based on synthetic data: Robustness to noise in offer sets

In this section, we describe the results of an extensive simulation study, the main purpose of which is to characterize the performance of CTC models relative to the classical RUM models under various noise regimes. We find that CTC models are generally more robust to noise in the offer sets and outperform the classical RUM models when noise affects training and test offer sets differently.

To streamline the analysis of this simulation study, we assume two ground truth models of demand: the classical MNL model, and the more general rank list model. Given the similarity of
the insights obtained, here we report results for the MNL ground truth model and defer results based on the rank-based ground truth model to Section A4.1 in the Appendix.

In our simulations, customers have perfect information of the offer set and consider all the items on offer. Given the offer set \( S \), the customer chooses product \( a_j \) with probability \( v_j / \left(1 + \sum_{i \in S} v_i\right) \), where the parameter \( v_i > 0 \) is the “weight” or the attraction value corresponding to product \( a_i \). The modeler observes customer choices, but does not observe the offer sets perfectly. In the presence of such noise, we compare the predictions of an MNL model against an ICS model, both fitted to the choice observations, to understand the conditions under which one outperforms the other one.

In our setup, the benchmark MNL model does not suffer from model misspecification but does suffer from noise in the offer sets. The ICS model, on the other hand, suffers from both model misspecification and noise in the definition of the offer sets, but it is meant to be better prepared to circumvent the latter.

Our main finding is that the ICS model significantly outperforms the ground-truth MNL model when the noise is asymmetric between the training and test data sets. In other words, if product availability is hard to predict (because it might look different from the training data), then models based on the consider-then-choose framework outperform classical RUM models.

5.1. Synthetic data generation process

We assume that we have \( n = 15 \) items in the product universe. For each product \( a_j \), we sample its nominal utility \( u_j \) uniformly at random from the interval \([1, 2]\) and set its MNL weight \( v_j = \exp(u_j) \). We normalize the MNL weight of the outside option to 1. We parameterize the level of noise in the definition of the offer set using two parameters: the noise exposure parameter, \( \gamma \in [0, 1] \), and the noise intensity parameter, \( \eta \in [0, 1] \). The noise exposure parameter determines if a product is exposed to noise, and the noise intensity parameter specifies the conditional noise level, as described below. Then, for given values of \( \gamma \) and \( \eta \), and realized parameters \( v \) of the MNL model, the data-simulation procedure proceeds as follows:

1. We sample 100 offer sets, \( \{S_m\}_{m=1}^{100} \), uniformly at random, which correspond to the underlying ‘true’ offer sets.
2. For each offer set \( S_m \), we generate 10,000 sales transactions according to the MNL model with parameter values \( v \).
3. We generate a single exposure set \( S_X \) containing products exposed to noise by including each product \( a_j \in N \) with probability \( \gamma \).
4. We generate noisy offer set observations by modifying each ‘true’ offer set \( S_m \) by adding extra products from the exposure set \( S_X \) with probability \( \eta \); specifically, we obtain the noisy offer set \( \tilde{S}_m \) by adding to the set \( S_m \) each product \( a_j \in S_X \setminus S_m \) with probability \( \eta \).
We generated both training and test datasets for 200 different combinations of $\gamma$ and $\eta$: $\gamma \in \{0.05, 0.1, ..., 1\}$ and $\eta \in \{0.1, 0.2, ..., 1\}$. For each combination of noise parameters, we generated 100 realizations of MNL parameters. Next, for each MNL ground truth realization, using the procedure above, we generated both training and test synthetic instances of the same size. Each instance consists of 100 randomly generated offer sets, and for each offer set, 10,000 transactions, giving a total of 1,000,000 transaction records. All our algorithms were coded in Python (version 2.7.2) using Gurobi (version 7.0) as the optimization engine, and run on a 3.0Ghz processor with 16GB of RAM.

Our simulation set up is designed to capture not only different noise intensities but also different degrees of noise asymmetry between the training and test datasets. The noise exposure parameter $\gamma$ captures this asymmetry as follows. Letting $S^T_X$ and $S^X_T$ denote the training and test exposure sets, respectively, the cardinality of the symmetric difference $(S^T_X \setminus S^X_T) \cup (S^X_T \setminus S^T_X)$ captures the number of products that are exposed to noise in only one of the datasets. In expectation, this cardinality is equal to $2\gamma(1-\gamma)n$ because the probability that a product is exposed to noise in only one of the data sets is equal to $\gamma(1-\gamma) + (1-\gamma)\gamma = 2\gamma(1-\gamma)$. Therefore, the degree of asymmetry in the level of noise between the training and testing datasets is highest when $\gamma = 0.5$ and lowest when $\gamma = 0$ or $\gamma = 1$. Furthermore, the offer sets are perfectly observed in both the training and test datasets when $\gamma = 0$ or $\eta = 0$, and the noise is perfectly symmetric when $\gamma = 1$. In practice, noise may be asymmetric when making future demand forecasts because promotion strategies, customer preferences, or replenishment processes change.

5.2. Prediction scores

We evaluate our models on two standard metrics, the mean absolute percentage error (MAPE) and the root mean square error (RMSE), defined as follows (in percentage points):

$$\text{MAPE} = \frac{100}{|N|} \sum_{a_j \in N} \left| \frac{n_j - \hat{n}_j}{10 + n_j} \right| \quad \text{RMSE} = \frac{100}{\sum_{a_j \in N} n_j} \sqrt{\frac{1}{|N|} \sum_{a_j \in N} (n_j - \hat{n}_j)^2},$$

where $n_j$ denotes the observed sales for $a_j \in N$ in the test dataset and $\hat{n}_j$ denotes our prediction. We make predictions on noisy test offer sets, so $\hat{n}_j = 10000 \cdot \sum_{m=1}^{100} f(a_j, \tilde{S}_m)$, where $\tilde{S}_m$ is the noisy test offer set and $f(a_j, \tilde{S}_m)$ is the probability that product $a_j$ will be purchased from offer set $\tilde{S}_m$ under the fitted model. Note that we add 10 in the denominator of MAPE to deal with undefined instances, and we divide the RMSE score by the total number of observed sales in the test data set to make it a relative metric so that it is more interpretable.

Intuitively, both scores quantify the power of the model combinations to predict the market shares for each product, with lower scores indicating a better prediction accuracy.
5.3. Results and discussion

In Figure 2 we present a heatmap of the prediction scores under an MNL model fitted to the training data where each column corresponds to a particular noise intensity $\eta$ and each row corresponds to a particular noise exposure $\gamma$. We focus on the MAPE and RMSE prediction scores in the left and right panels, respectively. Recall that the MNL model is the ground truth model for this simulation study. As expected, the MNL model captures the ground-truth choice probabilities almost exactly when $\gamma = 0.05$ and $\eta = 0.1$, i.e., there is only a small amount of noise added to the sales transaction data. However, the MNL prediction scores worsen with higher noise intensity for a given level of noise exposure. Interestingly, it can also be seen that MNL prediction scores are not monotonic with respect to the noise exposure level, i.e., the scores first increase and then decrease with the noise exposure level, for a given noise intensity. The fitted MNL model performs the worst at noise exposure level $\gamma = 0.5$, which as noted above, corresponds to the highest degree of noise asymmetry between the training and test data sets, whereas the performance is relatively good even at high noise intensity levels when the degree of asymmetry is low ($\gamma$ is close to 0 or 1).

To provide a quantitative understanding of the variation of the model performance with respect to the two noise parameters $\gamma$ and $\eta$, we carry out the following linear regression:

$$Y_{ik} = \beta_0 + \beta_1 \cdot \eta_i + \beta_2 \cdot \eta_i^2 + \beta_3 \cdot \text{Asymm}_k + \beta_4 \cdot \text{Shared}_k + \varepsilon_{ik},$$

(20)

where the index $i$ stands for the noise intensity $\eta_i \in \{0.1, 0.2, ..., 1\}$, and the index $k$ stands for the noise exposure $\gamma_k \in \{0.05, 0.1, ..., 1\}$. The outcome variable $Y_{ik}$ is the MAPE prediction score of the corresponding cell. The covariate Asymm$_k$ is the probability that an item in the product universe is exposed to noise only in the test dataset or only in the training dataset, and the covariate Shared$_k$ is the probability that an item in the product universe is exposed to noise both in the test and training datasets. Note that Asymm$_k = \gamma_k (1 - \gamma_k) + (1 - \gamma_k) \gamma_k = 2\gamma_k (1 - \gamma_k)$ and Shared$_k = \gamma_k^2$, where $\gamma_k$ is the noise exposure. The first two terms in the regression capture dependence on the noise intensity level $\eta_i$ and the last two terms capture dependence on the noise exposure level $\gamma_k$.

We add a quadratic term in the noise intensity level to capture any potential non-linear response to the noise intensity level. We also separate out the dependence on $\gamma$ into degree of asymmetry and degree of overlap in noise.

The results for the regression (20) are reported in the last column in Table 1. It follows from there that the noise intensity $\eta$ deteriorates the predictive performance of the MNL model in a non-linear way, with the coefficient for the linear term being positive and the coefficient for the quadratic term being negative. The variables Asymm$_k$ and Shared$_k$ are positively correlated with the MAPE score, i.e., the prediction performance of the MNL model worsens as the number of
Figure 2  Heatmap of the prediction scores under MNL model where each column corresponds to a particular noise intensity $\eta$, and each row corresponds to a particular noise exposure $\gamma$. We focus on the MAPE and RMSE scores in left and right panels, respectively. The lower the score, the better.

items in the product universe that are exposed to the noise increases. Interestingly, the variable $\text{Asymm}$ has more than seven times higher magnitude than the variable $\text{Shared}$ which indicates that the benchmark (i.e, MNL model) struggles the most in making accurate predictions when the impact of the noise is asymmetric between the training and test sales transactions. Note that the independent variables in the regression model (20), included as Model (5) in Table 1, explain most of the variation in the MAPE score under the MNL model, i.e., $R_{adj}^2 = 0.93$.

<table>
<thead>
<tr>
<th></th>
<th>Model (1) Score</th>
<th>Model (2) Score</th>
<th>Model (3) Score</th>
<th>Model (4) Score</th>
<th>Model (5) Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Intensity}$</td>
<td>27.336***</td>
<td></td>
<td></td>
<td></td>
<td>42.079***</td>
</tr>
<tr>
<td></td>
<td>(17.997)</td>
<td></td>
<td></td>
<td></td>
<td>(14.246)</td>
</tr>
<tr>
<td>$\text{Intensity}^2$</td>
<td>22.929***</td>
<td></td>
<td></td>
<td>-13.403***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(15.755)</td>
<td></td>
<td></td>
<td>(-5.122)</td>
<td></td>
</tr>
<tr>
<td>$\text{Asymm}$</td>
<td></td>
<td>34.662***</td>
<td></td>
<td></td>
<td>39.320***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(8.677)</td>
<td></td>
<td></td>
<td>(28.338)</td>
</tr>
<tr>
<td>$\text{Shared}$</td>
<td>-2.590</td>
<td></td>
<td>-5.385***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-1.143)</td>
<td></td>
<td>(8.015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. Observations:</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>R-squared:</td>
<td>0.621</td>
<td>0.556</td>
<td>0.275</td>
<td>0.007</td>
<td>0.929</td>
</tr>
<tr>
<td>Adj. R-squared:</td>
<td>0.619</td>
<td>0.554</td>
<td>0.272</td>
<td>0.002</td>
<td>0.928</td>
</tr>
</tbody>
</table>

$t$ statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1  Regression models where the dependent variable is the MAPE score under the MNL model.
The regression results indicate that it is not the noise per se, but the asymmetry in noise that is hurting model performance. The reason is that when noise is symmetric, model estimates are biased but the bias is in the correct direction. For example, if a product is frequently stocked out but the model does not know about it, then the model attributes low sales to a low attraction value as opposed to a stockout, resulting in an MNL weight that is biased downward. If the product continues to be frequently stocked out in the test data set, then the model should continue to forecast low sales, which would happen with a downward biased MNL weight. On the other hand, if the stockout frequency is asymmetric and, say, reduces in the test data set (perhaps because of better replenishment), then the MNL model forecasts would be incorrectly biased downward.

In Figure 3, we compare the performances of the MNL and the ICS models under the different noise regimes. Each cell presents the improvement obtained by the ICS model over the MNL model, computed as the difference in the corresponding prediction scores, so that higher values are better for the ICS model. Unsurprisingly, the MNL model is significantly better than the ICS model when the level of noise is small (small values of both $\gamma$ and $\eta$) because the ICS model is misspecified. But, at higher values of noise intensity $\eta$ and values of $\gamma$ close to 0.5, we see that the ICS model is better than the MNL model. Table 2 presents the results from regressing the improvement scores according to regression equation (20). As above, we find that the benefits of the ICS model are most significant when the noise is asymmetric between the training and test offer sets. Although not shown here, the pattern of the heatmap for the ICS model looks similar to that of the MNL model in Figure 2. However, the ICS model is more robust to noise (both the intensity and the degree of asymmetry) and its performance does not deteriorate as much, because of which it outperforms the MNL model in asymmetric and high noise regimes.

The results in this section are robust to different asymmetric scenarios of noise as well (see Figure A10 in Appendix A4.1). We also show that qualitative results remain the same if we compare the ICS model with the more competitive LC-MNL benchmark, which subsumes the data generating MNL model (see details in Appendix A4.1).

In order to shed some light on the underlying mechanism that drives the superior performance of ICS over the MNL model, we present a stylized study in Appendix A4.2. The main insight we obtain is that the one dimensional cannibalization property of the ICS model makes it robust to offer set noise. This property states that the presence or absence of lower ranked products does not affect the demand for higher ranked products. As a result, offer set noise mainly impacts the generally lower-ranked products, limiting the overall error rates. See Appendix A4.2 for details.
Figure 3  Heatmap of the prediction scores improvements of the ICS model versus MNL, where each column corresponds to a particular noise intensity $\eta$ and each row corresponds to a particular noise exposure $\gamma$. We focus on the MAPE and RMSE scores in left and right panels, respectively. The higher the score, the better.

<table>
<thead>
<tr>
<th></th>
<th>Model (1)</th>
<th>Model (2)</th>
<th>Model (3)</th>
<th>Model (4)</th>
<th>Model (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Intensity}$</td>
<td>$4.162^{***}$</td>
<td>$3.025^{***}$</td>
<td>$7.311^{***}$</td>
<td>$-2.966^{***}$</td>
<td>$2.539^{***}$</td>
</tr>
<tr>
<td>$\text{Intensity}^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(-9.055)</td>
<td>(4.811)</td>
</tr>
<tr>
<td>$\text{Asymm}$</td>
<td></td>
<td></td>
<td>$7.311^{***}$</td>
<td>$11.530^{***}$</td>
<td>$2.539^{***}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(7.112)</td>
<td>(25.421)</td>
<td>(22.207)</td>
</tr>
<tr>
<td>$\text{Shared}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{const}$</td>
<td>$-2.966^{***}$</td>
<td>$-1.842^{***}$</td>
<td>$-3.108^{***}$</td>
<td>$-1.588^{***}$</td>
<td>$-11.042^{***}$</td>
</tr>
<tr>
<td></td>
<td>(-9.055)</td>
<td>(-7.449)</td>
<td>(-8.281)</td>
<td>(-6.332)</td>
<td>(-36.333)</td>
</tr>
<tr>
<td>No. Observations:</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>R-squared:</td>
<td>0.239</td>
<td>0.161</td>
<td>0.203</td>
<td>0.105</td>
<td>0.874</td>
</tr>
<tr>
<td>Adj. R-squared:</td>
<td>0.235</td>
<td>0.156</td>
<td>0.199</td>
<td>0.100</td>
<td>0.871</td>
</tr>
</tbody>
</table>

$t$ statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2  Regression models where the dependent variable is the MAPE score improvement of the ICS model over the MNL model.

6. Case study on retailing

Our analysis so far has been based on synthetic data, establishing that CTC models outperform classic choice models when offer sets are noisy and the noise is asymmetric between the training and test offer sets. We now present a real-world case study where such conditions are met.
For our study, we use the household purchase panel and store data from the IRI Academic Dataset (Bronnenberg et al. (2008)). This panel dataset keeps track of the household purchase histories for grocery and drug store chains, collected from the two largest Behavior Scan markets in the US over the years 2001-2011. Because we now have access to panel data, we implement a variant of the EM algorithm for the estimation of the parameters of the GCS model, as described in Appendix A3.3.2.

The purpose of this empirical study is threefold: (i) provide various real-world scenarios based on the IRI dataset where we are likely to face significant noise in the offer set definitions when making the long-term demand predictions, (ii) investigate the prediction performance of choice models under different noise regimes, e.g., quantify the improvement of the GCS model over the latent class MNL (LC-MNL) model under several real world scenarios with various noise intensities, and (iii) compare the GCS model studied in this paper with the more restricted ICS model.

Our main findings are as follows: (a) the improvements of GCS versus the benchmark LC-MNL are higher for scenarios in which the offer sets have a high level of noise, (b) the predictive performance of the GCS model is robust to the noise level in the offer sets, and (c) the GCS model significantly outperforms the ICS model in prediction accuracy, indicating that the independent consideration set model is indeed restrictive.

In this section, we present all the comparisons with respect to the GCS model. It is natural to wonder if the more general CTC model, which allows for heterogeneity in customer preference orders, offers additional gains in prediction performance. We also carried out a similar analysis with the CTC model; see Appendix A3.5 for details. From our results, we can not claim dominance of GCS or CTC; that is, the GCS model dominates the CTC model on the RMSE metric whereas the CTC model outperforms the GCS model on the MAPE metric. Based on this finding, we focus our analysis and discussions mainly on the more parsimonious GCS model.

6.1. Data preprocessing

The dataset consists of weekly sales transactions. We analyze a total of 20 categories, presented in Table A1. We focus on sales transaction data from calendar year 2007. For every store visit, we are given the following information: the Universal Product Code (UPC) and price of the purchased item, a binary indicator if the product is on price or display promotion, the purchased quantity, the customer ID, the store ID, and the week when the purchase was made. Since we are not given explicit information about the subset of items offered to each individual upon her store visit, we follow existing literature (e.g., Jagabathula and Vulcano (2018)) and construct this subset by aggregating all the transactions made in a particular store within a given category during a particular week. We also aggregate items with the same vendor code into a single product due to
data sparsity and divide the sales transaction data into two parts: the training set, which consists of the first 26 weeks of sales observations; and the test set, which consists of the last 26 weeks of sales observations. Note that we exploit the panel data structure in the predictive performance analyses below.

6.2. Benchmark models
We compare CTC models with the benchmark, LC-MNL choice model with $K$ latent classes. In this model, each customer belongs to one unobservable class, and customers from class $h \in \{1, 2, ..., K\}$ make purchases according to the MNL model associated with that class. The model is described by the parameters of the MNL model characterizing each class and by the prior probabilities of customers belonging to each of the classes. Once the model parameters are estimated, we make transaction-level predictions for each customer by averaging the predictions from $K$ single-class models, weighted by the posterior probability of class-membership. Similarly to the GCS model, we estimated the model for $K = 1, 2, ..., 5$, and report the best performance measure from these 5 variants, for each of the performance metrics introduced in Section 5.2. Because we have panel data, we make individual-level predictions and consequently, the number of predicted sales $\hat{n}_j$ for product $a_j$ is now calculated as $\hat{n}_j = \sum_{i \in U} \sum_{t=1}^{T_i} f_i(a_j, \tilde{S}_t)$, where $U$ is the set of all customers, $T_i$ is the number of transactions of customer $i$, and $f_i(a_j, \tilde{S}_t)$ is the predicted probability that customer $i$ purchases item $a_j$ from noisy offer set estimate $\tilde{S}_t$.

6.3. Results and discussion
Demand predictions or forecasts using choice models implicitly involve two steps: (a) forecasting the offer set and (b) predicting the demand for each product given the forecasted offer set. Most existing literature on choice models has only focused on the second step, implicitly assuming that the future offer set is accurately specified. In practice, one must also forecast future offer sets and these forecasts often contain errors. In our study, we follow these two steps explicit to study the impact of forecast errors in the first step on the overall accuracy of predictions.

6.3.1. Different data aggregations towards offer set forecasting. We consider three different prediction tasks that naturally arise in practical retail contexts and which differ in the level of difficulty of forecasting the offer sets.

1. Short-term forecasts. Often, store managers want to make short- or immediate-term forecasts, such as for the next week, to help with inventory and promotion planning. For these forecasts, the manager has a reasonably accurate estimate of the product assortment or offer set.
2. **Long-term forecasts.** To be successful in major strategic and investment decisions, a store manager must also make long-term demand forecasts, say, over the next quarter or the next year. For making these forecasts, the retailer often does not have a good sense of how the product assortment evolves over the forecasting horizon because of product replenishment and stock outs, which manifest as errors in offer set forecasts.

