Inferring Consideration Sets from Sales Transaction Data

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

Understanding consumer preferences is critical when optimizing prices and assortments in retail operations, and when matching supply and demand in online platforms. In pursuing such an objective, a key input is the set of products that are both available and considered, from which a customer makes a choice. In practice, product availability is often hard to predict because it is subject to market conditions or other operational factors and considered products are unobserved. In this paper we propose a methodology to identify consideration sets, defined as those that are available and considered, from sales transaction records in a data driven way. We assume that customers are boundedly rational and make purchases in a two-stage process. First, they sample their consideration set and then purchase the most preferred item therein.

Our contribution to the literature is two-fold. Theoretically, we address the problem of identifiability of consider-then-choose models from data. Since calibrating this class of choice models is a hard problem, we propose a framework to effectively estimate them and infer consideration sets. The methodology to model the consideration set formation is founded on machine learning techniques that can account for nonlinear-in-parameter utilities in a tractable way. Then we apply the proposed framework on synthetic data and two real datasets: one from a big retail chain and the other one from a car-sharing platform. We observe that accounting for consideration sets can significantly boost the predictive performance in comparison with classical choice-based demand benchmarks, particularly in cases when the assortment of available products is not clearly defined. Our findings suggest that consider-then-choose models tend to be rather robust to the degree of ambiguity in the consideration set definition, and their relative performance with respect to state-of-the-art choice-based demand models improves with this ambiguity.

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

History: Original version: June 2019. This revision: January 2021.

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 models of demand, 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. (2020)) 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 (e.g., Farias et al. (2013)) and the exponomial (e.g., Alptekino˘glu 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). This standard assumption overlooks two problems.

First, the set of available alternatives is usually unobserved and often arbitrarily defined. For instance, in retailing, offer sets are commonly constructed from past transaction data in an ad hoc way assuming that a product is available at a particular time if it appeared in another transaction close to that time. The set of available products is formed by taking the union of all observable transactions within a period (say, a day) (see Villas-Boas and Winer (1999)). Such an assumption is reasonable for frequently purchased products but in general it results in a noisy approximation of the true offer set. Even if the data does provide an explicit description of the category offer set, its reliability may be questionable in light of potential inventory inaccuracies, a well-studied phenomenon in retailing (e.g., DeHoratius and Raman (2008)). In the context of online sharing platforms, the definition of the offer set as everything on offer within a radius rests on the arbitrary definition of the magnitude of such radius (e.g., all cars within 0.2 or 0.5 miles from the location of a user intending to book?).

More importantly, for demand forecasting, the modeler is not able to observe future availability sets (which will depend on operational planning decisions, stockouts, inventory management policies, etc.). This issue is particularly acute in online sharing platforms, where availability sets are determined by individual providers on the market, whose decisions are hard to predict.
Second, and concurrently, consideration sets of consumers or users are also unobserved. Take any reasonable definition of offer set just discussed. Is an individual really considering all products on offer at the moment of making a choice? For instance, in the case of retailing, is a customer selecting from the coffee category really evaluating the full assortment or is he only considering decaf options? In the case of car sharing platforms, is an agent evaluating all cars available within a 0.2-mile radius, or is he only considering compact cars? In the marketing field, a common and empirically validated model is to assume that when confronted with a purchasing decision, consumers engage in a two-stage decision making process in which they first evaluate some of the available products to form a smaller set of interest and then make the choice (e.g., Howard and Sheth (1969), Alba and Chattopadhyay (1985), Hauser and Wernerfelt (1990)).

Due to these two problems, when relying on a transaction-based data source, a fair and cautious modeling claim to make is that a purchased item (or booking in a platform) was preferred to all non-purchased items (or non booked alternatives) that were both available and considered simultaneously. In this paper, we contribute to the literature by proposing a choice-based demand framework that mitigates the negative impact of the unobservability of both the offer set and the consideration set, with the ultimate goal of making demand predictions.

1.1. Summary of results

We model customer preferences through a consider-then-choose model of demand. To fix ideas, we describe the setup in the context of retailing. Take a particular category of products and its associated full assortment—defined as the set of all possible products offered at some point during a predetermined time window (e.g., a year). Customers facing the category are characterized by two elements: (i) a preference list over the full set of products, and (ii) a distribution over subsets of the full assortment (i.e., over the consideration sets). Both elements are common across all customers. In each choice instance, a customer samples a consideration set 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.

Three comments about the model basis are in order. First, naturally, the source for calibrating the model is sales transaction data. In particular, for the portion of the data used to train the model, we assume that the availability is known. Mathematically, this assumption is without loss of generality because otherwise the sampled subset can be thought of as capturing propensity of being both considered and available. Second, our consideration set definition is product-based; i.e., every product in the full assortment is described by a single parameter that specifies its propensity to be considered (Manzini and Mariotti (2014), Gallego and Li (2017)).

1 This is different from the feature-based argument to define the consideration set, where products are screened along all of their features and are only included in the consideration set if their attributes are within pre-specified acceptable ranges that may be heterogeneous across individuals (e.g., see Jagabathula and Rusmevichientong (2017)).
can potentially depend on product features. Third, in order to be able to identify our model, we make the simplifying assumption that the customer preferences are fully specified by a unique ranking over the products in the product category (Masatlioglu et al. (2012); Manzini and Mariotti (2014); Aouad et al. (2020)). As a result, the heterogeneity of customers and stochasticity of choice is captured in the model through the consideration set formation.

The main goal of our proposal is to predict purchases by inferring consideration sets from sales transaction data. Specifically, we make the following contributions:

- **A general consider-then-choose model.** We study a general consider-then-choose (GCC) model. We fix the choice rule to be consistent with a single preference ordering but allow the distribution over consideration sets to be unrestricted. The key distinction of our work from the literature is that we allow the distribution over consideration sets to be general, for instance, by capturing correlations across different propensities. As such, our model subsumes the model of Manzini and Mariotti (2014) and other behavioral effects like the *inertia in choice* or *short-term brand loyalty* (Jeuland (1979)).

- **Identification conditions for the GCC model.** Theoretically, we derive necessary and sufficient conditions for a collection of observed purchase frequencies to be consistent with an underlying GCC model. We then show that the GCC model is indeed identifiable from sales transactions data. In particular, we provide a closed-form expression for computing the distribution over consideration sets from observed purchase frequencies. We also show that when consideration sets are of size at most $k$, the consideration set distribution can be recovered from choice probabilities under offer sets of size at most $k$.

- **Methodology to estimate the parameters of the GCC model.** We study the estimation problem starting from a particular case of the GCC model where the propensities of the different products items are independent (like in Manzini and Mariotti (2014)). We propose a mixed integer non-linear programming (MINLP) formulation of the maximum likelihood estimation (MLE) problem to estimate model parameters from sales transaction data. We show that this MINLP can be solved by constructing a sequence of MILPs obtained by successive linearizations of the original problem. It follows from existing results that this procedure converges to the global maximum. The procedure also provides a bound on the optimality gap at every step. Then, we propose an expectation-maximization (EM) algorithm in order to estimate the parameters of the GCC model.

- **Numerical experiments: Robustness to noise in offer sets.** 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

---

2 The inertia in choice is a behavioral bias that captures the fact that when facing frequently purchased consumer goods, customers tend to stick to the same option rather than evaluate all products on offer in each store visit.
with competitive benchmarks (e.g., MNL and rank-based models). We verify this conjecture by explicitly incorporating the noise factor in the generation of synthetic data instances and testing the predictive performance of the choice models evaluated under different noise scenarios.

- **Empirical analysis:** better demand predictions for the retail industry and online platforms. We compare choice models under several real world scenarios in retailing where we are likely to face significant noise in the offer set definitions. Our findings suggest that the relative performance of our GCC model over the benchmarks improves as we move to scenarios with higher noise level. We also analyze a dataset obtained from an online car-sharing platform. This experiment also provides evidence that consider-then-choose models are robust to noise in the offer set definition relative to classical choice models.

By modeling the propensity of consideration of items with a linear-in-attributes utility, we note a significant improvement of GCC in terms of predictive performance. Our framework also allows us the flexibility to model the propensity of consideration of products using a utility function that have a non-linear dependence on product attributes (e.g., via decision trees and random forests). We show that these non-linear models can be calibrated using existing packages and result in a further boost in prediction performance. For instance, after calibrating and testing the models with the industry partner dataset we obtain that the random forest-based consider-then-choose model outperforms the benchmark on standard metrics: by 43.3% on the mean absolute percentage error (MAPE) score, and by 53.7% on the root mean squared error (RMSE) metric.

The remainder of this paper is organized as follows. In Section 2 we present the related literature. A review of the Manzini-Mariotti model and the introduction of our extension, that we label **general consider-then-choose** (GCC) model, is discussed in Section 3 jointly with some identifiability results. Section 4 describes the data model and the methodology we propose to estimate the GCC model. 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). Section 8 introduces a multiclass version of the GCC model and some preliminary empirical results. Finally, our concluding remarks are discussed in Section 9.

### 2. Literature review

Our model is 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).

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.

As an extension of the Manzini-Mariotti model, we assume here that there is a distribution function over all possible subsets in the product universe, and every time the customer makes a purchase, she first samples a consideration set according to this distribution, and then chooses the most preferred available item. Our generalization relaxes the independent assumption of the attention parameters, significantly expanding the explanatory power of the model. We distinguish ourselves from existing work by focusing on calibrating such a general model and empirically testing its performance on real-world data.

We conclude this section by noticing that the ideas of identifying consideration sets from sales transaction data can be extended to the identification of competition sets. Empirical studies on this topic include, e.g., Lederman et al. (2014) and Li et al. (2018).
3. General consider-then-choose (GCC) model

In this section we formally describe the model and explore its location within the broad random utility model (RUM) class. We show that GCC belongs to RUM but is narrower in its scope. Then, we present theoretical results on the conditions under which our model can be identified from aggregated sales transaction data.

3.1. Model description

We consider a universe $N$ of products $\{a_1, a_2, \ldots, a_n\}$. We let $a_0$ denote the ‘no-purchase’ or the ‘outside’ option. Customers arrive to the store sequentially, and 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 $a_0$. For now, we assume that the set $S$ is observable. 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: $P_j(S) \geq 0$ for all $a_j \in S^+$ and $\sum_{a_j \in S^+} P_j(S) = 1$ 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 customer choices follow a two-stage consider-then-choose model. In the first stage, the customer forms a consideration set $C \subseteq N$ and in the second stage, the customer chooses either a product from the choice set $S \cap C$ of products or the outside option $a_0$, when the offered set of products is $S$. 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 because either she has strong unobserved preferences (which prevent her from ever buying certain products) or cognitive overload prevents her from evaluating all the products on offer before choosing. Throughout, we assume that the offer set $S$ is perfectly observed. As noted above, this assumption is without loss of generality because otherwise the sampled set $C$ can be thought of as capturing the set of products that are both considered and available.

The model is specified by two mathematical objects: a distribution $\lambda: 2^N \rightarrow [0, 1]$ over consideration sets such that $\sum_{C \subseteq N} \lambda(C) = 1$, and a choice rule that specifies which product is chosen from the subset $S \cap C$. In each choice instance, a randomly drawn customer from the population samples a consideration set $C$ according to $\lambda$ and then chooses a product from the set $(S \cap C)^+$ according to the choice rule. The most general model would accommodate any distribution $\lambda$ and stochastic preference rule, but such a general model cannot be identified from transaction data because it would subsume the RUM modeling framework which can not be identified for $n \geq 4$ (Sher et al.)
Figure 1  Choice process. There are \( n = 5 \) items in the product universe plus no purchase option 0. A customer represented by a full ranking \( \sigma \) samples consideration set \( C = \{1, 2, 3\} \) according to some pre-specified distribution \( \lambda \). 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 our model.

(2011)). Therefore, we restrict the degrees of freedom, by assuming that customers’ choice rule in the second stage is described by a single preference ordering over the products.

The preference ordering or ranking of the products in \( N \) is described by a bijective ranking function \( \sigma: N \rightarrow \{1, \ldots, n\} \) specifying a preference rank \( \sigma(a_j) \) for each product \( a_j \). Assuming that lower-valued ranks are preferred over higher-valued ranks, ranking \( \sigma \) indicates that product \( a \) is preferred over product \( b \) if and only if \( \sigma(a) < \sigma(b) \). The preference ordering \( \sigma \) induces an antireflexive, antisymmetric, and transitive preference relation \( \succ_\sigma \), defined as \( a \succ_\sigma b \) if and only if \( \sigma(a) < \sigma(b) \).

Under this choice rule, the customer chooses the most preferred product according to \( \sigma \) from the subset \( S \cap C \); that is, the customer chooses the product \( \arg \min \{\sigma(a_j) : a_j \in S \cap C\} \) if \( S \cap C \neq \emptyset \) and the outside option \( a_0 \) otherwise. Figure 1 illustrates the choice process for a particular store visit given a full rankings \( \sigma \), distribution over consideration sets \( \lambda \), and an offer set \( S \). Note that we are implicitly assuming that the outside option is the least preferred product in the ranking \( \sigma \), so the customer always makes a purchase whenever \( S \cap C \) is non-empty. This assumption is without loss of generality (wlog) because otherwise the customer will never purchase a product \( a_j \) that is less preferred than \( a_0 \), which is always available in \((S \cap C)^+\). In such case, product \( a_j \) could just be eliminated from the universe \( N \). With this restriction on the choice rule, we show below that the model can be identified from the transaction data, even without any further assumptions on \( \lambda \).

