Demand Estimation Under Uncertain Consideration Sets

Autoría ditelliana: Vulcano, Gustavo
Otros autores: Jagabathula, Srikanth, Mitrofanov, Dmitry
Fecha de publicación: 2023
Versión final Publicada en: Operations Research (ISSN: 0030-364X)
El presente documento es una versión aceptada del artículo publicado

¿Cómo citar este artículo?
Jagabathula, S., Mitrofanov, D., & Vulcano, G. (2023). Demand Estimation Under Uncertain Consideration Sets. Operations Research. https://doi.org/10.1287/opre.2022.0006

Demand Estimation under Uncertain Consideration Sets

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

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

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

Key words

History

Over the last two decades there has been growing interest in the operations management (OM) academic eld and in the industry practice to incorporate sophisticated demand models, which provide high-quality inputs for critical tasks such as inventory management, dynamic pricing and assortment planning. Examples of such models include the multinomial logit (MNL), the nested logit, the mixed logit and the latent class MNL, which are traditional in the marketing and economics elds but novel in terms of their applicability in operational contexts. These models have been widely studied resulting in the development of speci c 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 model (e.g., Farias et al. (2013)) and the exponomial model (e.g., Alptekinoglu 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 di erent 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 a platform) selected one option from a collection of alternatives or *offer set*, which is usually de ned as a full category assortment in the retail case, or the full set of available options within an arbitrary radius in spatial choice models (e.g., car sharing). Given the transaction data, most choice models are trained assuming that the chosen option is preferred over all the other products on o er. However, customers may not consider everything on o er. For instance, in retailing, a customer selecting from the co ee category may not evaluate the full assortment and might consider only a subset of products (e.g., decaf) in a choice instance. In online platforms, even if we arbitrarily de ne the o er set as the cars available within a 0.2-mile radius, an agent may evaluate only compact cars. If we ignore these consideration sets, we make the incorrect inference that the chosen product is preferred over products not even considered, leading to model bias. To deal with this issue, the so-called *consider-then-choose* (CTC) models have been proposed in the literature. These models posit that customers sample a consideration set and then choose the most preferred product from the intersection of the o er set and the consideration set (e.g., Howard and Sheth (1969), Alba and Chattopadhyay (1985), Hauser and Wernerfelt (1990)).

CTC models have been studied quite extensively in the marketing literature (e.g., see Roberts and Lattin (1997)). They also gained recent popularity within the OM literature to make assortment and pricing decisions (e.g., see Feldman et al. (2019); Aouad et al. (2020)).

Despite their richness, CTC models are often di cult to t in practice because customers consideration sets are not observed we know the customer choice and the o er set, but the consideration set itself could be any subset of the full category containing the chosen product. When consideration sets are not observed, CTC models may not be identi able and it is unclear what, if any, predictive advantage they o er over simply assuming that customers consider everything on o er as classic models do. Existing literature is mostly silent on these issues. It o ers broad empirical evidence that customers do indeed form consideration sets and studies speci c instances of CTC models. But it mostly takes it as given that rms should t CTC models over classic choice models when consideration sets are not observed.

In this paper, we systematically study a very general class of CTC models when we only observe customer choices and o er sets. Theoretically, we observe that the general CTC model class is equivalent to the random utility maximization (RUM) class of models (which assumes that customers consider everything on o er), indicating that CTC models span the same modeling scope as classic choice models. We also show that CTC models in general are not identi able from transaction data alone, but some restricted versions of them indeed are.

To empirically evaluate the CTC models, we develop techniques to estimate them from transaction data. Our numerical analysis on both synthetic and real-world data shows that CTC models outperform classic choice models when o er sets are not observed perfectly (i.e., they are noisy) and the noise is asymmetric between the training and test data. Here, the o er set *noise* associated with a product refers to a product being erroneously recorded as o ered or stocked out when in fact it was not. We say noise is *asymmetric* if the degree of the o er set noise associated with a product di ers between training and test datasets (e.g., because factors a ecting stockouts vary over time). On the other hand, when the noise is symmetric, CTC models have comparable performance to classic models, providing little to no predictive advantage over them.

Noise in o er sets is quite common in practice. In retailing, for instance, o er set descriptions are often unreliable because of potential inventory inaccuracies (e.g., DeHoratius and Raman (2008)). For example, Kang and Gershwin (2005) analyze the accuracy of inventory records of a global retailer and nd that only 51% of them match actual inventory on average, with the worst store experiencing more than 67% mismatch. In online sharing platforms, *offer* or *availability* sets are not even clearly de ned. For example, in the context of a peer-to-peer car-sharing platform, an o er set can be de ned as all the available car listings on the platform or more reasonably, as all

the car listings within a radius of the customer s location. Either de nition is arbitrary and prone to errors.

Furthermore, the so-called test o er set inputs used for predictions tend to be noisier (i.e., more di erent from the actual ones) than the training o er sets. Choice models must be given o er sets to make predictions but the rm faces a high degree of uncertainty about future o er sets. For example, a retailer is uncertain about next week s o er set because existing products may be depleted, stocked-out products may be replenished, or new products might be introduced. Future o er sets are even more uncertain for online sharing platforms because availability is determined in real-time not by the platform but by individual providers in the market whose decisions are di cult to predict. These observations make the sequence (i) consider, and (ii) choose, particularly important when modeling customers choices in practice.

Unlike existing literature which studies speci c instances of CTC models, we study the following general class. The population preferences are characterized by a joint distribution over rank lists of the full set of products and over subsets of the product universe (i.e., over the consideration sets). This joint distribution is common across all customers. In each choice instance, a customer samples a consideration set and a preference list, and purchases the most preferred product in the *choice set*, which results from the intersection between the sampled consideration set and the exhibited o er set, or does not purchase at all if neither of the considered products is o ered.

We make the following remarks about our model. First, we note that we use a product-based as opposed to a feature-based consideration set de nition. In a product-based consideration set de nition, the model directly speci es the probability of consideration for each product. By contrast, a feature-based de nition assumes that customers screen on features, considering only those products whose attributes are within pre-speci ed acceptable ranges (e.g., see Jagabathula and Rusmevichientong (2017)). Product-based consideration sets are more general and tractable to analyze. They can also readily subsume feature-dependence, as we illustrate in this paper. Second, a key distinguishing aspect of our model is that we allow the distribution over consideration sets to be general, in contrast to the bulk of the existing literature which has generally restricted it to belong to speci c classes. Third, while the joint distribution can be general in principle, for the purpose of estimation, we approximate it with an appropriately de ned mixture distribution.

The main goal of our proposal is to characterize business environments under which the use of CTC models may provide higher quality estimates to describe the choice behavior of the customer basis, compared to RUM-based estimates. Speci cally, we make the following contributions:

- *Statistical properties of CTC models.* Recall that the RUM class is equivalently described by a distribution over product preference lists (Block and Marschak (1960); Strauss (1979); Farias et al. (2013)), so that customers choose the most preferred product among the available ones according to the sampled ranking list.

  We first show that the CTC model class is indeed equivalent to the RUM class, in that the set of choice probabilities induced by the CTC model class is the same as those induced by the RUM class (see Proposition 1). Yet, despite the equivalent explanatory power, CTC may be more natural in terms of its practical usage when accommodating consideration set formation becomes relevant (e.g., in a latent way, when customers do not consider all the items on offer; or in an explicit way, when having access to surveys where customers reveal their consideration sets).

  Given the equivalency to the RUM class, CTC models are not fully identifiable. But what is somewhat surprising is that the marginal distribution over consideration sets is uniquely identified (see Proposition 3). We also investigate cases when consideration sets are small (see Corollary 1 and Proposition 4), and show that we can compute the marginal distribution over consideration sets efficiently. Furthermore, if we restrict all the customers to choose using the same preference list, then the ranking is also identifiable (see Proposition 5). We also establish conditions to verify whether the observed data are consistent with a specific instance of the CTC model class.
- *Methodology to estimate the parameters of CTC models.* For estimation, we approximate the general CTC model class with a mixture of what we call independent consideration set (ICS) models (see Section 4.1). An ICS model is specified by a single ranking and an independent distribution over consideration sets in which a customer samples a consideration set by including each product independently with a certain probability. We propose an expectation-maximization (EM) algorithm to estimate the mixture. For each mixture component, we propose an outer-approximation algorithm to ensure convergence to the maximum likelihood estimate.
- *Numerical experiments on synthetic data:* Because the consideration set formation is explicitly modeled in consider-then-choose type of frameworks, it is likely that its predictive performance is robust to the noise in the definition of the offer sets in comparison with competitive benchmarks (e.g., MNL and ranking-based models). We verify this conjecture by explicitly adding noise, erroneously including or excluding products from the offer sets. We find that the CTC models outperform classic models when the noise is asymmetric between the training and test offer sets (e.g., the sets of items that are exposed to noise in both training and test sets minimally intersect). Their performance on the other hand is comparable when noise is

symmetric. This result shows that when using choice models in practice, it is not su cient to choose the model that provides the best predictive performance on the historical data set, but it is important to also understand how the prediction task might di er from the training task in the data generation process (e.g., product availability).

- *Empirical analysis: better demand predictions for the retail industry and online platforms.* To support our ndings on synthetic data, we compare choice models under several real-world scenarios in retailing and online platforms where we are likely to face signi cant noise in the o er set de nitions. On real-world grocery transaction data, we nd that the relative performance of CTC models over the benchmarks improves as the level of noise in the test o er sets increases, for instance, when making long-term predictions. On a data set obtained from an online car-sharing platform, we show that CTC models outperform classic models because of the signi cant uncertainty in future availability sets. We also show that CTC models o er the exibility to use machine learning models, such as decision trees and random forests, in modeling the consideration set distribution. These models are more interpretable and lead to better prediction accuracy than standard choice-based demand models (up to 53.7% improvement in the root mean squared error metric).

We emphasize that CTC models have traditionally been used to account for uncertainty in what the customers actually considered. In practice, decision-makers also face uncertainty about what the customers were actually o ered. To the best of our knowledge, we are the rst to highlight the uncertainty in o er sets and show that CTC models can successfully deal with both types of uncertainty.

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

Consider-then-choose (CTC) models are built upon the key concept of *consideration sets*. This notion has recently gained attention in the OM literature, but is well studied in the marketing and psychology elds, 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 et al. (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 that 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 about 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 co ee. 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 aforementioned papers provide a foundation for the belief that the distribution over consideration sets in the CTC model, which encompasses all possible subsets, may be sparse in reality and hence, tractable to estimate.

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 justi cation 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 are 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 heuristics (Tversky (1972); Montgomery and Svenson (1976); Hauser (2014); Hogarth and Karelaia (2005)). Therefore, the theoretical groundwork from the elds of marketing and psychology strongly suggests that the set of products or services that people consider is highly unlikely to be the same as the set of o ered products.

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 o er 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 de ned by two elements: a collection of consideration sets and a collection of customer types represented by rank lists. Di erent constraints in the de nition of these collections lead to di erent special versions of the model (e.g., limiting the number of features that consumers use to lter 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 e cient.

Feldman and Topaloglu (2018) consider the assortment optimization problem under the MNL model when consideration sets for di erent 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 o er set. Jagabathula and Rusmevichientong (2017) propose a model where, rst, customers consider the set of products with prices less than a threshold, and then choose the most preferred product from the set considered. They develop a tractable nonparametric expectation maximization (EM) algorithm to t the model to transaction data and design an e cient algorithm to determine the pro t-maximizing combination of o er sets and prices. Jagabathula and Vulcano (2018) propose a framework to estimate individual consumer preferences under some heuristic rules used by consumers to form their consideration sets (e.g., they consider only products under promotion jointly with the ones purchased in the previous store visit). In spite of the recent attention caught by CTC models in the OM eld, most of the existing papers mainly focus on various optimization problems under the CTC choice rule, ignoring the question of when this type of model can stand out and outperform classical benchmarks. In contrast, in this paper, we want to better understand the application area of CTC models and develop a methodology to calibrate these models from sales transaction data.