3. **Warehouse forecasts.** Another scenario is when the warehouse of the retail chain distributes products to the stores and makes centralized decisions on the inventory level in the warehouse. In this case, the warehouse is likely to make predictions at the centralized level without knowing the up-to-date information on product assortments in every store.

In each of these scenarios, the retailer has access to the same purchase observations to train models on. It is the first step, that of forecasting the offer sets, that these scenarios differ on. As we shall see, the CTC models and the classic choice models differ on their ability to deal with noise in offer set forecasts.

We simulate these scenarios using the purchase observations we have as follows. As mentioned above, we split the purchase transactions into training and test sets. In the training set, the purchased product and the corresponding offer set are known. Because of the way the offer set was inferred, it may contain errors, but we train all the models disregarding any potential errors. Note that this process reflects the standard way in which choice models are trained in the literature and in practice. In addition to a trained choice model, the retailer must build what we call an ‘OSForecaster’ or an offer set forecaster. We abstract away from details of how such a forecaster might be constructed and instead peek in our test dataset to simulate one. Specifically, let $S_{rt}$ denote the set of offered products at store $r$ and period $t$ in the test period. Further, let $T_{test}$ denote the set of test time periods.

For short-term forecasts, we provide the retailer with the number of purchasing customers at each store $r$ and period $t$, and the retailer must predict the sales for each of the products. In practice, this will entail making sales predictions for the next week for each of the stores. To make this prediction, the retailer must first forecast an offer set for each store and time period during the test horizon. To reflect the fact that the retailer makes few errors in forecasting the next week’s offer set, we assume that the short-term OSForecaster returns the ‘true’ offer set $S_{rt}$ when queried with a store $r$ and time period $t$. The predictions of the OSForecaster are more complex for the other two scenarios. In particular, to reflect the difficulty of making such offer set forecasts, we assume that the OSForecaster provides not a point forecast, but a distribution over possible offer sets. We describe below how we use the test data to construct these distributions.

For long-term forecasts, we provide the retailer with the total number of purchasing customers at each store over the 26 weeks ($\approx$ 6 months) comprising the entire test horizon, and the retailer must
predict the sales for each of the products. To make these predictions, the retailer must first forecast an offer set to use for each store. We assume that when queried with a store \( r \), the OSForecaster returns a uniform distribution over 20 assortments, each of which is constructed as follows: we construct a random collection \( \mathcal{A} \) of offer sets by including offer set \( S_{rt} \) for each \( t \in T_{test} \) with probability 0.5 and then obtain one offer set by taking the union of all sets in the collection \( \mathcal{A} \). This construction is designed to reflect the possibility that the retailer is able to forecast some of the stock out and replenishment events, but not all.

The procedure we use for warehouse forecasts is similar. The retailer must generate sales forecasts for each period \( t \), across all the stores in the corpus, for which the retailer must forecast an offer set for each period \( t \). We assume that when queried with a period \( t \), the OSForecaster returns a uniform distribution over 20 assortments, each of which is constructed as follows: we construct a random collection \( \mathcal{A} \) of offer sets by including offer set \( S_{rt} \) for each store \( r \) with probability 0.5 and then obtain one offer set by taking the union of all sets in the collection \( \mathcal{A} \). This construction is designed to reflect the possibility that the retailer has information on the offer sets at some of the stores but not all.

### 6.3.2. Performance assessment.

We evaluate all three scenarios in terms of the accuracy of predicting the total sales of each product across the entire test horizon, as described in Section 6.2. When the OSForecaster outputs a distribution, our predictions will be averaged over all the offer sets in the distribution. In Figure 4 we present scatter plots of the improvements of the GCS model versus the LC-MNL model across 20 product categories under the three forecast scenarios discussed above. In the left and right panels we measure the predictive performance of the models under the MAPE and RMSE metrics, respectively, as defined in (19) (see Appendix A4.3 for alternative definitions of these metrics). We observe that GCS outperforms LC-MNL for around half of product categories for short-term forecasts, and for almost all product categories under the second and third forecast types. Note that we have dots located to the right of pluses with crosses being in between, under both MAPE and RMSE scores and across most of the product categories. It reveals that the improvement of GCS over LC-MNL across product categories increases when we switch from short- to long-term and from long-term to warehouse forecasts.

Figure 5 exhibits MAPE (left panel) and RMSE (right panel) scores of the GCS and LC-MNL models, averaging across 20 product categories, for the three different scenarios. We observe that the performance of the LC-MNL model deteriorates once we shift from short to long-term and from long-term to warehouse forecasts. On the other hand, the predictive performance of the GCS model only moderately decreases once we switch to the noisy scenarios, i.e., the performances stays rather flat for all three scenarios. From the panels in Figure 5 we observe that the improvements
of GCS over LC-MNL are -5.6% (-0.027%), 5.3% (0.081%), and 33.5% (0.56%) under the first, second, and third scenarios, respectively, based on the MAPE (RMSE) score.

The above observations show that the GCS model does not offer much in terms of performance gains when the test offer sets can be forecast accurately. It is only when one makes significant errors in forecasting the test offer sets that we see a deterioration in the performance of the LC-MNL model, while the performance of the GCS model remains robust to noise.

Turning back to Figure 4, we notice that the improvement of GCS over LC-MNL varies across product categories for a given scenario. To better explain this variation, we regress the improvement of GCS over LC-MNL for each category against the noise intensity, which captures how much the
Figure 5  Average prediction scores over 20 product categories under GCS and LC-MNL choice models represented by dashed bars and solid bars respectively. We focus on MAPE and RMSE scores in left and right panels respectively. The lower the score the better.

Forecasted offer set differs from the true offer set. We only consider the scenarios of long-term and warehouse forecasts for this analysis because the noise intensity for short-term forecasts is zero by definition. We define noise intensity at the transaction-level and aggregate the metric across all the transactions. For each transaction $\tau$ in the test set, let $S_\tau$ denote the true offer set and $\tilde{S}_\tau$ denote the forecasted offer set. For the transaction occurring at store $r$ in period $t$, the forecasted offer set $\tilde{S}_\tau$ takes the value $\text{OSForecaster}(r)$ and $\text{OSForecaster}(t)$ for long-term and warehouse predictions, respectively. We then use the following natural definition for noise intensity:

$$\text{noise intensity} = \frac{1}{\# \text{ of test transactions}} \sum_\tau \mathbb{E} \left[ \frac{|\tilde{S}_\tau \setminus S_\tau|}{|\tilde{S}_\tau|} \right],$$

where the expectation is with respect to the distribution of the offer set predictions $\tilde{S}_\tau$. Intuitively, for long-term forecasts, the noise intensity captures how much the offer set at a store varies over time; if there are very few stock outs, then the offer set remains stable and the noise intensity is close to zero, but if there are many stock outs then the offer set varies a lot and the noise intensity is large. Similarly, for warehouse forecasts, the noise intensity captures how much the offer sets vary across stores. The left and right panels in Figure 6 illustrate the regression under the long-term and warehouse forecast scenarios, respectively. We see a clear positive correlation between the improvement of GCS over LC-MNL and noise intensity in both panels suggesting the improvement becomes more significant with higher noise intensity in the product category.

In sum, our results all indicate that the CTC models have an advantage over the classic choice models when we expect the forecasted offer sets to be noisy. We make a few remarks. First, it is
natural to wonder if our results are sensitive to the specific OSForecaster model we have used. To alleviate this concern, we repeat the above analysis with a ‘black-box’ noise model for the OSForecaster, where we generate the forecasted offer set $\tilde{S}_{rt}$ for each store $r$ and test time period $t$ by randomly adding products to the ‘true’ offer set $S_{rt}$, as done in the simulation study. The qualitative insights continue to hold; see Appendix A4.4. Our study also highlights the need to invest efforts into forecasting future offer sets more accurately; few, if any, such studies are available in the existing literature, indicating a natural direction for future work.

Second, we highlight the following peculiarity of our study: for long-term and warehouse forecast scenarios, the models are trained on noiseless offer sets but then tested on noisy offer sets. This is clearly in contrast to standard practice where efforts are made to train models on the same setups that they are tested on. Wouldn’t it then be better to ignore the the fact that we have noiseless offer sets and instead train our models also on ‘forecasted’ offer sets? In Appendix A4.5, we repeat our analysis by training all the models on offer sets obtained by applying the OSForecaster (as described above) to the training data. We find that the performance of the GCS model improves a bit but the LC-MNL model improves quite a lot, but the LC-MNL model is not able to completely close the gap. We interpret this result as follows. If the underlying noise process is known, then it makes sense to incorporate in the training process. But in practice, the noise process cannot

---

2 We thank an anonymous referee for suggesting this study.
be fully modeled because the factors affecting offer set noise (e.g., stock out and replenishment processes) vary over time, in which case the CTC models offer some protection against noise.

We conclude this section by comparing the predictive performance of the GCS model against a single-class ICS model; see Figure 7. We observe that the GCS model outperforms the ICS model on the MAPE metric by 18.5%, 18.4%, and 6.3% under the short-term, long-term, and warehouse forecast scenarios, respectively, on average, across 20 product categories. This finding suggests that a mixture of product-form consideration set distributions captures customer heterogeneity better.

7. Case study on the car sharing dataset

The issue of noise in future offer sets is particularly acute for online platforms because product availability is determined by the market in real time and is often hard to predict. In this section, we apply our consider-then-choose framework to a dataset from an industry partner, which runs an online peer-to-peer car sharing platform. Our main finding is that consider-then-choose frameworks significantly outperform classical RUM models for predicting demand in these business environments.

In the rest of the section, we first provide some background information on our industry partner. Then, we describe the data and present our modeling assumptions. We incorporate the product feature information into the choice models in order to gain insights about consideration set formation. Then, we calibrate different variations of consider-then-choose and benchmark models from the platform data and compare their predictive performance.

7.1. Industry partner and data analysis

Our industry partner is an online, peer-to-peer car sharing service that enables drivers to rent cars from private car owners, and owners to rent out their cars. The company offers its users a
smartphone application to match car owners with renters on-demand. Car owners can use the
application to list their vehicles by posting the picture of the vehicle and providing its detailed
characteristics. In addition, car owners set the availability of their cars, hourly or daily prices, and
potential conditions for sharing them. Every listed car has a device installed into it so that the
renters are able to locate and unlock cars through the same application. As a car renter, the user
of the platform can easily search for the cars nearby and book the available alternative by entering
the license number and credit card information.

For the empirical analysis in this section we use a historical dataset including a sample of the
rentals completed in a major US city over a period of two years. Each observation in the dataset is
a rental (i.e., a renter who booked the listed car from a particular location given the set of available
alternatives on a specific day/time). Our dataset includes 26.8K rentals from around five hundred
car providers. For each rental, we have access to several observable features, such as car owner ID,
hourly rental price, car access (i.e., open or closed), car location hours (i.e., 24 hours or restricted),
car location type (i.e., garage, street, surface lot, or valet), car brand (e.g, BMW, Tesla, MINI),
car type (i.e., economy, standard, full size, SUV, trucks, luxury), car age, and some other various
binary car features such as transmission, premium wheels, power seats, bluetooth/wireless, leather
interior, sunroof/moonroof, premium sound, power windows, GPS navigation system, roof rack,
tinted windows. In Appendix A5.1 we examine the extent to which various features specified above
(e.g., hourly rental price) impact the consideration set structure of renters. A detailed summary of
the data is provided in Table A5 in Appendix A5. We split the dataset into two parts: the first 80%,
in-sample, rental observations, and the remaining 20%, out of sample transactions.

7.2. Modeling assumptions
The dataset consists of the rental request observations such that for every transaction we know
which car was reserved and we can infer the set of available cars, listed in the online platform
at the time of the request, with their characteristics. The offer sets are approximately built by
aggregating all listed and available cars within 0.3 miles distance from the location of the car which
was in fact rented, defining tuples of the form \((a_{jt}, S_t)\), where \(a_{jt}\) is the chosen car and \(S_t\) is the
set of cars available at the reservation time \(t\).

In general, in order to calibrate feature-based consider-then-choose models (e.g., GCS) with our
dataset, we need to estimate two types of parameters: the ranking \(\sigma\) over all cars listed in the online
platform (514 cars in total), and the parameters associated with the consideration set formation
of renters. In order to simplify the estimation procedure, we assume that the ranking \(\sigma\) is known a
priori. Specifically, the cars are ranked according to their popularity among renters, defined as the
number of times the vehicle was rented over the training dataset. Modeling the second stage choice
process this way, we do not parameterize the ranking $\sigma$ which implies that the cars are assumed to have the same attributes over time, set at their average values. However, according to our dataset, this assumption is justified (see Appendix A5.2.1 for details).

7.3. Feature-based predictive accuracy results

Next, we conduct an out-of-sample prediction testing of the models to quantify the performance of consider-then-choose models versus the benchmark while taking into account the car attributes. Overall, we compare the predictive performance of the LC-MNL benchmark with three variants of the CTC framework: GCS, Decision Tree-based ICS (DT-ICS) and Random Forest-based ICS (RF-ICS), on the accuracy of two prediction measures: MAPE and RMSE (see Section 5.2), where lower scores stand for better prediction. Note that we estimate GCS model by a mixture of L-ICS models with a specific preference order.

In order to optimize strategic and marketing decisions, the online platform needs to make long-term (or medium-term) demand forecasts for the cars listed on the online application. In the real world settings, the company can not rely on the accurate data on car availabilities over time for the distant future, i.e., we can not test prediction power of choice models by using the offer sets from the test dataset described above. Instead, the company might divide the city into several geographical areas and make predictions based on the aggregate assortment of cars listed in each area over the test time horizon. For our case study we divide the city in 42 equal-spaced areas and estimate the assortments of cars by taking the superset of all the cars on offer over the entire horizon captured by the hold-out data for each area. Note that in this way we have 42 different offer sets (each corresponding to a particular area) while making predictions.

In Figure 8, we present the prediction performance results of the models based on MAPE (left panel) and RMSE (right panel) scores, averaged across all car brands. The MAPE score of consider-then-choose models exhibit an improvement of 16.7%, 23.4%, and 43.3% over LC-MNL for GCS, DT-ICS and RF-ICS, respectively. We also observe that consider-then-choose models obtain improvements of 6.2%, 10.9%, and 53.7% over LC-MNL for GCS, DT-ICS and RF-ICS, respectively, based on RMSE metrics.

Figure 9 exhibits MAPE scores computed for every brand separately under the RF-ICS and LC-MNL models, where the brands are ordered according to their popularity (i.e., percentage of the total number of reservations in the training dataset coming from every brand), e.g., Honda is the most popular brand while Mercury is the least popular brand in the dataset. We note that the RF-ICS model outperforms the benchmark LC-MNL model almost consistently across all the brands. The panels also illustrate that MAPE scores vary significantly across brands both for RF-ICS and LC-MNL models. To further analyze this variation, in Figure 10 we regressed the improvement
Figure 8  Prediction results under Consider-then-Choose (CTC) and LC-MNL models with features based on MAPE and RMSE scores, aggregated over car brands.

of RF-ICS over LC-MNL against the popularity of brands (left panel), and the improvement of RF-ICS over LC-MNL against the MAPE score of the LC-MNL model (right panel). We observe a clear positive correlation between MAPE score improvements and popularity of brands, which indicates that we can better predict the demand for more popular brands. We can also see a clear positive correlation between the improvements and MAPE score under LC-MNL model, allowing us to conclude that consider-then-choose type of models are especially relevant in prediction tasks, i.e., CTC models dominate LC-MNL, when LC-MNL model provides a relatively poor prediction performance. Being robust to the noise, consider-then-choose models (and in particular, RF-ICS) provide significantly better predictive performance under these circumstances. Note that these insights are consistent with our numerical study based on the synthetic dataset in Section 5.

The results above indicate that consider-then-choose models forecast customer choices considerably better than the traditional LC-MNL model under both RMSE and MAPE scores. First of all, accounting for the consideration set formation with the linear-in-parameters GCS model with logistically distributed error term, we can better predict the choices of customers. This improvement can be attributed to the effectiveness of consider-then-choose models to alleviate the noise impact on the offer set definition from sales transaction data. Moreover, we can further boost the predictive performance of the CTC models by modeling the consideration set formation in a nonlinear-in-parameters way, with decision trees or random forests. After calibrating DT-ICS and RF-ICS models we can get some insights of how customers form their consideration sets. In particular, Figure A17 in Appendix A5 illustrates an instance of the decision tree obtained after fitting the DT-ICS model.
8. Conclusion

In this paper, we analyze the importance of modeling customer choices by accounting for unobserved consideration sets. Even though consider-then-choose (CTC) models are gaining popularity in the OM field, it is not clear from the existing literature when CTC models can outperform the classical models in the prediction performance if companies collect only transaction data. We show how one can effectively estimate the general class of CTC models and also address the problem of identifying such models when relying only on sales transaction data. We check the performance of the proposed methodology under a synthetic data setting where we control for different levels of noise and asymmetries of noise between the training and the testing data. Next, we apply CTC models on two real world contexts: a retail operation and a car-sharing platform.

Our empirical results suggest that the predictive performance of CTC models significantly outperforms state-of-the-art RUM-based benchmarks widely used in the marketing, economics and more recently, OM literature, when there is noise in the data describing the offer sets. Moreover, we show that the relative improvement of consider-then-choose models in predictive performance...
becomes even more significant when there exist asymmetries between the accuracy of the description of the offer sets used to train the model, and the accuracy of the description of the offer sets used in the hold-out sample to derive forecasts. These results make our methodology promising for researchers interested in choice-based demand estimation, particularly for cases where the offer sets are not fully observable or whose definition is hard to anticipate looking forward.

References


Electronic copy available at: https://ssrn.com/abstract=3410019


Electronic copy available at: https://ssrn.com/abstract=3410019


Wright P, Barbour F (1977) *Phased decision strategies: Sequels to an initial screening* (Graduate School of Business, Stanford University).
A1. Preliminaries on Consider-then-Choose models

For completeness, we summarize the relevant notation from the main body and also introduce additional notation. We consider a universe $N$ of $n$ products $\{a_1, a_2, \ldots, a_n\}$. We let $a_0$ denote the ‘no-purchase’ or the ‘outside’ option. A customer is presented with a subset $S \subseteq N$ of products and the customer chooses either one of the products in $S$ or the outside option $a_0$. We let $P_j(S)$ denote the probability that a customer chooses product $a_j \in S$ and $P_0(S)$ the probability that the customer chooses the outside option. We use $S^+$ to denote the set $S \cup \{a_0\}$. Let $\lambda: 2^N \to [0,1]$ define a distribution over consideration sets such that $\sum_{C \subseteq N} \lambda(C) = 1$. The preference relation $\succ$ specifies a rank ordering $\sigma$ over $n+1$ items which consist of the products in $N$ plus ‘no-purchase’ option $a_0$ with $\sigma(a_i)$ denoting the preference rank of product $a_i$. The lower the rank of the product, the higher the preference, so that a customer’s ranking $\sigma$ indicates that product $a$ is preferred to product $b$ if and only if $\sigma(a) < \sigma(b)$, or equivalently $a \succ \sigma b$. We assume that there is a distribution $\mu: \mathcal{S}_n \to [0,1]$ over $\mathcal{S}_n$, which is the set of all full rankings or permutations of products in $N^+$ with cardinality $(n+1)!$.

To simplify the exposition, we also let $\bar{X} := N \setminus X$, $X^+ := X \cup \{a_0\}$, and $P_i(X) = \Pr(a_i|X^+)$. 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 $A, B$.

A2. Proofs of Technical Results

Proof of Proposition 1: We prove this result by construction. Consider any CTC model with the underlying joint distribution $\nu(\cdot, \cdot)$ over consideration sets and preference orders. Let $K$ denote the number of tuples $\{\sigma, C\}$ where $\nu(\sigma, C) > 0$ such that $\nu(\sigma_h, C_h) > 0$ for $h \in \{1, \ldots, K\}$. Then, we can define ICS model with parameters $\sigma_h$ and $\theta_h$ for every $h \in \{1, \ldots, K\}$ such that $\theta_{hk} = 1$ if item $a_k$ belongs to the consideration set $C_h$ and 0, otherwise. Finally, it is straightforward to verify that the mixture of these $K$ ICS models, where probability of each segment $h$ is equal to $\nu(\sigma_h, C_h)$, results into the same choice frequencies as CTC model. $\square$

Proof of Proposition 2: First, we argue that if choice data is consistent with $RUM$ then it is also consistent with CTC since if we restrict CTC model in the way that a consideration set of each customer is equal to an offer set (i.e, $C = S$) then CTC model will induce the same choice frequencies as $RUM$. Then, it remains to prove that if choice data is consistent with CTC then
it is also consistent with RUM. Therefore, in order to complete the proof it is sufficient to argue that for every joint distribution function over rankings and consideration sets $\nu(\cdot, \cdot)$ there is a distribution function over rankings $\mu(\cdot)$ which induce the same choice data.