We refer to this model in which the distribution \( \lambda \) over consideration sets is unrestricted but the choice rule is restricted to a single preference ordering \( \succ \) as the general consider-then-choose (GCC) model. It follows from our description above that the choice probability \( \mathbb{P}_j(S) \) under the GCC model is given by

\[
\mathbb{P}_j(S) = \begin{cases} 
\sum_{C \subseteq N} \lambda(C) \cdot I[a_j \in S \cap C] \cdot I[a_j \succ_\sigma a_k \forall a_k \in S \cap C, a_k \neq a_j], & \text{if } a_j \in S \\
\sum_{C \subseteq N} \lambda(C) \cdot I[S \cap C = \emptyset], & \text{if } a_j = a_0,
\end{cases}
\]  

(1)

where \( I[A] \) 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, that is, \( A = \emptyset \), is always satisfied.
3.2. Connection to the RUM class

To understand how our model is related to existing models, we first ask if the GCC model belongs to the general random utility maximization (RUM) class (Block et al. (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, these models assume that in each choice instance, customers sample utility values for the products according to some joint distribution and choose the offered product that ranks highest. Equivalently, the RUM class of models is described by a distribution over preference orderings of products (Strauss (1979); Farias et al. (2013)), so that a customer samples a preference ordering according to the distribution and chooses the most preferred offered product according to the sampled preference list. It is straightforward to show that GCC model belongs to the class of RUM models, but the RUM models is a strict superset of the GCC class. We formally state this result below.

**Proposition 1.** The GCC choice model is a special case of the RUM choice rule, that is, \( \text{GCC} \subset \text{RUM} \), but \( \text{GCC} \neq \text{RUM} \).

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 GCC class.

While prior literature has studied consider-then-choose models, almost all the references have imposed restrictions on the structure of \( \lambda \), perhaps with more flexible assumptions on the consumer preferences. We differ from the literature (e.g., Manzini and Mariotti (2014)) by allowing \( \lambda \) to be very general, but with the restriction that the choice rule is fully described by a single preference list.

3.3. Identification of the GCC model

In what follows, we study some properties about the identifiability of the GCC model. For the first set of results, we will assume that we do have access to data consistent with an underlying GCC model. Later, in Proposition 5, we derive necessary and sufficient conditions for a collection of observed purchase frequencies to be indeed consistent with an underlying GCC model.

For now, suppose that purchases are generated according to an underlying GCC model. We provide different necessary conditions that would allow to recover the GCC parameters of the data generating process from a properly defined collection of choice probabilities \( P_j(S) \) and offer sets \( S \subseteq N \); here, we are ignoring any finite sample issues and assuming that the exact choice probabilities are known.

We start the description of the identification process by focusing on recovering the unique underlying rank list \( \sigma \). In fact, this preference order can be recovered independently of \( \lambda \) using the choice probabilities under all offer sets of size at most two; that is, \( \{ P_j(S); a_j \in S, |S| \leq 2 \} \). Specifically, we can state the following result:
Proposition 2. Suppose that the collection of choice probabilities \( \{P_j(S) : a_j \in S, |S| \leq 2\} \) are consistent with an underlying GCC model. Then, if \( P_i(\{a_i\}) > P_i(\{a_i, a_j\}) \), for all \( 1 \leq i, j \leq n, i \neq j \); we have that \( \sigma(a_j) < \sigma(a_i) \).

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 GCC probability definition in (1) that \( P_i(\{a_i\}) \leq P_i(\{a_i, a_j\}) \), which is a contradiction.

The remaining building block of the GCC model is the identification of the distribution \( \lambda \). In what follows we present a collection of results that provide different necessary conditions on the specification of the underlying choice probabilities that allow such identification. The first result only requires the specification of the no purchase probabilities, but over all possible offer sets.

Proposition 3. Suppose that the collection of choice probabilities \( \{P_0(S) : S \subseteq N\} \) are consistent with an underlying GCC model. Then, we have that

\[
\lambda(C) = \sum_{X \subseteq C} (-1)^{|C| - |X|} P_0(N \setminus X).
\]

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 \( g(X) = \sum_{Y \subseteq X} f(Y) \), then the inclusion-exclusion principle states that \( f(Y) = \sum_{X \subseteq Y} (-1)^{|Y| - |X|} g(X) \). Our result then follows from replacing \( f(Y) \) with \( \lambda(C) \) and defining \( g(X) = P_0(N \setminus X) = \sum_{C \subseteq X} \lambda(C) \). 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 of the 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 in the GCC model is bounded above by \( k \), with \( k < n \), it follows from Proposition 3 that to recover \( \lambda \), we need choice probabilities under offer sets of size \( n - k \) and larger. Recall that \( |N| = n \), and we are using \( k \) as \( k = \max_{X \subseteq N} |X| \).

Corollary 1. In a GCC model, suppose that customers sample consideration sets of size at most \( k \) for some \( 1 \leq k \leq n \); that is, \( \lambda(C) = 0 \) whenever \( |C| > k \). Then, the distribution over consideration sets \( \lambda \) can be identified using choice probabilities under offer sets of size \( n - k \) or larger, that is, from the collection \( \{P_0(S) : |S| \geq n - k\} \).

When the consideration sets are small, Corollary 1 shows 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 4. In a GCC model, suppose that customers sample consideration sets of size at most $k$ for some $1 \leq k \leq n$; that is, $\lambda(C) = 0$ whenever $|C| > k$. Let $\{P_0(S) : S \subseteq N, |S| \leq k\}$ be a collection of choice probabilities that are consistent with such a GCC 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 quite 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 $k$.

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

Proposition 5. The collection of choice probabilities $\{P_j(S) : a_j \in S^+, S \subseteq N\}$ is consistent with a GCC model with unique parameters $\sigma$ and consideration distribution $\lambda$ such that $\lambda(C) > 0$ for all $|C| \leq 3$ 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 be 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 $P_1(S \setminus \{a_2\}) \overset{=}{>} P_1(S)$, then it must be that $P_1(S' \setminus \{a_2\}) \overset{=}{>} P_1(S')$.

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 5 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 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 5 is presented in Appendix A2. As it can be observed therein, establishing necessity is straightforward, but establishing sufficiency is challenging.

### 4. Data model and estimation methodology

We begin this section by focusing on the special case of the GCC model where products enter consideration sets independently (a la Manzini-Mariotti). We provide a mixed integer nonlinear programming (MINLP) formulation of the maximum likelihood estimation (MLE) problem to calibrate this model. Then we show that solving the MINLP can be reduced to solving a sequence of mixed integer linear programs (MILPs) and propose two algorithms (an outer-approximation and a cutting plane, see Appendix A3) in order to implement the solution process. We continue with a MINLP formulation of the MLE problem to infer GCC model parameters from sales transaction data and propose an EM algorithm to estimate them.

Next, we show a way to model consideration set formation with covariates, such as product features and price. In particular we focus on three widely used methods in machine learning to describe consideration sets of customers: logistic-based, decision tree-based, and random forest-based consideration set models. We show that logistic-based, consider-then-choose model can be calibrated using an outer-approximation algorithm (see Appendix A3). We conclude the section by presenting two score metrics used in the paper to assess the prediction performance of the models followed by a description of the benchmark.

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

#### 4.1. Independent consider-then-choose (ICC) model

We first consider a special case of the GCC model in which the consideration set distribution has a product form. Specifically, we assume that customers build a consideration set by tossing a coin for each product $a_j$ and decide to independently include it with probability $\theta_j \in [0, 1]$. It then follows that $\lambda(C) = \prod_{a_j \in C} \theta_j \prod_{a_j \notin C} (1 - \theta_j)$. We call this model the independent consider-then-choose (ICC) model. This was indeed the model studied in Manzini and Mariotti (2014). It can be shown that under the ICC model, the probability of choosing product $a_j$ from offer set $S$ is

$$\Pr_j(S) = \theta_j \prod_{a_i \in S; a_i \succ a_j} (1 - \theta_i).$$

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

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

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

$$
\max_{\theta, \delta} \mathcal{L}(\theta, \delta) \quad (3)
$$

s.t.:

1. $\delta_{jk} + \delta_{kj} = 1, \ \forall \ j, k, \ j < k$,
2. $\delta_{jk} + \delta_{kp} + \delta_{pj} \leq 2, \ \forall \ j, k, p, \ j \neq k \neq p$,
3. $0 \leq \theta_j \leq 1, \ \forall \ j$,
4. $\delta_{jk} \in \{0, 1\}, \ \forall \ j, k$, 

where constraints (4) and (5) ensure that $\delta$ indeed represents a total order. In particular, the set of constraints (4) ensures that either $a_j$ is preferred over $a_k$ or vice versa, and the set of constraints (5) imposes the total ordering among any three products.

**4.1.1. Estimation methodology.** 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

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

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 $(\delta, \tau, \theta)$:

$$
\max_{\theta, \tau, \delta} \mathcal{L}(\theta, \tau) \quad (7)
$$

s.t.:

1. $\tau_{kj} \leq \theta_k, \ \forall \ j, k$,
2. $\tau_{kj} \leq \delta_{kj}, \ \forall \ j, k$,
3. $\tau_{kj} \geq \theta_k + \delta_{kj} - 1, \ \forall \ j, k$,
4. $\tau_{kj} \geq 0, \ \forall \ j, k$,
5. $\delta_{jk} + \delta_{kj} = 1, \ \forall \ j, k, \ j \leq k$,
6. $\delta_{jk} + \delta_{kp} + \delta_{pj} \leq 2, \ \forall \ j, k, p, \ j \neq k \neq p$,
7. $0 \leq \theta_j \leq 1, \ \forall \ j$,
8. $\delta_{jk} \in \{0, 1\}, \ \forall \ j, k$,
where linear constraints (8)-(11) ensure that $\tau_{kj} = \delta_{kj} \theta_k$, $\forall j,k$, given that $\delta$ is a binary variable, and constraints (12) and (13) ensure again a total order on $\delta$. We reformulate the MLE problem to have a linear objective function:

$$\max_{\theta, \tau, \delta, \mu} \mu$$

subject to $(\delta, \tau, \theta)$ satisfy (8) − (15),

$$\mu \leq \mathcal{L}(\theta, \tau).$$  \hspace{1cm} (17)

Note that if we know the ranking $\sigma$, then optimization problem (16) reduces to solving a globally concave maximization problem with a unique, closed form solution given by

$$\theta_j = \frac{\sum_{t=1}^{T} \mathbb{1}[a_{jt} = a_j]}{\sum_{t=1}^{T} \mathbb{1}[a_{jt} = a_j] + \sum_{t=1}^{T} \mathbb{1}[a_j \in S_t, a_j > a_{jt}]}.$$  \hspace{1cm} (18)

Next we show how to apply the outer-approximation method Duran and Grossmann (1986) to solve the optimization problem (16)-(17). 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 (17) 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 (17)) 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 (17) at a point $(x, y)$, i.e.,

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

Then we define the following optimization problem

$$\max_{\theta, \tau, \delta, \mu} \mu$$

subject to $(\delta, \tau, \theta)$ satisfy (8) − (15),

$$C[x, y] \forall (x, y) \in \mathcal{A}$$

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

where we have replaced constraint (17) with the finite collection of linear constraints (20) at points in a set $\mathcal{A}$ that is incrementally built. It follows from the convexity of constraint (17) that every
point that satisfies constraint (17) also satisfies the collection of constraints (20) for every finite set \( A \). We thus obtain an outer approximation. We also add bound constraints \( \mu_L \) and \( \mu_U \) on 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 ICC model is provided below.

**ICC 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 (18) 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 (18) given \( \sigma^{(i)} \) for all \( i \in \{1, 2, ..., m\} \).

**Step 3** Set \( \tau_k^{(i)} := \delta_k^{(i)} \theta_k^{(i)} \), 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 (19) with set \( A = \{(\theta^{(k)}, \tau^{(k)})\} \) and obtain solution \( \delta^{(i)}, \mu^{(i)} \). Note that, at each iteration, we add only one constraint to the optimization problem solved in the previous iteration.
- Update \( \theta^{(i)} \) using equation (18) given \( \delta^{(i)} \). Set \( \tau_k^{(i)} := \delta_k^{(i)} \theta_k^{(i)} \), for all \( j, k \).
- Set \( \mu_L := \max_{0 \leq i \leq m} \{ \mathcal{L}(\theta^{(i)}, \tau^{(i)}), \mu_L \} \) and \( \mu_U = \mu^{(i)} \).

Endwhile

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

**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 ICC 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 A7 in Appendix A3 that the optimality gap of the outer approximation algorithm (1) to calibrate the ICC 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.
4.2. GCC model

For the GCC model, the associated log-likelihood function is given by

\[ L(\delta, \lambda) = \sum_{t=1}^{T} \log \left( \sum_{C \subseteq N} \lambda(C) \cdot \mathbf{1}[a_{j_t} \in S_t \cap C] \cdot \mathbf{1}[\delta_{j_t k} = 1, \ \forall a_k \in S_t \cap C, k \neq j_t] \right). \]  

(21)

The likelihood estimation (MLE) problem can be formulated as follows:

\[
\max_{\delta, \lambda} \mathcal{L}(\delta, \lambda) \\
\text{subject to } \delta \text{ satisfies (4) – (6),} \\
\lambda(C) \geq 0, \ \forall C \subseteq N, \\
\sum_{C \subseteq N} \lambda(C) = 1.
\]

The first set of constraints guarantees that either product \( a_j \) is preferred to product \( a_k \) in the rank list or product \( a_k \) is preferred to product \( a_j \). The second and third sets of constraints ensure the validity of the probability distribution function \( \lambda \) over consideration sets.