Our proposal here follows a di erent 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 de ned by the realization of the *attention parameter* of every alternative. This attention parameter is equivalent to our *propensity parameter* when the o er 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 rst evaluated by Gallego and Li (2017), who verify in a case study in the airline industry that its ability to t booking data outperforms both the MNL and mixtures of MNLs in most of the markets evaluated. They also show that the related assortment optimization problem runs in polynomial time even with capacity constraints. In our

paper, we study a more general class of CTC models and focus on understanding their predictive power in di erent operations contexts.

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

We consider a universe of products { }, in addition to the no-purchase or outside option (we use the terms no-purchase option and outside option interchangeably throughout). We use to denote the set ∪{ }.

Customers arrive at the store sequentially. In each choice instance, a customer is presented with a subset ⊆ of products and chooses either one of the products in or the outside option . We let ℙ ( ) denote the probability that a customer chooses product ∈ and ℙ ( ) the probability that the customer chooses the outside option. Our goal is to model this choice process through a stochastic model that speci es all the choice probabilities {ℙ ( ): ∈ ⊆ }, where we use to denote the set ∪{ }. We assume that the choice probabilities satisfy the standard probability laws: ℙ ( ) ≥ 0 for all ∈ and $\sum_{j\in\ +}$ ℙ ( ) = 1 for all ⊆ .

To explicitly account for the fact that customers may not consider all products on o er before making a choice, we assume that their choice behavior follows a two-stage *consider-then-choose* (CTC) model. In the rst stage, the customer forms a consideration set ⊆ , and her own preference is realized (we describe this mechanism below), and in the second stage, selects either a product from the *choice set* ∩ or the outside option . In other words, the customer picks one element from the extended choice set (∩) = (∩)∪{ }.

In this model, for a product to be purchased, it must be both *offered and considered*. The seller restricts customers choices by deciding the set of products to o er. But the customer further restricts her choices to just the ones in her consideration set . Two usual reasons to justify this approach are: (i) the customer has strong unobserved preferences (which prevent her from ever buying certain products), or (ii) the cognitive overload that prevents her from evaluating all the products on o er before choosing.

To describe customer preferences, let $\mathscr{S}$ denote the set with cardinality ! of all full rankings or permutations of products in in which the outside option is ranked at the bottom. The *preference ordering* or *ranking* of the products in is described by a bijective ranking function : → {1 + 1} specifying a preference rank ( ) for each product . The preference ordering induces an antire exive, antisymmetric, and transitive preference relation ≻ , de ned

n a_0

C $\{a_1, a_2, a_3\}$ σ

ν σ, C S $\{a_1, a_2, a_5\}$ $S \cap C^{+}$ $\{a_0, a_1, a_2\}$

C σ a_1 a_1

as $\succ$ if and only if ( ) ( ). Since we are xing the position of the outside option to be at the bottom of each ranking, we have that ( ) = +1 for all $\in \mathscr{S}$. Customer preferences are described by a joint probability distribution $: \mathscr{S} \times 2$ $\rightarrow [0\ 1]$, where 2 denotes the collection of all subsets of the set . In each choice instance, when confronted with an o er set , a customer samples a ranking and a consideration set $\subseteq$ with probability ( ) and chooses the product $\arg\min\{$ ( ): $\in$ ($\cap$) $\}$.

Figure 1 illustrates the choice process for a particular store visit given a joint distribution function over consideration sets and full rankings from where the customer samples a consideration set = $\{$ $\}$ and a ranking = ( ), and where we are representing the ranking as a tuple with the products listed in the order of their preference. The seller exhibits set = $\{$ $\}$. In this choice instance, item is purchased. If the sampled consideration set had been = $\{$ $\}$ instead, then the customer would have gone for the no-purchase option.

Formally, the choice probability $\mathbb{P}$ ( ) under this model is given by

$$\mathbb{P}(\ ) = \sum_{\in \mathscr{S}_n} \sum_{\subseteq} (\quad) \cdot \mathbf{I}[\ \in (\ \cap\ )\ ] \cdot \mathbf{I}[\ \succ \quad \forall \quad \in (\ \cap\ ) \quad \neq\ ] \tag{1}$$

where $\mathbf{I}[\ ]$ is the standard indicator function taking the value 1 if condition is satis ed, and the value 0 otherwise. We further assume that the empty condition $= \varnothing$ is always satis ed. We say that choice data ($\mathbb{P}$ ( ): $\in$ $\in \mathcal{S}$), for some collection of subsets $\mathcal{S} \subseteq 2$, is *consistent* with an underlying CTC model if there exists a distribution $(\cdot\ \cdot)$ that satis es equation (1) for all $\in$ and $\in \mathcal{S}$.

The sampling of the ranking as part of the choice process resembles the general random utility maximization (RUM) framework (Block and Marschak, 1960). The RUM class is the most

studied choice-based demand class in the literature and includes popular models such as the MNL, the nested logit (NL), and the mixture of MNLs (MMNL) models. At the core, it assumes that customers sample utility values for products from some underlying joint distribution and choose the product with the highest utility. A natural question that may arise is how RUM and CTC classes are related. We establish the following result.

PROPOSITION 1. *The collection of choice probabilities* $\{\mathbb{P}(\ ):\ \in\ \subseteq\ \}$ *is consistent with an underlying RUM model if and only if it is also consistent with an underlying CTC model.*

The proof of the proposition is rather straightforward and is provided in Appendix A2 for the sake of completeness. The above result shows that the CTC model provides an equivalent but alternative parameterization of the RUM class and hence o ers the same explanatory power. Yet, as we show below, this parameterization allows for a more compact representation of customers preferences, particularly, when they do indeed form consideration sets. It also allows us to use product and customer-level features to explicitly explain how customers form consideration sets and use data sources other than transaction data, such as survey data, to help infer customers consideration sets. Finally, it also seems particularly appropriate for some speci c business contexts, as we will illustrate in subsequent sections.

The CTC model as stated above is not amenable to estimation, and therefore, we parameterize it as a mixture of what we call *independent consideration set* (ICS) models. As shown below, this parameterization is without loss of generality. A single-class ICS model is de ned by a *single* preference ordering and a product form consideration set distribution. As above, the outside option is ranked at the bottom of the preference list; that is, $(\ ) = \ +1$. Further, each product $\in$ is associated with an inclusion probability (or *propensity*) parameter $\in [0\ 1]$, which denotes the probability that a customer includes product in her consideration set. Customers make product inclusion decisions independently of each other, and therefore, the probability of sampling consideration set $\subseteq$ is $(\ ) = \prod_{\ j\in} \prod_{\ j\in\ \setminus} (1-\ )$. The ICS model was studied in Manzini and Mariotti (2014), and it can be shown that the probability of choosing product $\in$ from o er set is

$$\mathbb{P}(\ ) = \begin{cases} \prod_{\ i\in} (1-\ ) & \text{if} \quad = \\ \prod_{\ i\in\ \ i\succ_\sigma\ j} (1-\ ) & \text{otherwise} \end{cases}$$

where the rst expression corresponds to the probability of the event that none of the o ered products are considered (which results in $(\ \cap\ ) = \{\ \}$, and hence, the selection of the outside option), and the second expression corresponds to the probability of the event that product is considered but none of the products preferred over are. Finite mixtures of these models provide more modeling exibility, and in fact, subsume the entire class of CTC models, as shown in the proposition below.

PROPOSITION 2. *Every CTC model with joint probability distribution* $(\cdot\ \cdot)$ *over consideration sets and full rankings can be represented as a finite mixture of* $$ *ICS models with preference orders* $$ *, propensity parameters* $$ *, and mixture weights* $(\quad)$ *for* $\in\{1\quad\}$*, and for some value of* $$ *.*

The proof of this result is rather intuitive and the argument is as follows. Consider any CTC model with underlying joint distribution $(\cdot\ \cdot)$ over consideration sets and preference lists. Let $$ denote the support size of the distribution $(\cdot\ \cdot)$, so that $(\quad)\quad 0$ for some collection of tuples $\{(\quad):\ =1\quad\}$. Then, we can de ne $$ independent ICS models with parameters $$ and $$, for every $\in\{1\quad\}$ such that $\quad=1$ if item $$ belongs to the consideration set $$, and 0 otherwise. The mixing distribution over the $$ classes is de ned by $(\quad)$ such that $(\quad)=(\quad)$.

Our next goal is to study the problem of identifying customer consideration sets from sales transaction data alone.

In this section, we derive various conditions under which the CTC models are identi able. To this end, our goal is to investigate if we can uniquely infer the parameters of the underlying CTC model from the collection of choice probabilities $\{\mathbb{P}(\ ):\ \in\quad\subseteq\ \}$.

Our rst results shows that the marginal distribution over consideration sets, de ned as $(\ )=\sum_{\in\mathscr{S}_n}(\quad)$, is uniquely identi able from the observed choice probabilities alone. Speci cally, we have the following result:

PROPOSITION 3. *Suppose that a collection of choice probabilities* $\{\mathbb{P}(\ ):\ \subseteq\ \}$ *are consistent with an underlying CTC model* $(\cdot\ \cdot)$*. Then, the marginal distribution* $(\cdot)$ *over consideration sets, defined as* $(\ )=\sum_{\in\mathscr{S}_n}(\quad)$*, is uniquely identified:*

$$(\ )=\sum_{\subseteq}(-1)^{|\ |-|\ |}\,\mathbb{P}(\ \setminus\ )\quad\forall\quad\subseteq$$

This is a rather surprising result. Intuitively, it seems that consideration sets cannot be identi ed from choice observations alone: if all we know is that a customer chose product $$ from o er set $$, then the customer may have considered any subset of products containing product $$, all the way from the singleton set $\{\ \}$ to the entire product universe $$; we would not have any basis to select one subset over the other. While this intuition is correct for all the products in $$, the choice of the outside option reveals more.

To see this, note that in the de nition of the CTC model class, we x the position of the outside option to be at the bottom (i.e., at position $+1$) of all the preference lists. This does not restrict

the generality of the CTC model class (cf. Proposition 1), yet provides us with valuable information on the customers consideration sets. Speci cally, because the outside option is the least preferred option, a customer chooses the outside option *only if* the sampled consideration set is disjoint from the o ered set. In other words, if is the o er set, then a customer choosing the outside option must have sampled a consideration set $\subseteq \setminus$. We, therefore, have that $\mathbb{P}( ) = \sum_{\subseteq \setminus} ( )$. For example, when $=$, it follows that a customer will choose the outside option only if they sample the empty set as the consideration set; therefore, $\mathbb{P}( ) = (\emptyset)$. Similarly, if $= \setminus\{ \}$, then the customer chooses the outside option only if their sampled consideration set is $\{ \}$ or the empty set, which implies that $\mathbb{P}(\setminus\{ \}) = (\{ \}) + (\emptyset)$. From this, we can back out the value of $(\{ \})$. More generally, we use a particular form of the inclusion-exclusion principle stated in Graham (1995) to back out the consideration set distribution. For any nite set , if $: 2 \to \mathbb{R}$ and $: 2 \to \mathbb{R}$ are two real-valued set functions de ned on the subsets of such that $( ) = \sum_{\subseteq} ( )$ for every $\subseteq$, then the inclusion-exclusion principle states that $( ) = \sum_{\subseteq} (-1)^{| |-| |} ( )$ for every $\subseteq$. Our result then follows from replacing $( )$ with $( )$ and de ning $( ) = \mathbb{P}(\setminus) = \sum_{\subseteq} ( )$. For completeness, we provide an alternative proof of this result from the rst principles in Appendix A2.

Although the marginal distribution over consideration sets is identi able, the marginal distribution over rankings itself is not. This follows from an existing result in (Sher et al., 2011), which shows that a general distribution over rankings (which is one of the building blocks of our CTC class) is not uniquely identi able from choice probabilities alone for ≥ 4. Therefore, the mixture of ICS models is also non-identi able given the result stated in Proposition 2.