In what follows below we consider the RUM model defined by the distribution $\mu'(\cdot)$ over the rankings where all the preference lists satisfy the following property – all the items in a preference list which are less preferred than no purchase option are ranked according to their index, i.e., for all $a_i$ and $a_j$ such that $a_0 > a_i$ and $a_0 > a_j$ we have that $a_i > a_j$ iff $i < j$. Then, it is straightforward to verify that choice frequencies generated by CTC choice model with distribution $\nu(\cdot, \cdot)$ can be also induced by RUM model with distribution $\mu'(\cdot)$ such that probability to sample a preference order $\sigma'$ is equal to $\sum_{\sigma \in \mathcal{S}} \sum_{C \subseteq N} \nu(\sigma, C) \mathbb{I}[\sigma_C = \sigma']$, where $\sigma_C$ is the preference order obtained from $\sigma$ when we reposition the items that are not in $C$ by putting them at the bottom of a preference list and rank them according to their index, i.e., for all $a_i, a_j \in C$ we have that $a_i \succ_{\sigma_C} a_j$ iff $a_i \succ_{\sigma} a_j$ and for all $a_i, a_j \notin C$ we have that $a_i \succ_{\sigma_C} a_j$ iff $i < j$. □

**Proof of Proposition 3:** To complete the proof of this proposition it is sufficient to show the example of the sales transaction data when two different marginal distributions over rankings and considerations sets result into the same choice frequencies. Let us first assume that we have only one item in the product universe plus the no-purchase option, i.e., $N = \{a_0, a_1\}$ and it follows from the sales transaction data that $P_1(\{a_1\}) = 1/2$. Then, we consider the following two distribution functions over the rankings ($\mu$ and $\mu'$) and consideration sets ($\lambda$ and $\lambda'$) such that $\mu$ (resp. $\mu'$) and $\lambda$ (resp. $\lambda'$) are independent: (i) distribution function $\mu$ over rankings is specified by $\mu(\{a_1, a_0\}) = 2/3$ and $\mu(\{a_0, a_1\}) = 1/3$ and distribution function $\lambda$ over subsets is specified by $\lambda(\{a_1, a_0\}) = 3/4$ and $\lambda(\{a_0\}) = 1/4$; and (ii) distribution function $\mu$ over rankings is specified by $\mu(\{a_1, a_0\}) = 3/4$ and $\mu(\{a_0, a_1\}) = 1/4$ and distribution function $\lambda$ over subsets is specified by $\lambda(\{a_1, a_0\}) = 2/3$ and $\lambda(\{a_0\}) = 1/3$. It is straightforward to verify that both combinations of $\mu$, $\lambda$ and $\mu'$, $\lambda'$ can explain the sales transactions data specified above.

□

**Lemma A1.** For any sets $Z \subseteq N$ and $Y \subseteq Z$, and the function $f : 2^N \to \mathbb{R}$, we have

$$\sum_{P \subseteq Y} \sum_{X \subseteq P} (-1)^{|P| - |X|} \cdot f(Z \setminus X) = f(Z \setminus Y). \quad (A1)$$

**Proof:** First consider the inclusion-exclusion principle stated by Graham (1995) in the following form. Let $N$ be a finite set and $g : 2^N \to \mathbb{R}$ be a real-valued function defined on the subsets of $N$. Define the function $h : 2^N \to \mathbb{R}$ by $h(X) := \sum_{Y \subseteq X} g(Y)$, then $g(X) := \sum_{Y \subseteq X} (-1)^{|X| - |Y|} h(Y)$. 

Electronic copy available at: https://ssrn.com/abstract=3410019
Then we show that the lemma follows from the stated above inclusion-exclusion principle. Let $g(X) := f(Z \setminus X)$, and $h(P) := (-1)^{|P|} \sum_{X \subseteq P} (-1)^{|X|} \cdot g(X)$, which implies that

$$h(P) \cdot (-1)^{|P|} = \sum_{X \subseteq P} (-1)^{|X|} \cdot g(X),$$

by invoking the inclusion-exclusion principle we obtain that

$$(-1)^{|Y|} \cdot g(Y) = \sum_{P \subseteq Y} h(P) = \sum_{P \subseteq Y} (-1)^{|P|} \sum_{X \subseteq P} (-1)^{|X|} \cdot g(X) = \sum_{P \subseteq Y} \sum_{X \subseteq P} (-1)^{|P| - |X|} \cdot g(X)$$

$$= \sum_{P \subseteq Y} \sum_{X \subseteq P} (-1)^{|P| - |X|} \cdot f(Z \setminus X).$$

\[ \square \]

**Lemma A2.** The combinatorial identity below is valid

$$- \sum_{\beta = 0}^{\min(r,u)} C_w^\beta \cdot \left( \sum_{\alpha = r+1-\beta}^{u-\beta} (-1)^\alpha \cdot C_{\alpha}^{u-\beta} \right) = \begin{cases} 1, & \text{if } w = u, \\ 0, & \text{if } w < u, \end{cases} \tag{A2}$$

where $r < w$ when $w = u$.

**Proof:** Let us consider two cases:

**Case 1:** $w = u$. In this case $r < w$ by invoking the assumptions of the lemma.

$$\begin{align*}
- \sum_{\beta = 0}^{r} C_w^\beta \cdot \left( \sum_{\alpha = r+1-\beta}^{u-\beta} (-1)^\alpha \cdot C_{\alpha}^{u-\beta} \right) &= - \sum_{\beta = 0}^{r} C_w^\beta \cdot \left( \sum_{\alpha = r+1-\beta}^{u-\beta} (-1)^\alpha \cdot C_{\alpha}^{u-\beta} \right) \\
&= - \sum_{\beta = 0}^{r} \sum_{\alpha = r+1-\beta}^{u-\beta} (-1)^\alpha \cdot \frac{u!}{\alpha! \cdot (u - \alpha - \beta)!} = 1,
\end{align*}$$

where the last equality is proved by induction on $s = u - r$:

**Base case:** $s = 1$.

$$\begin{align*}
- \sum_{\beta = 0}^{r} \sum_{\alpha = r+1-\beta}^{u-\beta} (-1)^\alpha \cdot \frac{u!}{\alpha! \cdot (u - \alpha - \beta)!} &= - \sum_{\beta = 0}^{r} (-1)^{u-\beta} \cdot \frac{u!}{\beta!} = (-1)^{u+1} \cdot \sum_{\beta = 0}^{\beta} (-1)^{\beta} \cdot \frac{u!}{\beta!} = (-1)^{u+1} \cdot (1 - 1)^u = 1.
\end{align*}$$

**Induction hypothesis:** $s = p$.

**Induction step:** $s = p + 1$.

$$\begin{align*}
- \sum_{\beta = 0}^{u-p-1} \sum_{\alpha = u-p-\beta}^{u-\beta} (-1)^\alpha \cdot \frac{u!}{\alpha! \cdot (u - \alpha - \beta)!} \quad \text{[since } r = u - p - 1]\ \\
&= - \sum_{\beta = 0}^{u-p} \sum_{\alpha = u-p-\beta}^{u-\beta} (-1)^\alpha \cdot \frac{u!}{\alpha! \cdot (u - \alpha - \beta)!} + \sum_{\alpha = 0}^{p} (-1)^\alpha \cdot \frac{u!}{\alpha! \cdot (u-p)! \cdot (p-\alpha)!}.
\end{align*}$$
the last equality is proved by induction on \( s \).

Assuming that the result holds for \( p \), the induction step is as follows:

**Induction step:**

\[
= - \sum_{\beta=0}^{u-p} \sum_{\alpha=u-p-\beta}^{u-\beta} (-1)^{\alpha} \cdot \frac{u!}{\alpha! \beta! (u - \alpha - \beta)!} + \frac{u!}{p! (u-p)!} \cdot \sum_{\alpha=0}^{p} (-1)^{\alpha} \cdot \frac{p!}{\alpha! (p-\alpha)!}
\]

\[
= - \sum_{\beta=0}^{u-p} \sum_{\alpha=u-p-\beta}^{u-\beta} (-1)^{\alpha} \cdot \frac{u!}{\alpha! \beta! (u - \alpha - \beta)!} - \sum_{\beta=0}^{u-p} (-1)^{u-p-\beta} \cdot \frac{u!}{(u-p-\beta)! \beta! p!}
\]

\[
= 1 - \sum_{\beta=0}^{u-p} (-1)^{u-p-\beta} \cdot \frac{u!}{(u-p-\beta)! \beta! p!}, \quad \text{[by induction hypothesis, } r = u-p]\]

\[
= 1 + (-1)^{u-p-1} \cdot \frac{u!}{(u-p)!} \cdot \sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot \frac{(u-p)!}{(u-p-\beta)! \beta!}
\]

\[
= 1.
\]

**Case 2:** \( w < u \), the last equality is proved by induction on \( s = u-r \):

**Base case:** \( s = 1 \). Then \( r = u - 1 \geq w \), so that \( \min(r, w) = w \). And we have that

\[
- \sum_{\beta=0}^{w} C_w^{\beta} \cdot \left[ \sum_{\alpha=r+1}^{u-\beta} (-1)^{\alpha} C_{u-\beta}^{\alpha} \right] = - \sum_{\beta=0}^{w} C_w^{\beta} \cdot \left[ \sum_{\alpha=u-p-\beta}^{u-\beta} (-1)^{\alpha} C_{u-\beta}^{\alpha} \right]
\]

\[
= (-1)^{1+u} \sum_{\beta=0}^{w} (-1)^{\beta} C_w^{\beta} = 0.
\]

**Induction hypothesis:** \( s = p \).

**Induction step:** \( s = p + 1 \).

**Condition 1:** \( u-p > w \). Then \( \min(u-p, w) = w \) and \( \min(u-p-1, w) = w \). We have that

\[
- \sum_{\beta=0}^{w} C_w^{\beta} \cdot \left[ \sum_{\alpha=u-p+1}^{u-\beta} (-1)^{\alpha} C_{u-\beta}^{\alpha} \right] = - \sum_{\beta=0}^{w} C_w^{\beta} \cdot \left[ \sum_{\alpha=u-p-\beta}^{u-\beta} (-1)^{\alpha} C_{u-\beta}^{\alpha} \right] \quad [\text{since } r = u-p-1]
\]

\[
= - \sum_{\beta=0}^{w} C_w^{\beta} \cdot \left[ \sum_{\alpha=u-p+1}^{u-\beta} (-1)^{\alpha} C_{u-\beta}^{\alpha} \right] - \sum_{\beta=0}^{w} C_w^{\beta} \cdot (-1)^{u-p-\beta} \cdot C_{u-\beta-p}^{u-\beta-p} \quad [\text{by induction hypothesis, } r = u-p]
\]

\[
= (-1)^{1+u-p} \cdot \sum_{\beta=0}^{w} (-1)^{\beta} \cdot C_w^{\beta} \cdot C_{u-\beta-p}^{u-\beta-p} = (-1)^{1+u-p} \cdot \frac{w!}{p!} \cdot \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)!}{\beta! (w-\beta)! (u-p-\beta)!}.
\]

Now it is sufficient to show that \( \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)!}{\beta! (w-\beta)! (u-p-\beta)!} = 0 \). We prove it by induction on \( p \). For \( p = 0 \), it follows that \( \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)!}{\beta! (w-\beta)! (u-p-\beta)!} = \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{1}{\beta! (w-\beta)!} = \frac{1}{w!} \sum_{\beta=0}^{w} (-1)^{\beta} \cdot C_w^{\beta} = 0 \).

Assuming that the result holds for \( p = m \), we prove it for \( p = m + 1 \):

\[
\sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)!}{\beta! (w-\beta)! (u-m-1-\beta)!} = \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)! (u-m-\beta)}{\beta! (w-\beta)! (u-m-\beta)!}.
\]
\begin{align*}
&= (u - m) \cdot \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u - \beta)! \beta}{\beta! (w - \beta)! (u - m - \beta)!} - \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u - \beta)! \beta}{\beta! (w - \beta)! (u - m - \beta)!} \\
&= - \sum_{\beta=1}^{w} (-1)^{\beta} \cdot \frac{(u - \beta)! \beta}{\beta! (w - \beta)! (u - m - \beta)!}, \quad \text{[by induction hypothesis, } p = m] \\
&= - \sum_{\beta=1}^{w} (-1)^{\beta} \cdot \frac{(u - \beta)! (w - \beta)! (u - m - \beta)!}{(\beta - 1)! ((w - 1 - \beta)! ((u - 1) - m - \beta)!} \\
&= \sum_{\beta=0}^{w-1} (-1)^{\beta} \cdot \frac{(u - 1 - \beta)!}{\beta! ((w - 1) - \beta)! ((u - 1) - m - \beta)!} = 0, \quad \text{[by induction hypothesis, } p = m].
\end{align*}

Condition 2: $u - p \leq w$. Then $\min(u - p, w) = u - p$ and $\min(u - p - 1, w) = u - p - 1$. We have that

\begin{align*}
&= - \sum_{\beta=0}^{u-p-1} C_{\beta}^{w} \cdot \left[ \sum_{\alpha=u-p-\beta}^{u-\beta} (-1)^{\alpha} C_{\alpha}^{u-\beta} \right] = - \sum_{\beta=0}^{u-p} C_{\beta}^{w} \cdot \left[ \sum_{\alpha=u-p-\beta}^{u-\beta} (-1)^{\alpha} C_{\alpha}^{u-\beta} \right] + \sum_{\alpha=0}^{p} (-1)^{\alpha} C_{\alpha}^{p} \\
&= - \sum_{\beta=0}^{u-p} C_{\beta}^{w} \cdot \left[ \sum_{\alpha=u-p-\beta}^{u-\beta} (-1)^{\alpha} C_{\alpha}^{u-\beta} \right] \\
&= - \sum_{\beta=0}^{u-p} C_{\beta}^{w} \cdot \left[ \sum_{\alpha=u-p+1-\beta}^{u-\beta} (-1)^{\alpha} C_{\alpha}^{u-\beta} \right] - \sum_{\beta=0}^{u-p} C_{\beta}^{w} \cdot (-1)^{u-p-\beta} \cdot C_{u-\beta-\beta-p}^{u-\beta-p} \\
&= - \sum_{\beta=0}^{u-p} C_{\beta}^{w} \cdot (-1)^{u-p-\beta} \cdot C_{u-\beta-\beta-p}^{u-\beta-p}, \quad \text{[by induction hypothesis, } r = u - p] \\
&= (-1)^{u-p+1} \sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot C_{\beta}^{w} \cdot C_{u-\beta-p}^{u-\beta-p} = (-1)^{1+u-p} \cdot \frac{w!}{p!} \sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot \frac{(u - \beta)!}{\beta! (w - \beta)! (u - p - \beta)!}.
\end{align*}

Now it is sufficient to prove that $\sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot \frac{(u - \beta)!}{\beta! (w - \beta)! (u - p - \beta)!} = 0$. We prove it by induction on $u - w$. For $u - w = 0$, it follows that $\sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot \frac{(u - \beta)!}{\beta! (w - \beta)! (u - p - \beta)!} = \sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot \frac{1}{\beta! (u - p - \beta)!} = \frac{1}{(u - p)!} \sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot C_{\beta}^{u-p} = 0$. Assuming that the result holds for $u - w = m$, we prove it for $u - w = m + 1$:

\begin{align*}
&= \sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot \frac{(u - \beta)!}{\beta! (u - m - 1 - \beta)! (u - p - \beta)!} = \sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot \frac{(u - \beta)! (u - m - \beta)}{\beta! (u - m - \beta)! (u - p - \beta)!} \\
&= (u - m) \cdot \sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot \frac{(u - \beta)!}{\beta! (u - m - \beta)! (u - p - \beta)!} - \sum_{\beta=0}^{u-p} (-1)^{\beta} \cdot \frac{(u - \beta)! \beta}{\beta! (u - m - \beta)! (u - p - \beta)!}.
\end{align*}
\[= \sum_{\beta=0}^{u-p} (-1)^\beta \cdot \frac{(u-\beta)! \beta}{\beta! (u-m-\beta)! (u-p-\beta)!}, \quad \text{[by induction hypothesis, } w = u-m]\]
\[= \sum_{\beta=1}^{u-p} (-1)^\beta \cdot \frac{(u-\beta)! \beta}{\beta! (u-m-\beta)! (u-p-\beta)!}\]
\[= \sum_{\beta=1}^{u-p} (-1)^\beta \cdot \frac{(u-\beta)!}{(\beta-1)! (u-m-\beta)! (u-p-\beta)!}\]
\[= \sum_{\beta=0}^{u-p-1} (-1)^\beta \cdot \frac{(u-1-\beta)!}{\beta! (u-m-1-\beta)! (u-p-1-\beta)!}\]
\[= \sum_{\beta=0}^{u-p-1} (-1)^\beta \cdot \frac{((u-1)-\beta)!}{\beta! (u-1-m-\beta)! ((u-1)-p-\beta)!}\]
\[= 0, \quad \text{[by induction hypothesis].}\]

**Proof of Proposition 4:** For every \( C \subseteq N \) we define boolean functions \( \chi_C : 2^N \rightarrow \mathbb{R} \) and \( \psi_C : 2^N \rightarrow \mathbb{R} \) by

\[\chi_C(X) = (-1)^{|C|} \cdot 1[C \subseteq X],\]
\[\psi_C(X) = (-1)^{|X|} \cdot 1[X \subseteq C],\]

where \( 1[A] \) is an indicator function which is equal to 1, if condition \( A \) is satisfied, and 0 otherwise. Then for all \( C_1, C_2 \subseteq N \) we claim that

\[\sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \begin{cases} 1, & \text{if } C_1 = C_2, \\ 0, & \text{otherwise}, \end{cases}\]  \hfill (A3)

First, we show that \( \sum_{X \subseteq N} \chi_C(X) \cdot \psi_C(X) = 1 \) for every \( C \subseteq N \):

\[\sum_{X \subseteq N} \chi_C(X) \cdot \psi_C(X) = \sum_{X \subseteq N} 1[C \subseteq X] \cdot (-1)^{|C|+|X|} 1[X \subseteq C] = (-1)^{|C|+|C|} = 1\]

Then we show that \( \sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = 1 \) for all \( C_1, C_2 \subseteq N \) s.t. \( C_1 \neq C_2 \):

\[\sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \sum_{X \subseteq N} 1[C_1 \subseteq X] \cdot (-1)^{|C_1|+|X|} 1[X \subseteq C_2]\]
\[= (-1)^{|C_1|} \sum_{X \subseteq N} (-1)^{|X|} 1[C_1 \subseteq X \subseteq C_2]\]
\[= (-1)^{|C_1|} \cdot (-1)^{|C_1|} \cdot \sum_{k=0}^{|C_2|-|C_1|} (-1)^k C_k^{n-|C_1|}, \text{ where } C_k^n = \frac{n!}{k!(n-k)!}\]

\[\text{[since the expression depends only on the cardinality of sets, the summation over the sets is reduced to the summation over the cardinality of sets]}
\[= (-1)^{|C_1|} \cdot (1-1)^{|C_2|-|C_1|} = 0.\]
Consequently, the probability to choose the \textit{“no purchase”} option $a_0$ from the offer set $\{N \setminus X\}^+$ is given by

$$
\mathbb{P}_0(N \setminus X) = \sum_{C \subseteq X} \lambda(C) = \sum_{C \subseteq N} \lambda(C) \cdot (-1)^{|C|} \cdot \chi_C(X) \cdot [C \subseteq X] \tag{A4}
$$