4.2.1. Estimation Methodology. To make estimation tractable, we approximate the general distribution \( \lambda \) using a finite mixture. Specifically, we assume that the underlying population consists of \( K \) customer segments, with segment \( h \) customers sampling consideration set \( C \) with probability

\[ \prod_{a_j \in C} \theta_{hj} \prod_{a_j \notin C} (1 - \theta_{hj}), \]

where \( \theta_{hj} \) is the probability to include item \( a_j \) in the consideration set by type \( h \) customers. Assuming that a fraction of \( \gamma_h \geq 0 \) customers belong to segment-\( h \), we obtain

\[ \lambda(C) = \sum_{h=1}^{K} \gamma_h \prod_{a_j \in C} \theta_{hj} \prod_{a_j \notin C} (1 - \theta_{hj}). \]

A key observation is that for sufficiently large \( K \) we can calibrate the GCC model by estimating the following three components: (1) segment probabilities \( \gamma_h \), (2) propensity parameters for each segment \( \theta_{hj} \), and (3) the ranking \( \sigma_h \). Therefore, this parameterization of the GCC model is a natural approach when we have sparsity in customer segments. In Appendix A3.3 we provide the detailed analysis of how to calibrate the GCC model with the EM algorithm. 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.

From here onward, we estimate the model for \( K = 1, 2, ..., 5 \) and report the best performance measure from these 5 variants, for every prediction metric that we introduce later in Section 4.4.

4.3. Consider-then-choose models with features

In consider-then-choose models, so far we have assumed that each product has a separate propensity parameter \( \theta_j = \Pr[a_j \in C] \). We now describe three ways to make propensity depend on product features, where \( x_{jk} \) is the observed \( k \)th feature of product \( a_j \):
• **Logistic-based consider-then-choose model.** We assume that customers have linear-in-parameters utility $U_j$ from considering product $a_j \in \mathcal{N}$, given by

$$U_j = \beta_0^j + \sum_k \beta_k x_{jk} + \varepsilon_j,$$

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.,

$$a_j \in C \text{ iff } U_j = \beta_0^j + \sum_k \beta_k x_{jk} + \varepsilon_j \geq 0.$$

Then the propensity of product $a_j$ is given by

$$\Pr[a_j \in C] = \frac{\exp(\beta_0^j + \sum_k \beta_k x_{jk})}{1 + \exp(\beta_0^j + \sum_k \beta_k x_{jk})}.$$

• **Decision tree-based consider-then-choose model.** Here it is assumed that individuals decide which items to consider based on a tree with leaves $m \in \{1, 2, \ldots, M\}$ 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:

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

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 the threshold value, 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 consider-then-choose model.** 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:

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

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

4.3.1. **Estimation methodology.** In a similar spirit to Section 4.1, we can formulate the maximum likelihood estimation problem for the single class, logistic-based consider-then-choose (L-CC) 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. Note that this corresponds to the feature-based ICC model. For the multiclass (GCC) version of the model, we use an estimation procedure
similar to the GCC model with no features, with the only difference that now customers from each segment \( h \) sample consideration sets according to the logistic-based consider-then-choose model which takes into account the feature representation of products.

On the other hand, the calibration of DT-CC and RF-CC 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 A4.2 for the details).

4.4. Prediction scores and benchmark

In this section we describe two score metrics used in our empirical study to quantify the prediction power of the choice models in question. Under each of these scores the main objective is to predict the product to be purchased at time \( t + 1 \) given the corresponding offer set \( S_{t+1} \). The first score we use is the MAPE, computed as follows

\[
\text{MAPE} = \frac{100\%}{|N|} \sum_{a_j \in N} \frac{|n_j - \hat{n}_j|}{10 + \hat{n}_j}, \quad \text{and} \quad \hat{n}_j = \sum_{t=1}^{T} f(a_j, S_t),
\]

where \( f(a_j, S_t) \) is the probability of choosing item \( a_j \) from offer set \( S_t \) for the transaction at time \( t \); \( n_j \) is the observed number of times product \( a_j \) was purchased during the time horizon of length \( T \). Thus, the estimate \( \hat{n}_j \) stands for the expected number of purchases of product \( a_j \). Note that we add 10 in the denominator to deal with undefined instances. The second score, RMSE, is given by

\[
\text{RMSE} = \frac{100\%}{T} \sqrt{\frac{1}{|N|} \sum_{a_j \in N} (n_j - \hat{n}_j)^2}.
\]

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.

We compare our 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 \( h \in \{1, 2, \ldots, K\} \) make purchases according to the MNL model associated with that class. The model is described by the parameters of the MNL characterizing each class and by the prior probabilities of customers belonging to each of the classes. Once the model parameters are estimated, we make customer-level
(or transaction-level) predictions by averaging the predictions from $K$ single-class models, weighted by the posterior probability of class-membership. Similarly to the GCC model, we estimated the model for $K = 1, 2, \ldots, 5$, and report the best performance measure from these 5 variants, for every performance metric that we introduced above.

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 demonstrate that choice models based on the consider-then-choose framework are more robust to “noise” in the definition of offer sets than their classical counterparts. Specifically, we consider the case when the offer sets are not perfectly observed and we focus our analysis on understanding when the consider-then-choose type of models are better equipped to handle offer set noise than classical RUM-based models.

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 A5 in the Appendix.

Customers have perfect information on the offer set $S$, and they consider all the items on offer. Given the offer $S$, the customer chooses product $a_j$ with probability $v_j / \left(1 + \sum_{a_i \in S} v_i\right)$, 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 the MNL model against the ICC model 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 ICC 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 ICC 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 consider-then-choose framework outperform classical RUM models.

5.1. Synthetic data generation process

We parameterize the level of noise in the definition of the offer set by two parameters: $\gamma$ and $\eta$. We define $\gamma \in [0, 1]$ as the noise depth such that each item in the product universe is exposed to this noise with probability $\gamma$, i.e., if $n$ is the number of items in the product universe then $\gamma n$ is the average number of items that are exposed to noise. Next, let $\eta$ be the noise intensity such
that $\eta$ is the conditional probability of an item to be in the offer set as a noisy observation given that the item is exposed to the noise. The higher $\gamma$ and $\eta$, the more noise is added to the dataset. Deterministic utility values for items $a_j \in N$ are randomly chosen from the interval $[1,2]$, i.e., $v_j \sim U[1,2]$, normalizing the nominal utility of the outside option at value 0. We assume that we have $n = 15$ items in the product universe. 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 randomly sample 100 offer sets, i.e., $\{S_m\}_{m=1}^{100}$ as basis for the sales transaction data.
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 100 sets $\{\bar{S}_m\}_{m=1}^{100}$ such that each set $\bar{S}_m$ is obtained by tossing a coin for each product $a_j \in N$ and deciding to include it in $\bar{S}_m$ with probability $\gamma$.
4. We transform the sales transactions data by adding extra items with probability $\eta$ for every offer set $S_m$ if these items belong to $\bar{S}_m$, i.e., we modify every offer set $S_m$ of sales transactions data such that every item $a_j \in N \setminus S_m$ is added into $S_m$ with probability $\eta$ if $a_j \in \bar{S}_m$. For instance, if $\eta = 0.5$ and $\bar{S} = N$ then for every transaction, characterized by the tuple $(a_j, S_m)$, we modify offer set $S_m$ by adding on average half of the items from the subset $N \setminus S_m$.

Using the procedure above, we generate both training and test synthetic instances to test the hypothesis that ICC choice model is rather robust to the noise in the definition of offer sets in comparison with the MNL model. Each instance consists of $100 \times 10,000 = 1,000,000$ of transaction records. We generate both training and test datasets for 200 different combinations of $\gamma$ and $\eta$: $\gamma \in \{0.05, 0.1, ..., 1\}$ and $\eta \in \{0.1, 0.2, ..., 1\}$. We simulated 100 different instances of sales transaction data for each parameter combination. 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.

5.2. Results and discussion

In Figure 2 we present a 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 depth $\gamma$. We focus on the MAPE and RMSE prediction scores in the left and right panels, respectively. Recall that MNL is the ground truth model for this simulation study. As expected, 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. We observe that the MNL prediction scores worsen with higher noise intensity for a given noise depth. Interestingly, it can also be seen that MNL prediction scores are not monotonic with respect to the noise depth, i.e., the scores first increase with higher noise depth and then decrease for a given noise intensity. To
better explain the variation of prediction scores by the noise depth and noise intensity variables in Figure 2, we run the following linear regression:

\[
\text{Score}_{ik} = \beta_0 + \beta_1 \cdot \eta_i + \beta_2 \cdot \eta_i^2 + \beta_3 \cdot \text{Asymm}_k + \beta_4 \cdot \text{Shared}_k + \varepsilon_{ik},
\]  

(22)

where the noise intensity \( \eta_i \in \{0.1, 0.2, ..., 1\} \), and where \( \text{Asymm}_k \) is the probability that an item in the product universe is exposed to noise only in the test data set or only in the training dataset for \( k \in \{1, 2, ..., 20\} \), and \( \text{Shared}_k \) is the probability that an item in the product universe is exposed to noise both in the test and training datasets for \( k \in \{1, 2, ..., 20\} \). Note that \( \text{Asymm}_k = \gamma_k (1 - \gamma_k) + (1 - \gamma_k) \gamma_k = 2 \gamma_k (1 - \gamma_k) \) and \( \text{Shared}_k = \gamma_k^2 \), where \( \gamma_k \) is the noise depth such that \( \gamma_k \in \{0.05, 0.1, ..., 1\} \). The results for the regression (22) are reported in the last column in Table 1. It follows from Table 1 that noise intensity deteriorates the predictive performance of the MNL model in non-linear way with the coefficient for the linear term being positive, and the coefficient for the quadratic term being negative. The variables \( \text{Asymm}_k \) and \( \text{Shared}_k \) are positively correlated with the MAPE score, i.e., the prediction performance of the MNL model worsens as the number of items in the product universe that are exposed to the noise increases. Interestingly, the variable \( \text{Asymm} \) has more than seven times higher economic significance 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.\(^3\) Note that the independent variables in the regression model (22) explain most of the variation in the MAPE score under the MNL model, i.e., \( R^2_{adj} = 0.93 \).

In Figure 3 we present the heatmap of the prediction scores improvements under ICC model versus MNL model, where each column corresponds to a particular noise intensity \( \eta \) and each row corresponds to a particular noise depth \( \gamma \). We focus on the MAPE and RMSE prediction scores in left and right panels, respectively. Non-surprisingly, ground-truth MNL model significantly outperforms ICC model when there is only a small amount of noise added to the sales transactions data, e.g., \( \gamma = 0.05 \) and \( \eta = 0.1 \). As expected, our model cannot capture the ground-truth choice probabilities exactly because of model misspecification.

Figure 3 reveals that the improvement of ICC prediction scores over MNL are not monotonic with respect to the noise depth (i.e., the improvements first increase with higher noise depth and then decrease for a given noise intensity) and noise intensity (i.e., the scores first increase with higher noise intensity and then decrease for a given noise depth). To better explain the variation

\(^3\) Note that the highest noise asymmetry is achieved when \( \gamma \) is equal to 0.5. We have the worst performance under MNL model for this level of the noise depth \( \gamma \), see Figure 2.
of prediction improvements by the noise depth and noise intensity variables in Figure 3, we run the following linear regression which is similar to the regression (22):

\[
\text{Score}_{\text{Impr}}_{ij} = \beta_0 + \beta_1 \cdot \text{Intensity}_i + \beta_2 \cdot \text{Intensity}_i^2 + \beta_3 \cdot \text{Asymm}_j + \beta_4 \cdot \text{Shared}_j + \varepsilon_{ij}.
\]  

(23)

The results for the regression (23) are presented in the last column in Table 2. It follows from the table that the improvement of ICC model over MNL increases with noise intensity non-linearly such that the coefficients for the linear and squared term are positive and negative respectively. Since the coefficient corresponding to the variable Asymm is more than twice higher than the
coefficient corresponding to the variable Shared, we conclude that the ICC model has a higher chance to outperform the benchmark in scenarios when the sets of items that are exposed to noise in training and test datasets do not intersect. Results in this section are robust to different asymmetric scenarios of noise as well (see Figure A6 in Appendix A6).

Interestingly, it can be inferred from Tables 1 and 2 that the dependent variables (i.e., Intensity, Asymm, and Shared) impact the improvement of ICC over MNL qualitatively in the same way they impact the MNL prediction scores. As a result, it can be stated that the ability of ICC model to outperform MNL model is higher in scenarios when MNL model struggles to provide accurate predictions.

6. Case study on retailing
In real world retailing settings, noise is likely to result in an estimate of the offer set that is a superset of the true offer set. This type of noise implies that we may not know the offer set exactly, but we could assess a superset of the true offer set. Two of the factors that sustain this assumption are stockout events that could not be identified in real time, and inaccuracy in inventory records (e.g., according to DeHoratius and Raman (2008), most of the discrepancies between inventory records and physical inventories reflect that the former are higher than the latter).

In this section we compare the predictive power of GCC, ICC, and the latent-class MNL (LC-MNL) benchmark models based on 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 47 markets across the US over the years 2001-2011.

Figure 3  Heatmap of the prediction scores improvements under ICC model versus MNL where each column corresponds to a particular noise intensity and each row corresponds to a particular noise depth. We focus on the MAPE and RMSE scores in left and right panels respectively.
Table 2 Regression models where the dependent variable is the MAPE score improvement of the ICC model over the MNL model.