Empirical evidence in the marketing literature suggests that the size of the consideration sets for most customers in di erent categories is relatively small, e.g., Hoyer (1984) concludes that the median number of laundry detergents that a consumer considers before making a purchase is one. When the size of consideration sets is bounded above by , with , it follows from Proposition 3 that to recover , we need choice probabilities under o er sets of size $-$ or larger.

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

1 n $\mathbb{P}_0$ S $\sum_{C \subseteq N \setminus S} \lambda\, C$ n S

When the consideration sets are small, Corollary 1 argues that it is su cient to collect choice probabilities for the no-purchase alternative from large o er sets. In many applications, however, rms cannot o er very large o er sets to their customers because of space constraints either in a physical store or on the relevant locations (e.g., top slots) of a website. The next proposition shows that when the consideration sets are of size at most , then the consideration set distribution can be recovered using choice probabilities of o er sets of size at most .

PROPOSITION 4. *Consider a CTC model in which customers sample consideration sets of size at most for some* $1 \leq \leq$. *Let* $\{\mathbb{P}(\ ): \subseteq \ | \ | \leq \}$ *be a collection of choice probabilities that are consistent with such a CTC model. Then, we have*

$$(\ ) = \sum_{\subseteq} \sum_{\supseteq \ \cup} (-1)^{\ | \ |-|} \ ^{|} \cdot \mathbf{I}[| \ \cup \ | \leq \ | \ |] \cdot \mathbb{P}(\ )$$

where denotes the symmetric difference $(\ \setminus \ ) \cup (\ \setminus \ )$.

The proof of the proposition is involved. It requires establishing several combinatorial identities. We present it in Appendix A2.

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

More precisely, we assume that the GCS model is de ned by a single preference ordering and a distribution $(\cdot)$ over consideration sets. As before, the outside option is ranked at the bottom of the preference list, and a customer samples a consideration set according to and then chooses the most preferred product according to from the set $(\ \cap \ )$. With this restriction, we show that the choice rule is also identi able:

PROPOSITION 5. *Suppose that the collection of choice probabilities* $\{\mathbb{P}(\ ): \in \ | \ | \leq 2\}$ *are consistent with an underlying GCS model. Then, for all* $1 \leq \leq$ *and* $\neq$, *we have that if* $\mathbb{P}(\{\ \}) \ \mathbb{P}(\{\ \})$, *then* $(\ ) \ (\ )$.

The argument is intuitive. Given that there is a single preference list shared by all the consumers, if, for any given alone, the presence of another lowers its probability of being chosen, it is because is preferred over it.

Unsurprisingly, the GCS model is not as rich as the CTC model class. In particular, it is a special case of the RUM choice rule.

PROPOSITION 6. *The GCS choice model is a special case of the RUM choice rule, that is,* $\subseteq$ *, but* $\neq$.

The proof of the proposition is provided in Appendix A2 and uses the result in Proposition 5. It exhibits an example of a choice model that belongs to the RUM class but not to the GCS class.

Going back to Proposition 5, we assumed therein that the collection of observed choice probabilities is consistent with an underlying GCS model. To verify if that is indeed the case, we establish here a set of necessary and sufficient conditions that the observed choice probabilities must satisfy.

PROPOSITION 7. *The collection of choice probabilities* $\{\mathbb{P}(\ ):\ \in\ \subseteq\ \}$ *is consistent with a GCS model with unique parameters* *and consideration set distribution* *such that* $(\ )$ 0 *for all* $|\ |\leq 3$ *and* $(\ )\geq 0$ *whenever* $|\ |$ 3 *if and only if it satisfies the following conditions:*

Condition 1. For all offer sets $\subseteq$ *and* $\in$ *such that* $\neq$ *: if* $\mathbb{P}(\ \setminus\{\ \})\neq\mathbb{P}(\ )$*, then it must hold that* $\mathbb{P}(\ \setminus\{\ \})=\mathbb{P}(\ )$.

Condition 2. For all offer sets $'\subseteq$ *and* $\in\ \cap\ '$ *such that* $\neq$ *:*
if $\mathbb{P}(\ \setminus\{\ \})$ $\mathbb{P}(\ )$*, then it must hold that* $\mathbb{P}(\ '\setminus\{\ \})$ $\mathbb{P}(\ ')$*; and*
if $\mathbb{P}(\ \setminus\{\ \})=\mathbb{P}(\ )$*, then it must hold that* $\mathbb{P}(\ '\setminus\{\ \})=\mathbb{P}(\ ')$.

Condition 3. For all offer sets $\subseteq$ *, we have that* $\sum_{\subseteq}(-1)^{|\ |-|\ |}\mathbb{P}(\ \setminus\ )\geq 0$ *with a strict inequality when* $|\ |\leq 3$.

Proposition 7 is similar to the set of conditions established in Manzini and Mariotti (2014) (see Theorem 1) for the case when the consideration set distribution has the product form due to the independence of the attention (or propensity) parameters. Our result consists of new conditions applied to a general consideration set distribution . Condition 1 is similar to the I-Asymmetry assumption in Manzini and Mariotti (2014), which states that either product influences the sales of product 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 influences the sales of product in one other set, then it must continue to do that in all the other sets. That is, the direction of influence is consistent across all the other sets. Condition 3 is a technical restriction to ensure the existence of a valid probability distribution function over the consideration sets. The strict inequality in Condition 3 is needed to ensure that the preference list over products in satisfies the transitivity requirement. The proof of Proposition 7 is presented in Appendix A2. As it can be observed therein, establishing necessity is straightforward, but establishing sufficiency requires significant work.

In this section, we propose techniques to estimate the general CTC model and its restricted versions presented above, from sales transaction data. The building block is an algorithm to t a single-class, ICS model, which we then extend using the expectation-maximization (EM) framework to t a nite mixture of them.

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

We highlight here that the CTC model is amenable to incorporating product features. In Appendix A3.1 we illustrate how to do it using three popular methods in machine learning logistic, decision tree, and random forest- regressions.

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

$$\mathscr{L}(\quad) = \sum \left[\log \quad _t + \sum_{k \in \atop / \ t} [\quad _t \log(1 - \quad)] \right]$$

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

$$\begin{aligned} & \max_{\boldsymbol{\theta}\, \boldsymbol{\delta}} \ \mathscr{L}(\quad) & & \\ \textbf{s.t.:} \quad & + \quad = 1 \quad \forall & & (2) \\ & + \quad + \quad \leq 2 \quad \forall \qquad \neq \quad \neq & & (3) \\ & \in \{0 \ 1\} \quad \forall & & (4) \\ 0 \leq & \quad \leq 1 \quad \forall & & (5) \end{aligned}$$

where constraints (2) and (3) ensure that indeed represents a total order. In particular, the set of constraints (2) ensures that either is preferred over or vice versa, and the set of constraints (3) imposes the total ordering among any three products.

4.1.1. Estimation methodology. To be able to solve the above problem, we reformulate it as follows. First, we introduce a new variable de ned as $=$, $\forall$ and rewrite the likelihood function in the following way

$$\mathscr{L}(\quad) = \sum \left[\log \quad _t + \sum_{k \in \atop / \ t} \log(1 - \quad _t) \right]$$

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

$$\begin{aligned} \max_{\boldsymbol{\theta}\,\boldsymbol{\tau}\,\boldsymbol{\delta}} \quad & \mathscr{L}(\quad) \\ \textbf{s.t.:} \quad & \leq \qquad \forall && (6)\\ & \leq \qquad \forall && (7)\\ & \geq \quad + \quad -1 \quad \forall && (8)\\ & \geq 0 \quad \forall && (9)\\ & (\quad) \text{ satisfy } (2)-(5) \end{aligned}$$

where linear constraints (6)-(9) ensure that $\quad = \quad, \forall \quad$, given that $\quad$ is a binary variable, and constraints (2) and (3) ensure again a total order on . We reformulate the MLE problem to have a linear objective function:

$$\max_{\boldsymbol{\theta}\,\boldsymbol{\tau}\,\boldsymbol{\delta}} \quad \text{subject to } (\qquad) \text{ satisfy } (2)-(5) \text{ and } (6)-(9) \tag{10}$$

$$\leq \mathscr{L}(\quad) \tag{11}$$

Note that if we know the ranking , then optimization problem (10) reduces to solving a globally concave maximization problem with a unique, closed form solution given by

$$= \frac{\sum \quad \mathbb{I}[\ _t = \ ]}{\sum \quad \mathbb{I}[\ _t = \ ] + \sum \quad \mathbb{I}[\ \in \qquad \succ \ _t]} \tag{12}$$

Next, we show how to apply the outer-approximation method of Duran and Grossmann (1986) to solve the optimization problem (10)-(11). The proposed algorithm e ectively exploits its structure, where we have linearity of the constraints involving the binary variables , and convexity of the non-linear constraint (11) which only depends on continuous variables and . 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 (11)) 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 $\mathcal{C}[\quad]$ denote a constraint which is a linear approximation of the constraint (11) at a point ( ), i.e.,

$$\mathcal{C}[\quad] := \left\{ \quad \leq \mathscr{L}(\quad) + \sum \frac{1}{\ _t} \cdot (\ _t - \ _t) + \sum \sum_{k \in \ / \ t} \frac{1}{\ _t - 1} \cdot (\ _t - \ _t) \right\}$$

Then we de ne the following optimization problem

$$\max_{\boldsymbol{\theta}\,\boldsymbol{\tau}\,\boldsymbol{\delta}} \quad (13)$$

$$\text{subject to } (\qquad) \text{ satisfy } (2)-(5) \text{ and } (6)-(9)$$

$$\mathcal{C}[\quad] \;\forall\, (\quad) \in \mathcal{A} \quad (14)$$

$$\leq \quad \leq$$

where we have replaced constraint (11) with the nite collection of linear constraints (14) at points in a set $\mathcal{A}$ that is incrementally built. It follows from the convexity of constraint (11) that every point that satis es constraint (11) also satis es the collection of constraints (14) for every nite set $\mathcal{A}$. We thus obtain an outer approximation. We also add bounds and to the log-likelihood function, which will be chosen in each iteration to tighten the interval containing the solution. This outer approximation de nes the optimization subproblem as an MILP. Because of the potentially many continuous points required for outer-approximation, we solve a sequence of MILPs to build up increasingly tight relaxations of the original MINLP. The algorithm to calibrate the ICS model is provided below.

ICS model calibration algorithm

Input Given sales transaction data, do:

Step 1 Sort products in decreasing number of sales and let $^{(0)}$ (i.e., $^{(0)}$) denote the corresponding ranking. Compute $^{(0)}$ using equation (12) given $^{(0)}$.

Step 2 Obtain all possible rankings $^{(1)}$ $^{(2)}$ $^{(m)}$, by swapping positions of any pair or two pairs of items in $^{(0)}$. Compute $^{(i)}$ using equation (12) given $^{(i)}$ for all $\in\{1\ 2\ \}$.

Step 3 Set $^{(i)}_{kj} := {}^{(i)}_{kj}\ {}^{(i)}_{k}$, $\forall$ and for all $\in\{0\ 1\ 2\ \}$. Recall that $^{(i)}$ is an alternative parameterization of the ranking $^{(i)}$. Set $_L := \min_{0\leq i\leq m} \mathscr{L}(\ ^{(i)}\ ^{(i)})$ and $_U := \infty$. Set := .

Step 4 While $|\ _U - \ _L|$ and running time did not exceed the limit, do:

- Set := $+1$.
- Solve optimization problem (13) with set $\mathcal{A}\{(\ ^{(k)}\ ^{(k)})\}|_{k=0}^{i-1}$ and obtain solution $^{(i)}$ $^{(i)}$. Note that in each iteration, we only add one constraint to the optimization problem solved in the previous iteration.
- Update $^{(i)}$ using equation (12) given $^{(i)}$.
- Set $^{(i)}_{kj} := {}^{(i)}_{kj}\ {}^{(i)}_{k}$, $\forall$.
- Set $_L :=$ $\{\mathscr{L}(\ ^{(i)}\ ^{(i)})\ _L\}$ and $_U =\ ^{(i)}$.