Then it follows that

$$
\sum_{X \subseteq C} (-1)^{|C| - |X|} \cdot \mathbb{P}_0(N \setminus X) = \sum_{X \subseteq N} \mathbb{P}_0(N \setminus X) \cdot (-1)^{|C|+|X|} [X \subseteq C] \tag{A5}
$$

$$
= (-1)^{|C|} \sum_{X \subseteq N} \mathbb{P}_0(N \setminus X) \cdot \psi_C(X) \tag{A5}
$$

$$
= (-1)^{|C|} \sum_{X \subseteq N} \sum_{C_1 \subseteq N} \lambda(C_1) \cdot (-1)^{|C_1|} \cdot \chi_{C_1}(X) \cdot \psi_C(X) \tag{A5}
$$

$$
= (-1)^{|C|} \sum_{C_1 \subseteq N} \lambda(C_1) \cdot (-1)^{|C_1|} \cdot \sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_C(X) \tag{A5}
$$

$$
= (-1)^{|C|} \lambda(C) \cdot (-1)^{|C|}, \quad \text{by Equation (A3)}
$$

Now it remains to prove the uniqueness of probability distribution function $\lambda$ obtained from purchasing transactions data under CTC choice model. Note that Equation (A4) relates probability distribution $\lambda$ over consideration sets to the choice frequencies $\mathbb{P}_0(N \setminus X)$ through the system of linear equations:

$$
\mathbb{P}_0(N \setminus X) = \sum_{C \subseteq N} \lambda(C) \cdot (-1)^{|C|} \cdot \chi_C(X), \quad \forall X \subseteq N \iff \mathbf{y} = A \cdot \lambda, \tag{A6}
$$

where $\mathbf{y} = (y_X)_{X \subseteq N}$ denotes the $|2^N| \times 1$ vector of choice fractions and $\lambda = (\lambda_C)_{C \subseteq N}$ denotes the $|2^N| \times 1$ vector that represents the probability distribution function over consideration sets. $A$ is the $|2^N| \times |2^N|$ matrix such that $A$’s entry corresponding to the row $X$ and column $C$ is equal to $(-1)^{|C|} \cdot \chi_C(X)$. Therefore, the relation between the choice frequencies and the underlying model can be represented in a compact form as $\mathbf{y} = A \cdot \lambda$. Then the proof of uniqueness of $\lambda$ reduces to showing that $\det(A) \neq 0$. From Equation (A5) we have

$$
\lambda(C) = (-1)^{|C|} \sum_{X \subseteq N} \mathbb{P}_0(N \setminus X) \cdot \psi_C(X), \quad \forall C \subseteq N \iff \lambda = B \cdot \mathbf{y},
$$
which establishes alternative linear relationship between choice frequencies \( \Pr_0(N \setminus X) \) and the model parameters \( \lambda \) in a compact form as \( \lambda = B \cdot y \), where \( B \) is the \( |2^N| \times |2^N| \) matrix such that \( B \)'s entry corresponding to the row \( C \) and column \( X \) is equal to \((-1)^{|C|} \cdot \psi_C(X)\). Therefore, we get

\[
\lambda = B \cdot y = B \cdot A \cdot \lambda. \quad \text{[by Equation (A6)]}
\]

\[\implies I = B \cdot A \implies \det(I) = \det(B) \cdot \det(A)\]

\[\implies 1 = \det(B) \cdot \det(A) \implies \det(A) \neq 0.\]

\[\square\]

**Proof of Proposition 5:** It follows from the proposition that

\[
\lambda(C) = \sum_{X \subseteq N} \sum_{Y \supseteq X \cup C} (-1)^{1+|Y|-|X \Delta C|} \cdot \Pr_0(X) \cdot \mathbf{I}[|X \cup C| \leq k < |Y|]
= \sum_{X \subseteq N} \sum_{Y \supseteq X \cup C} \Pr_0(X) \cdot (-1)^{1+|X \cap C|} \cdot (-1)^{|Y|-|X \cup C|} \cdot \mathbf{I}[|X \cup C| \leq k < |Y|]
= \sum_{X \subseteq N} \Pr_0(X) \cdot (-1)^{1+|X \cap C|} \cdot \mathbf{I}[|X \cup C| \leq k] \cdot \sum_{Y \supseteq X \cup C} (-1)^{|Y|-|X \cup C|} \cdot \mathbf{I}[|Y| > k]
\]

\[\text{[since the expression depends only on the cardinality of sets } Y, \text{ the summation over the sets } Y \text{ is reduced to the summation over the cardinality of sets } Y]\]

\[
= \sum_{X \subseteq N} \Pr_0(X) \cdot (-1)^{1+|X \cap C|} \cdot \mathbf{I}[|X \cup C| \leq k] \cdot \sum_{\alpha=k+1-|X \cup C|}^{n-|X \cup C|} (-1)^\alpha C_\alpha^n \cdot (-X \cup C),
\]

\[
\text{where } C_\alpha^n = \frac{n!}{k!(n-k)!}.
\]

For every \( C \subseteq N \) we define boolean functions \( \chi_C : 2^N \rightarrow \mathbb{R} \) and \( \psi_C : 2^N \rightarrow \mathbb{R} \) by

\[
\chi_C(X) = \mathbf{I}[C \subseteq \bar{X}, |C| \leq k],
\]

\[
\psi_C(X) = (-1)^{1+|X \cap C|} \cdot \mathbf{I}[|X \cup C| \leq k] \cdot \sum_{\alpha=k+1-|X \cup C|}^{n-|X \cup C|} (-1)^\alpha C_\alpha^n \cdot (-X \cup C).
\]

Restricting consideration sets and offer sets by the size of up to \( k \) (by assumption of proposition), we represent the probability to choose the “no purchase” option \( a_0 \) from the offer set \( X^+ \) through linear combination of boolean functions \( \chi_C(X) \) as follows:

\[
\Pr_0(X) = \sum_{C \subseteq N} \lambda(C) \cdot \mathbf{I}[C \subseteq \bar{X}, |C| \leq k] = \sum_{C \subseteq N} \lambda(C) \cdot \chi_C(X). \quad \text{(A7)}
\]

Then for all \( C_1, C_2 \subseteq N \) such that \(|C_1|, |C_2| \leq k < n\) we claim that

\[
\sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \begin{cases} 1, & \text{if } C_1 = C_2, \\ 0, & \text{otherwise.} \end{cases} \quad \text{(A8)}
\]
Consequently, it follows from the claim that

$$\sum_{X \subseteq N} P_0(X) \cdot (-1)^{1+|X\cap C|} \cdot I[|X \cup C| \leq k] \cdot \sum_{\alpha = k+1 - |X \cup C|}^{n-|X \cup C|} (-1)^\alpha c_\alpha^{n-|X \cup C|} \quad \text{(A9)}$$

$$= \sum_{X \subseteq N} P_0(X) \cdot \psi_C(X) = \sum_{X \subseteq N} \sum_{C_1 \subseteq N} \lambda(C_1) \cdot \chi_{C_1}(X) \cdot \psi_C(X)$$

$$= \sum_{C_1 \subseteq N} \lambda(C_1) \cdot \sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_C(X) = \lambda(C), \quad \text{by Equation (A8)}.$$  

Now to complete the proof of the proposition, it is sufficient to prove the claim and show the uniqueness of the solution. We prove the claim by considering two different cases.

**Case 1:** $C_2 \subseteq C_1$,

$$\sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \sum_{X \subseteq N} (-1)^1 \cdot I[|C_1| \leq k] \cdot I[X \cap C_1 = \emptyset] \cdot I[|X| \leq k - |C_2|] \times \sum_{\alpha = k+1 - |X|}^{n-|X|} (-1)^\alpha c_\alpha^{n-|X|}$$

$$\text{in this case, } X \cap C_1 = \emptyset, \ |X \cap C_1| = 0, \ |X \cap C_2| = 0, \text{ and } |X \cup C_2| = |X| + |C_2|$$

$$= - \sum_{X \subseteq N} I[|C_1| \leq k] \cdot I[X \cap C_1 = \emptyset] \cdot I[|X| \leq k - |C_2|] \times \sum_{\alpha = k - |C_2|}^{n-|C_2| - |X|} (-1)^\alpha c_\alpha^{n-|C_2| - |X|},$$

$$\text{[since the expression depends only on the cardinality of sets, the summation over the sets is reduced to the summation over the cardinality of sets]}

\begin{align*}
\min(k-|C_2|, n-|C_1|) &= \sum_{\beta = 0}^{\min(k-|C_2|, n-|C_1|)} \left( \sum_{\alpha = k - |C_2| + 1 - \beta}^{n-|C_2|-\beta} (-1)^\alpha c_\alpha^{n-|C_2|-\beta} \right), \quad \text{[where } \beta \text{ - cardinality of set } X] \\
&= \begin{cases} 
1, & \text{if } C_1 = C_2, \\
0, & \text{if } C_1 \subset C_2,
\end{cases}
\end{align*}$$

where the last equality follows by invoking Lemma A2, where $w = n - |C_1|$, $r = k - |C_2|$, and $u = n - |C_2|$.

**Case 2:** $C_2 \not\subseteq C_1$.

$$\sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \sum_{X \subseteq N} (-1)^{1+|X\cap C_2|} \cdot I\left[X \cap C_1 = \emptyset, |C_1| \leq k\right] \cdot I\left[|X \cup C_2| \leq k\right] \times \sum_{\alpha = k+1 - |X \cup C_2|}^{n-|X \cup C_2|} (-1)^\alpha c_\alpha^{n-|X \cup C_2|}$$

[since the expression depends only on the cardinality of sets, the summation over the sets is reduced to the summation over the cardinality of sets]
In this figure $C_1 \not\subset C_2$ and $n - |C_1| - |C_5| > 0$. We have that $C_5 = C_2 \setminus \{C_1 \cap C_2\}$, $Y = X \cap C_5$, $Z = X \setminus Y$, where $\gamma$ and $\beta$ correspond to the cardinalities of $Y$ and $Z$ respectively.

In this figure $C_1 \subset C_2$, then since $k < n$ by assumption of the proposition it follows that $n - |C_1| - |C_5| > 0$. We have that $C_5 = C_2 \setminus \{C_1 \cap C_2\}$, $Y = X \cap C_5$, $Z = X \setminus Y$, where $\gamma$ and $\beta$ correspond to the cardinalities of $Y$ and $Z$ respectively.

In this figure $C_1 \not\subset C_2$ and $n - |C_1| - |C_5| = 0$. We have that $C_5 = C_2 \setminus \{C_1 \cap C_2\}$, $X \subseteq C_5$, and $\gamma$ corresponds to the cardinality of $X$.
where the last equality follows since \(|C_5| > 0\), and \(\sum_{\gamma=0}^{|C_5|}(-1)^{\gamma} \cdot C_7^{[\gamma]} = 0\).

In order to complete the proof, we show the uniqueness of probability distribution function \(\lambda\) in our setting. First, note that Equation (A7) relates probability distribution \(\lambda\) over consideration sets to the choice frequencies \(Pr(a_0|X)\) through the system of linear equations:

\[
P_0(X) = \sum_{\gamma \subseteq N} \lambda(C) \cdot \chi_C(X), \quad \forall X \subseteq N \iff y = A \cdot \lambda,
\]

(A10)

where \(y = (y_X)_{X \subseteq N}\) denotes the \(|2^N| \times 1\) vector of choice fractions and \(\lambda = (\lambda_C)_{C \subseteq N}\) denotes the \(|2^N| \times 1\) vector that represents the probability distribution function over consideration sets. \(A\) is the \(|2^N| \times |2^N|\) 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 frequencies and the underlying model can be represented in a compact form as \(y = A \cdot \lambda\). Then the proof of uniqueness of \(\lambda\) reduces to showing that \(\det(A) \neq 0\). It follows from Equation (A9) that

\[
\lambda(C) = \sum_{X \subseteq N} P_0(X) \cdot \psi_C(X), \quad \forall C \subseteq N \iff \lambda = B \cdot y,
\]

which provides another relationship between choice frequencies \(P_0(X)\) and the model parameters \(\lambda\) in a linear form as \(\lambda = B \cdot y\), where \(B\) is the \(|2^N| \times |2^N|\) matrix such that \(B\)'s entry corresponding to the row \(C\) and column \(X\) is equal to \(\psi_C(X)\). Therefore, we get

\[
\lambda = B \cdot y = B \cdot A \cdot \lambda, \quad \text{[by Equation (A10)]}
\]

\[
\Rightarrow I = B \cdot A \Rightarrow \det(I) = \det(B) \cdot \det(A)
\]

\[
\Rightarrow 1 = \det(B) \cdot \det(A) \Rightarrow \det(A) \neq 0.
\]

\[\square\]

**Lemma A3.** Assume that for all consideration sets \(C \subseteq N\) we have that

\[
\sum_{X \subseteq C} (-1)^{|C| - |X|} P_0(N \setminus X) \geq 0,
\]
Lemma A4. If a sample of sales transaction data satisfies Conditions 1, 2, and 3, then for all consideration sets $C \subseteq S$ s.t. $S \subseteq N$ it follows that

$$\sum_{X \subseteq C} (-1)^{|C| - |X|} \mathbb{P}_0(S \setminus X) \geq 0,$$

with strict inequality for consideration sets of the size up to three, i.e., if $|C| \leq 3$.

Proof: Suppose that $C \subseteq S$ and $S \subseteq N$. Let $\bar{S}$ denote $N \setminus S$. We can now establish the following chain of equalities:

$$\sum_{B \subseteq C} (-1)^{|C| - |B|} \mathbb{P}_0(S \setminus B) = \sum_{B \subseteq C} (-1)^{|C| - |B|} \cdot \mathbb{P}_0(\{N \setminus \bar{S} \setminus B\})$$

$$= \sum_{B \subseteq C} \sum_{A \subseteq S} \sum_{D \subseteq A} (-1)^{|A| - |D|} \cdot (-1)^{|C| - |B|} \cdot \mathbb{P}_0(\{N \setminus B\} \setminus D)$$

$$= \sum_{A \subseteq S} \sum_{D \subseteq A} \sum_{B \subseteq C} (-1)^{|C| - |B|} \cdot \mathbb{P}_0(\{N \setminus B\} \setminus D)$$

$$= \sum_{A \subseteq S} \sum_{X \in \langle A \cup C \rangle} (-1)^{|C| + |A| - |X|} \cdot \mathbb{P}_0(N \setminus X) \quad \text{[where } X = D \cup B, \text{since } A \cap C = \emptyset \text{]}$$

$$= \sum_{A \subseteq S} \sum_{X \in \langle A \cup C \rangle} (-1)^{|C \cup A| - |X|} \cdot \mathbb{P}_0(N \setminus X) \quad \text{[since } A \cap C = \emptyset \text{]}$$

$$= \sum_{X \subseteq C'} \sum_{X \subseteq C'} (-1)^{|C'| - |X|} \cdot \mathbb{P}_0(N \setminus X) \quad \text{[where } C' = A \cup C \text{]}$$

$$\geq 0, \text{ with strict inequality when } |C| \leq 3, \quad \text{[by assumptions of the Lemma,}}$$

since

$$\sum_{X \subseteq C'} (-1)^{|C'| - |X|} \cdot \mathbb{P}_0(N \setminus X) \geq 0 \text{ with strict inequality when } |C'| \leq 3.$$
$S_2 = S \cup \{a_2\}$ and $a_2 \notin S$. Next, assume, by contradiction, that $\mathbb{P}_1(S_2 \setminus \{a_2\}) < \mathbb{P}_1(S_2)$. Consequently, by Condition 2 it follows that $\mathbb{P}_1(\{a_1\}) < \mathbb{P}_1(\{a_1, a_2\})$. Then by Condition 1 we have that

$$\mathbb{P}_2(\{a_2\}) = \mathbb{P}_2(\{a_1, a_2\}). \tag{A11}$$

It now follows that

$$\begin{align*}
\mathbb{P}_1(\{a_1\}) - \mathbb{P}_1(\{a_1, a_2\}) &= \left(1 - \mathbb{P}_0(\{a_1\})\right) - \left(1 - \mathbb{P}_0(\{a_1, a_2\}) - \mathbb{P}_2(\{a_1, a_2\})\right), \quad \text{[by standard probability property]} \\
&= \left(1 - \mathbb{P}_0(\{a_1\})\right) - \left(1 - \mathbb{P}_0(\{a_1, a_2\}) - \mathbb{P}_2(\{a_2\})\right), \quad \text{[by Equation (A11)]} \\
&= \left(1 - \mathbb{P}_0(\{a_2\}) - \mathbb{P}_0(\{a_1, a_2\})\right), \quad \text{[by standard probability property]} \\
&= 1 - \mathbb{P}_0(\{a_1\}) - \mathbb{P}_0(\{a_2\}) + \mathbb{P}_0(\{a_1, a_2\}) > 0, \quad \text{[by Condition 3 and Lemma A3, when } C = S = \{a_1, a_2\}, \text{]} \\
\end{align*}$$

which contradicts to $\mathbb{P}_1(\{a_1\}) < \mathbb{P}_1(\{a_1, a_2\})$. Then we have

$$\begin{align*}
\mathbb{P}_1(S_2) &\leq \mathbb{P}_1(S_2 \setminus \{a_2\}) = \mathbb{P}_1(S), \quad \text{[note that } |S| - |S_1| = k] \\
&\leq \mathbb{P}_1(S_1), \quad \text{[by induction hypothesis].}
\end{align*}$$

Therefore, the result now follows by induction. \(\square\)

**Lemma A5.** Consider $a_1, a_2 \in S$. $S \subseteq N$, $a_1 \neq a_2$. Then GCS choice model, with strict preference list $\sigma$ and distribution over consideration sets $\lambda$ where $\lambda(C) > 0$ if $|C| \leq 3$, implies the following list of implications:

a) $\mathbb{P}_1(S \setminus \{a_2\}) > \mathbb{P}_1(S) \iff a_2 > a_1$, and $\forall S' \subseteq N$ s.t. $a_1, a_2 \in S'$: $\mathbb{P}_1(S' \setminus \{a_2\}) > \mathbb{P}_1(S')$,

b) $\mathbb{P}_1(S \setminus \{a_2\}) = \mathbb{P}_1(S) \iff a_1 > a_2$, and $\forall S' \subseteq N$ s.t. $a_1, a_2 \in S'$: $\mathbb{P}_1(S' \setminus \{a_2\}) = \mathbb{P}_1(S')$,

c) $\mathbb{P}_1(S \setminus \{a_2\}) \neq \mathbb{P}_1(S) \iff \mathbb{P}_2(S \setminus \{a_1\}) = \mathbb{P}_2(S)$.

**Proof:** a) Suppose that $\mathbb{P}_1(S \setminus \{a_2\}) > \mathbb{P}_1(S)$. Assume, by contradiction, that $a_1 > a_2$. Then it can be inferred from purchase probability definition under GCS, see Equation (1), that $\mathbb{P}_1(S \setminus \{a_2\}) = \mathbb{P}_1(S)$, which leads to contradiction. As a result, we have that $a_2 > a_1$ since preferences are strict and asymmetric. Then $\forall S' \subseteq N$ s.t. $a_1, a_2 \in S'$ we establish that

$$\begin{align*}
\mathbb{P}_1(S' \setminus \{a_2\}) - \mathbb{P}_1(S') &\geq \lambda(\{a_1, a_2\}), \quad \text{[by Equation (1)]} \\
&> 0, \quad \text{[by Assumption that } \lambda(C) > 0 \text{ if } |C| \leq 3].}
\end{align*}$$
b) Suppose that $P_1(S \setminus \{a_2\}) = P_1(S)$. Assume, by contradiction, that $a_2 \succ a_1$. Then it follows that

$$P_1(S \setminus \{a_2\}) - P_1(S) \geq \lambda(\{a_1, a_2\}), \quad \text{[by Equation (1)]}$$

$$> 0, \quad \text{[by Assumption that } \lambda(C) > 0 \text{ if } |C| \leq 3].$$

which contradicts the assumption above. As a result, we have that $a_1 \succ a_2$, since preferences are strict and asymmetric. Then by Equation (1) we have that $\forall S' \subseteq N$ s.t. $a_1, a_2 \in S'$: $P_1(S' \setminus \{a_2\}) = P_1(S')$.

c) Suppose that $P_1(S \setminus \{a_2\}) \neq P_1(S)$. Then it is straightforward to verify that $P_1(S \setminus \{a_2\}) > P_1(S)$, since the following inequality holds from the Lemma A4: $P_1(S \setminus \{a_2\}) \geq P_1(S)$. Consequently, invoking the implication from part a), we have $a_2 \succ a_1$, and by Equation (1) we obtain that $P_2(S \setminus \{a_1\}) = P_2(S)$. □

Proof of Proposition 7: It follows directly from the proof of Proposition 2 that $GCS \subseteq RUM$. Then, it remains to show that RUM model class is not a specific case of GCS model class. To this end, we provide a particular example of RUM model class resulting in customers’ choice frequencies that are inconsistent with GCS choice rule.