<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><strong>Intensity</strong></td>
<td>4.162***</td>
<td>16.621***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.883)</td>
<td>(17.214)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Intensity^2</strong></td>
<td>3.025***</td>
<td>-11.327***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(6.157)</td>
<td>(-13.241)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Asymm</strong></td>
<td>7.311***</td>
<td>11.530***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(7.112)</td>
<td>(25.421)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Shared</strong></td>
<td>2.539***</td>
<td>4.878***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.811)</td>
<td>(22.207)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>const</strong></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>
</tbody>
</table>

No. Observations: 200 200 200 200 200  
R-squared: 0.239 0.161 0.203 0.105 0.874  
Adj. R-squared: 0.235 0.156 0.199 0.100 0.871

\( t \) statistics in parentheses  
* \( p < 0.1 \), ** \( p < 0.05 \), *** \( p < 0.01 \)

For this experiment, we calibrate the GCC model by applying the EM algorithm with panel data where we rank items in the product universe according to their popularity in the sales transaction data for every category (see Appendix A3.3.3 for details).

Overall, the main purpose of this empirical study is threefold: (i) provide various real-world scenarios based on the IRI dataset when 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 GCC over LC-MNL under several real world scenarios with various noise intensities, and (iii) compare the GCC model proposed in this paper with the more restricted ICC model.

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

6.1. Data analysis  
The dataset consists of weekly sales transactions. We analyze a total of 20 categories, presented in Table A5. 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 construct this subset by aggregating all the transactions made in a particular store within a given category during a particular week. We aggregate items with the same vendor code into a single product due to data sparsity. We 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.

6.2. Results and discussion

We evaluate the demand prediction models based on the IRI dataset under three different scenarios in the retail setting. Each of these scenarios is defined by different ways of treating the holdout sample data over the last 26 weeks of the year. For the three cases, the training dataset is the same, defined by the first 26 weeks of the year.

1. Short-term forecasts. In this case we assume that the retailer has accurate information on the product assortments. To this end we make predictions by feeding the models with accurate offer sets obtained from the test dataset. This is reasonable when the retailer wants to predict demand in the near future (e.g., next week).

2. Long-term forecasts. To be successful in major strategic and investment decisions, the retailer needs accurate long-term demand forecasts. In this case, the robustness of the demand prediction models to the noise in the offer set definitions is very crucial as retailers usually do not have accurate information on product stock-outs for the distant future. To this end, we assume that a store manager makes demand predictions given the overall assortment of products in their store. We obtain the product assortments for each category-store combination by taking the union of all the offer sets over the test time horizon. Note that under this aggregation we have only one offer set for each category-store combination while making predictions over the holdout sample.

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. Instead, the warehouse would know the estimate of the product assortments across all the stores over a week. We model this scenario by feeding the prediction task for a given category with the union of product assortments over all the stores in the retail chain, on a weekly basis. Note that this way we have one offer set for each category-product-week combination while making predictions, leading to 26 different offer sets for each category-product combination. Each offer set is the same across all stores in a given week.

We expect the noise in the test offer sets to be higher for long-term and warehouse forecasts when compared to short-term forecasts. In Figure 4 we present scatter plots of the improvements of GCC
model versus LC-MNL model across 20 product categories under the three scenarios discussed above: (1) short-term forecasts, represented by pluses, (2) long-term forecasts, represented by crosses, and (3) warehouse forecasts, represented by dots. In the left and right panels we measure the predictive performance of the models using the MAPE and RMSE metrics, respectively. We observe that GCC 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 GCC 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 GCC 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 GCC 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 GCC over LC-MNL are -5.6% (-0.027%), 3.7% (0.061%), and 68.4% (0.973%) under the first, second, and third scenarios, respectively, based on the MAPE (RMSE) score.

Turning back to Figure 4, we notice that the improvement of GCC over LC-MNL varies across product categories for a given scenario. To better explain this variation, we regress the improvement of GCC over LC-MNL for each category against the noise intensity. We measure the noise intensity in the following way

\[
\text{noise intensity} = \frac{1}{T} \sum_{t=1}^{T} \frac{|\tilde{S}_t| - |S_t|}{|S_t|},
\]

where \( S_t \) is the offer set from the actual sales transactions data and \( \tilde{S}_t \) is a noisy offer set in transaction \( t \), and \( T \) is the number of observations in the holdout sample. Intuitively, the noise intensity under the long-term forecasts scenario is equal to the average percentage of the items that are stocked out in a store for a given category. The noise intensity under the warehouse forecasts scenario is equal to the average percentage of the items that are unavailable in a store over the total assortment of items that are available in the retail chain for a given week. 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 GCC over LC-MNL and noise intensity in both panels suggesting the improvement becomes more significant with higher noise intensity in the product category.
Figure 4 Scatter plots of the prediction score improvements (absolute difference) of GCC model over LC-MNL across 20 product categories. 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.

In Figure 7 we compare the prediction performance of GCC against ICC based on the MAPE score across 20 product categories. We observe that the relative improvement of GCC over ICC is 18.5%, 18.1%, and 8.9% under the first, second, and third noise scenarios, respectively, on average, across 20 product categories.

We emphasize two major findings here. First, we observe that the relative predictive performance of GCC over LC-MNL improves with the level of noise in the forecast offer sets. This finding is consistent with the analysis in Section 5 based on the synthetic dataset. Moreover, the improvements vary significantly across product categories and noise regimes. Second, we find that GCC significantly outperforms the ICC model in terms of the prediction performance as it can better capture heterogeneity of customer preferences.
Figure 5  The average prediction scores over 20 product categories under GCC 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.

Figure 6  Scatter plots and linear regressions of the MAPE score improvement of GCC vs. LC-MNL over the noise intensity across 20 product categories. Improvements are defined as the difference between two scores. In the left and right panels we focus on the long-term and warehouse forecast scenarios, respectively.

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 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 books 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),

Figure 7 Scatter plots of the MAPE scores of 20 product categories under the GCC vs. ICC models. The left, middle, and right panels correspond to the cases of short-term, long-term, and warehouse forecast scenarios, respectively. The lower the score the better.
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 A4.1 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 A6 in Appendix A6. 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., GCC) 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 A4.2.1 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 our model class: GCC, Decision Tree-based CC (DT-CC) and Random Forest-based CC (RF-CC), on the accuracy of two prediction measures: MAPE and RMSE (see Section 4.4), where lower scores stand for better prediction.

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 in the hold-out data sample 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 GCC, DT-CC and RF-CC, respectively. We also observe that consider-then-choose models obtain improvements of 6.2%, 10.9%, and 53.7% over LC-MNL for GCC, DT-CC and RF-CC, respectively, based on RMSE metrics.

Figure 9 exhibits MAPE scores computed for every brand separately under the RF-CC 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-CC model outperforms the benchmark LC-MNL model more or less consistently across all the brands. The panels also illustrate that MAPE scores vary significantly across brands both for RF-CC and LC-MNL models. To further analyze this variation, in Figure 10 we regressed the improvement of RF-CC over LC-MNL against the popularity of brands (left panel), and the improvement of RF-CC 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
Figure 9 MAPE scores of 32 brands under RF-CC and LC-MNL models.

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., CC models dominate LC-MNL, when LC-MNL model provides a relatively bad prediction performance. Being robust to the noise, consider-then-choose models (and in particular, RF-CC) 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 GCC 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 CC models by modeling the consideration set formation in a nonlinear-in-parameters way, with decision trees or random forests. We take it as a strong supportive evidence
for the validity of inferring consideration sets from transaction data with consider-then-choose models. After calibrating DT-CC and RF-CC models we can get some insights of how customers form their consideration sets. In particular, Figure A9 in Appendix A6 illustrates an instance of the decision tree obtained after fitting the DT-CC model.

8. Model extension and future research

The GCC model proposed in this paper might suffer from its one-directional cannibalization property, i.e., preferences of individuals are characterized by a unique ranking. Even though the same one directional cannibalization helps in reducing the impact of noise in the offer set, the GCC model might be worse off when there is no noise in the sales transaction data and cannibalization in the data generation process is bi-directional. This could be the case for products which are not vertically differentiated, where it is difficult to impose a natural preference order among them. In order to circumvent this, we propose a general consideration-then-general choice (GCGC) model as an extension of the GCC model where we allow customers to be heterogeneous in their preferences.

First, 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. According to this model, before making a choice, customers sample both a consideration set $C \subset N$ and a preference order $\sigma \in \mathcal{S}_n$ from the joint distribution over consideration sets and rankings. Then, customers choose the most preferred alternative in $C$ in accordance with the preference order $\sigma$. This model can be estimated in a similar way as the GCC model via an EM algorithm (see Appendix A3.4 for details) by dividing customers into $K$ segments such that for every segment $h \in \{1, ..., K\}$, a customer considers an
arbitrary subset of items $C \subseteq N$ with likelihood $\lambda(C) = \prod_{a_j \in C} \theta_{hj} \prod_{a_j \notin C} (1 - \theta_{hj})$ (recall that $\theta_{hj}$ is the probability to include item $a_j$ in the consideration set in segment $h$) and make choices according to the preference order $\sigma_h$. In this regard, GCGC is treated as a mixture of ICCs. As the number $K$ increases, this heuristic calibration would become more accurate. If we have sparsity in customer segments, then using this estimation technique with a relatively small $K$ would provide the exact calibration of the GCGC model.

We compare the prediction performance of the GCC model versus GCGC on the IRI academic dataset under different noise scenarios in the retail industry (see Section 6). We estimated the GCC model and five variants of the GCGC model, for $K = 1, 2, ..., 5$, and report the best performance measure from the best $K$ for MAPE and RMSE metrics. Figure A10 in Appendix A6 illustrates MAPE (left panel) and RMSE (right panel) scores of the GCC and the best out of $K$ GCGC models, averaging across 20 product categories, for the three different noise regimes. We observe that the predictive performance of both models is rather robust to the noise in the offer sets, i.e., the performances stay rather flat for all three noise regimes. From the panels in Figure A10 we can not claim dominance of GCC or GCGC, i.e., GCC dominates GCGC based on the RMSE metric whereas GCGC outperforms GCC based on the MAPE metric. Additional analysis of improvement across categories presented in Figure A11 in Appendix A6 reveals the same insight.

An important area of future research would be to further explore the GCGC model and its application in different scenarios, e.g., find the optimal number of segments in the GCGC model to strike a good balance between capturing heterogeneity of customers in their preferences and robustness to noise in sales transaction data, address the problem of identifiability of GCGC model (i.e., what is the maximum value of $K$ for the model to be identifiable?), and propose a more efficient way to estimate the model for large-scale problems.

9. Conclusion

In this paper we propose a customer-centric method to identify consideration sets from sales transaction data. Motivated by behavioral and psychological aspects of customers, the vast majority of existing papers focusing on consideration set definition impose a prior belief on the consideration set formation (e.g., screening rules, trade-off between expected cost and benefit of search, etc.). Different from this line of research, our approach is completely data-driven.

In the spirit of the consider-then-choose framework, we assume that customers make purchasing decisions in two stages. In the first stage, a boundedly rational consumer forms a consideration set, which is usually a small subset of the items in the product category. In the second stage, she evaluates all products in her consideration set and purchases the most preferred available one. In this paper we propose an effective way to model the consideration set formation in the first
stage using two popular machine learning methods: decision trees and random forests. Even though consideration sets are unobservable, our modeling approach allows us to infer the most likely subset of items considered by each individual, depending on past purchasing transactions. After checking the good performance of our methodology under a synthetic data setting, we apply it to two real world settings: a retail operation and a car-sharing platform. Our empirical results suggest that the predictive performance of consider-then-choose models significantly outperforms state-of-the-art benchmarks widely used in the marketing and economics literature. Moreover, we show that the relative improvement of consider-then-choose models in predictive performance becomes even more significant with increased noise in the consideration set definition embedded in the data. 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.

References


Wright P, Barbour F (1977) *Phased decision strategies: Sequels to an initial screening* (Graduate School of Business, Stanford University).
Inferring consideration sets from sales transactions data

Appendix

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 $\lambda: 2^N \rightarrow [0,1]$ 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 $\mu: \mathcal{S}_n \rightarrow [0,1]$ 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 $A \sqcup B$ denote $\{a \cup b : a \in A, b \in B\}$ for any sets $A, B$.

A2. Proofs of Technical Results

Proof of Proposition 1: First, we argue that GCC class of models describing customer choice behavior is consistent with RUM. Indeed, it is straightforward to verify that GCC choice model with underlying preference order $\sigma$ is equivalent to the rank-based (i.e., RUM) model where all the preference lists are obtained from a common permutation $\sigma$. In particular, consider the RUM model with the probability over rankings $\mu$ such that $\forall C \subseteq N, \mu(\sigma(C)) = \lambda(C)$, and $\mu(X) = 0$ for $X \notin \{\sigma(C) : C \subseteq N\}$, where $\sigma(C)$ is the full preference order with items in $C$ at the top positions consistently with $\sigma$ (e.g., the most preferred item in $C$ is placed on the top position, the worst preferred item in $C$ is placed on $k$-th position, where $k$ is the cardinality of $C$) followed by the items that are not in $C$. It’s easy to verify that the RUM with the defined distribution over rankings $\mu$ results into the same probabilities of purchases as GCC model. Then, it remains to show that RUM model class is not a specific case of GCC model class. To this end, we provide a particular
example of RUM model class resulting in customers’ choice frequencies that are inconsistent with GCC 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:

\[
\mathbb{P}_1(\{ a_1, a_2 \}) = \mu_1, \quad \mathbb{P}_1(\{ a_1 \}) = 1 \Rightarrow a_2 \text{ is preferred to } a_1, \text{ by GCC definition},
\]

\[
\mathbb{P}_2(\{ a_1, a_2 \}) = 1 - \mu_1, \quad \mathbb{P}_2(\{ a_2 \}) = 1 \Rightarrow a_1 \text{ is preferred to } a_2, \text{ by GCC definition}.
\]

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

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

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

**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 \( g(X) := \sum_{Y \subseteq X} (-1)^{|X| - |Y|} h(Y) \).

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

\[
h(P) \cdot (-1)^{|P|} = \sum_{X \subseteq P} (-1)^{|X|} \cdot g(X), \quad \text{by invoking the inclusion-exclusion principle we obtain that}
\]

\[
(-1)^{|Y|} \cdot g(Y) = \sum_{P \subseteq Y} (-1)^{|Y| - |P|} \cdot h(P) \cdot (-1)^{|P|}, \quad \text{which implies that}
\]

\[
f(Z \setminus Y) = g(Y) = \sum_{P \subseteq Y} h(P) = \sum_{P \subseteq Y} (-1)^{|P|} \sum_{X \subseteq P} (-1)^{|X|} \cdot g(X) = \sum_{P \subseteq Y} \sum_{X \subseteq P} (-1)^{|P| - |X|} \cdot g(X)
\]

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

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

\[
\sum_{\beta=0}^{\min(r,u)} \binom{w}{\beta} \cdot \sum_{\alpha=r+1-\beta}^{u-\beta} (-1)^{\alpha} \binom{u}{\alpha} = \begin{cases} 1, & \text{if } w = u, \\ 0, & \text{if } w < u, \end{cases} \tag{A2}
\]

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.