Endwhile

Step 5 Find solution $^* =\ ^{(i^*)}$ and $^* =\ ^{(i^*)}$ where $^* := \operatorname{argmax}_{0\leq k\leq i} \mathscr{L}(\ ^{(k)}\ ^{(k)})$.

Step 6 Stop.

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

Empirically, we analyze the performance of the proposed algorithm to estimate the ICS model on the IRI Academic dataset (to be described later in Section 6). We limited its running time to 3 hours, and the precision was set to $=1 \ -6$. It follows from Figure A4 in Appendix A3 that the optimality gap of the outer approximation algorithm (1) to calibrate the ICS model is 3.3% on average over 20 product categories, determined by the time limit, which indicates that this algorithm provides quite a reasonable performance in our setting. We also implemented a cutting plane method as an alternative algorithm (details also provided in Appendix A3), but the results were of slightly lower quality.

To make estimation tractable, we represent the general distribution $(\cdot \ \cdot)$ using a nite mixture of ICS models and then apply the EM algorithm. In a nutshell, this method starts from arbitrary parameter estimates $\mathbf{x}$. Then, it computes the conditional expected value of the log-likelihood function $\mathrm{E}[\log \mathscr{L}(\mathbf{x})|\mathbf{x} \ ]$ (the E , expectation, step). Next, the resulting expected log-likelihood function is maximized to compute new estimates $\mathbf{x}$ (the M , maximization, step), and both steps are repeated within a loop until convergence or a time limit is reached, to get a sequence of estimates $\{\mathbf{x} \quad =1 \ 2 \ \}$.

The log-likelihood function to be maximized can be represented in the following way:

$$\log \mathscr{L}(\quad) = \sum \log \Big(\sum \quad {}_{t} \prod_{\substack{j \in \ t \\ j \succ_h \ j_t}} (1 - \quad) \Big)$$

where ≥ 0 is the weight of the class , s.t. $\sum \quad = 1$; denotes the set of o ered items at time ; ${}_{t}$ denotes the product purchased at time ; and denotes the number of transactions.

We next brie y outline the E-step and M-step of every iteration and how we start the algorithm in the context of our CTC estimation problem. Further speci c details are relegated to Appendix A3.5.

Initialization: we randomly allocate sales transaction to one of the classes which allows us to compute initial mixing point probabilities , and estimate initial parameters: and for all $\in \{1 \quad \}$. To this end, we use the outer approximation algorithm for the ICS model on each class.

Then, we iterate over the sequence of the following E- and M-steps:

E-step: we compute the probability of every transaction at time to come from a segment customer, based on the parameter estimates (¯ ¯ ¯) of the previous iteration, and the transaction data.

M-step: rst, we update the mixing distribution ( ) of every segment ∈{1 2 } based on the E-step membership probabilities, and then optimize the resulting conditional expected value of the log-likelihood function by using the outer approximation algorithm on each segment. This way we obtain and for all ∈{1 }.

Even though it is well acknowledged that the convergence of EM algorithms is not guaranteed a priori, it was veri ed consistently in all our experiments.

A key observation is that we can calibrate the CTC model by estimating the following parameters at the segment level: (1) membership probability , (2) propensity parameter , and (3) ranking . Therefore, the approach is convenient when we have a small to moderate support of customer segments.

In a similar way, we can also calibrate the GCS model with the EM algorithm (see Appendix A3.4 for details). Note that it is very straightforward to build upon Proposition 2 and show that GCS model can also be represented as a mixture of ICS models sharing the same rank list .

For both the CTC and the GCS models, in our implementation, we consider a mixture of up to = 5 classes and report out-of-sample results for the number of mixtures that drove to best in-sample results.

In this section, we describe the results of an extensive simulation study, the main purpose of which is to characterize the performance of CTC models relative to the classical RUM models under various noise regimes. We nd that CTC models are generally more robust to noise in the o er sets and outperform the classical RUM models when noise a ects training and test o er sets di erently.

To streamline the analysis of this simulation study, we assume three ground truth, RUM-based models of demand: the classical MNL, the rank list, and the LC-MNL . To showcase the potential of CTC models, we start here from the restricted, single-class, ICS 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 and LC-MNL ground truth models to Appendix A4.1.

In our simulations, customers have perfect information of the o er sets and consider all the items on o er. Given the o er set , the customer chooses product with probability $\left(1+\sum_{i\in}\right)$, where the parameter 0 is the weight or the attraction value corresponding to product , and the 1 stands for the weight of the outside option. The modeler observes customer choices but

does not observe the o er sets perfectly. In fact, the assumed o er sets could be a superset of the true o er set, consistently with the common inventory inaccuracy problems. In the presence of such noise, we compare the predictions of an MNL model against an ICS model, both tted to the choice observations, to understand the conditions under which one outperforms the other one.

In our setup, the benchmark MNL model does not su er from model misspec cation but does su er from noise in the o er sets. The ICS model, on the other hand, su ers from both model misspeci cation and noise in the de nition of the o er sets, but it is designed to handle the latter more e ectively.

Our main nding is that the ICS model signi cantly 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 di erent from the training data), then models based on the consider-then-choose framework outperform classical RUM models.

We assume that we have $= 15$ items in the product universe. For each product , we sample its nominal utility uniformly at random from the interval [1 2] and set its MNL weight $= \exp( )$. We normalize the MNL weight of the outside option to 1. We parameterize the level of noise in the de nition of the o er set using two parameters: the *noise exposure* parameter, $\in [0\ 1]$, and the *noise intensity* parameter, $\in [0\ 1]$. The noise exposure parameter determines if a product is exposed to noise. Intuitively, the noise exposure parameter is designed to control the degree of asymmetry in the level of noise between the training and testing datasets. The noise intensity parameter speci es the conditional noise level, as described below. Next, for given values of and , and realized parameters of the MNL model, the data simulation proceeds as follows:

1. We sample 100 o er sets, $\{ \}|$, uniformly at random, which correspond to the underlying true o er sets.
2. For each o er set , we generate 10,000 sales transactions according to the MNL model with parameter values .
3. We generate a single exposure set containing products exposed to noise by including each product $\in$ with probability .
4. We generate noisy o er set observations by modifying each true o er set . To this end, we add extra products from the exposure set with probability ; speci cally, we obtain the noisy o er set by adding each product $\in \ \setminus$ with probability to the set .

We generated both training and test datasets for 200 di erent combinations of and : $\in \{0\ 05\ 0\ 1 \quad 1\}$ and $\in \{0\ 1\ 0\ 2 \quad 1\}$. First, we generated 100 realizations of MNL parameters.

Next, for given values and and for each MNL ground truth realization, using the procedure above, we generated both training and test synthetic instances of the same size. Each instance consists of 100 randomly generated o er sets, and for each o er set, 100,000 transactions, giving a total of 10,000,000 transaction records. All our algorithms were coded in Python (version 2.7.2) using Gurobi (version 7.0) as the optimization engine, and run on a 3.0Ghz processor with 16GB of RAM.

Our simulation setup is designed to capture not only di erent noise intensities but also di erent degrees of noise asymmetry between the training and test datasets. The noise exposure parameter captures this asymmetry as follows. Letting and denote the training and test exposure sets, respectively, the cardinality of the symmetric di erence (\ ) ∪ (\ ) captures the number of products that are exposed to noise in only one of the datasets. In expectation, this cardinality is equal to 2 (1 −) because the probability that a product is exposed to noise in only one of the data sets is equal to (1 −) + (1 −) = 2 (1 −). Therefore, the degree of asymmetry in the level of noise between the training and testing datasets is highest when = 0 5 and lowest when = 0 or = 1. Furthermore, the o er sets are perfectly observed in both the training and test datasets when = 0 or = 0, and the noise is perfectly symmetric when = 1. In practice, noise may be asymmetric when calibrating future demand forecasts because promotion strategies, product availability, or replenishment processes may change over time.

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

$$\text{MAPE} = \frac{100}{|\ \ |} \sum_{j \in} \frac{|\ \ - \ \ |}{10 +} \qquad \text{RMSE} = \frac{100}{\sum\limits_{j \in}} \sqrt{\frac{1}{|\ \ |} \sum_{j \in} (\ \ - \ \ )} \tag{15}$$

where denotes the observed sales for ∈ in the test dataset and denotes our prediction. We make predictions on noisy test o er sets, so = 10000 · ∑ ( ), where is the noisy test o er set and ( ) is the probability that product will be purchased from o er set under the tted model. Note that we add 10 in the denominator of MAPE to deal with unde ned instances, and we divide the RMSE score by the total number of observed sales in the test data set to make it a relative metric so that it is more interpretable.

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

In Figure 2 we present a heatmap of the prediction scores under an MNL model tted to the training data where each column corresponds to a particular noise intensity and each row corresponds to a particular noise exposure . We focus on the MAPE and RMSE prediction scores in the left and right panels, respectively. Recall that the MNL model is the ground truth model for this simulation study. As expected, the MNL model captures the ground-truth choice probabilities almost exactly when = 0 05 and = 0 1, i.e., there is only a small amount of noise added to the sales transaction data. However, the MNL prediction scores worsen with higher noise intensity for a given level of noise exposure. Interestingly, it can also be seen that MNL prediction scores are not monotonic with respect to the noise exposure level, i.e., the scores rst increase and then decrease with the noise exposure level, for a given noise intensity. The tted MNL model performs the worst at noise exposure level = 0 5, which as noted above, corresponds to the highest degree of noise asymmetry between the training and test data sets, whereas the performance is relatively good even at high noise intensity levels when the degree of asymmetry is low (is close to 0 or 1).

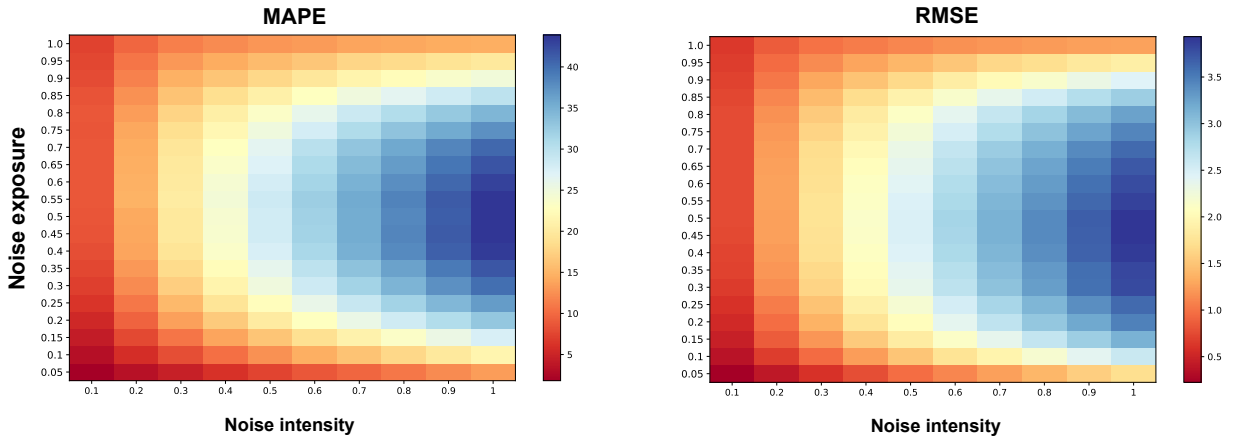
To provide a quantitative understanding of the variation of the model performance with respect to the two noise parameters and , we carry out the following linear regression:

$$= + \cdot + \cdot + \cdot \text{Asymm} + \cdot \text{Shared} + \tag{16}$$

where the index stands for the noise intensity $\in \{0\ 1\ 0\ 2 \quad 1\}$, and the index stands for the noise exposure $\in \{0\ 05\ 0\ 1 \quad 1\}$. The outcome variable is the MAPE prediction score of the corresponding cell.

The rst two terms in the regression capture dependence on the noise intensity level and the last two terms capture dependence on the noise exposure level . We add a quadratic term in the noise intensity level to capture any potential non-linear response to the noise intensity level. We also separate out the dependence on into degree of asymmetry and degree of overlap in noise. The covariate Asymm is the probability that an item in the product universe is exposed to noise only in the test dataset or only in the training dataset, and the covariate Shared is the probability that an item in the product universe is exposed to noise both in the test and training datasets. Note that Asymm $= (1 -) + (1 -) = 2 (1 -)$ and Shared $=$, where is the noise exposure.