Let $N$ denote the universe of two items plus the “no purchase” option $a_0$, i.e., $N = \{a_1, a_2\}$. Then let $\mu: \mathcal{L}_3 \to [0, 1]$ denote a specification of RUM class such that customers sample either preference list $\sigma_1 = \{a_1, a_2, a_0\}$ with probability $\mu_1 \in (0, 1)$ or preference list $\sigma_2 = \{a_2, a_1, a_0\}$ with probability $1 - \mu_1$. Consequently, probability distribution function $\mu$ over preference lists results in the following choice frequencies:

$$P_1(\{a_1, a_2\}) = \mu_1, \quad P_1(\{a_1\}) = 1 \Rightarrow a_2 \text{ is preferred to } a_1, \text{ by GCS definition},$$

$$P_2(\{a_1, a_2\}) = 1 - \mu_1, \quad P_2(\{a_2\}) = 1 \Rightarrow a_1 \text{ is preferred to } a_2, \text{ by GCS definition}.$$

These choice frequencies are inconsistent with GCS model class, which only allows a unique preference order of products, i.e., according to GCS choice rule either product $a_1$ is preferred to product $a_2$ or product $a_2$ is preferred to product $a_1$.

□

Proof of Proposition 8: Necessity: if purchasing transactions data is consistent with GCS choice model with strict preference list $\sigma$ and distribution over consideration sets $\lambda$ where $\lambda(C) > 0$ if $|C| \leq 3$, then we claim that three axioms Condition 1, Condition 2, and Condition 3 are satisfied. First, it follows from Proposition 4 that Condition 3 is satisfied. Then Condition 1 and Condition 2 are satisfied by Lemma A5.
Sufficiency: we claim that the choice rule that satisfies Condition 1, Condition 2, and Condition 3 is a GCS choice model with the strict preference list $\sigma$ where no purchase option is the least preferred item, and probability distribution function $\lambda$ over consideration sets such that $\lambda(C) > 0$ if $|C| \leq 3$.

Define a binary relation $\delta_{ij}$ between products $a_i, a_j \subseteq N$, $a_i \neq a_j$, where $\delta_{ij} = 1$ if $P_j(S \setminus \{a_i\}) > P_j(S)$ for some $S \subseteq N$ s.t. $a_i, a_j \in S$ (note, by Condition 2 it implies that $P_j(S \setminus \{a_i\}) > P_j(S)$ for all $S \subseteq N$ s.t. $a_i, a_j \in S$), and zero otherwise. We claim that $\delta_{ij}$ is complete, asymmetric, and transitive binary relation.

First, we prove that this binary relation is complete, i.e., either $\delta_{ij} = 1$ or $\delta_{ji} = 1$. Suppose that $P_j(S \setminus \{a_i\}) \leq P_j(S)$ for some $S \subseteq N$, i.e., $\delta_{ij} = 0$. Then it follows from the Lemma A4 that $P_j(S \setminus \{a_i\}) = P_j(S)$. Moreover, by Condition 2 we have that $P_j(\{a_j\}) = P_j(\{a_i, a_j\})$. We can now establish the following chain of equalities:

\[
P_i(\{a_i\}) - P_i(\{a_i, a_j\})
\]

\[
= \left(1 - P_0(\{a_i\})\right) - \left(1 - P_0(\{a_i, a_j\}) - P_j(\{a_i, a_j\})\right), \quad \text{[by standard probability property]}
\]

\[
= \left(1 - P_0(\{a_i\})\right) - \left(1 - P_0(\{a_i, a_j\}) - P_j(\{a_j\})\right), \quad \text{[by Condition 2, see above]}
\]

\[
= \left(1 - P_0(\{a_i\})\right) - \left(P_0(\{a_j\}) - P_0(\{a_i, a_j\})\right), \quad \text{[by standard probability property]}
\]

\[
= 1 - P_0(\{a_i\}) - P_0(\{a_j\}) + P_0(\{a_i, a_j\}) > 0, \quad \text{[by Condition 3 and Lemma A3, where $C = S = \{a_i, a_j\}$]},
\]

which concludes that $\delta_{ji} = 1$. Therefore, completeness of binary relation $\delta_{ij}$ now follows.

Second, we establish that the defined binary relation $\delta$ is asymmetric, i.e., if $\delta_{ij} = 1$ then $\delta_{ji} = 0$. Suppose that $P_j(S \setminus \{a_i\}) > P_j(S)$ for some $S \subseteq N$, i.e., $\delta_{ij} = 1$. Then by Condition 1 we have that $P_i(S \setminus \{a_j\}) = P_i(S)$ (note, by Condition 2 we have that for all $S' \subseteq N$ s.t. $a_i, a_2 \in S'$: $P_i(S' \setminus \{a_j\}) = P_i(S')$), which further implies that $\delta_{ji} = 0$. As a result, asymmetry of binary relation $\delta_{ij}$ now follows.

Third, we show the transitivity of binary relation $\delta$, i.e., if $\delta_{ij} = 1$ and $\delta_{jk} = 1$ then $\delta_{ik} = 1$ for all $a_i, a_j, a_k \in N$. Assume by contradiction that binary relation $\delta$ is not transitive. To this end, there exist $a_i, a_j, a_k \in N$ such that $\delta_{ij} = 1$, $\delta_{jk} = 1$, $\delta_{ik} = 0$ with the following list of implications:

\[
\delta_{ij} = 1 \Rightarrow P_j(S \setminus \{a_i\}) > P_j(S), \quad [\text{for some } S \subseteq N ]
\]

\[
\Rightarrow P_j(\{a_j, a_k\}) > P_j(\{a_i, a_j, a_k\}), \quad \text{[by Condition 2]}
\]

\[
\Rightarrow P_i(\{a_i, a_k\}) = P_i(\{a_j, a_i, a_k\}), \quad \text{[by Condition 1]}
\]

\[
\Rightarrow P_i(\{a_i\}) = P_i(\{a_j, a_i\}), \quad \text{[by Condition 2], (A12)}
\]

\[
\Rightarrow P_i(\{a_i\}) = P_i(\{a_j, a_k\}), \quad \text{[by Condition 2], (A13)}
\]
\[
\delta_{jk} = 1 \Rightarrow P_k(S \setminus \{a_j\}) > P_k(S), \quad \text{[for some } S \subseteq N]\]
\[
\Rightarrow P_k(\{a_i, a_k\}) > P_k(\{a_i, a_j, a_k\}), \quad \text{[by Condition 2]}\]
\[
\Rightarrow P_j(\{a_i, a_j\}) = P_j(\{a_i, a_j, a_k\}), \quad \text{[by Condition 1]} \quad (A14)\]
\[
\Rightarrow P_j(\{a_j\}) = P_j(\{a_j, a_k\}), \quad \text{[by Condition 2]}, \quad (A15)\]
\[
\delta_{jk} = 0 \Rightarrow P_k(S \setminus \{a_i\}) \leq P_k(S), \quad \text{[for some } S \subseteq N]\]
\[
\Rightarrow P_k(S \setminus \{a_i\}) = P_k(S), \quad \text{[by Lemma A4]} \quad (A16)\]
\[
\Rightarrow P_k(\{a_k\}) = P_k(\{a_i, a_k\}), \quad \text{[by Condition 2]} \quad (A17)\]

Using the property of the choice rule, i.e., \( \forall S \subseteq N : \sum_{a_i \in S^+} P_r(S) = 1 \), for offer sets \( S_1 = \{a_i, a_j\}, S_2 = \{a_j, a_k\}, S_3 = \{a_i, a_k\}, \) and \( S_4 = \{a_i, a_j, a_k\} \) we further establish the following list of implications:

For \( S_1 = \{a_i, a_j\} : P_i(S_1) + P_j(S_1) + P_0(S_1) = 1 \)
\[
\Rightarrow P_i(\{a_i\}) + P_j(S_1) + P_0(S_1) = 1, \quad \text{[by Equation (A13)]}\]
\[
\Rightarrow P_i(\{a_i\}) + P_j(S_1) + P_0(S_1) = 1, \quad \text{[by Equation (A14)]}\]
\[
\Rightarrow P_j(S_4) = P_0(\{a_i\}) - P_0(S_1), \quad \text{[by standard probability property]}. \quad (A18)\]

For \( S_2 = \{a_j, a_k\} : P_j(S_2) + P_k(S_2) + P_0(S_2) = 1 \)
\[
\Rightarrow P_j(\{a_j\}) + P_k(S_2) + P_0(S_2) = 1, \quad \text{[by Equation (A15)]}\]
\[
\Rightarrow P_j(\{a_j\}) + P_k(S_4) + P_0(S_2) = 1, \quad \text{[by Equation (A16)]}\]
\[
\Rightarrow P_k(S_4) = P_0(\{a_j\}) - P_0(S_2), \quad \text{[by standard probability property]}. \quad (A19)\]

For \( S_3 = \{a_i, a_k\} : P_k(S_3) + P_i(S_3) + P_0(S_3) = 1 \)
\[
\Rightarrow P_k(\{a_k\}) + P_i(S_3) + P_0(S_3) = 1, \quad \text{[by Equation (A17)]}\]
\[
\Rightarrow P_k(\{a_k\}) + P_i(S_4) + P_0(S_3) = 1, \quad \text{[by Equation (A12)]}\]
\[
\Rightarrow P_i(S_4) = P_0(\{a_k\}) - P_0(S_3), \quad \text{[by standard probability property]}. \quad (A20)\]

For \( S_4 = \{a_i, a_j, a_k\} : P_i(S_4) + P_j(S_4) + P_k(S_4) + P_0(S_4) = 1 \)
\[
\Rightarrow 0 = P_0(\emptyset) - P_0(\{a_i\}) - P_0(\{a_j\}) - P_0(\{a_k\}) + P_0(S_1) + P_0(S_2)
\[ + \mathbb{P}_0(S_3) - \mathbb{P}_0(S_4), \quad \text{since } \mathbb{P}_0(\emptyset) = 1, \text{ and by Equations (A18)-(A20)} \]
\[ > 0, \quad \text{by Condition 3 and Lemma A3, where } C = S = S_4, \]

which leads to contradiction. Therefore, the preference relation \( \delta \) is transitive. Since we proved that binary relation \( \delta \) is complete, asymmetric, and transitive, it specifies strict preference list \( > \) over products in \( N \), s.t. \( a_i > a_j \) iff \( \delta_{ij} = 1 \). In addition, it immediately follows from the axioms that \( a_0 \) is the least preferred item in the product universe according to the preference list \( > \), i.e., for all \( a_i \in N \) we have that \( \delta_{0i} = 0 \):

\[ \mathbb{P}_0(\emptyset) - \mathbb{P}_0(\{a_i\}) > 0, \quad \text{by Condition 3 and Lemma A3, where } C = S = \{a_i\}, \]

which implies that \( \delta_{0i} = 0 \) by definition.

Next, we prove that

\[ \mathbb{P}_r(S) = \mathbb{P}_0(S' \setminus \{a_r\}) - \mathbb{P}_0(S'), \quad \forall \ a_r \in S \quad \text{s.t. } S \subseteq N, \]

where \( S' \) is the set of products that consists of product \( a_r \) and all the items in \( S \) that are preferred to item \( a_r \), i.e., \( S' = \{a_j \in S : a_j > a_r\} \cup \{a_r\} \). The argument is proved by induction on the cardinality \( k \) of the offer set \( S \), i.e., \( k = |S| \). For the base case, \( k = 1 \), we have \( \mathbb{P}_r(\{a_r\}) = 1 - \mathbb{P}_0(\{a_r\}) = \mathbb{P}_0(\emptyset) - \mathbb{P}_0(\{a_r\}) \). Suppose the result follows for \( k \leq p \), then we prove it for \( k = p + 1 \). We consider two cases.

**Case 1**: product \( a_r \) is not the least preferred item in \( S \). In other words, there exists \( a_j \in S \) s.t. \( a_r > a_j \). Then by definition of the binary relation \( \delta \) we have that \( \mathbb{P}_j(S \setminus \{a_r\}) > \mathbb{P}_j(S) \), and the result now follows:

\[ \mathbb{P}_r(S) = \mathbb{P}_r(S \setminus \{a_j\}), \quad \text{[by Condition 1]} \]
\[ = \mathbb{P}_0(S' \setminus \{a_r\}) - \mathbb{P}_0(S'), \quad \text{[by induction hypothesis, and note that } a_j \notin S' \text{ since } a_r > a_j]. \]

**Case 2**: product \( a_r \) is the least preferred item in \( S \). Consider offer set \( S = \{a_r, a_1, a_2, \ldots, a_{p-1}\} \) such that \( \text{w.l.o.g. } a_{p-1} > \ldots > a_2 > a_1 > a_r \). Assuming \( a_r \in S \), we can now establish the following chain of equalities:

\[ \mathbb{P}_r(S) = 1 - \mathbb{P}_0(S) - \sum_{i=1}^{p-1} \mathbb{P}_i(S) \]
\[ = -\mathbb{P}_0(S) + \mathbb{P}_0(\emptyset) - \sum_{i=1}^{p-1} \mathbb{P}_i(\{a_r, a_1, a_2, \ldots, a_{p-1}\}) \]
\[= -P_0(S) + P_0(\emptyset) - \sum_{i=1}^{p-1} P_i(\{a_i, a_{i+1}, \ldots, a_{p-1}\}) \] 
by Condition 1
\[= -P_0(S) + P_0(\emptyset) - \sum_{i=1}^{p-1} \left( P_0(\{a_{i+1}, \ldots, a_{p-1}\}) - P_0(\{a_i, a_{i+1}, \ldots, a_{p-1}\}) \right),\]
by induction hypothesis
\[= -P_0(S) + P_0(\{a_1, a_2, \ldots, a_{p-1}\}) \]
\[= P_0(\{a_1, a_2, \ldots, a_{p-1}\}) - P_0(\{a_r, a_1, a_2, \ldots, a_{p-1}\}) = P_0(S' \setminus \{a_r\}) - P_0(S').\]

Let us denote two particular sets \( \hat{S} \) and \( \bar{S}' \) as follows: \( \hat{S} = N \setminus \{S' \setminus \{a_r\}\} \), \( \bar{S}' = N \setminus S' \). We can now establish the following chain of equalities:

\[ P_r(S) = P_0(S' \setminus \{a_r\}) - P_0(S') \]
\[= P_0(S' \setminus \{a_r\}) + \left( \sum_{C \subseteq S} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot P_0(N \setminus X) - P_0(N \setminus \hat{S}) \right) - P_0(S') \]
by invoking Lemma A1, where \( Z = N, Y = \hat{S}, P = C \), and \( f(Z \setminus Y) = P_0(N \setminus \hat{S}) \]
\[= \sum_{C \subseteq S} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot P_0(N \setminus X) - P_0(S') \quad \text{since } N \setminus \hat{S} = S' \setminus \{a_r\} \]
\[= \sum_{C \subseteq S} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot P_0(N \setminus X) - \sum_{C \subseteq S'} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot P_0(N \setminus X) \quad \text{since } N \setminus \bar{S}' = S' \]
by invoking Lemma A1, where \( Z = N, Y = \bar{S}', P = C \), and \( f(Z \setminus Y) = P_0(N \setminus \bar{S}') \]
\[= \sum_{C \subseteq S} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot P_0(N \setminus X) - \sum_{C \subseteq S'} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot P_0(N \setminus X) \quad \text{since } \hat{S} = \bar{S}' \cup \{a_r\} \]
\[= \sum_{C \subseteq (S' \cup \{a_r\})} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot P_0(N \setminus X) - \sum_{C \subseteq S'} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot P_0(N \setminus X) \]
\[= \sum_{C \subseteq (S' \cup \{a_r\})} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot P_0(N \setminus X) \]
\[= \sum_{C \subseteq (S' \cup \{a_r\})} \lambda(C), \quad \text{where } \lambda(C) = \sum_{X \subseteq C} (-1)^{|C|-|X|} P_0(N \setminus X) \]
\[= \sum_{C \subseteq N} \lambda(C) \cdot I[a_r \in C] \cdot I[C \in (\bar{S}') \setminus \{a_r\}] \]
\[= \sum_{C \subseteq N} \lambda(C) \cdot I[a_r \in C] \cdot I[a_r \succ a_k \forall a_k \in S \cap C, a_k \neq a_r] \]
\[= \sum_{C \subseteq N} \lambda(C) \cdot I[a_r \in S \cap C] \cdot I[a_r \succ a_k \forall a_k \in S \cap C, a_k \neq a_r] \quad \text{since we assume that } a_r \in S, \]
otherwise the choice probability is 0 ,
which is exactly the equation to compute the probability to purchase \( a_r \in S \) for the offer set \( S \subseteq N \) under GCS choice model. As a result, we also have \( P_0(S) = \sum_{C \subseteq N} \lambda(C) \cdot I[S \cap C = \emptyset] \) because of the standard probability law, i.e., \( P_0(S) = 1 - \sum_{a_r \in S} P_r(S) \). Note that above chain of equations specifies probability distribution function \( \lambda \) over consideration sets. Moreover, it follows from Proposition 4 that \( \lambda \) is defined uniquely. In order to complete the proof, we show that the preference relation \( \succ \) is also defined uniquely. Suppose, by contradiction, there is another strict preference order \( \succ' \) such that \( \succ' \neq \succ \) and \( P_r(\cdot, \lambda) = P_r(\cdot, \lambda) \). Therefore there exist items \( a_i, a_j \in N \) s.t. \( a_i \succ a_j \) and \( a_j \succ' a_i \). By definition of GCS choice rule, we have

\[
P_i(\{a_i, a_j\}, \lambda) = \sum_{C \subseteq N} I[a_i \in C] \cdot \lambda(C),
\]

\[
P_i(\{a_i, a_j\}, \lambda) = \sum_{C \subseteq N} I[a_i \in C] \cdot I[a_j \notin C] \cdot \lambda(C).
\]

As a result, we can establish now the following chain of inequalities:

\[
P_i(\{a_i, a_j\}, \lambda) - P_i(\{a_i, a_j\}, \lambda) \geq \lambda(\{a_i, a_j\}) > 0, \quad \text{by Condition 3},
\]

which contradicts to \( P_r(\cdot, \lambda) = P_r(\cdot, \lambda) \). □

A3. MINLP Formulation and Estimation Methodologies of Consider-then-Choose models

We start this section by providing the MINLP formulation for the logistic based consider-then-choose model. Then we describe the outer-approximation algorithm which is used to calibrate different variants of consider-then-choose models followed by the empirical validation of this algorithm. We finish this section by describing the EM algorithm to calibrate the GCS and CTC models.

A3.1. MINLP Formulation: Logistic-based Consider-then-Choose model.