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

where the last equality is proved by induction on \( s = u - r \):

Base case: \( s = 1 \).

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

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

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

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

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

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

$$= - \sum_{\beta=0}^{w} C_{\beta}^{w} \cdot \sum_{\alpha=u-p+1-\beta}^{u-\beta} (-1)^{\alpha} C_{\alpha}^{u-\beta} - \sum_{\beta=0}^{w} C_{\beta}^{w} \cdot (-1)^{u-p-\beta} \cdot C_{u-p-\beta}^{u-\beta}$$

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

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

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 $\sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{1}{\beta!(w-\beta)!(u-p-\beta)!} = \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{1}{\beta!(w-\beta)!(u-m-\beta)!} = 0$. Assuming that the result holds for $p = m$, we prove it for $p = m + 1$:

$$\sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)!}{\beta!(w-\beta)!(u-m-\beta)!} = \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)!}{\beta!(w-\beta)!(u-m-\beta)!}$$

$$= (u - m) \cdot \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)!}{\beta!(w-\beta)!(u-m-\beta)!} - \sum_{\beta=0}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)!}{\beta!(w-\beta)!(u-m-\beta)!}$$

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

$$= - \sum_{\beta=1}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)!}{\beta!(w-\beta)!(u-m-\beta)!}$$

$$= - \sum_{\beta=1}^{w} (-1)^{\beta} \cdot \frac{(u-\beta)!}{(\beta-1)!(w-\beta)!(u-m-\beta)!}$$

$$= \sum_{\beta=0}^{w-1} (-1)^{\beta} \cdot \frac{(u-1-\beta)!}{\beta!(w-\beta)!(u-m-\beta)!}$$

$$= \sum_{\beta=0}^{w-1} (-1)^{\beta} \cdot \frac{(u-1-\beta)!}{\beta!(w-1-\beta)!(u-m-\beta)!}$$

$$= 0, \quad \text{[by induction hypothesis, } p = m].$$

**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

$$- \sum_{\beta=0}^{u-p-1} C_{\beta}^{w} \cdot \sum_{\alpha=u-p-\beta}^{u-\beta} (-1)^{\alpha} C_{\alpha}^{u-\beta} = - \sum_{\beta=0}^{u-p} C_{\beta}^{w} \cdot \sum_{\alpha=u-p-\beta}^{u-\beta} (-1)^{\alpha} C_{\alpha}^{u-\beta} + \sum_{\alpha=0}^{p} (-1)^{\alpha} C_{\alpha}^{p}$$
Now it is sufficient to prove that
\[
\sum_{\beta=0}^{u-p}(1-\beta)\cdot C^w_{\beta}\cdot C^u_{u-\beta}\cdot C^u_{u-\beta-p} = 0.
\]

\[\sum_{\beta=0}^{u-p}(1-\beta)\cdot C^u_{u-\beta}\cdot C^u_{u-\beta-p}\;
\text{by induction hypothesis, } r = u-p\]

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

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

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

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

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

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

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

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

\[
= 0, \quad \text{by induction hypothesis}.
\]

Proof of Proposition 3: 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

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

\[
\psi_C(X) = (-1)^{|X|} I[X \subseteq C],
\]
Then we show that
\[ \sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \begin{cases} 1, & \text{if } C_1 = C_2, \\ 0, & \text{otherwise}, \end{cases} \tag{A3} \]

First, we show that \( \sum_{X \subseteq N} \chi_C(X) \cdot \psi_C(X) = 1 \) for every \( C \subseteq N \):
\[
\sum_{X \subseteq N} \chi_C(X) \cdot \psi_C(X) = \sum_{X \subseteq N} \mathbb{I}[C \subseteq X] \cdot (-1)^{|C|+|X|} \mathbb{I}[X \subseteq C] = (-1)^{|C|+|C|} = 1
\]

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

[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]
\[
= (-1)^{2|C_1|} \cdot (1 - 1)^{|C_2| - |C_1|} = 0.
\]

Consequently, the probability to choose the “no purchase” option \( a_0 \) from the offer set \( \{N \setminus X\}^+ \) is given by
\[
\mathbb{P}_0(N \setminus X) = \sum_{C \subseteq X} \lambda(C) = \sum_{C \subseteq N} \lambda(C) \cdot (-1)^{2|C|} \cdot \mathbb{I}[C \subseteq X] \tag{A4}
\]
\[
= \sum_{C \subseteq N} \lambda(C) \cdot (-1)^{|C|} \cdot \chi_C(X).
\]

Then it follows that
\[
\sum_{X \subseteq C} (-1)^{|C| - |X|} \cdot \mathbb{P}_0(N \setminus X) = \sum_{X \subseteq N} \mathbb{P}_0(N \setminus X) \cdot (-1)^{|C|+|X|} \mathbb{I}[X \subseteq C] \tag{A5}
\]
\[
= (-1)^{|C|} \cdot \sum_{X \subseteq N} \mathbb{P}_0(N \setminus X) \cdot \psi_C(X) \\
= (-1)^{|C|} \cdot \sum_{X \subseteq N} \sum_{C_1 \subseteq N} \lambda(C_1) \cdot (-1)^{|C_1|} \cdot \chi_{C_1}(X) \cdot \psi_C(X) \\
= (-1)^{|C|} \cdot \sum_{C_1 \subseteq N} \lambda(C_1) \cdot (-1)^{|C_1|} \cdot \sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_C(X) \\
= (-1)^{|C|} \cdot \lambda(C) \cdot (-1)^{|C|}, \tag{by Equation (A3)}
\]
\[
= \lambda(C).
\]
Now it remains to prove the uniqueness of probability distribution function $\lambda$ obtained from purchasing transactions data under GCC choice model. Note that Equation (A4) relates probability distribution $\lambda$ over consideration sets to the choice frequencies $P_0(N \setminus X)$ through the system of linear equations:

$$P_0(N \setminus X) = \sum_{C \subseteq N} \lambda(C) \cdot (-1)^{|C|} \cdot \chi_C(X), \quad \forall X \subseteq N \iff y = A \cdot \lambda,$$

(A6)

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 $(-1)^{|C|} \cdot \chi_C(X)$. Therefore, the relation between the choice frequencies and the underlying model can be represented in a compact form as $y = A \cdot \lambda$. Then the proof of uniqueness of $\lambda$ reduces to showing that $\det(A) \neq 0$. From Equation (A5) we have

$$\lambda(C) = (-1)^{|C|} \cdot \sum_{X \subseteq N} P_0(N \setminus X) \cdot \psi_C(X), \quad \forall C \subseteq N \iff \lambda = B \cdot y,$$

which establishes alternative linear relationship between choice frequencies $P_0(N \setminus X)$ 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

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

$$\implies I = B \cdot A \implies \det(I) = \det(B) \cdot \det(A)$$

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

\[ \Box \]

**Proof of Proposition 4:** It follows from the proposition that

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

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

$$= \sum_{X \subseteq N} P_0(X) \cdot (-1)^{1+|X \cap C|} \cdot I[|X \cup C| \leq k] \cdot \sum_{Y \supseteq X \cup C} (-1)^{|Y|-|X\cup C|} \cdot I[|Y| > k]$$

[since the expression depends only on the cardinality of sets $Y$, the summation over the sets $Y$ is reduced to the summation over the cardinality of sets $Y$]

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

[where $C^n_{\alpha} = \frac{n!}{k!(n-k)!}$]
For every $C \subseteq N$ we define boolean functions $\chi_C : 2^N \to \mathbb{R}$ and $\psi_C : 2^N \to \mathbb{R}$ by

$$\chi_C(X) = \mathbf{1}[C \subseteq X, |C| \leq k],$$

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

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:

$$P_0(X) = \sum_{C \subseteq N} \lambda(C) \cdot \mathbf{1}[C \subseteq X, |C| \leq k] = \sum_{C \subseteq N} \lambda(C) \cdot \chi_C(X). \quad (A7)$$

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

$$\sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \begin{cases} 1, & \text{if } C_1 = C_2, \\ 0, & \text{otherwise}. \end{cases} \quad (A8)$$

Consequently, it follows from the claim that

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

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

$$= \sum_{C_1 \subseteq N} \lambda(C_1) \cdot \sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \lambda(C), \quad \text{by Equation (A8).}$$

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: $C_2 \subseteq C_1$.**

$$\sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \sum_{X \subseteq N} (-1)^{1} \cdot \mathbf{1}[|C_1| \leq k] \cdot \mathbf{1}[X \cap C_1 = \emptyset] \cdot \mathbf{1}[|X| \leq k - |C_2|]$$

$$\times \sum_{\alpha = k+1-|X \cup C_2|}^{n-|X|} (-1)^{\alpha} C^n_{\alpha} \cdot |X - |C_2| |$$

[in this case, $X \cap C_1 = \emptyset$, $|X \cap C_1| = 0$, $|X \cap C_2| = 0$, and $|X \cup C_2| = |X| + |C_2|$]

$$= - \sum_{X \subseteq N} \mathbf{1}[|C_1| \leq k] \cdot \mathbf{1}[X \cap C_1 = \emptyset] \cdot \mathbf{1}[|X| \leq k - |C_2|] \cdot \sum_{\alpha = k+|C_2|+1-|X|}^{n-|C_2|} (-1)^{\alpha} C^n_{\alpha} \cdot |X| - |C_2|,$$

[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]
Figure A1. In this figure $C_1 \not\subseteq C_2$ and $n - |C_1| - |C_2| > 0$. 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.

Figure A2. 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.

Figure A3. 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 $X$.

$$
\sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \sum_{X \subseteq N} (-1)^{1+|X \cap C_2|} \cdot \mathbf{1}[X \cap C_1 = \emptyset, |C_1| \leq k] \cdot \mathbf{1}[|X \cup C_2| \leq k]
$$

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: $C_2 \not\subset C_1$. 

$$
\sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \sum_{X \subseteq N} (-1)^{1+|X \cap C_2|} \cdot \mathbf{1}[X \cap C_1 = \emptyset, |C_1| \leq k] \cdot \mathbf{1}[|X \cup C_2| \leq k]
$$
A10

\[
\sum_{\alpha=k+1-|X\cup C_2|} \left( -1 \right)^{\alpha} C^{n-|X\cup C_2|}_\alpha \times \sum_{\alpha=0}^{n-|X\cup C_2|} \left( -1 \right)^{\alpha} C^{n-|X\cup C_2|}_\alpha \]

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

\[
\begin{cases}
\sum_{\gamma=0}^{|C_5|} (-1)^{1+\gamma} \cdot C^{C_5}_\gamma \cdot \sum_{\beta=0}^{|C_2|, n-|C_1|-|C_5|} C^{n-|C_1|-|C_5|}_\beta \cdot \sum_{\alpha=k+1-|C_2|} (-1)^{\alpha} C^{n-|C_2|}_\alpha \cdot \left( \sum_{\alpha=0}^{n-|C_2|} \left( -1 \right)^{\alpha} C^{n-|C_2|}_\alpha \right), & \text{if } C_1 \subset C_2 \\
\sum_{\gamma=0}^{|C_5|} (-1)^{1+\gamma} \cdot C^{C_5}_\gamma \cdot \sum_{\beta=0}^{|C_2|, n-|C_1|-|C_5|} C^{n-|C_2|}_\beta \cdot \sum_{\alpha=k+1-|C_2|} (-1)^{\alpha} C^{n-|C_2|}_\alpha \cdot \left( \sum_{\alpha=0}^{n-|C_2|} \left( -1 \right)^{\alpha} C^{n-|C_2|}_\alpha \right), & \text{if } C_1 \nsubseteq C_2 \text{ and } n-|C_1|-|C_5| > 0, \text{ see Figure A1}
\end{cases}
\]

where \( C_5 = C_2 \setminus \{ C_1 \cap C_2 \} \)

\[
\begin{cases}
\sum_{\gamma=0}^{|C_5|} (-1)^{\gamma} \cdot C^{C_5}_\gamma \cdot \sum_{\beta=0}^{|C_2|, n-|C_1|-|C_5|} C^{n-|C_1|-|C_5|}_\beta \cdot \sum_{\alpha=k+1-|C_2|} (-1)^{\alpha} C^{n-|C_2|}_\alpha \cdot \left( \sum_{\alpha=0}^{n-|C_2|} \left( -1 \right)^{\alpha} C^{n-|C_2|}_\alpha \right), & \text{if } C_1 \nsubseteq C_2 \text{ and } n-|C_1|-|C_5| > 0, \\
\sum_{\gamma=0}^{|C_5|} (-1)^{\gamma} \cdot C^{C_5}_\gamma \cdot \sum_{\beta=0}^{|C_2|, n-|C_1|-|C_5|} C^{n-|C_2|}_\beta \cdot \sum_{\alpha=k+1-|C_2|} (-1)^{\alpha} C^{n-|C_2|}_\alpha \cdot \left( \sum_{\alpha=0}^{n-|C_2|} \left( -1 \right)^{\alpha} C^{n-|C_2|}_\alpha \right), & \text{if } C_1 \subset C_2, \\
\sum_{\gamma=0}^{|C_5|} (-1)^{\gamma} \cdot C^{C_5}_\gamma \cdot \sum_{\alpha=k+1-|C_2|} (-1)^{\alpha} C^{n-|C_2|}_\alpha, & \text{if } C_1 \nsubseteq C_2 \text{ and } n-|C_1|-|C_5| = 0, \\
\end{cases}
\]

= 0,

where the last equality follows since \(|C_5| > 0\), and \( \sum_{\gamma=0}^{|C_5|} (-1)^{\gamma} \cdot C^{C_5}_\gamma = 0 \).