The results for the regression (16) are reported in the last column of Table 1. It follows from there that the noise intensity deteriorates the predictive performance of the MNL model in a non-linear way, with the coe cient for the linear term being positive and the coe cient for the quadratic term being negative. The variables Asymm (i.e., degree of noise asymmetry) and Shared (i.e, the degree



η γ

of the shared noise) are positively correlated with the MAPE score, which also implies that the prediction performance of the MNL model worsens as the number of items in the product universe that are exposed to noise increases. Interestingly, the coe cient of the variable has more than seven times higher magnitude than the coe cient of the variable , even though both variables have a comparable range of possible values (e.g., varies from 0 to 0.5 and varies from 0 to 1), which indicates that the benchmark (i.e, MNL model) struggles the most in making accurate predictions when the impact of the noise is asymmetric between the training and test sales transactions. Note that the independent variables in the regression model (16), included as Model (5) in Table 1, explain most of the variation in the MAPE score under the MNL model, i.e., $= 0\ 93$.

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

In Figure 3, we compare the performances of the MNL and the ICS models under di erent noise regimes. Each cell presents the improvement obtained by the ICS model over the MNL model, computed as the di erence in the corresponding prediction scores, so that higher values are better

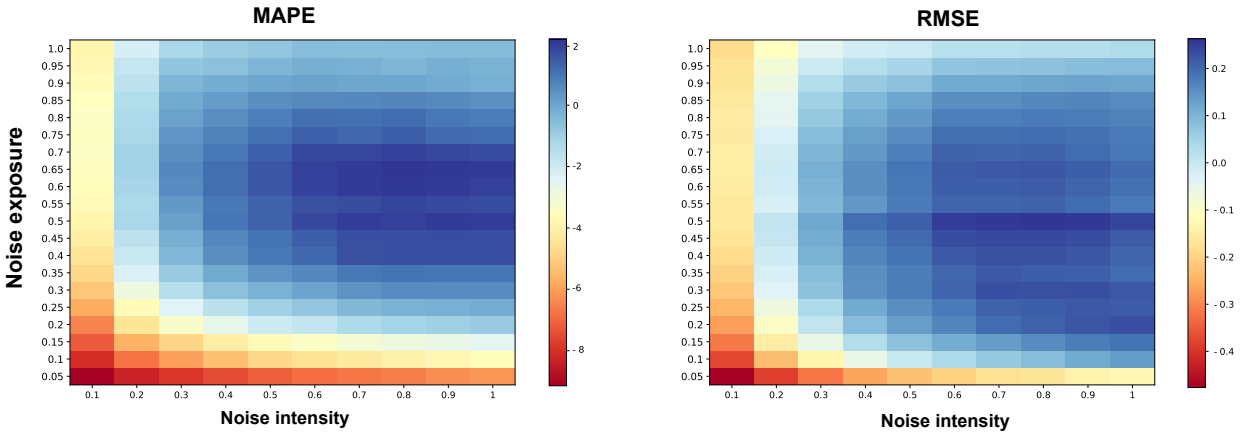
	Model (1) Score	Model (2) Score	Model (3) Score	Model (4) Score	Model (5) Score
	27.336***				42.079***
	(17.997)				(14.246)
^{2}		22.929***			-13.403***
		(15.755)			(-5.122)
			34.662***		39.320***
			(8.677)		(28.338)
				-2.590	5.385***
				(-1.143)	(8.015)
	6.260***	12.467***	9.770***	22.224***	-11.694***
	(6.643)	(17.021)	(6.698)	(20.645)	(-12.579)
No. Observations:	200	200	200	200	200
R-squared:	0.621	0.556	0.275	0.007	0.929
Adj. R-squared:	0.619	0.554	0.272	0.002	0.928

t
* $p <$. ** $p <$. *** $p <$.

for the ICS model. Unsurprisingly, the MNL model is sign cantly better than the ICS model when the level of noise is small (small values of both and) because the ICS model is misspeci ed. But, at high values of noise intensity , and values of close to 0 5, we see that the ICS model outperforms the MNL model. Table 2 presents the results from regressing the improvement scores according to regression equation (16). As above, we nd that the bene ts of the ICS model are most signi cant when the noise is asymmetric between the training and test o er sets. Although not shown here, the pattern of the heatmap for the ICS model looks similar to that of the MNL model in Figure 2. However, the ICS model is more robust to noise (both the intensity and the degree of asymmetry) and its performance does not deteriorate as much, because of which it outperforms the MNL model in asymmetric and high noise regimes.

We highlight that the results in this section are robust to di erent asymmetric scenarios of noise as well (see Figure A10 in Appendix A4.1).

In order to shed some light on one potential mechanism that drives the superior performance of ICS over the MNL model, we present a stylized study in Appendix A4.2. The main insight we obtain is that the one-dimensional cannibalization property of the ICS model makes it robust to o er set noise. This property states that the presence or absence of lower-ranked products does not a ect the demand for higher-ranked products. As a result, o er set noise mainly impacts the lower-ranked products, limiting the overall error rates. See Appendix A4.2 for details.



η γ

	Model (1) Impr.	Model (2) Impr.	Model (3) Impr.	Model (4) Impr.	Model (5) Impr.
	4.162***				16.621***
	(7.883)				(17.214)
2		3.025***			-11.327***
		(6.157)			(-13.241)
			7.311***		11.530***
			(7.112)		(25.421)
				2.539***	4.878***
				(4.811)	(22.207)
	-2.966***	-1.842***	-3.108***	-1.588***	-11.042***
	(-9.055)	(-7.449)	(-8.281)	(-6.332)	(-36.333)
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

* $p <$. ** $p <$. *** $p <$.

Our analysis so far has been based on synthetic data, establishing that even the restricted ICS model outperforms classic choice models when o er sets are uncertain and the noise is asymmetric between the training and test o er sets. We now present a real-world case study where such conditions are met.

For our study here, we use the household purchase panel and store data from the IRI Academic Dataset (Bronnenberg et al. (2008)). This panel dataset keeps track of the household purchase histories for grocery and drug store chains, collected from the two largest Behavior Scan markets in the US over the years 2001-2011. We also use the GCS model, more general than the previous ICS, as a baseline to conduct our analysis. Because we now have access to panel data, we also implement a variant of the EM algorithm for the estimation of the parameters of the GCS model, as described in Appendix A3.4.2.

The purpose of this empirical study is threefold: (i) provide various real-world scenarios based on the IRI dataset where we are likely to face signi cant noise in the o er set de nitions when making the long-term demand predictions, (ii) investigate the prediction performance of choice models under di erent noise regimes, e.g., quantify the improvement of the GCS model over the latent class MNL (LC-MNL) model on panel data under several real-world scenarios with various noise intensities, and (iii) compare the GCS model studied in this paper with the more restricted ICS model of Manzini and Mariotti (2014).

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

In this section, we present all the comparisons with respect to the GCS model (see the GCS model estimation details in Appendix A3.4.2). It is natural to wonder if the more general CTC model (see Section 4.2 above for the CTC model estimation framework), which allows for heterogeneity in customer preference orders, o ers additional gains in prediction performance. To this end, we carried out a similar analysis with the CTC model; see Appendix A3.6 for details. From our results, we can not claim dominance of GCS or CTC; that is, the GCS model dominates the CTC model under the rst noise generation process whereas the CTC model outperforms the GCS model under the second noise generation process. Based on this nding, we focus our analysis and discussions mainly on the more parsimonious GCS model.

The dataset consists of weekly sales transactions. We analyze a total of 20 categories, presented in Table A1. We focus on sales transaction data from the 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 o ered to each individual upon her store visit, we follow existing literature (e.g., Jagabathula and Vulcano (2018)) and construct this subset by aggregating all the transactions made in a particular store within a given category during a particular week. We also aggregate items with the same vendor code (i.e., items that are associated with the same brand name) into a single product due to data sparsity and divide the sales transaction data into two parts: the training set, which consists of the rst 26 weeks of sales observations; and the test set, which consists of the last 26 weeks of sales observations. Note that we exploit the panel data structure in the predictive performance analyses below.

We compare the GCS model with the benchmark: the LC-MNL choice model with latent classes. In this model, each customer belongs to one unobservable class, and customers from class $\in$ {1 2 } make purchases according to the MNL model associated with that class. The model is described by the parameters of the MNL model characterizing each class and by the prior probabilities of customers belonging to each of the classes. Once the model parameters are estimated, we make transaction-level predictions for each customer by averaging the predictions from single-class models, weighted by the posterior probability of class membership. Similarly to the GCS model, we estimated the model for $= 1\ 2$ 5, and report the best performance measure from these 5 variants, for each of the performance metrics introduced in Section 5.2. Because we have panel data, we make individual-level predictions and consequently, the number of predicted sales for product is now calculated as $= \sum_{\in} \sum^{i}$ ( ), where is the set of all customers, is the number of transactions of customer , and ( ) is the predicted probability that customer purchases item from noisy o er set estimate . Note that we do not have access to the no-purchase observations in our dataset and thus our prediction metrics do not include them.

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

6.3.1. Different data aggregations towards offer set forecasting. We consider three di erent prediction tasks that naturally arise in practical retail contexts and which di er in the level of di culty of forecasting the o er sets.

1. *Short-term forecasts.* Often, store managers want to make short- or immediate-term forecasts, say, for the next week, to help with inventory and promotion planning. For these forecasts, the manager has a reasonably accurate estimate of the product assortment or o er set.
2. *Long-term forecasts.* To be successful in major strategic and investment decisions, a store manager must also make long-term demand forecasts, say, over the next quarter or the next year. For making these forecasts, the retailer often does not have a good sense of how the product assortment evolves over the forecasting horizon because of product replenishment and stockouts, which manifest as errors in o er set forecasts.
3. *Warehouse forecasts.* Another scenario is when the warehouse of the retail chain distributes products to the stores and makes centralized decisions on the inventory level in the warehouse. In this case, the warehouse is likely to make predictions at the centralized level without knowing the up-to-date information on product assortments in every store.

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

We simulate these scenarios using the purchase observations we have as follows. As mentioned above, we split the purchase transactions into training and test sets. In the training set, the purchased product and the corresponding o er set are known. Because of the way the o er set was inferred, it may contain errors, but we train all the models disregarding any potential errors. Note that this process re ects the standard way in which choice models are trained in the literature and in practice. In addition to a trained choice model, the retailer must build what we call an OSForecaster or an *offer set forecaster*. We abstract away from details of how such a forecaster might be constructed and instead peek in our test dataset to simulate one. Speci cally, let denote the set of o ered products at store and period in the test horizon. Further, let $\mathcal{T}$ denote the set of test time periods.