In this part of the section we formulate the maximum likelihood estimation problem for the logistic-based consider-then-choose model, and then simplify it in such a way so that we can apply the outer-approximation algorithm in Section A3.2 in order to calibrate it. Recall that \( \delta_{kj}, \forall j, k, k \neq j \), is binary linear ordering variable such that \( \delta_{kj} = 1 \) if product \( a_k \) goes before product \( a_j \) in the preference list \( \succ \) (or, equivalently, \( \sigma \)), and \( \delta_{kj} = 0 \) otherwise. The data log-likelihood function under this model is given by

\[
\mathcal{L}(\beta, \delta) = \sum_{t=1}^{T} \left[ \log \frac{e^{\beta X_{jt}}}{1 + e^{\beta X_{jt}}} + \sum_{a_k \in S_t: k \neq j_t} \delta_{kj_t} \log \frac{1}{1 + e^{\beta X_{kt}}} \right],
\]
and the ML problem can be represented in the following way

\[
\max_{\beta, \delta} \mathcal{L}(\beta, \delta) \quad (A21)
\]

s.t.: \(\delta_{jk} + \delta_{kj} = 1, \quad \forall \ j, k, \ j \leq k,\)

\[\delta_{jk} + \delta_{kp} + \delta_{pj} \leq 2, \quad \forall \ j, k, p \ j \neq k \neq p,\]

\[\delta_{jk} \in \{0,1\}, \quad \forall \ j, k,\]

where the constraints ensure that \(\delta\) indeed represents a total order. In particular, the first set of constraints ensures that either \(a_j\) is preferred over \(a_k\) or vice versa, and the second set of constraints imposes the total ordering among any three products. To simplify the likelihood function, we introduce a new variable \(\tau\) defined as \(\tau_{ikj} = \delta_{kj} \beta_i, \ \forall \ i, j, k\) and rewrite the likelihood function in the following way

\[
\mathcal{L}(\beta, \delta, \tau) = \sum_{t=1}^{T} \left[ \log \frac{e^{\beta x_{jt}}}{1 + e^{\beta x_{jt}}} + \sum_{a_k \in S_t; k \neq j_t} (\delta_{kjt} - 1) \log \left( \frac{1}{2} \right) + \sum_{a_k \in S_t; k \neq j_t} \log \frac{1}{1 + e^{\sum_i \tau_{ikjt} x_{ikt}}} \right],
\]

since if \(\delta_{kjt} = 1\) we have that \(\tau_{ikjt} = \beta_i, \ \forall \ i,\) and

\[
\delta_{kjt} \log \frac{1}{1 + e^{\beta x_{kt}}} = \log \frac{1}{1 + e^{\beta x_{kt}}} = \log \frac{1}{1 + e^{\sum_i \tau_{ikjt} x_{ikt}}} = \log (\frac{1}{2}) + \log \frac{1}{1 + e^{\sum_i \tau_{ikjt} x_{ikt}}},
\]

if \(\delta_{kjt} = 0\) we have that \(\tau_{ikjt} = 0, \ \forall \ i,\) and

\[
\delta_{kjt} \log \frac{1}{1 + e^{\beta x_{kt}}} = 0 = - \log \left( \frac{1}{2} \right) + \log \frac{1}{1 + e^{0}} = \log (\frac{1}{2}) + \log \frac{1}{1 + e^{\sum_i \tau_{ikjt} x_{ikt}}},
\]

Let \(M\) be the value of the largest component in vector \(\beta, \ i.e., \ M = \max_i \beta_i.\) And we also define \(\mathcal{L}(\beta, \tau)\) in the following way

\[
\mathcal{L}(\beta, \tau) = \sum_{t=1}^{T} \left[ \log \frac{e^{\beta x_{jt}}}{1 + e^{\beta x_{jt}}} + \sum_{a_k \in S_t; k \neq j_t} \log \frac{1}{1 + e^{\sum_i \tau_{ikjt} x_{ikt}}} \right].
\]

We can then formulate the MLE problem in terms of the variables \((\delta, \beta, \tau)\):

\[
\max_{\beta, \tau, \delta} \mathcal{L}(\beta, \tau) + \sum_{a_k \in S_t; k \neq j_t} (\delta_{kjt} - 1) \log \left( \frac{1}{2} \right) \quad (A22)
\]

s.t.: \(\tau_{ikj} \leq \beta_i, \ \forall \ i, j, k,\)

\(\tau_{ikj} \leq M \delta_{kj}, \ \forall \ i, j, k,\)

\(\tau_{ikj} \geq \beta_i + M \delta_{kj} - M, \ \forall \ i, j, k,\)

\(\tau_{kj} \geq -M \delta_{kj}, \ \forall \ j, k,\)

\(\delta_{jk} + \delta_{kj} = 1, \ \forall \ j, k, j \leq k,\)

\(\delta_{jk} + \delta_{kp} + \delta_{pj} \leq 2, \ \forall \ j, k, p \ j \neq k \neq p,\)

\(\delta_{jk} \in \{0,1\}, \ \forall \ j, k,\)
where the first four sets of linear constraints ensure that $\tau_{ikj} = \delta_{kj} \beta_i, \forall i, j, k$, given that $\delta$ is a binary variable.

A3.2. Estimation Methodology: Outer-approximation algorithm

The high level idea behind outer-approximation algorithm is to approximate the convex constraint in MINLP by the set of linear constraints. This way we can solve the MINLP by solving the sequence of MILPs. In particular, we start with a feasible solution of MINLP. Then, we linearize the convex constraint at the previously obtained feasible solution. The next step is to solve MILP with a linearized convex constraint and obtain an additional point to linearize the convex constraint and continue iteratively. Note that, at each iteration, we add only one additional constraint to the optimization problem, solved at previous iteration. More formally, suppose that the problem is to solve the MINLP (P) defined below.

$$\begin{align*}
\max_{\theta, \tau, \delta, \mu, \mu_1} & \quad \mu \\
\text{s.t.:} & \quad \mu_1 \leq \mathcal{L}(\theta, \tau), \\
& \quad A\theta + B\tau + C\delta \leq 0, \\
& \quad \mu = \mu_1 + E\delta, \\
& \quad \mu_L \leq \mu \leq \mu_U, \\
& \quad \delta_{jk} \in \{0, 1\}, \quad \forall j, k,
\end{align*}$$

where $\mathcal{L}(\theta, \tau)$ is a concave function; $\theta$, $\tau$, $\mu$, and $\mu_1$ are continuous decision variables; $\delta$ is a binary variable; $A$, $B$, $C$, and $E$ are constant vectors; and $\mu_L$ and $\mu_U$ are lower and upper bounds of $\mu$, respectively. Note that optimization problems (14) and (A22) have a similar structure and can be represented as the above mentioned MINLP (P) without loss of generality.

Before we provide the details of the outer-approximation algorithm, let us define the linear constraint $D(\theta^i, \tau^i)$ that is added to the optimization problem at the $i$th iteration for given $(\theta^i, \tau^i)$, i.e., we linearize the convex constraint at $(\theta^i, \tau^i)$:

$$D(\theta^i, \tau^i) = \{\theta, \tau : \mathcal{L}(\theta^i, \tau^i) + \frac{\partial \mathcal{L}(\theta^i, \tau^i)}{\partial \theta} \theta + \frac{\partial \mathcal{L}(\theta^i, \tau^i)}{\partial \tau} \tau = \mu_1, \quad \mu \in \mathbb{R}^1\}. $$

The broad idea of the outer-approximation algorithm is that at the $i$th iteration, we substitute convex constraint $\mu_1 \leq \mathcal{L}(\theta, \tau)$ with a set of linear constraints $\Omega^i = \{D(\theta^k, \tau^k)\}_{k=0}^i$. 

Next we define two subproblems $S(\delta^i)$ and $M^i$ of the optimization problem (P) that are used to describe the Algorithm 1, which exploits the outer-approximation technique. First, let us define the concave subproblem $S(\delta^i)$ for given $\delta^i$ (i.e., if $\delta^i$ is known) in the following way

$$\begin{align*}
\max_{\theta, \tau, \delta, \mu, \mu_1} & \quad \mu \\
\text{s.t.:} & \quad \mu_1 \leq \mathcal{L}(\theta, \tau), \\
& \quad A\theta + B\tau + C\delta \leq 0, \\
& \quad \mu = \mu_1 + E\delta, \\
& \quad \mu_L \leq \mu \leq \mu_U, \\
& \quad \delta_{jk} \in \{0, 1\}, \quad \forall j, k
\end{align*}$$

Electronic copy available at: https://ssrn.com/abstract=3410019
\[ s.t.: \mu_1 \leq \mathcal{L}(\theta, \tau), \]
\[ A\theta + B\tau + C\delta^i \leq 0, \]
\[ \mu = \mu_1 + E\delta^i. \]

Note that solving MINLP (P) reduces to solving concave subproblem above if \( \delta \) is known. Second, let us define the MILP subproblem \( M^i \) for given \( \Omega^i, \mu^i_L, \) and \( \mu^i_U \) as follows

\[
\begin{align*}
\max_{\theta, \tau, \delta, \mu^i} & \quad \mu \\
\text{s.t.:} & \quad (\theta, \tau, \mu_1) \in \Omega^i, \\
& \quad A\theta + B\tau + C\delta \leq 0, \\
& \quad \mu = \mu_1 + E\delta, \\
& \quad \mu_L \leq \mu \leq \mu_U, \\
& \quad \delta_{jk} \in \{0, 1\}, \quad \forall \, j, k.
\end{align*}
\]

Note that solving MINLP (P) reduces to solving MILP subproblem if we approximate convex constraint \( \mu_1 \leq \mathcal{L}(\theta, \tau) \) with a set of linear constraints \( \Omega^i \).

Now we can formally apply the outer-approximation method (Duran and Grossmann, 1986) to solve the optimization problem (P), see Algorithm 1. The proposed algorithm effectively exploits the structure of the optimization problem (P) where we have a linearity of the binary variables and convexity of the non-linear constraint which only depends on continuous variables. In order to linearize the optimization problem, we use the outer-approximation of a convex set by intersection of its collection of supporting half-spaces. To this end, the outer approximation defines the optimization problem \( (M^i) \) as MILP. Because of the potentially many continuous points required for outer-approximation, we solve a sequence of MILPs to build up increasingly tight relaxation of the original MINLP. Overall, the proposed Algorithm 1 consists of solving a finite sequence of convex problems \( (S(\delta^i)) \) and relaxed versions of a MILP \( (M^i) \).

**A3.2.1. Outer-approximation algorithm vs. cutting plane algorithm** Note that Algorithm 1 to solve optimization problem (P) requires the solution of both convex optimization problem \( (S(\delta^i)) \) and MILP \( (M^i) \). The solution of the convex optimization problem \( (S(\delta^i)) \) in each iteration might be super computationally intensive, while in solving the MILP \( (M^i) \) the computational work, on the other hand, might be more moderate, because for every iteration \( i \) we need to solve the MILP problem \( (M^i) \) which is the previous MILP problem \( (M^{i-1}) \) with only one additional linear constraint added. Therefore, we propose to use the cutting plane algorithm to solve the
Algorithm 1 Outer-Approximation algorithm for optimization problem (P)

1: procedure OUTER-APPROXIMATION\( (P) \)
2: \( \Omega^0 = \mathbb{R}^n \times \mathbb{R}^m, \mu_L = -\infty, \mu_U = \infty, i = 1 \)
3: Select arbitrary \( \delta^1 \), i.e., it can be arbitrary full ranking
4: while \( |\mu_U - \mu_L| > \varepsilon \) do
5:   Solve concave subproblem \( S(\delta^i) \) such that \( \mu_L = \mu^* \) (i.e., the optimal objective function of \( S(\delta^i) \)), and \( (\theta^i, \tau^i) = (\theta^*, \tau^*) \) (i.e., the optimal solution of \( S(\delta^i) \))
6:   Set \( \Omega^i = \Omega^{i-1} \cap D(\theta^i, \tau^i) \)
7: Solve MILP subproblem \( M^i \) such that \( \mu_U = \mu^* \) (i.e., the optimal objective function of \( M^i \)), and \( (\theta^i, \tau^i, \delta^i) = (\theta^*, \tau^*, \delta^*) \) (i.e., the optimal solution of \( M^i \))
8: \( i = i + 1 \)
9: return \( (\theta^i, \tau^i, \delta^i) \).

MINLP in this case (Westerlund and Pettersson, 1995), which would require the solution of only the finite sequence of MILP problem \( (M^i) \), see Algorithm 2. Note that Algorithm 2 is identical to the Algorithm 1 except that we skip the Step 5 in the cutting plane Algorithm 2. Even though the main iteration loop of Algorithm 1 is, generally, more efficient, we have global convergence for both Algorithms 1 and 2.

Algorithm 2 Cutting plane algorithm for optimization problem (P)

1: procedure CUTTING PLANE\( (P) \)
2: \( \Omega^0 = \mathbb{R}^n \times \mathbb{R}^m, \mu_L = -\infty, \mu_U = \infty, i = 1 \)
3: Select arbitrary \( \delta^1 \), i.e., it can be arbitrary full ranking
4: Select arbitrary \( \lambda^1 \), i.e., it can be arbitrary distribution over consideration sets
5: Set \( \tau^1 = \delta^1 \cdot \lambda^1, \mu_U^0 = -\infty, \mu_U^1 = \infty \)
6: while \( |\mu_U^i - \mu_U^{i-1}| > \varepsilon \) do
7:   Set \( \Omega^i = \Omega^{i-1} \cap D(\theta^i, \tau^i) \)
8: Solve MILP subproblem \( M^i \) such that \( \mu_U = \mu^* \), \( \mu_U^i = \mu^* \) (i.e., the optimal objective function of \( M^i \)), and \( (\theta^i, \tau^i, \delta^i) = (\theta^*, \tau^*, \delta^*) \) (i.e., the optimal solution of \( M^i \))
9: \( i = i + 1 \)
10: return \( (\theta^i, \tau^i, \delta^i) \).

A3.2.2. Empirical validation of the algorithms. In this section, we analyze the performance of outer-approximation algorithm (1) and cutting plane algorithm (2) to estimate ICS model with IRI Academic dataset (see data descriptive statistics in Table A1). We limited the running
Figure A4  Results of applying the outer-approximation algorithm to estimate ICS model with IRI Academic dataset. We present the number of iterations completed in the algorithm, the upper and lower bounds of the objective functions, and the average computational time in minutes for each iteration at the left, middle, and right panels respectively. Note that in the middle column the absolute value of the optimality gap is equal to the distance between cross and dot.

Figure A5  Results of applying the cutting plane algorithm to estimate ICS model with IRI Academic dataset. We present the number of iterations completed in the algorithm, the upper and lower bounds of the objective functions, and the average computational time in minutes for each iteration at the left, middle, and right panels respectively. Note that in the middle column the absolute value of the optimality gap is equal to the distance between cross and dot.

time of the algorithms by 3 hours, and the precision was set to 1e-6. It follows from Figure A4 that the optimality gap of the outer approximation algorithm (1) to calibrate the ICS model is 3.3% on average over 20 product categories. On the other hand, it is shown in Figure A5 that the optimality gap of the cutting plane algorithm (2) to calibrate the ICS model is 4.5% on average over 20 product categories. Following these findings, we apply outer-approximation algorithm (1) to calibrate ICS model in our analysis as it provides significantly faster convergence to the optimal solution, which is consistent with previous studies.
A3.3. GCS Estimation Methodology: EM Algorithm

In this section we present the EM algorithm to calibrate the GCS model. We provide two versions of this algorithm that can be applied with the aggregate-level and individual-level sales transaction data.

A3.3.1. Estimation with aggregate level data. The log-likelihood function to calibrate the GCS model, after we reparametrize it by dividing all the transactions into $K$ segments, is given by

$$
\log \mathcal{L}(\theta; \gamma, \sigma) = \sum_{t=1}^{T} \log \left( \sum_{h=1}^{K} \gamma_h \prod_{a_j \in S_t, a_j \not\in \sigma} (1 - \theta_{h,j}) \right),
$$

(A23)

where $\gamma_h \geq 0$ is the weight of the class $h$ a priory, s.t. $\sum_{h=1}^{K} \gamma_h = 1$; $S_t$ denotes the set of offered items at time $t$; $a_{jt}$ denotes the product purchased at time $t$; $T$ denotes the time horizon.

Non-surprisingly, the above likelihood function is nonconcave. In order to alleviate the complexity of solving the MLE problem directly, we use the Expectation Maximization (EM) algorithm. First, let us outline the main principles of EM procedure. We start with arbitrary initial parameter estimates $\hat{x}^{(0)}$. Then, we compute the conditional expected value of the log-likelihood function
E[log L(x)|x^{(0)}] (the “E”, expectation, step). Next, the resulting expected log-likelihood function is maximized to compute new estimates \( \hat{x}^{(1)} \) (the “M”, maximization, step), and we repeat the algorithm until convergence to get a sequence of estimates \( \{\hat{x}^{(q)}\}, q = 1, 2, \ldots \). We further describe the E-step and M-step of every iteration and how we start the algorithm in the context of our estimation problem.

**Initialization:** we initialize the EM with a random allocation of observations to one of the \( K \) classes, resulting in an initial allocation \( D_1, D_2, \ldots, D_K \), which form a partition of the collection of all the transactions. Then we set \( \gamma_h^{(0)} = |D_h|/(\sum_{d=1}^K |D_d|) \). Then \( \sigma^{(0)} \) (i.e., \( \gamma^{(0)} \)) and \( \theta_h^{(0)} \) for all \( h \in \{1, \ldots, K\}, a_j \in N^+ \) are obtained by solving the following optimization problem:

\[
\max_{\sigma, \theta_h} \sum_{t \in D_h} \left( \log \theta_{h,j,t} + \sum_{a_j \in S_t: a_j \succ_{\sigma} a_{jt}} \log(1 - \theta_{h,j}) \right),
\]

which is solved by using the outer approximation algorithm in Section A3.2.

**E-step:** we compute \( P_h^{(q)} \), which is the membership probability of every transaction at time \( t \) to belong to the segment \( h \) (i.e., \( t \in \Gamma_h \), where \( \Gamma_h \) is the set of transactions in class \( h \)) based on the parameter estimates \( \{\sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)}\} \) and the purchasing transactions data \((a_{jt}, S_t)_{t=1}^T\):

\[
P_h^{(q)} = \text{Pr}\left( t \in \Gamma_h \big| \sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)}, (a_{jt}, S_t)_{t=1}^T \right)
\]

\[
= \text{Pr}\left( t \in \Gamma_h \big| \sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)}, (a_{jt}, S_t) \right) \text{ [independence of purchases]}
\]

\[
= \frac{\text{Pr}\left( (a_{jt}, S_t) \big| \sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)} \right) \cdot \text{Pr}\left( t \in \Gamma_h \big| \sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)} \right)}{\sum_{r=1}^K \text{Pr}\left( (a_{jr}, S_t) \big| \sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)}, t \in \Gamma_r \right) \cdot \text{Pr}\left( t \in \Gamma_r \big| \sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)} \right)} \text{ [Bayes’ theorem]}
\]

\[
= \frac{\gamma_h^{(q-1)} \left[ \theta_h^{(q-1)} \prod_{a_j \in S_t: a_j \succ_{\gamma^{(q-1)} a_{jt}}} (1 - \theta_h^{(q-1)}) \right]}{\sum_{r=1}^K \gamma_r^{(q-1)} \left[ \theta_r^{(q-1)} \prod_{a_j \in S_t: a_j \succ_{\gamma^{(q-1)} a_{jt}}} (1 - \theta_r^{(q-1)}) \right]}.
\]

As a result, conditional expected value of the log-likelihood function is given by

\[
\sum_{h=1}^K \sum_{t=1}^T P_h^{(q)} \log \left( \theta_{h,j,t} \prod_{a_j \in S_t: a_j \succ_{\sigma} a_{jt}} (1 - \theta_h^{(q-1)}) \right).
\]
M-step: first, we update class membership probabilities for every segment $h \in \{1, 2, ..., K\}$:

$$
\gamma^{(q)}_h = \frac{\sum_{t=1}^{T} P^{(q)}_{ht}}{T},
$$

and then optimize the conditional expected value of the log-likelihood function, obtained in the previous step, in terms of $\theta$ and $\sigma$:

$$
\max_{\sigma, \theta} \sum_{h=1}^{K} \sum_{t=1}^{T} P^{(q)}_{ht} \log \left( \theta_{h,jt} \prod_{a_j \in S_t: a_j \succ_{\sigma} a_{jt}} (1 - \theta_{hj}) \right),
$$

which is solved using outer-approximation algorithm in Section A3.2.

A3.3.2. Estimation with panel data. In the EM algorithm above we assumed access to the aggregate level sales transaction data (i.e., sales transaction data without access to the customer tags). The EM algorithm is updated in the following way if we have access to the individual-level sales transaction data with $m$ customers:

Initialization: we initialize the EM with a random allocation of individuals to one of the $K$ classes, resulting in an initial allocation $D_1, D_2, \ldots, D_K$, which form a partition of the collection of all the individuals. Then we set $\gamma^{(0)}_h = |D_h|/(\sum_{d=1}^{K} |D_d|)$. Then $\sigma$ (i.e., $\succ_{\sigma}$) and $\theta^{(0)}_{hj}$ for all $h \in \{1, ..., K\}, a_j \in N^+$ are obtained by solving the following optimization problem:

$$
\max_{\sigma, \theta} \sum_{i \in D_h} \left( \log \theta_{h,jit} + \sum_{a_j \in S_{it}: a_j \succ_{\sigma} a_{jt}} \log(1 - \theta_{hj}) \right),
$$

which is solved by using the outer approximation algorithm in Section A3.2.

E-step: we compute $P^{(q)}_{hi}$, which is the membership probability of every individual $i$ to belong to the segment $h$ based on the parameter estimates $\{\sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)}\}$ and the purchasing transactions data $(a_{jit}, S_{it})|_{t=1}^{T_i}$:

$$
P^{(q)}_{hi} = \frac{\gamma^{(q-1)}_h \prod_{t=1}^{T_i} \left( \theta^{(q-1)}_{h,jit} \prod_{a_j \in S_{it}: a_j \succ_{\sigma} a_{jt}} (1 - \theta^{(q-1)}_{hj}) \right)}{\sum_{r=1}^{K} \gamma^{(q-1)}_r \prod_{t=1}^{T_i} \left( \theta^{(q-1)}_{r,jit} \prod_{a_j \in S_{it}: a_j \succ_{\sigma} a_{jt}} (1 - \theta^{(q-1)}_{rj}) \right)}.
$$

M-step: first, we update class membership probabilities for every segment $h \in \{1, 2, ..., K\}$:

$$
\gamma^{(q)}_h = \frac{\sum_{i=1}^{m} P^{(q)}_{hi}}{m},
$$
and then optimize the conditional expected value of the log-likelihood function, obtained in the previous step, in terms of $\theta$ and $\sigma$:

$$
\max_{\sigma, \theta} \sum_{i=1}^{m} \sum_{h=1}^{K} P_{hi}^{(q)} \sum_{t=1}^{T_i} \log \left( \theta_{h,j_{it}} \prod_{a_j \in S_{hj}} (1 - \theta_{hj}) \right),
$$

which is solved using outer-approximation algorithm in Section A3.2.