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:

\[
P_0(X) = \sum_{C \subseteq N} \lambda(C) \cdot \chi_C(X), \ \forall \ X \subseteq N \iff \ y = A \cdot \lambda, \quad (A10)
\]

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

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

$$\implies I = B \cdot A \implies \det(I) = \det(B) \cdot \det(A)$$

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

□

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

$$\sum_{X \subseteq C} (-1)^{|C|-|X|} P_0(N \setminus X) \geq 0,$$

with strict inequality for consideration sets of the size up to three, i.e., if $|C| \leq 3$, then for all consideration sets $C \subseteq S$ s.t. $S \subseteq N$ it follows that

$$\sum_{X \subseteq C} (-1)^{|C|-|X|} P_0(S \setminus X) \geq 0,$$

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:

$$\sum_{B \subseteq C} (-1)^{|C|-|B|} P_0(S \setminus B) = \sum_{B \subseteq C} (-1)^{|C|-|B|} \cdot P_0(\{N \setminus \bar{S} \setminus B\})$$

$$= \sum_{B \subseteq C} (-1)^{|C|-|B|} \cdot P_0(\{N \setminus B \setminus \bar{S}\})$$

$$= \sum_{B \subseteq C} \sum_{A \subseteq S} \sum_{D \subseteq A} (-1)^{|A|-|D|} \cdot (-1)^{|C|-|B|} \cdot P_0(\{N \setminus B \setminus D\})$$

[by invoking Lemma A1 for every $B \subseteq C$, where $Z = N \setminus B$, $Y = \bar{S}$, $P = A$, and $f(Z \setminus Y) = (-1)^{|C|-|B|} \cdot P_0(\{N \setminus B \setminus \bar{S}\})$]

$$= \sum_{A \subseteq \bar{S}} \sum_{D \subseteq A} \sum_{B \subseteq C} (-1)^{|C|+|A|-|D|-|B|} \cdot P_0(\{N \setminus B \setminus D\})$$

Electronic copy available at: https://ssrn.com/abstract=3410019
Article submitted to Management Science; manuscript no. MS-0001-1922.65

\[
= \sum_{A \subseteq S} \sum_{X \in \langle A \cup C \rangle} (-1)^{|C|+|A|-|X|} \cdot P_0(N \setminus X) \quad \text{[where } X = D \cup B, \text{ since } A \cap C = \emptyset]\]

\[
= \sum_{A \subseteq S} \sum_{X \subseteq A \cup C} (-1)^{|C|-|X|} \cdot P_0(N \setminus X) \quad \text{[since } A \cap C = \emptyset]\]

\[
= \sum_{A \subseteq S} \sum_{X \subseteq C'} (-1)^{|C'|-|X|} \cdot P_0(N \setminus X) \quad \text{[where } C' = A \cup C]\]

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

\[\text{since } \sum_{X \subseteq C'} (-1)^{|C'|-|X|} \cdot P_0(N \setminus X) \geq 0 \text{ with strict inequality when } |C'| \leq 3.\]

\[\Box\]

Lemma A4. If a sample of sales transaction data satisfies Conditions 1, 2, and 3, then for all \(a_1 \in S_1 \cap S_2\) where \(S_1, S_2 \subseteq N\) and \(S_1 \subseteq S_2\) we have that \(P_i(S_1) \geq P_i(S_2)\). 

Proof: Prove the result by induction on the \(n = |S_2| - |S_1|\). We consider \(a_1 \in S_1 \cap S_2\) and \(S_1 \subseteq S_2\). For the base case \(n = 0\) we have that \(S_1 = S_2\) and \(P_1(S_1) = P_1(S_2)\). Assume that the result holds for \(n = k\), i.e., \(S_2 = S\) and \(|S| - |S_1| = k\). Then we prove it for \(n = k + 1\). Let us suppose w.l.o.g. that \(S_2 = S \cup \{a_2\}\) and \(a_2 \not\in S\). Next, assume, by contradiction, that \(P_1(S_2 \setminus \{a_2\}) < P_1(S_2)\). Consequently, by Condition 2 it follows that \(P_1(\{a_1\}) < P_1(\{a_1, a_2\})\). Then by Condition 1 we have that

\[P_2(\{a_2\}) = P_2(\{a_1, a_2\}). \quad (A11)\]

It now follows that

\[P_1(\{a_1\}) - P_1(\{a_1, a_2\}) = (1 - P_0(\{a_1\})) - (1 - P_0(\{a_1, a_2\}) - P_2(\{a_1, a_2\})), \quad \text{[by standard probability property]}\]

\[= (1 - P_0(\{a_1\})) - (1 - P_0(\{a_1, a_2\}) - P_2(\{a_2\})), \quad \text{[by Equation (A11)]}\]

\[= (1 - P_0(\{a_1\})) - (P_0(\{a_2\}) - P_0(\{a_1, a_2\})), \quad \text{[by standard probability property]}\]

\[= 1 - P_0(\{a_1\}) - P_0(\{a_2\}) + P_0(\{a_1, a_2\}) > 0, \quad \text{[by Condition 3 and Lemma A3, when } C = S = \{a_1, a_2\}\],

which contradicts to \(P_1(\{a_1\}) < P_1(\{a_1, a_2\})\). Then we have

\[P_1(S_2) \leq P_1(S_2 \setminus \{a_2\}) = P_1(S), \quad \text{[note that } |S| - |S_1| = k]\]

\[\leq P_1(S_1), \quad \text{[by induction hypothesis]}\]

Therefore, the result now follows by induction. \(\Box\)
Lemma A5. Consider \(a_1, a_2 \in S, S \subseteq N\), \(a_1 \neq a_2\). Then GCC 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) \(P_1(S \setminus \{a_2\}) > P_1(S) \implies a_2 \succ a_1\), and \(\forall S' \subseteq N\) \(s.t. a_1, a_2 \in S'\): \(P_1(S' \setminus \{a_2\}) > P_1(S')\),

b) \(P_1(S \setminus \{a_2\}) = P_1(S) \implies a_1 \succ a_2\), and \(\forall S' \subseteq N\) \(s.t. a_1, a_2 \in S'\): \(P_1(S' \setminus \{a_2\}) = P_1(S')\),

c) \(P_1(S \setminus \{a_2\}) \neq P_1(S) \implies P_2(S \setminus \{a_1\}) = P_2(S)\).

Proof: a) Suppose that \(P_1(S \setminus \{a_2\}) > P_1(S)\). Assume, by contradiction, that \(a_1 \succ a_2\). Then it can be inferred from purchase probability definition under GCC, see Equation (1), that \(P_1(S \setminus \{a_2\}) = P_1(S)\), which leads to contradiction. As a result, we have that \(a_2 \succ a_1\) since preferences are strict and asymmetric. Then \(\forall S' \subseteq N\) \(s.t. a_1, a_2 \in S'\) we establish that

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

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

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

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

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

which contradicts to 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 5: Necessity: if purchasing transactions data is consistent with GCC 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 3 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 GCC 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:

\[
P_i(\{a_i\}) - P_i(\{a_i, a_j\})
\]
\[
= \left(1 - P_0(\{a_i\}) \right) - \left(1 - P_0(\{a_i, a_j\}) - P_0(\{a_i, a_j\}) \right), \text{ [by standard probability property]}
\]
\[
= \left(1 - P_0(\{a_i\}) \right) - \left(1 - P_0(\{a_i, a_j\}) - P_0(\{a_j\}) \right), \quad \text{[by Condition 2, see above]}
\]
\[
= \left(1 - P_0(\{a_i\}) \right) - \left(P_0(\{a_j\}) - P_0(\{a_i, a_j\}) \right), \text{ [by standard probability property]}
\]
\[
= 1 - P_0(\{a_i\}) - P_0(\{a_j\}) + P_0(\{a_i, a_j\}) > 0, \quad \text{[by Condition 3 and Lemma A3, where $C = S = \{a_i, a_j\}$]},
\]

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 $P_i(S \setminus \{a_i\}) = P_i(S)$ (note, by Condition 2 we have that for all $S' \subseteq N$ s.t. $a_1, a_2 \in S'$: $P_i(S' \setminus \{a_j\}) = P_i(S')$), 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:

\[
\delta_{ij} = 1 \Rightarrow P_j(S \setminus \{a_i\}) > P_j(S), \text{ [for some $S \subseteq N$]}
\]
\[
\Rightarrow P_j(\{a_j, a_k\}) > P_j(\{a_i, a_j, a_k\}), \quad \text{[by Condition 2]}
\]
\[
\Rightarrow P_i(\{a_i, a_k\}) = P_i(\{a_i, a_j, a_k\}), \quad \text{[by Condition 1]} \tag{A12}
\]
\[
\Rightarrow P_i(\{a_i\}) = P_i(\{a_j, a_k\}), \quad \text{[by Condition 2]}, \tag{A13}
\]
\[
\delta_{jk} = 1 \Rightarrow P_k(S \setminus \{a_j\}) > P_k(S), \text{ [for some $S \subseteq N$]}
\]
\[
\Rightarrow P_k(\{a_i, a_k\}) > P_k(\{a_i, a_j, a_k\}), \quad \text{[by Condition 2]}
\]
For $\forall S \subseteq N: \sum_{a \in S^+} \mathbb{P}_r(S) = 1$, 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 $S_1 = \{a_i, a_j\}$: $\mathbb{P}_i(S_i) + \mathbb{P}_j(S_1) + \mathbb{P}_0(S_1) = 1$

\[ \Rightarrow \mathbb{P}_i(\{a_i\}) + \mathbb{P}_j(S_1) + \mathbb{P}_0(S_1) = 1, \quad \text{by Equation (A13)} \]

\[ \Rightarrow \mathbb{P}_i(\{a_i\}) + \mathbb{P}_j(S_1) + \mathbb{P}_0(S_1) = 1, \quad \text{by Equation (A14)} \]

\[ \Rightarrow \mathbb{P}_j(S_1) = \mathbb{P}_0(\{a_i\}) - \mathbb{P}_0(S_1), \quad \text{[by standard probability property].} \quad (A18) \]

For $S_2 = \{a_j, a_k\}$: $\mathbb{P}_j(S_2) + \mathbb{P}_k(S_2) + \mathbb{P}_0(S_2) = 1$

\[ \Rightarrow \mathbb{P}_j(\{a_j\}) + \mathbb{P}_k(S_2) + \mathbb{P}_0(S_2) = 1, \quad \text{by Equation (A15)} \]

\[ \Rightarrow \mathbb{P}_j(\{a_j\}) + \mathbb{P}_k(S_2) + \mathbb{P}_0(S_2) = 1, \quad \text{by Equation (A16)} \]

\[ \Rightarrow \mathbb{P}_k(S_2) = \mathbb{P}_0(\{a_j\}) - \mathbb{P}_0(S_2), \quad \text{[by standard probability property].} \quad (A19) \]

For $S_3 = \{a_i, a_k\}$: $\mathbb{P}_k(S_3) + \mathbb{P}_i(S_3) + \mathbb{P}_0(S_3) = 1$

\[ \Rightarrow \mathbb{P}_k(\{a_k\}) + \mathbb{P}_i(S_3) + \mathbb{P}_0(S_3) = 1, \quad \text{by Equation (A17)} \]

\[ \Rightarrow \mathbb{P}_k(\{a_k\}) + \mathbb{P}_i(S_3) + \mathbb{P}_0(S_3) = 1, \quad \text{by Equation (A12)} \]

\[ \Rightarrow \mathbb{P}_i(S_3) = \mathbb{P}_0(\{a_k\}) - \mathbb{P}_0(S_3), \quad \text{[by standard probability property].} \quad (A20) \]

For $S_4 = \{a_i, a_j, a_k\}$: $\mathbb{P}_i(S_4) + \mathbb{P}_j(S_4) + \mathbb{P}_k(S_4) + \mathbb{P}_0(S_4) = 1$

\[ \Rightarrow 0 = \mathbb{P}_0(\emptyset) - \mathbb{P}_0(\{a_i\}) - \mathbb{P}_0(\{a_j\}) - \mathbb{P}_0(\{a_k\}) - \mathbb{P}_0(S_1) + \mathbb{P}_0(S_2) \]

\[ + \mathbb{P}_0(S_4) - \mathbb{P}_0(S_4), \quad \text{[since } \mathbb{P}_0(\emptyset) = 1, \text{ and by Equations (A18)-(A20)} \]

\[ > 0, \quad \text{[by Condition 3 and Lemma A3, where } C = S = S_4],} \]
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 \( \succ \) over products in \( N \), s.t. \( a_i \succ a_j \) 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 \( \succ \), i.e., for all \( a_i \in N \) we have that \( \delta_{0i} = 0 \):

\[
P_0(\emptyset) - P_0(\{a_0\}) > 0, \quad \text{[by Condition 3 and Lemma A3, where } C = S = \{a_0\}]\]

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

Next, we prove that

\[
P_r(S) = P_0(S' \setminus \{a_r\}) - P_0(S'), \quad \forall a_r \in S \quad \text{s.t. } S \subseteq N,
\]

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., \( S' = \{a_j \in S : a_j \succ a_r\} \cup \{a_r\} \). 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 \( P_r(\{a_r\}) = 1 - P_0(\{a_r\}) = P_0(\emptyset) - P_0(\{a_r\}) \). 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. \( a_r \succ a_j \). Then by definition of the binary relation \( \delta \) we have that \( P_j(S \setminus \{a_r\}) > P_j(S) \), and the result now follows:

\[
P_r(S) = P_r(S \setminus \{a_j\}), \quad \text{[by Condition 1]}
\]

\[
= P_0(S' \setminus \{a_r\}) - P_0(S'), \quad \text{[by induction hypothesis,}
\]

and note that \( a_j \notin S' \) since \( a_r \succ a_j \).