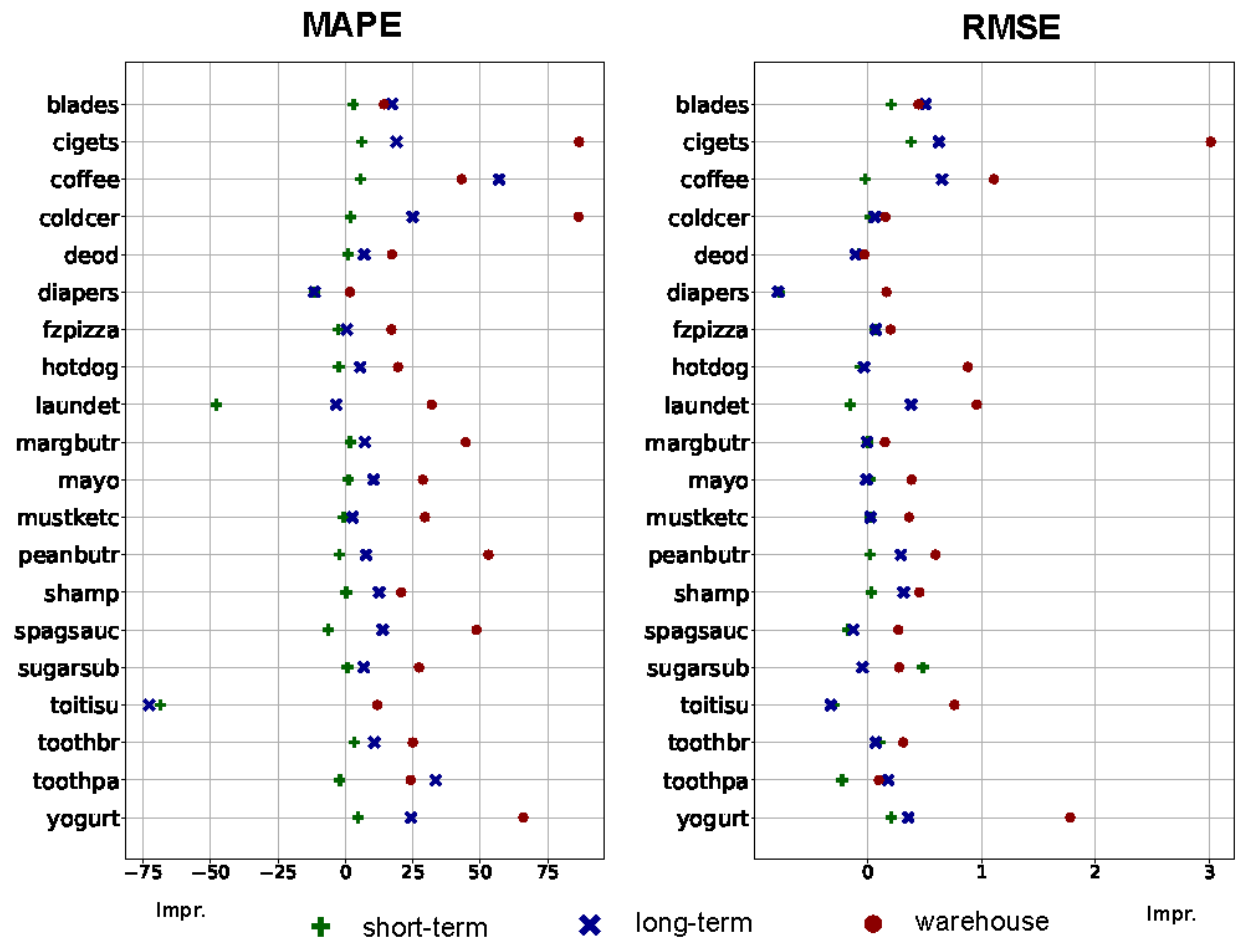
For short-term forecasts, we provide the retailer with the number of purchasing customers at each store and period , and the retailer must predict the sales for each of the products. In practice, this will entail making sales predictions for the next week for each of the stores. To make this prediction, the retailer must rst forecast an o er set for each store and time period during the test horizon. To re ect the fact that the retailer makes few errors in forecasting the next week s o er set, we assume that the short-term OSForecaster returns the true o er set when queried

with a store and time period . The predictions of the OSForecaster are more complex for the other two scenarios. In particular, to re ect the di culty of making such o er set forecasts, we assume that the OSForecaster provides not a point forecast, but a distribution over possible o er sets. We describe below how we use the test data to construct these distributions.

For long-term forecasts, we provide the retailer with the total number of purchasing customers at each store over the 26 weeks ($\approx$6 months) comprising the entire test horizon, and the retailer must predict the sales for each of the products. To make these predictions, the retailer must rst forecast an o er set to use for each store. We assume that when queried with a store , the OSForecaster returns a uniform distribution over 20 assortments, each of which is constructed as follows: we rst construct a random collection $\mathcal{A}$ of o er sets by including o er set for each $\in \mathcal{T}$ with probability 0 5 and then obtain one o er set by taking the union of all sets in the collection $\mathcal{A}$. Note that every o er set in the OSForecaster is obtained from sampling a new random collection $\mathcal{A}$ of o er sets. This construction is designed to re ect the possibility that the retailer is able to forecast some of the stock out and replenishment events, but not all.

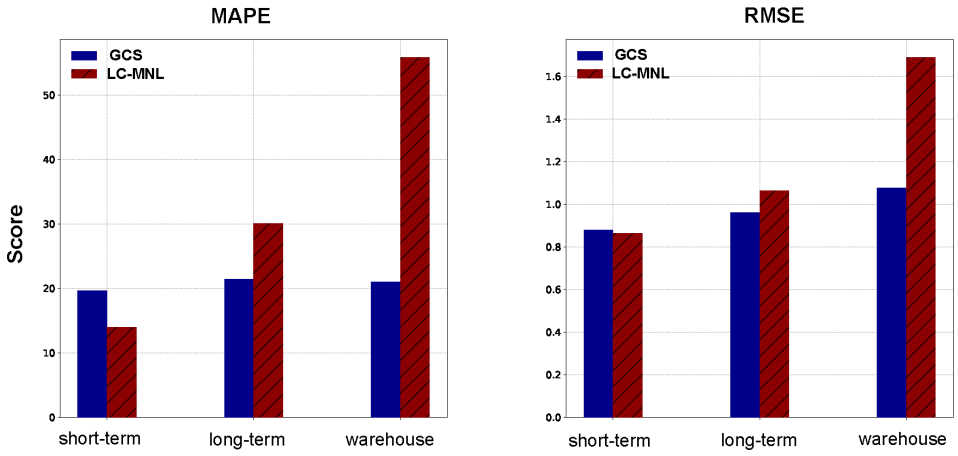
The procedure we use for warehouse forecasts is similar. The retailer must generate sales forecasts for each period , across all the stores in the corpus, for which the retailer must forecast an o er set for each period . We assume that when queried with a period , the OSForecaster returns a uniform distribution over 20 assortments, each of which is constructed as follows: we construct a random collection $\mathcal{A}$ of o er sets by including o er set for each store with probability 0 5, and then obtain one o er set by taking the union of all sets in the collection $\mathcal{A}$. This construction is designed to re ect the possibility that the retailer has information on the o er sets at some of the stores but not all.

6.3.2. Performance assessment. We evaluate all three scenarios in terms of the accuracy of predicting the total sales of each product across the entire test horizon, as described in Section 5.2. When the OSForecaster outputs a distribution, our predictions will be averaged over all the o er sets in the distribution. In Figure 4 we present scatter plots of the improvements of the GCS model versus the LC-MNL model across 20 product categories under the three forecast scenarios discussed above. In the left and right panels we measure the predictive performance of the models under the MAPE and RMSE metrics, respectively, as de ned in (15) (see Appendix A4.3 for alternative de nitions of these metrics and corresponding results). We observe that GCS outperforms LC-MNL for around half of product categories for short-term forecasts, and for almost all product categories under the second and third forecast types. Note that we have dots located to the right of pluses with crosses being in between, under both MAPE and RMSE scores and across most of the product categories. It reveals that the improvement of GCS over LC-MNL across product



categories increases when we switch from short- to long-term, and from long-term to warehouse forecasts.

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



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

Turning back to Figure 4, we notice that the improvement of GCS over LC-MNL varies across product categories for a given scenario. To better explain this variation, we regress the improvement of GCS over LC-MNL for each category against the noise intensity, which captures how much the forecasted o er set di ers from the true o er set. We only consider the scenarios of long-term and warehouse forecasts for this analysis because the noise intensity for short-term forecasts is zero by de nition. We de ne noise intensity at the transaction level and aggregate the metric across all the transactions. For each transaction, in the test set, let denote the true o er set and denote the forecasted o er set. For the transaction occurring at store in period , the forecasted o er set takes the value OSForecaster( ) and OSForecaster( ) for long-term and warehouse predictions, respectively. We then use the following natural de nition for noise intensity:

$$\text{noise intensity} = \frac{1}{\#\text{ of test transactions}} \sum \mathbb{E}\left[\frac{| \quad \setminus \quad |}{| \quad |}\right]$$

where the expectation is with respect to the distribution of the o er set predictions . Intuitively, for long-term forecasts, the noise intensity captures how much the o er set at a store varies over time; if there are very few stockouts, then the o er set remains stable and the noise intensity is close to zero, but if there are many stock outs then the o er set varies a lot and the noise intensity

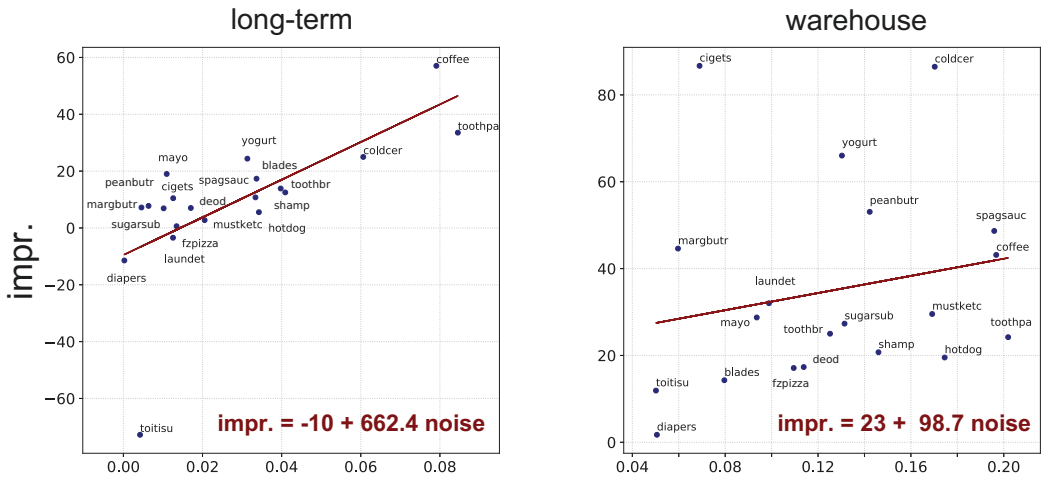


Figure 6 Scatter plots and linear regressions of the MAPE score improvement of GCS vs. LC-MNL over the noise intensity in the definition of offer sets across 20 product categories. Improvements are defined as the difference between two scores. In the left and right panels, we focus on the long-term and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

is large. Similarly, for warehouse forecasts, the noise intensity captures how much the offer sets vary across stores. The left and right panels in Figure 6 illustrate the regression under the long-term and warehouse forecast scenarios, respectively. We see a clear positive correlation between the improvement of GCS over LC-MNL and noise intensity in both panels, suggesting the improvement becomes more significant with higher noise intensity in the product category.

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

Second, we highlight the following peculiarity of our study: for long-term and warehouse forecast scenarios, the models are trained on noiseless offer sets but then tested on noisy offer sets. This is clearly in contrast to standard practice where efforts are made to train models on the same setups that they are tested on. Would not it then be better to ignore the fact that we have noiseless offer sets and instead train our models also on 'forecasted' offer sets? In Appendix A4.5, we repeat

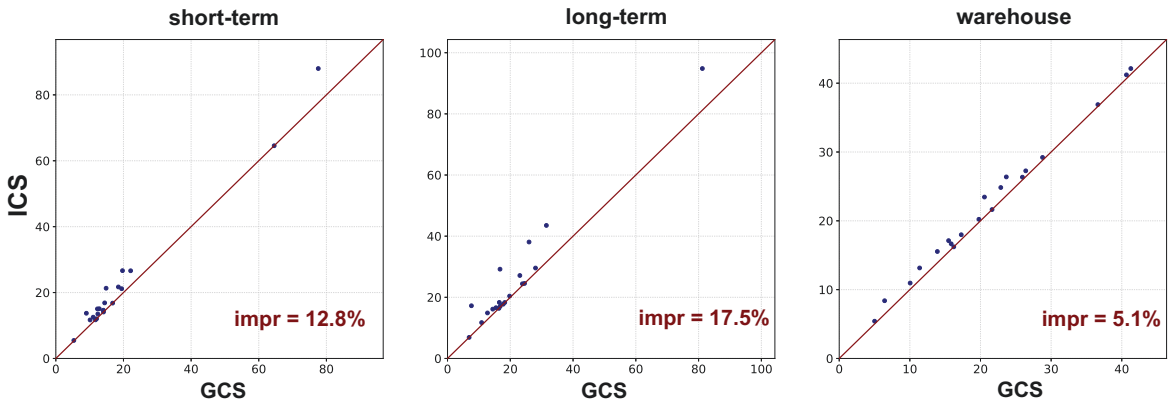


Figure 7 Scatter plots of the MAPE scores of 20 product categories under the ICS vs. GCS models. The left, middle, and right panels correspond to the cases of short-term, long-term, and warehouse forecast scenarios, respectively. Lower is better; therefore, GCS outperforms ICS for points above the 45° line.

our analysis by training all the models on offer sets obtained by applying the OSForecaster (as described above) to the training data[2]. We observe that the GCS model's performance increases slightly, but the improvement in the LC-MNL model's performance is significant. However, despite the substantial improvement in the LC-MNL model's performance, it is still unable to fully bridge the gap. We interpret this result as follows. *If* the underlying noise process is known, then it makes sense to incorporate it in the training process. But in practice, the noise process cannot be fully modeled because the factors affecting offer set noise (e.g., stock out and replenishment processes) vary over time, in which case the CTC models offer some protection against noise.