**A3.3.3. EM algorithm heuristics.** Note that the proposed EM algorithm might become computationally challenging for the large-scaled problems as we need to run an outer-approximation algorithm for every $q$th iteration. Alternatively, we might further assume that the preference order $\sigma$ (i.e., $\succ_{\sigma}$) over items in the product universe is known, e.g., we can rank the products according to their popularity in the sales transaction data or we can estimate the ranking from calibrating single class ICS model (see Section 4.1). In this case the "M" step for $q$th iteration in the EM algorithm reduces to solving a globally concave maximization problem with a unique, closed form solution (i.e., we don’t need to apply outer-approximation algorithm) given by:

$$
\theta_{hj}^{(q)} = \frac{\sum_{t=1}^{T_i} P_{hi}^{(q)} [a_{j_{it}} = a_j]}{\sum_{t=1}^{T_i} P_{hi}^{(q)} [a_{j_{it}} = a_j] + \sum_{t=1}^{T_i} P_{hi}^{(q)} [a_j \in S_t, a_j \succ_{\sigma} a_{j_{it}}]},
$$

which can be applied with aggregate level data (see Section A3.3.1), and

$$
\theta_{hj}^{(q)} = \frac{\sum_{i=1}^{m} \sum_{t=1}^{T_i} P_{hi}^{(q)} [a_{j_{it}} = a_j]}{\sum_{i=1}^{m} \sum_{t=1}^{T_i} P_{hi}^{(q)} [a_{j_{it}} = a_j] + \sum_{i=1}^{m} \sum_{t=1}^{T_i} P_{hi}^{(q)} [a_j \in S_t, a_j \succ_{\sigma} a_{j_{it}}]},
$$

which can be applied with panel data (see Section A3.3.2).

**A3.4. CTC Model: Estimation Methodology**

The CTC (i.e., general consideration - then - general choice) is the broadest class of consider-then-choose type of models where customers have heterogeneous preferences and consideration sets, i.e., before making a choice customers sample their preference order $\sigma$ over the items in the product universe and the subset of items $C$ to consider from the general distributions over product rankings and consideration sets respectively.

**A3.4.1. Estimation with aggregate level data.** Similarly to the Section A3.3, we calibrate CTC model by dividing transactions into $K$ segments such that customers in segment $h$ sample their consideration sets based on the attention parameters $\theta_h$ and have their preferences characterized by the ranking $\sigma_h$. Then the log-likelihood function can be represented in the following way

$$
\log \mathcal{L}(\theta, \gamma, \sigma) = \sum_{t=1}^{T} \sum_{h=1}^{K} \gamma_h \theta_{h,j_{ht}} \prod_{a_j \in S_{ht} \mid a_j \succ_h a_{j_{ht}}} (1 - \theta_{hj}),
$$

(A24)
where \( \gamma_h \geq 0 \) is the weight of the class \( h \), s.t. \( \sum_{h=1}^{K} \gamma_h = 1 \); \( S_t \) denotes the set of offered items at time \( t \); \( a_{jt} \) denotes the product purchased at time \( t \); \( T \) denotes the time horizon. Conceptually, we can obtain all the parameters of the CTC model (i.e., distributions over the preference lists and consideration sets) by maximizing the log-likelihood function above for a sufficiently large \( K \).

Next we provide the initialization of the EM algorithm to calibrate CTC model followed by the "E" and "M" steps of every iteration.

**Initialization:** we randomly allocate sales transaction to one of the \( K \) classes, resulting in an initial allocation \( D_1, D_2, \ldots, D_K \), which form a partition of the collection of all the transactions. Consequently, we set \( \gamma_h(0) = |D_h|/(\sum_{d=1}^{K} |D_d|) \). Then \( \sigma_h \) (i.e., \( \succ_h \)) and \( \theta_{hj}(0) \) for all \( h \in \{1, \ldots, K\} \), \( a_j \in N^+ \) are obtained by solving the following optimization problem:

\[
\max_{\sigma_h, \theta_h} \sum_{t=1}^{T} \left( \log \theta_{h,j_t} + \sum_{a_j \in S_t: a_j \succ_h a_{jt}} \log(1 - \theta_{hj}) \right),
\]

which is solved by using the outer approximation algorithm for the ICS model in Section A3.2.

**E-step:** we compute \( P_{ht}^{(q)} \), which is the membership probability of every transaction at time \( t \) to belong to the segment \( h \) based on the parameter estimates \( \{\sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)}\} \) and the purchasing transactions data \( (a_{jt}, S_t)^T \):

\[
P_{ht}^{(q)} = \frac{\gamma_h^{(q-1)} \left( \prod_{a_j \in S_t: a_j \succ_h a_{jt}} (1 - \theta_{hj}^{(q-1)}) \right)}{\sum_{r=1}^{K} \gamma_r^{(q-1)} \left( \prod_{a_j \in S_t: a_j \succ_h a_{jt}} (1 - \theta_{rj}^{(q-1)}) \right)}.
\]

**M-step:** first, we update class membership probabilities for every segment \( h \in \{1, 2, \ldots, K\} \):

\[
\gamma_h^{(q)} = \frac{\sum_{t=1}^{T} P_{ht}^{(q)}}{T},
\]

and then optimize the conditional expected value of the log-likelihood function, obtained in the previous step, in terms of \( \theta_h \) and \( \sigma_h \) for all \( h \in \{1, \ldots, K\} \):

\[
\max_{\sigma_h, \theta_h} \sum_{t=1}^{T} P_{ht}^{(q)} \log \left( \theta_{h,j_t} \prod_{a_j \in S_t: a_j \succ_h a_{jt}} (1 - \theta_{hj}) \right),
\]

which is solved by using the outer approximation algorithm for the ICS model in Section A3.2.

Note that the proposed EM algorithm we need to apply the outer-approximation algorithm for every iteration. In order to reduce the computation time for the large-scaled problems we might solve the optimization problem at "M"-step by ranking the products according to their popularity.
for each segment $h$. This way we can obtain the preference order $\sigma^{(q)}_h$ for every segment $h$ for $q$-th iteration. In this case the "M" stem in the EM algorithm reduces to solving a globally concave maximization problem with a unique, closed form solution given by:

$$\theta^{(q)}_{hj} = \frac{\sum_{t=1}^{T} P^{(q)}_{ht}[a_{jt} = a_j]}{\sum_{t=1}^{T} P^{(q)}_{ht}[a_{jt} = a_j] + \sum_{t=1}^{T} P^{(q)}_{ht}[a_j \in S_t, a_j > h^{(q)}_j a_{jt}]}.$$  

A3.4.2. Estimation with panel data. We update the EM algorithm above in the following way:

**Initialization:** we randomly allocate individuals to one of the $K$ classes, resulting in an initial allocation $D_1, D_2, \ldots, D_K$. Consequently, we set $\gamma^{(0)}_h = |D_h|/(\sum_{d=1}^{K} |D_d|)$. Then $\sigma_h$ (i.e., $\succ_h$) and $\theta^{(q)}_{hj}$ for all $h \in \{1, \ldots, K\}$, $a_j \in N^+$ are obtained by solving the following optimization problem:

$$\max_{\sigma_h, \theta_h} \sum_{i \in D_h} \left( \log \theta_{h,jit} + \sum_{a_j \in S_{it}: a_j \succ_h a_{jit}} \log(1 - \theta_{h,j}) \right),$$

which is solved by using the outer approximation algorithm for the ICS model in Section A3.2.

**E-step:** we compute $P^{(q)}_{hi}$, which is the membership probability of every individual $i$ to belong to the segment $h$ based on the parameter estimates $\{\sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)}\}$ and the purchasing transactions data $(a_{jt}, S_t)_{t=1}^{T}$:

$$P^{(q)}_{hi} = \frac{\gamma^{(q-1)}_h \prod_{t=1}^{T_i} \left[ \theta^{(q-1)}_{h,jit} \prod_{a_j \in S_{it}: a_j \succ_h a_{jit}} (1 - \theta^{(q-1)}_{h,j}) \right]}{\sum_{r=1}^{K} \prod_{t=1}^{T_i} \left[ \gamma^{(q-1)}_r \left( \theta^{(q-1)}_{r,jit} \prod_{a_j \in S_{it}: a_j \succ_h a_{jit}} (1 - \theta^{(q-1)}_{r,j}) \right) \right]}.$$  

**M-step:** first, we update class membership probabilities for every segment $h \in \{1, 2, \ldots, K\}$:

$$\gamma^{(q)}_h = \frac{\sum_{i=1}^{m} P^{(q)}_{hi}}{m},$$

and then optimize the conditional expected value of the log-likelihood function, obtained in the previous step, in terms of $\theta_h$ and $\sigma_h$ for all $h \in \{1, \ldots, K\}$:

$$\max_{\sigma_h, \theta_h} \sum_{i=1}^{m} P^{(q)}_{hi} \sum_{t=1}^{T_i} \log \left( \theta_{h,jit} \prod_{a_j \in S_{it}: a_j \succ_h a_{jit}} (1 - \theta_{h,j}) \right),$$

which is solved by using the outer approximation algorithm for the ICS model in Section A3.2.
Figure A6  Scatter plots of the prediction score improvements of GCS model over CTC across 20 product categories. Improvements are defined as the difference between two scores. We focus on the MAPE and RMSE scores in left and right panels respectively. We illustrate three scenarios in each panel: (1) short-term forecasts represented by pluses, (2) long-term forecasts represented by crosses, and (3) warehouse forecasts represented by dots.

Figure A7  The average prediction scores over 20 product categories under GCS and CTC choice models represented by dashed bars and solid bars respectively. We focus on MAPE and RMSE scores in left and right panels respectively. The lower the score the better.
A3.5. Comparison between CTC and GCS models

In this section, we compare the prediction performance of GCS with CTC based on IRI dataset. It follows from Figures A6 and A7 that GCS provides higher prediction accuracy than CTC based on the RMSE metric whereas CTC outperforms GCS based on the MAPE metric. As a result, we cannot claim dominance of GCS or CTC.

A4. Robustness check

A4.1. Study based on synthetic data: robustness to the ground truth model and benchmarks

In this section, we summarize the results of the extensive synthetic experiments conducted in order to check the robustness of the simulation results reported in Section 5, when we use the rank-based model as ground truth instead of the MNL. Recall that in Section 5, we compared the predictions of the MNL model against ICS model to understand the conditions under which the MNL benchmark outperforms the consider-then-choose model in the presence of noise in the offer sets.

The set up of the experiments in this section is identical to the one in Section 5. In the new set of experiments we simulate sales transaction data according to the rank-based model with fifteen customer types where type $i$ customers are making purchases according to the ranking $\sigma_i$, i.e., when faced with a given choice set customers are assumed to purchase the available option that ranks highest in their preference list. To this end, we randomly sample the set of fifteen rankings, each corresponding to a particular class, and also assume the equal probability of each class. In order to calibrate this rank-based model we exploit the EM algorithm proposed by van Ryzin and Vulcano (2017). This algorithm relies on the assumption that the set of rankings is known. Therefore, we initialize this algorithm with thirty preference orders such that fifteen of them are the rankings from the ground truth and the remaining fifteen rankings are sampled randomly. Moreover, we start the EM algorithm with equal probability of sampling each preference list.

Similarly to Section 5, in Figures A8 and A9 we present the heatmaps of the prediction scores under the rank-based model, and the prediction scores improvements of the ICS model versus the rank-based model, respectively. And in Tables A2 and A3, we report the results for the regression (20), where the dependent variables are the prediction scores and the improvement scores, respectively. The main insights remain the same – the results of this extensive simulation study demonstrate that choice models based on the consider-then-choose framework are more robust to noise in offer sets than their classical counterparts, i.e., ICS model outperforms the ground truth rank-based model under noisier regimes.

With the objective of benchmarking the ICS model with a more competitive discrete choice model, we resort to the LC-MNL. As it was mentioned above, we estimate the LC-MNL model for
Figure A8  Heatmap of the prediction scores under rank-based model where each column corresponds to a particular noise intensity and each row corresponds to a particular noise exposure. We focus on the MAPE and RMSE scores in left and right panels respectively. The lower the score the better.

<table>
<thead>
<tr>
<th></th>
<th>Model (1)</th>
<th>Model (2)</th>
<th>Model (3)</th>
<th>Model (4)</th>
<th>Model (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intensity</td>
<td>38.341***</td>
<td>48.526***</td>
<td>32.639***</td>
<td>-9.259***</td>
<td>30.882***</td>
</tr>
<tr>
<td></td>
<td>(23.312)</td>
<td>(15.904)</td>
<td>(20.358)</td>
<td>(-3.425)</td>
<td>(5.470)</td>
</tr>
<tr>
<td>Intensity$^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asymm</td>
<td></td>
<td>42.516***</td>
<td>4.827*</td>
<td>13.450***</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(29.661)</td>
<td>(16.57)</td>
<td>(19.378)</td>
<td></td>
</tr>
<tr>
<td>Shared</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>const</td>
<td>7.072***</td>
<td>15.593***</td>
<td>17.891***</td>
<td>26.428***</td>
<td>-13.927***</td>
</tr>
</tbody>
</table>

No. Observations: 200 200 200 200 200
R-squared: 0.733 0.677 0.131 0.014 0.955
Adj. R-squared: 0.732 0.675 0.127 0.009 0.954

$t$ statistics in parentheses
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A2  Regression models where the dependent variable is the MAPE score under the rank-based model.

$K = 1, 2, ..., 5$, classes and report the best performance measure from these five variants. Figure A11 illustrates the heatmaps of the MAPE and RMSE prediction score improvements of the ICS model versus the LC-MNL. As expected, Figure A11 confirms that the LC-MNL is a more competitive benchmark. However, our qualitative results remain the same: the ICS model outperforms the LC-MNL model, which subsumes the ground truth MNL model, under sufficiently noisy regimes.
Figure A9  Heatmap of the prediction scores improvements under ICS model versus rank-based model where each column corresponds to a particular noise intensity and each row corresponds to a particular noise exposure. We focus on the MAPE and RMSE scores in left and right panels respectively.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intensity</strong></td>
<td>16.211***</td>
<td>13.646***</td>
<td>9.837***</td>
<td>12.304***</td>
<td>23.908***</td>
</tr>
<tr>
<td></td>
<td>(16.814)</td>
<td>(14.968)</td>
<td>(0.697)</td>
<td>(-20.567)</td>
<td>(13.640)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-4.506)</td>
<td>(8.231)</td>
<td>(-18.832)</td>
<td>(31.147)</td>
</tr>
<tr>
<td><strong>Asymm</strong></td>
<td>1.991</td>
<td>1.991</td>
<td>1.991</td>
<td>1.991</td>
<td>1.991</td>
</tr>
<tr>
<td></td>
<td>(0.697)</td>
<td>(0.697)</td>
<td>(0.697)</td>
<td>(0.697)</td>
<td>(0.697)</td>
</tr>
<tr>
<td><strong>Shared</strong></td>
<td>-12.304***</td>
<td>-8.641***</td>
<td>-4.050***</td>
<td>-6.917***</td>
<td>-22.532***</td>
</tr>
<tr>
<td></td>
<td>(-20.567)</td>
<td>(-18.832)</td>
<td>(-3.884)</td>
<td>(-12.178)</td>
<td>(-40.841)</td>
</tr>
<tr>
<td><strong>const</strong></td>
<td></td>
<td></td>
<td>-6.917***</td>
<td></td>
<td>-22.532***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-4.506)</td>
<td></td>
<td>(-31.147)</td>
</tr>
</tbody>
</table>

- No. Observations: 200
- R-squared: 0.588, 0.531, 0.002, 0.255, 0.933
- Adj. R-squared: 0.586, 0.528, -0.003, 0.251, 0.931

Table A3  Regression models where the dependent variable is the MAPE score improvement of the ICS model over the rank-based model.

A4.2. Study based on synthetic data: Explanation about the superior performance of the ICS model in noisy regimes.

In this section, we provide more details on how the one directional cannibalization property of the ICS model helps it to outperform the ground-truth MNL model in the simulation study in Section 5. To streamline our analysis, we focus on the specific case of the simulation study in Section 5 where $\gamma = 0.5$ and $\eta = 1$, which corresponds to the case of maximum asymmetry between
Figure A10 Heatmap of the prediction scores improvements under the ICS model versus MNL where each column corresponds to a particular noise intensity $\eta$ and each row corresponds to a particular noise exposure $\gamma$ in the test dataset. For every scenario noise exposure in the training dataset is 10% lower than in the test dataset. We focus on the MAPE and RMSE scores in the left and right panels, respectively.

Figure A11 Heatmap of the prediction scores improvements under ICS model versus LC-MNL where each column corresponds to a particular noise intensity $\eta$ and each row corresponds to a particular noise exposure $\gamma$. We focus on the MAPE and RMSE scores in the left and right panels, respectively.

the training and testing datasets (i.e., the probability that a product is only offered in either the training or the testing dataset is 0.5, and every product in the exposure set will be added to either the training or the testing datasets). As it was mentioned in Section 5, the one directional cannibalization property of the ICS model provides robustness to the noise in the offer sets because it alleviates its impact on the demand prediction for the higher ranked items (i.e., the presence or absence of lower ranked products in the offer set does not affect the demand prediction for higher
ranked products). We confirm this intuition by estimating the following regression specification:

\[ \text{ScoreImpr}_{jc} = \text{Instance}_c + \beta \cdot \text{Ranking}_{jc} + \epsilon_{jc}, \tag{A25} \]

where \( j \) corresponds to item \( a_j \in N \) and \( c \) corresponds to a specific instance, i.e., to each of the 100 generated instances. \( \text{ScoreImpr}_{jc} \) is the improvement of item \( a_j \) demand forecasting error in instance \( c \) obtained by the ICS model over MNL, i.e., \( \text{ScoreImpr}_{jc} = \text{Score}^{ICS}_{jc} - \text{Score}^{MNL}_{jc} \), where we rely on two ways to measure the item’s demand forecasting accuracy: (1) absolute percentage error (\( APE_j \) = \( \frac{100 \vert n_j - \hat{n}_j \vert}{10 + n_j} \)), where \( n_j \) denotes the observed sales for \( a_j \) in the test dataset and \( \hat{n}_j \) denotes our prediction and (2) demand prediction error (\( DPE_j \) = \( \frac{100}{\sum_{a_j \in N} n_j} \sum_{a_j \in N} (n_j - \hat{n}_j) \)). Note that the former and the latter prediction errors are factored into the computations of the MAPE and RMSE scores, respectively, and that they are expressed in percentage points, so that the score improvement is also expressed in percentage points. \( \text{Ranking}_{jc} \) is the position of item \( a_j \) in the ranking inferred from the ICS model calibration on instance \( c \). \( \text{Instance}_c \) is a categorical variable to control for the instance-level fixed effects. Table A4 presents the results from the regression specification (A25) where the improvement by the ICS model over MNL is measured using \( APE \) score in columns (1) and (3), while the improvement by the ICS model over MNL using the \( DPE \) score is presented in columns (2) and (4). In columns (1) and (2) we do not control for the instance-level fixed effects, while in columns (3) and (4) we do. It follows from Table A4 that the coefficient of the \( \text{ranking} \) variable is negative and statistically significant, with a magnitude that implies a relevant decrease in \( \text{ScoreImpr}_{jc} \) for the low ranked products, which confirms the intuition that the ICS model is especially competitive when predicting the market shares of the top ranked items because those items are immune to the presence or absence of lower ranked products in the offer sets. Note that this finding is consistent across all four columns in Table A4 (with and without instance level fixed effects, under both prediction scores). Note that the \( R^2 \) in all the columns in Table A4 is relatively small which is not unexpected given that we only have a single covariate in the regression models and there might be other factors (in addition to the \( \text{Ranking} \) variable) which could explain the variation in the outcome variable \( \text{ScoreImpr} \).