**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 w.l.o.g. \( a_{p-1} \succ \ldots \succ a_2 \succ a_1 \succ a_r \). Assuming \( a_r \in S \), we can now establish the following chain of equalities:

\[
P_r(S) = 1 - P_0(S) - \sum_{i=1}^{p-1} P_i(S)
\]

\[
= -P_0(S) + P_0(\emptyset) - \sum_{i=1}^{p-1} P_i(\{a_r, a_1, a_2, \ldots, a_{p-1}\})
\]

\[
= -P_0(S) + P_0(\emptyset) - \sum_{i=1}^{p-1} P_i(\{a_i, a_{i+1}, \ldots, a_{p-1}\}), \quad \text{[by Condition 1]}
\]

\[
= -P_0(S) + P_0(\emptyset) - \sum_{i=1}^{p-1} \left( P_0(\{a_i, a_{i+1}, \ldots, a_{p-1}\}) - P_0(\{a_i, a_{i+1}, \ldots, a_{p-1}\}) \right),
\]

Electronic copy available at: https://ssrn.com/abstract=3410019
Let us denote two particular sets \( \hat{S} \) and \( \hat{S}' \) as follows: \( \hat{S} = N \setminus \{ S' \setminus \{ a_r \} \} \), \( \hat{S}' = N \setminus S' \). We can now establish the following chain of equalities:

\[
P_r(S) = P_0(S' \setminus \{ a_r \}) - P_0(S') \\
= P_0(S' \setminus \{ a_r \}) + \left( \sum_{C \subseteq \hat{S}} \sum_{X \subseteq C} (-1)^{|C| - |X|} \cdot P_0(N \setminus X) - P_0(N \setminus \hat{S}) \right) - P_0(S') \\
\quad \text{[by invoking Lemma A1, where } Z = N, Y = \hat{S}, P = C, \text{and } f(Z \setminus Y) = P_0(N \setminus \hat{S}) \text{]}
\]

\[
= \sum_{C \subseteq \hat{S}} \sum_{X \subseteq C} (-1)^{|C| - |X|} \cdot P_0(N \setminus X) - \left( \sum_{C \subseteq S'} \sum_{X \subseteq C} (-1)^{|C| - |X|} \cdot P_0(N \setminus X) \right) - P_0(N \setminus \hat{S}') - P_0(S') \\
\quad \text{[by invoking Lemma A1, where } Z = N, Y = \hat{S}', P = C, \text{and } f(Z \setminus Y) = P_0(N \setminus \hat{S}') \text{]}
\]

\[
= \sum_{C \subseteq \hat{S}} \sum_{X \subseteq C} (-1)^{|C| - |X|} \cdot P_0(N \setminus X) - \sum_{C \subseteq S'} \sum_{X \subseteq C} (-1)^{|C| - |X|} \cdot P_0(N \setminus X) \quad \text{[since } N \setminus \hat{S}' = S' \text{]}
\]

\[
= \sum_{C \in (S' \cup \{ a_r \})} (-1)^{|C| - |\hat{S}'|} \cdot P_0(N \setminus X) - \sum_{C \subseteq S'} \sum_{X \subseteq C} (-1)^{|C| - |X|} \cdot P_0(N \setminus X) \quad \text{[since } \hat{S} = \hat{S}' \cup \{ a_r \} \text{]}
\]

\[
= \sum_{C \subseteq (S' \cup \{ a_r \})} \lambda(C), \quad \text{where } \lambda(C) = \sum_{X \subseteq C} (-1)^{|C| - |X|} \cdot P_0(N \setminus X)
\]

\[
= \sum_{C \subseteq S'} \lambda(C) \cdot I[a_r \in C] \cdot I[C \in (S' \cup \{ a_r \})]
\]

\[
= \sum_{C \subseteq S'} \lambda(C) \cdot I[a_r \in C] \cdot I[a_r \succ a_k \forall a_k \in S' \cap C, a_k \neq a_r]
\]

\[
= \sum_{C \subseteq S'} \lambda(C) \cdot I[a_r \in S' \cap C] \cdot I[a_r \succ a_k \forall a_k \in S' \cap C, a_k \neq a_r] \quad \text{[since we assume that } a_r \in S, \text{ otherwise the choice probability is 0],}
\]

which is exactly the equation to compute the probability to purchase \( a_r \in S \) for the offer set \( S \subseteq N \) under GCC 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 3 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 \( P_i(\cdot)_{\succ',\lambda} = P_i(\cdot)_{\succ,\lambda} \). 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 GCC choice rule, we have

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

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

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

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

which contradicts to \( P_i(\cdot)_{\succ',\lambda} = P_i(\cdot)_{\succ,\lambda} \). \( \square \)

**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 GCC and GCGC 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

\[
\mathcal{L}(\beta, \delta) = \sum_{t=1}^{T} \left[ \log \frac{e^{\beta X_{jt}}}{1 + e^{\beta X_{jt}}} + \sum_{a_k \in S_t: k \neq j} \delta_{kj} \log \frac{1}{1 + e^{\beta X_{kt}}}, \right],
\]

and the ML problem can be represented in the following way

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

\[\text{s.t.:} \quad \delta_{jk} + \delta_{kj} = 1, \quad \forall j, k, j \leq k, \]

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

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

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

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

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

$$\delta_{kjt} \log \frac{1}{1 + e^{\beta X_{kt}}} = \log \frac{1}{1 + e^{\beta X_{kt}}} = \log \frac{1}{1 + e^{\sum_i \tau_{ikjt} X_{ikt}}} = (\delta_{kjt} - 1) \log \left( \frac{1}{2} \right) + \log \frac{1}{1 + e^{\sum_i \tau_{ikjt} X_{ikt}}}.$$

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

$$\delta_{kjt} \log \frac{1}{1 + e^{\beta X_{kt}}} = 0 = -\log \left( \frac{1}{2} \right) + \log \frac{1}{1 + e^{0}} = (\delta_{kjt} - 1) \log \left( \frac{1}{2} \right) + \log \frac{1}{1 + e^{\sum_i \tau_{ikjt} X_{ikt}}}.$$

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

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

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

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

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

$$\tau_{ikj} \leq M \delta_{kj}, \quad \forall i, j, k,$$

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

$$\tau_{kj} \geq -M \delta_{kj}, \quad \forall j, k,$$

$$\delta_{jk} + \delta_{kj} = 1, \quad \forall j, k, \quad j \leq k,$$

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

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

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 (16) 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 \((\theta^i, \tau^i)\):

\[
D(\theta^i, \tau^i) = \{ \theta, \tau : \mathcal{L}(\theta^i, \tau^i) + \frac{\partial \mathcal{L}(\theta^i, \tau^i)}{\partial \theta} \theta + \frac{\partial \mathcal{L}(\theta^i, \tau^i)}{\partial \tau} \tau - \mu_1 \geq 0, \ \mu \in \mathbb{R}^1 \}.
\]

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

\[
\begin{align*}
\max_{\theta, \tau, \mu, \mu_1} & \quad \mu \\
\text{s.t.:} & \quad \mu_1 \leq \mathcal{L}(\theta, \tau), \\
& \quad A\theta + B\tau + C\delta^i \leq 0, \\
& \quad \mu = \mu_1 + E\delta^i.
\end{align*}
\]
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

$$\max_{\theta, \tau, \delta, \mu, \mu_1} \mu$$

s.t.: $(\theta, \tau, \mu_1) \in \Omega^i$,

$$A\theta + B\tau + C\delta \leq 0,$$

$$\mu = \mu_1 + E\delta,$$

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

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

Note that solving MINLP (P) reduces to solving MILP subproblem if we approximate convex constraint $\mu_1 \leq L(\theta, \tau)$ 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 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 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 \( (\theta^i, \tau^i, \delta^i) \).

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 ICC model with IRI Academic dataset. We limited the running time of the algorithms by 3 hours, and the precision was set to 1e-6. It follows from Figure A7 that the optimality gap of the outer approximation algorithm (1) to calibrate the ICC model is 3.3% on average over 20 product categories. On the other hand, it is shown in Figure A8 that the optimality gap of the cutting plane algorithm (2) to calibrate the ICC model is 4.5% on average over 20 product categories. Following these findings, we apply outer-approximation algorithm (1) to calibrate ICC model in our analysis as it provides significantly faster convergence to the optimal solution, which is consistent with previous studies.
A3.3. GCC Estimation Methodology: EM Algorithm

In this section we present the EM algorithm to calibrate the GCC 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 GCC model, after we reparametrize it by dividing all the transactions into $K$ segments, is given by

$$\log L(\theta, \gamma, >_\sigma) = \sum_{t=1}^{T} \log \left( \sum_{h=1}^{K} \gamma_h \theta_{h,j_t} \prod_{a_j \in S_t: a_j >_\sigma a_{j_t}} (1 - \theta_{h,j_t}) \right),$$ (A23)

where $\gamma_h \geq 0$ is the weight of the class $h$ a priori, s.t. $\sum_{h=1}^{K} \gamma_h = 1$; $S_t$ denotes the set of offered items at time $t$; $a_{j_t}$ 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 $E[\log L(x)|\hat{x}^{(0)}]$ (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 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., $>_\sigma^{(0)}$) and $\theta_{h,j}^{(0)}$ for all $h \in \{1, \ldots, K\}$, $a_j \in N^+$ are obtained by solving the following optimization problem:

$$\max_{\sigma, \theta_{h,j}} \sum_{t \in D_h} \log \theta_{h,j_t} + \sum_{a_j \in S_t: a_j >_\sigma a_{j_t}} \log (1 - \theta_{h,j_t}),$$

which is solved by using the outer approximation algorithm 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 $h$ (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 $(a_{j_t}, S_t)_{t=1}^{T}$:

$$P_{ht}^{(q)} = \Pr \left( t \in \Gamma_h | \sigma^{(q-1)}, \theta^{(q-1)}, \gamma^{(q-1)}, (a_{j_t}, S_t)_{t=1}^{T} \right)$$
As a result, conditional expected value of the log-likelihood function is given by
\[\sum_{r=1}^{K} \frac{\gamma_{r}^{(q-1)} \prod_{a_j \in S_r} (1 - \theta_{r,j})}{\gamma_{h}^{(q-1)} \prod_{a_j \in S_t} (1 - \theta_{h,j})}\]

**M-step:** first, we update class membership probabilities for every segment \(h \in \{1, 2, ..., 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 \(\theta\) and \(\sigma:\n\max_{\sigma,\theta} \sum_{h=1}^{K} \sum_{t=1}^{T} P_{ht}^{(q)} \log \left( \theta_{h,j} \prod_{a_j \in S_t: a_j \succ \sigma a_{ht}} (1 - \theta_{h,j}) \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 $h \in \{1, \ldots, K\}$, $a_j \in N^+$ are obtained by solving the following optimization problem:

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

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

**E-step:** we compute $P_{hi}^{(q)}$, 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 $(a_{jit}, S_{it})_{t=1}^{T_i}$:

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

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

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

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

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

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 ICC 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:

$$
\theta^{(q)}_{hj} = \frac{\sum_{t=1}^{T} P_{ht}^{(q)} \mathbb{I}[a_{jt} = a_j]}{\sum_{t=1}^{T} P_{ht}^{(q)} \mathbb{I}[a_{jt} = a_j] + \sum_{t=1}^{T} P_{ht}^{(q)} \mathbb{I}[a_j \in S_t, a_j \succ_{\sigma} a_{jt}]},
$$

Electronic copy available at: https://ssrn.com/abstract=3410019
which can be applied with aggregate level data (see Section A3.3.1), and

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

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

### A3.4. GCGC Model: Estimation Methodology

The GCGC (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 GCGC 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

\[
\log L(\theta, \gamma, \sigma) = \sum_{t=1}^{T} \log \left( \sum_{h=1}^{K} \gamma_h \theta_{h,j_t} \prod_{a_j \in S_t; a_j \sigma h a_{j_t}} (1 - \theta_{hj}) \right),
\]

(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 \( t \); \( a_{j_t} \) denotes the product purchased at time \( t \); \( T \) denotes the time horizon. Conceptually, we can obtain all the parameters of the GCGC model (i.e., distributions over the preference lists and considerations sets) by maximizing the log-likelihood function above for a sufficiently large \( K \).