We conclude this section by comparing the predictive performance of the GCS model against a single-class ICS model; see Figure 7. We observe that the GCS model outperforms the ICS model on the MAPE metric by 12.8%, 17.5%, and 5.1% under the short-term, long-term, and warehouse forecast scenarios, respectively, on average, across 20 product categories. This finding suggests that a mixture of product-form consideration set distributions captures customer heterogeneity better, even when preferences are governed by a single rank list.

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.

[2] We thank an anonymous referee for suggesting this study.

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

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 a 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 cars nearby and book the available alternative by entering the license number and credit card information.

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

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 other 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, de ning tuples of the form ( ), where is the chosen car and is the set of cars available at the reservation time .

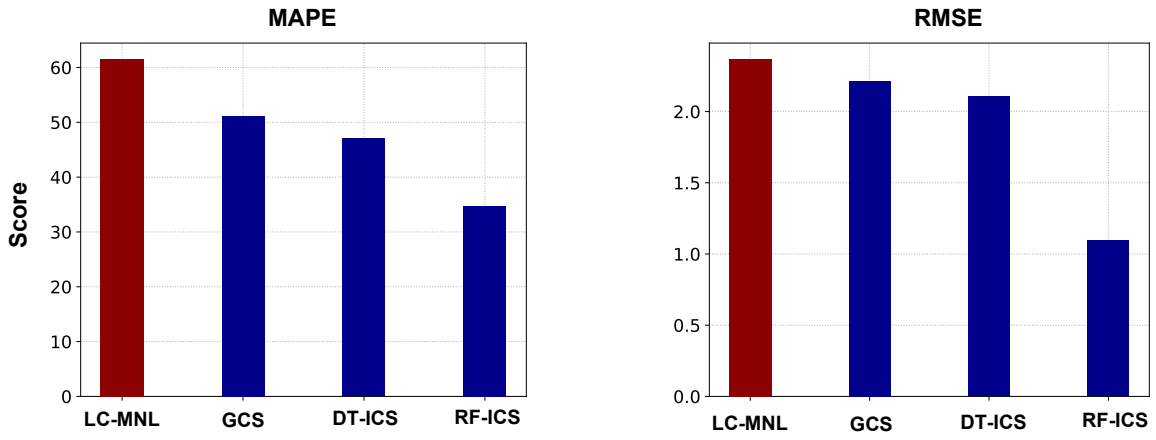
In general, in order to calibrate feature-based consider-then-choose models with our dataset, we need to estimate two types of parameters: the ranking over all cars listed in the online platform (514 cars in total), and the parameters associated with the consideration set formation of renters. In order to simplify the estimation procedure, we assume that the ranking is known a priori. Speci cally, the cars are ranked according to their popularity among renters, de ned 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 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 justi ed (see Appendix A5.2.1 for details).

Based on this popularity-based single ranking , we estimate three variants of the CTC framework that account for product features with the purpose of characterizing the distribution over consideration sets: i) a GCS model de ned as a mixture of logistic-based ICS (L-ICS) models, ii) a decision tree-based ICS (DT-ICS), and iii) a random forest-based ICS (RF-ICS). Details about the estimation of the L-ICS, DT-ICS, and RF-ICS models can be found in Appendix A3.1.

Our benchmark RUM model is the classic and competitive, feature-based LC-MNL model. For both GCS and LC-MNL, we tried up to $= 5$ classes to nd the optimal mixing distribution and report out-of-sample results for the number of mixtures that drove to best in-sample results.

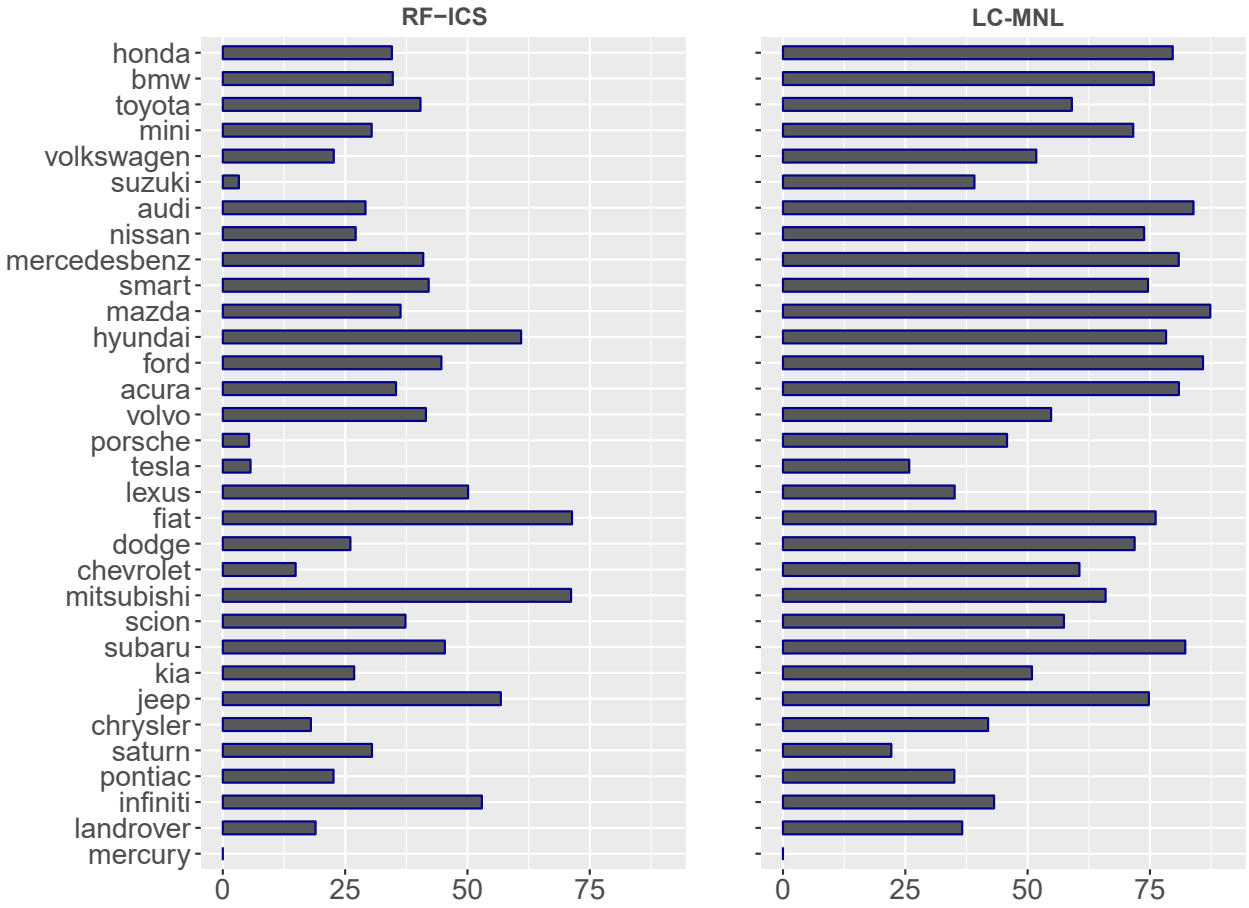
Next, we conduct an out-of-sample prediction testing of the analyzed models on the accuracy of two prediction measures: MAPE and RMSE (see Section 5.2), where lower scores stand for better prediction.

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 real-world settings, the company can not rely on accurate data on car availabilities over time for the distant future, i.e., we can not test the prediction power of choice models by using the o er 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 o er over the entire horizon captured by the hold-out data for each area. Note that in this way we have 42 di erent o er 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 exhibits an improvement of 16.7%, 23.4%, and 43.3% over LC-MNL for GCS, DT-ICS and RF-ICS, respectively. We also observe that consider-then-choose models obtain improvements of 6.2%, 10.9%, and 53.7% over LC-MNL for GCS, DT-ICS, and RF-ICS, respectively, based on RMSE metrics.

Figure 9 exhibits MAPE scores computed for every brand separately under the RF-ICS and LC-MNL models, where the brands are ordered according to their popularity (i.e., percentage of the total number of reservations in the training dataset coming from every brand), e.g., Honda is the most popular brand while Mercury is the least popular brand in the dataset. We note that the RF-ICS model outperforms the benchmark LC-MNL model almost consistently across all the brands. The panels also illustrate that MAPE scores vary signi cantly across brands both for RF-ICS and LC-MNL models. To further analyze this variation, in Figure 10 we regressed the improvement of RF-ICS over LC-MNL against the popularity of brands (left panel), and the improvement of RF-ICS over LC-MNL against the MAPE score of the LC-MNL model (right panel). We observe a clear positive correlation between MAPE score improvements and the popularity of brands, which indicates that we can better predict the demand for more popular brands. We can also see a clear positive correlation between the improvements and MAPE score under the LC-MNL model, allowing us to conclude that consider-then-choose type of models are especially relevant in prediction tasks, i.e., CTC models dominate LC-MNL, when the LC-MNL model provides a relatively poor prediction performance. Being robust to the noise, consider-then-choose models (and in particular, RF-ICS) provide signi cantly better predictive performance under these circumstances. Note that these insights are consistent with our numerical study based on the synthetic dataset in Section 5.



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

In this paper, we analyze the importance of modeling customer choices by accounting for unobserved consideration sets. Even though consider-then-choose (CTC) models are gaining popularity in the

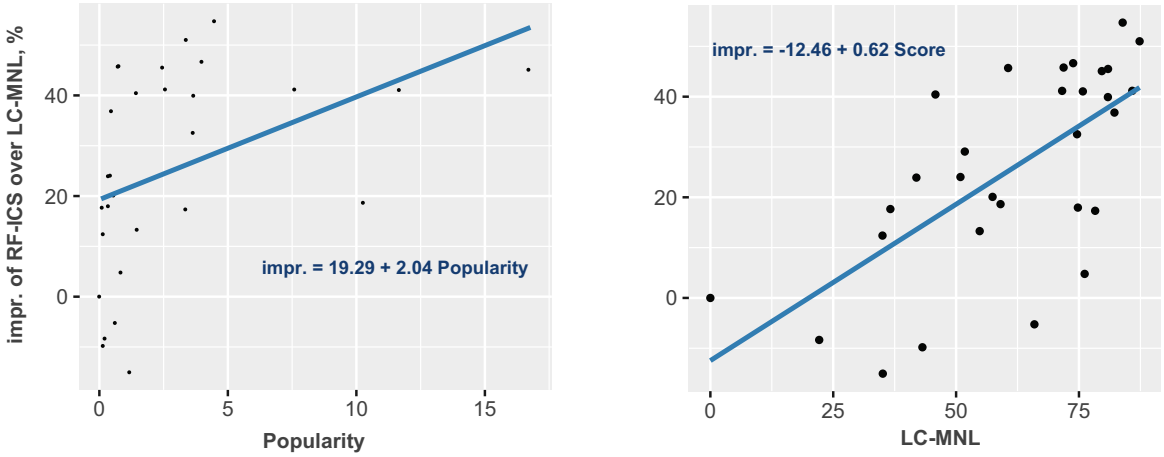


Figure 10 Left panel: scatter plot and linear regression of the percentage improvement of RF-ICS versus LC-MNL over brand popularity (i.e., percentage of total number of reservations in the training dataset), across 32 car brands. Right panel: scatter plot and linear regression of the percentage improvement of RF-ICS versus LC-MNL over LC-MNL prediction accuracy, across 32 car brands. In both panels, the prediction accuracy is measured by the MAPE score aggregated for every brand.