### A4.3. Robustness to the forecasting error metrics

In this section, we show the robustness of our prediction results in Section 6 to different specifications of the prediction metrics. To this end, we focus on the variations of RMSE and MAPE metrics when aggregating predictions either over a one week intervals (i.e., RMSE\(_T\) and MAPE\(_T\)
### Table A4

Regression models where the dependent variable is the APE (in Models (1) and (3)) or the DPE (in Models (2) and (4)) score improvement by the ICS model over the MNL model.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ranking</td>
<td>-0.408***</td>
<td>-0.024***</td>
<td>-0.408***</td>
<td>-0.024***</td>
</tr>
<tr>
<td>const</td>
<td>5.313***</td>
<td>0.317***</td>
<td>(-3.669)</td>
<td>(-3.536)</td>
</tr>
</tbody>
</table>

| Instance FE: | No | No | Yes | Yes |
| No. Observations: | 1500 | 1500 | 1500 | 1500 |
| R-squared: | 0.011 | 0.011 | 0.011 | 0.011 |

\* \( p < 0.1 \), \*\* \( p < 0.05 \), \*\*\* \( p < 0.01 \)

Equation A26:

\[
\text{MAPE}_T = 100 \frac{\sum_{t=1}^{T} \sum_{a_j \in S_t} |n_{jt} - \hat{n}_{jt}|}{\sum_{t=1}^{T} \sum_{a_j \in S_t} n_{jt}} \frac{10 + n_{jt}}{|S_t|},
\]

Equation A27:

\[
\text{RMSE}_T = 100 \sqrt{\frac{\sum_{t=1}^{T} \sum_{a_j \in S_t} (n_{jt} - \hat{n}_{jt})^2}{\sum_{t=1}^{T} \sum_{a_j \in S_t} (\sum_{a_j \in S_t} n_{jt})^2}},
\]

where \( S_t \) is the set of items that where offered to customers in week \( t \), \( T \) is the total number of weeks in the test dataset, \( n_{jt} \) is the observed number of times product \( a_j \) was purchased in week \( t \), and \( \hat{n}_{jt} \) is the predicted number of times product \( a_j \) to be purchased in week \( t \). In the same spirit we can compute RMSE and MAPE metrics when aggregating predictions over each offer set as follows

Equation A28:

\[
\text{MAPE}_{\tilde{S}} = 100 \frac{\sum_{m=1}^{M} \sum_{a_j \in \tilde{S}_m} |n_{jm} - \hat{n}_{jm}|}{\sum_{m=1}^{M} \sum_{a_j \in \tilde{S}_m} n_{jm}} \frac{10 + n_{jm}}{|\tilde{S}_m|},
\]

Equation A29:

\[
\text{RMSE}_{\tilde{S}} = 100 \sqrt{\frac{\sum_{m=1}^{M} \sum_{a_j \in \tilde{S}_m} (n_{jm} - \hat{n}_{jm})^2}{\sum_{m=1}^{M} \sum_{a_j \in \tilde{S}_m} (\sum_{a_j \in \tilde{S}_m} n_{jm})^2}},
\]

where \( \tilde{S}_m \) is the \( m \)'th offer set, \( M \) is the total number of different offer sets in the test dataset, \( n_{jm} \) is the observed number of times product \( a_j \) was purchased under the offer set \( m \), and \( \hat{n}_{jm} \) is the predicted number of times product \( a_j \) to be purchased under the offer set \( m \). Alternatively, we could aggregate predictions based on the ‘true’ offer sets in the following way

Equation A30:

\[
\text{MAPE}_{S} = 100 \frac{\sum_{m=1}^{M} \sum_{a_j \in N} |n_{jm} - \hat{n}_{jm}|}{\sum_{m=1}^{M} \sum_{a_j \in N} n_{jm}} \frac{10 + n_{jm}}{|S_m|},
\]
Figure A12  The relative improvement of GCS model over LC-MNL model in prediction performance based on the IRI dataset. RMSE and MAPE metrics are aggregated over a one week intervals and computed using equations (A26) and (A27).

\[ \text{RMSE}_S = 100 \sqrt{\frac{1}{\sum_{m=1}^{M'} |S_m| \sum_{m=1}^{M'} \sum_{a_j \in N} (n_{jm} - \hat{n}_{jm})^2}{\left( \sum_{a_j \in N} n_{jm} \right)^2}}, \quad (A31) \]

where \( S_m \) is the \( m \)'th ‘true’ offer set, \( M' \) is the total number of different ‘true’ offer sets in the test dataset.

In Figure A12, we exhibit improvements of GCS over LC-MNL under MAPE\(_T\) (left panel) and RMSE\(_T\) (right panel), averaging across 20 product categories, for the three different scenarios. From these panels we observe that the GCS significantly improves over LC-MNL once we shift from short to long-term and from long-term to warehouse forecasts based on the MAPE\(_T\) and RMSE\(_T\) scores. Figure A13 (resp. Figure A14) presents qualitatively the same results when we compute prediction metrics based on the MAPE\(_S\) (resp. MAPE\(_S\)) and RMSE\(_S\) (resp. RMSE\(_S\)) scores. As a result, we conclude that the direction of our results in Section 6 are robust to the definition of the prediction metrics.

A4.4. Case study on retailing: different noise generation process

In this section, we show the robustness of our prediction results in Section 6 to an alternative noise generation process. To this end, we add the noise only to the test dataset and we do it in a similar way to the simulation study presented in Section 5. Figure A15 illustrates the predictive performance of GCS and LC-MNL models when \( \eta \) is equal to 0.5 and \( \gamma \) takes three different values (0.1, 0.3, and 0.5). It follows from Figure A15 that our results are qualitatively the same under this alternative noise regime – the relative predictive performance of GCS over LC-MNL improves with the level of noise (measured by \( \gamma \)) in the offer sets.
Figure A13  Relative improvement of the GCS model over the LC-MNL model in prediction performance based on the IRI dataset. RMSE and MAPE metrics are aggregated over noisy offer sets and computed using equations (A28) and (A29).

Figure A14  Relative improvement of the GCS model over the LC-MNL model in prediction performance based on the IRI dataset. RMSE and MAPE metrics are aggregated over ‘true’ offer sets and computed using equations (A30) and (A31).

Figure A15  Average prediction scores over 20 product categories under GCS and LC-MNL choice models represented by dashed bars and solid bars respectively. The offer set noise is present only in the test dataset where we set $\eta$ to be 0.5 and $\gamma$ take three different values.
Figure A16  Average prediction scores over 20 product categories under GCS and LC-MNL choice models represented by dashed bars and solid bars respectively. The offer set noise is present both in the training and test datasets.

A4.5. Case study on retailing: both training and test datasets are exposed to the same offer set noise generation process

In this section, we show the robustness of our prediction results in Section 6 to the scenario when we use the same noise generation process to alter both the training and test datasets. Otherwise, the set up of the experiments is the same as in Section 6. Figure A16 illustrates that the findings are qualitatively the same under the updated set up – the relative predictive performance of GCS over LC-MNL improves with the level of noise in the offer sets.

A5. Case Study on the Car Sharing Dataset

In this section we calibrate Logistic-based ICS (L-ICS) and MNL models accounting for car features and discuss the modeling assumptions based on the car sharing dataset (see data descriptive statistics in Table A5). We also provide explanatory analysis of choice models in order to gain insights about consideration set formation of renters using the car feature information. In addition, we address the problem of a potential price endogeneity in our empirical explanatory analysis. We argue that in our setting we are unlikely to have any price endogeneity problems while calibrating the models. We conclude this section by presenting Figure A17 which illustrates an instance of the decision tree obtained after fitting the DT-CC model.

A5.1. Explanatory analysis

We start this section by calibrating the single class L-ICS model with features to examine the extent to which various variables impact the consideration set structure. Assuming that the cars are ranked according to their popularity among renters (see Section 7.2), the problem of fitting L-ICS model is the one of estimating the coefficients $\beta$. The car features available to the renters through the
online platform are divided into three groups: (1) car brand; (2) car location type and accessibility, including car access (i.e., open or closed), car location hours (i.e., 24 hours or restricted), car location type (i.e., garage, street, surface lot, or valet); and the third group including (3a) car type (i.e., economy, standard, fullsize, SUV, trucks, luxury), and (3b) car features: hourly price, car age, and some other various binary car features such as transmission, premium wheels, power seats, bluetooth/wireless, leather interior, sunroof/moonroof, premium sound, power windows, GPS navigation system, roof rack, tinted windows. Assuming that the error terms $\varepsilon_j$ are logistically distributed, we estimate the $\beta$ vector using logistic regression analysis.

The results for the L-ICS model appear in the first column of Table A6. In the middle column, the table lists the average marginal effects (AME) of the L-ICS model when all the covariates are at their mean. Then we also calibrate the usual linear-in-parameters MNL model where the utility from reserving the car alternative $j$ is represented with linear in parameters function $U_j$, i.e., $U_j = \beta^T x_j + \varepsilon_j$. On the right, the Table A6 presents the estimates of the MNL model parameters. However, the interpretation of the $\beta$ vector for L-ICS and MNL models is different. The parameters of the L-ICS model listed in the left column of Table A6 shows the estimated impact of exogenously imposed changes in car features on consideration set formation. Rather, the parameters of the MNL model in the right column of Table A6 show the influence of car features on the customer’s choices, i.e., revealed preferences. Notably, quite a few coefficients (17 out of 53) estimated based on L-ICS and MNL models are not aligned, i.e., the covariates that increase (or decrease) the likelihood of considering the car under L-ICS model might not necessarily increase (or decrease) the likelihood of booking the car under the MNL model, e.g., the utility of the renter from considering the car brand Jeep is higher by 0.98 ($t = 4.9, p < 0.01$) than the utility from considering the baseline brands while the utility of the renter from reserving the same brand under the MNL model is lower by 0.93 ($t = -6.2, p < 0.01$) than the utility from reserving baseline brands. Also some of the covariates (7 out of 53) that are statistically significant in explaining the choice of renters under MNL model might be statistically insignificant under L-ICS model, e.g., the utility from choosing a car parked in the street is lower ($t = -8.33, p < 0.01$) than the utility from choosing the car located in the valet parking area while the discrepancy between these two parking location types are insignificant ($t = 1.36, p > 0.10$) under the L-ICS model. The price and the car age coefficients are statistically significant and negative for both L-ICS and MNL models. However, the impact of additional $\$1$ increase in the car hourly rental price on the utility from considering the vehicle is equivalent to the car being 0.52 years older, while the impact of additional $\$1$ increase in the car hourly rental price on the utility from booking the vehicle under the MNL model is equivalent to the car being 3.75 years older. According to these findings, the car age plays relatively more
<table>
<thead>
<tr>
<th>Brands</th>
<th>Mean</th>
<th>Std%</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acura</td>
<td>2.52%</td>
<td>15.68%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Audi</td>
<td>4.54%</td>
<td>20.82%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>BMW</td>
<td>11.73%</td>
<td>32.18%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Buick</td>
<td>0.21%</td>
<td>4.61%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Chevrolet</td>
<td>0.79%</td>
<td>8.84%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Chrysler</td>
<td>0.41%</td>
<td>6.37%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Dodge</td>
<td>0.82%</td>
<td>9%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Fiat</td>
<td>0.9%</td>
<td>9.42%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Ford</td>
<td>2.63%</td>
<td>16.01%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Honda</td>
<td>16.77%</td>
<td>37.36%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Hyundai</td>
<td>3.42%</td>
<td>18.16%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Infiniti</td>
<td>0.21%</td>
<td>4.61%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Jeep</td>
<td>0.41%</td>
<td>6.39%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Kia</td>
<td>0.49%</td>
<td>6.95%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Land Rover</td>
<td>0.17%</td>
<td>4.14%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Lexus</td>
<td>1.24%</td>
<td>11.06%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Mazda</td>
<td>3.44%</td>
<td>18.22%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Mercedes Benz</td>
<td>3.73%</td>
<td>18.95%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Mercury</td>
<td>0.07%</td>
<td>2.59%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Mini</td>
<td>7.66%</td>
<td>26.59%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Mitsubishi</td>
<td>0.68%</td>
<td>8.21%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Nissan</td>
<td>4.05%</td>
<td>19.71%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Pontiac</td>
<td>0.21%</td>
<td>4.57%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Porsche</td>
<td>1.5%</td>
<td>12.14%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Saab</td>
<td>0.03%</td>
<td>1.73%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Saturn</td>
<td>0.28%</td>
<td>5.25%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Scion</td>
<td>0.62%</td>
<td>7.85%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Subaru</td>
<td>3.71%</td>
<td>18.89%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Suzuki</td>
<td>0.53%</td>
<td>7.29%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Smart</td>
<td>5.44%</td>
<td>22.69%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Tesla</td>
<td>1.42%</td>
<td>11.83%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Toyota</td>
<td>10.33%</td>
<td>30.43%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Volkswagen</td>
<td>7.54%</td>
<td>26.41%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Volvo</td>
<td>1.53%</td>
<td>12.28%</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Car types</th>
<th>Mean</th>
<th>Std%</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economy</td>
<td>14.83%</td>
<td>35.54%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Standard</td>
<td>48.83%</td>
<td>49.99%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Fullsize</td>
<td>19.56%</td>
<td>39.67%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>SUV</td>
<td>9.41%</td>
<td>29.2%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Trucks</td>
<td>3.31%</td>
<td>17.88%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Luxury</td>
<td>4.06%</td>
<td>19.74%</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Car location type and accessibility</th>
<th>Mean</th>
<th>Std%</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car access [open]</td>
<td>81.89%</td>
<td>38.51%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Car access hours [all hours]</td>
<td>93.52%</td>
<td>24.61%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Car location type [garage]</td>
<td>28.50%</td>
<td>45.14%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Car location type [street]</td>
<td>23.86%</td>
<td>42.62%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Car location type [surface lot]</td>
<td>43.85%</td>
<td>49.62%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Car location type [valet]</td>
<td>0.24%</td>
<td>4.84%</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Car features</th>
<th>Mean</th>
<th>Std%</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price (per hour)</td>
<td>8.63%</td>
<td>4.61%</td>
<td>2.0</td>
<td>300.0</td>
</tr>
<tr>
<td>Car age</td>
<td>5.32%</td>
<td>3.18%</td>
<td>-0.3</td>
<td>18.3</td>
</tr>
<tr>
<td>Transmission [automatic]</td>
<td>95.21%</td>
<td>21.35%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Premium wheels</td>
<td>29.38%</td>
<td>45.55%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Power seats</td>
<td>46.88%</td>
<td>49.90%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Bluetooth/wireless</td>
<td>33.74%</td>
<td>47.28%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Leather interior</td>
<td>53.56%</td>
<td>48.97%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Sunroof/moonroof</td>
<td>53.48%</td>
<td>49.88%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Premium sound</td>
<td>46.25%</td>
<td>49.86%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Power windows</td>
<td>92.90%</td>
<td>25.68%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>GPS navigation system</td>
<td>23.05%</td>
<td>42.11%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Roof rack</td>
<td>6.98%</td>
<td>25.48%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Tinted windows</td>
<td>13.24%</td>
<td>33.89%</td>
<td>0%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Table A5** Descriptive statistics, the car sharing dataset.
Table A6  Logistics-based Consider-then-Choose (L-ICS) and MNL model estimation results, the car sharing dataset. The baseline brands group consists of Buick, Chevrolet, Saab, and Saturn car brands. We aggregated these four car brands together because of the data sparsity.
important role during the formation of the consideration set in the L-ICS model compared to its role in the choice process under the MNL model.

Next we consider three types of car attributes (i.e., car brand, car location type and accessibility, and car type and features), with the objective of empirically verifying their impact on consideration set formation under the L-ICS model and on the choice probabilities under the MNL model. Models 1, 2, and 3 (both under L-ICS and MNL) incorporate all the covariates except car types and features, car location type and accessibility, and brands, respectively, e.g., Model 1 excludes car types and features while including all the other covariates. According to Table A7, the car type and features attributes are less statistically significant than car brand attributes under the L-ICS model, whereas the opposite effect takes place under the MNL model. These findings are robust to the various measures of statistical significance and goodness-of-fit presented in Table A7 such as LL, AIC, BIC, Likelihood Ratio (LR) statistics, and Wald statistics. Overall, it is implied that car location type and accessibility plays the least important role both for consideration set formation and for the final choice decision. The renters are likely to build their consideration sets based on car brands rather than on car properties, even though while evaluating alternatives towards choice customers are likely to pay more attention to car properties rather than to car brands.

### A5.2. Discussion of Model Estimation Assumptions

In this section we further discuss the assumptions imposed by the CTC models with features that we take into account when calibrating the models. And then we also address the problem of a potential price endogeneity in our empirical explanatory analysis. We argue that in our setting we are unlikely to have any price endogeneity problems estimating the models.

#### A5.2.1. Semiparametric approach.

Using the semiparametric approach in order to calibrate the two stage CTC model, we assume that renters form their consideration set taking into account car features. Then we assume that during the second stage renters choose the most preferred car among the considered ones according to the preference order $\sigma$ over the universe of car.
alternatives, which remains the same over time, i.e., the ranking is fixed over time. Modeling the second stage choice process this way, we do not parameterize the ranking $\sigma$ which implies that the cars are assumed to have the same attributes over time. In this subsection we justify this assumption according to our dataset.

We start by analyzing the variation of the hourly price parameter over car alternatives. In Table A8, we report that the average coefficient of variation (CV) of the hourly price across all the car alternatives is around 5% while owners of cars listed on average around two different values of the price. Moreover, the most frequently used value of the hourly price corresponds to 78% of the car rentals and the second most frequently used value of the hourly price corresponds to 16% of the car rentals. The low variation of the rental price is explained by the policies of the online platform, for the time span of the dataset, that allows the owners to choose the price by themselves, i.e., the platform as a central agent did not dynamically adjust the listed rental price to efficiently match demand and supply as opposed to many ride sharing platforms (e.g., Uber, Lyft) which optimize the price of the ride to match riders with drivers on-demand. Then in the same Table A8 we can also observe that more than 98% of car owners did not alter their car access (i.e., open or closed), access hours (i.e., 24 hours or restricted), and location type (i.e., garage, street, surface lot, or valet).

**A5.2.2. Price endogeneity problem.** Next, we want to address the concerns of potential price endogeneity in our empirical analysis. First of all, estimating the demand with personalized data significantly alleviates the price endogeneity problem since each renter has only a trivial influence on the number of cars supplied and the market rental price, while the empirical work with aggregate level transaction data is more likely to face a very sever endogeneity issues. Nevertheless, having access to individual consumer data is not always a big advantage because individuals’ demand could be correlated. For example, we might have unobservable demand or supply shocks

<table>
<thead>
<tr>
<th>Additional descriptive statistics:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of rentals</strong></td>
</tr>
<tr>
<td>26,914</td>
</tr>
<tr>
<td><strong>Number of car owners</strong></td>
</tr>
<tr>
<td>514</td>
</tr>
<tr>
<td><strong>Number of available alternatives (within 0.3 mile)</strong></td>
</tr>
<tr>
<td>5.7</td>
</tr>
<tr>
<td><strong>Rental duration (days)</strong></td>
</tr>
<tr>
<td>0.62</td>
</tr>
<tr>
<td><strong>Rental request in advance (days)</strong></td>
</tr>
<tr>
<td>1.24</td>
</tr>
<tr>
<td><strong>Price CV (averaging over car owners)</strong></td>
</tr>
<tr>
<td>0.053</td>
</tr>
<tr>
<td><strong>Average number of price modes</strong></td>
</tr>
<tr>
<td>2.33</td>
</tr>
<tr>
<td><strong>The most frequent price (percentage)</strong></td>
</tr>
<tr>
<td>0.78</td>
</tr>
<tr>
<td><strong>The second most frequent price (percentage)</strong></td>
</tr>
<tr>
<td>0.16</td>
</tr>
<tr>
<td><strong>Average number of car access modes</strong></td>
</tr>
<tr>
<td>1.07</td>
</tr>
<tr>
<td><strong>The most frequent car access (percentage)</strong></td>
</tr>
<tr>
<td>0.99</td>
</tr>
<tr>
<td><strong>Average number of car access hours modes</strong></td>
</tr>
<tr>
<td>1.03</td>
</tr>
<tr>
<td><strong>The most frequent car access hours (percentage)</strong></td>
</tr>
<tr>
<td>0.99</td>
</tr>
<tr>
<td><strong>Average number of car location type modes</strong></td>
</tr>
<tr>
<td>1.10</td>
</tr>
<tr>
<td><strong>The most frequent car location type (percentage)</strong></td>
</tr>
<tr>
<td>0.98</td>
</tr>
</tbody>
</table>

Table A8  Additional descriptive statistics, the car sharing dataset.
if a local convention was organized in a particular day that might shift the demand curve. In this case, we need to use instrumental variables to address the endogeneity problem. The natural approach in this case would be to use the typical Hausman-style instrument (Hausman, 1996), i.e., the average rental price of similar cars in other geographical locations. However, in our dataset we are highly unlikely to have any price endogeneity issues because the rental price variation of the listed cars is very insignificant as it was discussed above, i.e., the price does not react to any unobservable shocks (see Table A8).