Next we provide the initialization of the EM algorithm to calibrate GCGC 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 \( \theta_h^{(0)} \) for all \( h \in \{1, \ldots, K\} \), \( a_j \in N^+ \) are obtained by solving the following optimization problem:

\[
\max_{\sigma_h, \theta_h} \sum_{h \in D_h} \left( \log \theta_{h,j_t} + \sum_{a_j \in S_t; a_j \sigma_h a_{j_t}} \log(1 - \theta_{hj}) \right),
\]

which is solved by using the outer approximation algorithm for the ICC 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 $(a_{jt}, S_t)_{t=1}^T$:

$$P_{ht}^{(q)} = \frac{\gamma_{h}^{(q-1)} \prod_{a_j \in S_t} (1 - \theta_{h,j}^{(q-1)})}{\sum_{r=1}^{K} \gamma_{r}^{(q-1)} \prod_{a_j \in S_t} (1 - \theta_{r,j}^{(q-1)})}.$$  

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

$$\gamma_{h}^{(q)} = \frac{\sum_{t=1}^{T} P_{ht}^{(q)}}{T},$$

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 $h \in \{1, ..., K\}$:

$$\max_{\sigma_h, \theta_h} \sum_{t=1}^{T} P_{ht}^{(q)} \log \left( \theta_{h,jt} \prod_{a_j \in S_t} (1 - \theta_{h,j}) \right),$$

which is solved by using the outer approximation algorithm for the ICC 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 $\sigma_h^{(q)}$ 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:

$$\theta_{h,j}^{(q)} = \frac{\sum_{t=1}^{T} P_{ht}^{(q)} [a_{jt} = a_j]}{\sum_{t=1}^{T} P_{ht}^{(q)} [a_{jt} = a_j] + \sum_{t=1}^{T} P_{ht}^{(q)} [a_j \in S_t, a_j > h a_{jt}]}.$$

**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 $D_1, D_2, ..., D_K$. Consequently, we set $\gamma_h^{(0)} = |D_h|/(\sum_{d=1}^{K} |D_d|)$. Then $\sigma_h$ (i.e., $\succ_h$) and $\theta_{h,j}^{(0)}$ for all $h \in \{1, ..., K\}$, $a_j \in N^+$ are obtained by solving the following optimization problem:

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

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

**E-step:** we compute $P_{hi}^{(q)}$, 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 \( (a_{jt}, S_t)_{t=1}^{T_i} \):

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

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

\[
\gamma^{(q)}_h = \frac{\sum_{i=1}^{m} P^{(q)}_{ht}}{m},
\]

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 \( h \in \{1, ..., K\} \):

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

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

**A4. Case Study on the Car Sharing Dataset**

In this section we calibrate Logistic-based Consider-then-Choose (L-CC) and MNL models accounting for car features and discuss the modeling assumptions. 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.

**A4.1. Explanatory analysis**

We start this section by calibrating the single class L-CC 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-CC 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-CC model appear in the first column of Table A1. In the middle column, the table lists the average marginal effects (AME) of the L-CC 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 A1 presents the estimates of the MNL model parameters. However, the interpretation of the $\beta$ vector for L-CC and MNL models is different. The parameters of the L-CC model listed in the left column of Table A1 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 A1 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-CC and MNL models are not aligned, i.e., the covariates that increase (or decrease) the likelihood of considering the car under L-CC 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-CC 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-CC model. The price and the car age coefficients are statistically significant and negative for both L-CC 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 important role during the formation of the consideration set in the L-CC 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-CC model and on the choice probabilities under the MNL model. Models
Table A1 Logistics-based Consider-then-Choose (L-CC) 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.
Table A2  Statistical significance of three groups of car attributes: (1) car type and features, (2) car location type and accessibility, and (3) car brand.

<table>
<thead>
<tr>
<th>Excluded groups</th>
<th>Log-like</th>
<th>AIC</th>
<th>BIC</th>
<th>LR</th>
<th>Wald</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1  Car types and features</td>
<td>-38664.7</td>
<td>77403.3</td>
<td>77735.3</td>
<td>457.04</td>
<td>449.39</td>
</tr>
<tr>
<td>L-CC Model 2  Car location type and accessibility</td>
<td>-38552.1</td>
<td>77202.2</td>
<td>77641.9</td>
<td>231.94</td>
<td>231.67</td>
</tr>
<tr>
<td>Model 3  Brands</td>
<td>-38840.7</td>
<td>77729.4</td>
<td>77944.7</td>
<td>809.07</td>
<td>764.90</td>
</tr>
<tr>
<td>MNL Model 1  Car types and features</td>
<td>-35587.4</td>
<td>71246.8</td>
<td>71600.3</td>
<td>1492.05</td>
<td>1385.99</td>
</tr>
<tr>
<td>Model 2  Car location type and accessibility</td>
<td>-35205.7</td>
<td>70507.5</td>
<td>70978.8</td>
<td>728.75</td>
<td>705.53</td>
</tr>
<tr>
<td>Model 3  Brands</td>
<td>-35534.9</td>
<td>71115.8</td>
<td>71341.6</td>
<td>1387.03</td>
<td>1259.70</td>
</tr>
</tbody>
</table>

1, 2, and 3 (both under L-CC 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 A2, the car type and features attributes are less statistically significant than car brand attributes under the L-CC 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 A2 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.

A4.2. Discussion of Model Estimation Assumptions

In this section we further discuss the assumptions imposed by the CC 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.

A4.2.1. Semiparametric approach. Using the semiparametric approach in order to calibrate the two stage CC 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 A7, 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 A7 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).

A4.2.2. 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 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 A7).

A5. Robustness check: Study based on synthetic data
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 rank-based model as a ground truth instead of MNL. Recall that in Section 5, we compared the predictions of the MNL model against ICC model to understand the conditions under which the MNL benchmark outperforms the consider-then-choose model in the presence of the noise in the offer sets.

The set up of the experiments in this Section is identical to the one in Section 5. As it was mentioned above, 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 A4 and A5 we present the heatmaps of the prediction scores under the rank-based model and the prediction scores improvements under ICC model versus rank-based model, respectively. And in Tables A3 and A4, we report the results for the regressions (22) and (23), 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., ICC model outperforms ground truth rank-based model under noisier regimes.

Figure A4  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 depth. We focus on the MAPE and RMSE scores in left and right panels respectively. The lower the score the better.
Table A3  Regression models where the dependent variable is the MAPE score under the rank-based model.

<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>Intensity</td>
<td>38.341***</td>
<td></td>
<td></td>
<td></td>
<td>48.526***</td>
</tr>
<tr>
<td></td>
<td>(23.312)</td>
<td></td>
<td></td>
<td></td>
<td>(15.904)</td>
</tr>
<tr>
<td>Intensity^2</td>
<td></td>
<td>32.639***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(20.358)</td>
<td></td>
<td></td>
<td>(-3.425)</td>
</tr>
<tr>
<td>Asymm</td>
<td></td>
<td>30.882***</td>
<td></td>
<td></td>
<td>42.516***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(5.470)</td>
<td></td>
<td></td>
<td>(29.661)</td>
</tr>
<tr>
<td>Shared</td>
<td></td>
<td></td>
<td>4.827*</td>
<td></td>
<td>13.450***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1.657)</td>
<td></td>
<td>(19.378)</td>
</tr>
<tr>
<td>const</td>
<td>7.072***</td>
<td>15.593***</td>
<td>17.891***</td>
<td>26.428***</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>0.733</td>
<td>0.677</td>
<td>0.131</td>
<td>0.014</td>
<td>0.955</td>
</tr>
<tr>
<td>Adj. R-squared:</td>
<td>0.732</td>
<td>0.675</td>
<td>0.127</td>
<td>0.009</td>
<td>0.954</td>
</tr>
</tbody>
</table>

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

Figure A5  Heatmap of the prediction scores improvements under ICC model versus rank-based model where each column corresponds to a particular noise intensity and each row corresponds to a particular noise depth. We focus on the MAPE and RMSE scores in left and right panels respectively.

A6. Tables and Figures
Table A4  Regression models where the dependent variable is the MAPE score improvement of the ICC model over the rank-based model.

<table>
<thead>
<tr>
<th>Category Shorthand</th>
<th>Expanded Name</th>
<th># vend</th>
<th>OS size</th>
<th># cust.</th>
<th># trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 blades</td>
<td>Blades</td>
<td>10</td>
<td>4.18</td>
<td>703</td>
<td>1084</td>
</tr>
<tr>
<td>2 cigets</td>
<td>Cigarettes</td>
<td>18</td>
<td>7.14</td>
<td>452</td>
<td>2343</td>
</tr>
<tr>
<td>3 coffee</td>
<td>Coffee</td>
<td>73</td>
<td>19.80</td>
<td>3101</td>
<td>11526</td>
</tr>
<tr>
<td>4 colder</td>
<td>Cold cereal</td>
<td>45</td>
<td>17.66</td>
<td>4438</td>
<td>26701</td>
</tr>
<tr>
<td>5 deod</td>
<td>Deodorant</td>
<td>36</td>
<td>14.55</td>
<td>3145</td>
<td>2383</td>
</tr>
<tr>
<td>6 diapers</td>
<td>Diapers</td>
<td>8</td>
<td>3.30</td>
<td>337</td>
<td>919</td>
</tr>
<tr>
<td>7 fzpizza</td>
<td>Frozen pizza</td>
<td>47</td>
<td>15.50</td>
<td>3460</td>
<td>13431</td>
</tr>
<tr>
<td>8 hotdog</td>
<td>Hot dogs</td>
<td>44</td>
<td>16.81</td>
<td>3318</td>
<td>8886</td>
</tr>
<tr>
<td>9 laundet</td>
<td>Laundry detergent</td>
<td>24</td>
<td>10.08</td>
<td>3196</td>
<td>8698</td>
</tr>
<tr>
<td>10 margbutr</td>
<td>Margarine/Butter</td>
<td>19</td>
<td>10.35</td>
<td>3474</td>
<td>14596</td>
</tr>
<tr>
<td>11 mayo</td>
<td>Mayonnaise</td>
<td>19</td>
<td>6.86</td>
<td>3761</td>
<td>8676</td>
</tr>
<tr>
<td>12 mustketc</td>
<td>Mustard</td>
<td>60</td>
<td>17.07</td>
<td>3728</td>
<td>9238</td>
</tr>
<tr>
<td>13 peanbutr</td>
<td>Peanut butter</td>
<td>25</td>
<td>7.99</td>
<td>3153</td>
<td>8059</td>
</tr>
<tr>
<td>14 champ</td>
<td>Shampoo</td>
<td>81</td>
<td>18.74</td>
<td>1466</td>
<td>2884</td>
</tr>
<tr>
<td>15 spagsauc</td>
<td>Spaghetti/Italian sauce</td>
<td>74</td>
<td>17.85</td>
<td>3473</td>
<td>11879</td>
</tr>
<tr>
<td>16 sugarsub</td>
<td>Sugar substitutes</td>
<td>17</td>
<td>5.05</td>
<td>750</td>
<td>1406</td>
</tr>
<tr>
<td>17 toitsu</td>
<td>Toilet tissue</td>
<td>13</td>
<td>7.66</td>
<td>3760</td>
<td>14411</td>
</tr>
<tr>
<td>18 toothbr</td>
<td>Toothbrushes</td>
<td>52</td>
<td>15.86</td>
<td>1115</td>
<td>1810</td>
</tr>
<tr>
<td>19 toothpa</td>
<td>Toothpaste</td>
<td>38</td>
<td>12.05</td>
<td>2110</td>
<td>4482</td>
</tr>
<tr>
<td>20 yogurt</td>
<td>Yogurt</td>
<td>32</td>
<td>9.84</td>
<td>3766</td>
<td>24096</td>
</tr>
</tbody>
</table>

Table A5  Summary statistics of the data used in IRI case study.
Figure A6  Heatmap of the prediction scores improvements under ICC model versus MNL where each column corresponds to a particular noise intensity and each row corresponds to a particular noise depth in the test dataset. For every scenario noise depth in the training dataset is 10% lower than in the test dataset. We focus on the MAPE and RMSE scores in left and right panels respectively.

Figure A7  Results of applying the outer-approximation algorithm to estimate ICC 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 A8 Results of applying the cutting plane algorithm to estimate ICC 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 A9 Decision tree for consideration set formation of the renters based on the car sharing datasets.
<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>
<tr>
<td>Car types</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<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>
<tr>
<td>Car location type and accessibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<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>
<tr>
<td>Car features</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<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>49.87%</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 A6: Descriptive statistics, the car sharing dataset.
Additional descriptive statistics:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of rentals</td>
<td>26,791</td>
</tr>
<tr>
<td>Number of car owners</td>
<td>514</td>
</tr>
<tr>
<td>Number of available alternatives (within 0.3 mile)</td>
<td>5.7</td>
</tr>
<tr>
<td>Rental duration (days)</td>
<td>0.62</td>
</tr>
<tr>
<td>Rental request in advance (days)</td>
<td>1.24</td>
</tr>
<tr>
<td>Price CV (averaging over car owners)</td>
<td>0.053</td>
</tr>
<tr>
<td>Average number of price modes</td>
<td>2.33</td>
</tr>
<tr>
<td>The most frequent price (percentage)</td>
<td>0.78</td>
</tr>
<tr>
<td>The second most frequent price (percentage)</td>
<td>0.16</td>
</tr>
<tr>
<td>Average number of car access modes</td>
<td>1.07</td>
</tr>
<tr>
<td>The most frequent car access (percentage)</td>
<td>0.99</td>
</tr>
<tr>
<td>Average number of car access hours modes</td>
<td>1.03</td>
</tr>
<tr>
<td>The most frequent car access hours (percentage)</td>
<td>0.99</td>
</tr>
<tr>
<td>Average number of car location type modes</td>
<td>1.10</td>
</tr>
<tr>
<td>The most frequent car location type (percentage)</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table A7 Additional descriptive statistics, the car sharing dataset.

Figure A10 The average prediction scores over 20 product categories under GCC and GCGC 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.
Figure A11  Scatter plots of the prediction score improvements of GCC model over GCGC 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.