OM field, it is not clear from the existing literature when CTC models can outperform the classical models in the prediction performance if companies collect only transaction data. We show how one can effectively estimate a general class of CTC models and also address the problem of identifying such models when relying only on sales transaction data. We check the performance of the proposed methodology under a synthetic data setting where we control for different levels of noise and asymmetries of noise between the training and the testing data. Next, we apply CTC models to two real-world contexts: a retail operation and a car-sharing platform.

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

References

Alba J, Chattopadhyay A (1985) Effects of context and postcategory cues on recall of competing brands. *Journal of Marketing Research* 22(3):340–349.

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

Aouad A, Farias VF, Levi R (2020) Assortment optimization under consider-then-choose choice models. Forthcoming in *Management Science*.

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

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

Bronnenberg BJ, Kruger MW, Mela CF (2008) Database paper the iri marketing data set. *Marketing science* 27(4):745 748.

Campbell BM (1969) *The existence of evoked set and determinants of its magnitude in brand choice behavior*. Ph.D. thesis, Columbia University.

Davis JM, Gallego G, Topaloglu H (2014) Assortment optimization under variants of the nested logit model. *Operations Research* 62(2):250 273.

DeHoratius N, Raman A (2008) Inventory record inaccuracy: An empirical analysis. *Management Science* 54(4):627 641.

Duran MA, Grossmann IE (1986) An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical programming* 36(3):307 339.

Farias VF, Jagabathula S, Shah D (2013) A nonparametric approach to modeling choice with limited data. *Management Science* 59(2):305 322.

Feldman J, Paul A, Topaloglu H (2019) Assortment optimization with small consideration sets. *Operations Research* 67(5):1283 1299.

Feldman J, Topaloglu H (2015) Bounding optimal expected revenues for assortment optimization under mixtures of multinomial logits. *Production and Operations Management* 24(10):1598 1620.

Feldman J, Topaloglu H (2018) Capacitated assortment optimization under the multinomial logit model with nested consideration sets. *Operations Research* 66(2):380 391.

Gallego G, Li A (2017) Attention, consideration then selection choice model.

Graham RL (1995) *Handbook of combinatorics* (Elsevier).

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

Hauser JR (2014) Consideration-set heuristics. *Journal of Business Research* 67(8):1688 1699.

Hauser JR, Wernerfelt B (1990) An evaluation cost model of consideration sets. *Journal of consumer research* 16(4):393 408.

Hausman JA (1996) Valuation of new goods under perfect and imperfect competition. *The economics of new goods*, 207 248 (University of Chicago Press).

Hogarth RM, Karelaia N (2005) Simple models for multiattribute choice with many alternatives: When it does and does not pay to face trade-o s with binary attributes. *Management Science* 51(12):1860 1872.

Howard JA, Sheth JN (1969) The theory of buyer behavior. *New York* 63:145.

Hoyer WD (1984) An examination of consumer decision making for a common repeat purchase product. *Journal of consumer research* 11(3):822 829.

Jagabathula S, Rusmevichientong P (2017) A nonparametric joint assortment and price choice model. *Management Science* 63(9):3128 3145.

Jagabathula S, Subramanian L, Venkataraman A (2020) A conditional gradient approach for nonparametric estimation of mixing distributions. *Management Science* 66(8):3635 3656.

Jagabathula S, Vulcano G (2018) A partial-order-based model to estimate individual preferences using panel data. *Management Science* 64(4):1609 1628.

Kang Y, Gershwin SB (2005) Information inaccuracy in inventory systems: stock loss and stockout. *IIE transactions* 37(9):843 859.

Lee E, Lee B (2012) Herding behavior in online p2p lending: An empirical investigation. *Electronic Commerce Research and Applications* 11(5):495 503.

Lynch JG, Alba J, Hutchinson JW (1991) Memory and decision making. *Handbook of consumer behavior* 1 9.

Manzini P, Mariotti M (2014) Stochastic choice and consideration sets. *Econometrica* 82(3):1153 1176.

Montgomery H, Svenson O (1976) On decision rules and information processing strategies for choices among multiattribute alternatives. *Scandinavian Journal of Psychology* 17(1):283 291.

Murphy KP (2012) *Machine learning: a probabilistic perspective* (MIT press).

Newman JP, Ferguson ME, Garrow LA, Jacobs TL (2014) Estimation of choice-based models using sales data from a single rm. *Manufacturing & Service Operations Management* 16(2):184 197.

Ratchford BT (1982) Cost-bene t models for explaining consumer choice and information seeking behavior. *Management Science* 28(2):197 212.

Roberts JH, Lattin JM (1997) Consideration: Review of research and prospects for future insights. *Journal of Marketing Research* 406 410.

Sher I, Fox JT, Bajari P, et al. (2011) Partial identi cation of heterogeneity in preference orderings over discrete choices. Technical report, National Bureau of Economic Research.

Strauss D (1979) Some results on random utility models. *Journal of Mathematical Psychology* 20(1):35 52.

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

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

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

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

Vulcano G, Van Ryzin G, Ratli R (2012) Estimating primary demand for substitutable products from sales transaction data. *Operations Research* 60(2):313 334.

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

Westerlund T, Pettersson F (1995) An extended cutting plane method for solving convex minlp problems. *Computers & Chemical Engineering* 19:131 136.

Wright P, Barbour F (1977) *Phased decision strategies: Sequels to an initial screening* (Graduate School of Business, Stanford University).

Demand estimation under uncertain consideration sets
Appendix

Srikanth Jagabathula
NYU Stern School of Business, New York, NY, sjagabat@stern.nyu.edu

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

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

For completeness, we summarize the relevant notation from the main body and also introduce additional notation. We consider a universe of products { }. We let denote the no-purchase or the outside option. A customer is presented with a subset ⊆ of products and the customer chooses either one of the products in or the outside option . We let ℙ ( ) denote the probability that a customer chooses product ∈ and ℙ ( ) the probability that the customer chooses the outside option. We use to denote the set ∪{ }. Let : 2 → [0 1] de ne a distribution over consideration sets such that ∑ ⊆ ( ) = 1. The preference relation ≻ speci es a rank ordering over + 1 items which consist of the products in plus no-purchase option with ( ) denoting the preference rank of product . The lower the rank of the product, the higher the preference, so that a customer s ranking indicates that product is preferred to product if and only if ( ) ( ), or equivalently ≻ . We assume that there is a distribution : 𝒮 → [0 1] over 𝒮 , which is the set of all full rankings or permutations of products in with cardinality (+ 1)!.

To simplify the exposition, we also let := \ , := ∪{ }, and ℙ ( ) = Pr(|). Let ⟨ ⟩ denote the power set of , i.e., ⟨ ⟩ = 2 , and let ⊎ denote { ∪ : ∈ ∈ } for any sets .

Proof of Proposition 1: First, we argue that if choice data are consistent with an underlying RUM model, then they are also consistent with a CTC model. To this end, let the distribution (·) be a member of the RUM class. It de nes a distribution over the (+ 1)! preference lists of products

in , which includes all the products in and the outside option. We map each ranking in the support of (·) to the tuple ( ) as follows: (a) consists of all the products in that are preferred over the outside option; that is, = { ∈ : ( ) ( )} and (b) is the ranking obtained by moving the outside option to the last position (i.e., the (+1)th position) in ranking . We then de ne the CTC model (· ·) such that ( ) = ( ). It is straightforward to check that for any o er set , the ranking and the corresponding tuple ( ) result in the same choice. As a result, the choice probabilities under both (·) and (· ·) should match.

We are now left to prove the other direction. Consider a CTC model (· ·) that de nes a joint distribution over the preference lists of the products in that rank the outside option at the bottom, and the subsets of . We map each tuple ( ) in the support of (· ·) to the ranking over the products in by repositioning the products not in to be below the outside option in the ranking . More precisely, we have that ( ) ( ) if and only if ( ) ( ) whenever ∈ or ∈ . In addition ( ) ( ) whenever ∈ and ( ) ( ) whenever ∈ . We then de ne the RUM model (·) such that ( ) = ( ). Again, it is straightforward to check that both and the tuple ( ) result in the same choice from each o er set . It thus follows that the choice probabilities under both (· ·) and (·) match.

The result of the proposition now follows. □

We start this subsection by presenting Lemma A1 which shows how we can prove Proposition 3 by invoking a particular form of the inclusion-exclusion principle stated by Graham (1995). Then, we show how to prove Proposition 3 from rst principles. Finally, we present Lemma A2 followed by the proof of Proposition 4 which relies on the proof of the combinatorial identity in Lemma A2.

LEMMA A1. *For any sets* ⊆ *and* ⊆ *, and the function* : 2 → ℝ*, we have*

$$\sum_{\subseteq} \sum_{\subseteq} (-1)^{| |-| |} \cdot (\setminus) = (\setminus) \tag{A1}$$

Proof: First, consider the inclusion-exclusion principle stated by Graham (1995) in the following form. Let be a nite set and : 2 → ℝ be a real-valued function de ned on the subsets of . De ne the function : 2 → ℝ by ( ) := ∑ ⊆ ( ), then ( ) := ∑ ⊆ (−1)^{| |−| |} ( ).

Then we show that the lemma follows from the stated above inclusion-exclusion principle. Let ( ) := (\ ), and ( ) := (−1)^{| |} ∑ ⊆ (−1)^{| |} · ( ), which implies that

( ) · (−1)^{| |} = $\sum_{\subseteq} (-1)^{| |} \cdot ( )$ by invoking the inclusion-exclusion principle we obtain that

(−1)^{−| |} · ( ) = $\sum_{\subseteq} (-1)^{| |-| |} \cdot ( ) \cdot (-1)^{| |}$ which implies that

$$(\ \setminus\ ) = (\ ) = \sum_{\subseteq} (\ ) = \sum_{\subseteq} (-1)^{|\ |} \sum_{\subseteq} (-1)^{|\ |} \cdot (\ ) = \sum_{\subseteq} \sum_{\subseteq} (-1)^{|\ |-|\ |} \cdot (\ )$$
$$= \sum_{\subseteq} \sum_{\subseteq} (-1)^{|\ |-|\ |} \cdot (\ \setminus\ )$$

□

Proof of Proposition 3: For every $\subseteq$ we de ne boolean functions $: 2 \to \mathbb{R}$ and $: 2 \to \mathbb{R}$ by

$$(\ ) = (-1)^{|\ |} \cdot \mathbf{I}[\ \subseteq\ ]$$
$$(\ ) = (-1)^{|\ |} \cdot \mathbf{I}[\ \subseteq\ ]$$

where $\mathbf{I}[\ ]$ is an indicator function that is equal to 1, if condition is satis ed, and 0 otherwise. Then for all $\subseteq$ we claim that

$$\sum_{\subseteq} {}_1(\ ) \cdot {}_2(\ ) = \begin{cases} 1 & \text{if } = \\ 0 & \text{otherwise} \end{cases} \tag{A2}$$

First, we show that $\sum_{\subseteq} (\ ) \cdot (\ ) = 1$ for every $\subseteq$:

$$\sum_{\subseteq} (\ ) \cdot (\ ) = \sum_{\subseteq} \mathbf{I}[\ \subseteq\ ] \cdot (-1)^{|\ |\ |\ |} \mathbf{I}[\ \subseteq\ ] = (-1)^{|\ |\ |\ |} = 1$$

Then we show that $\sum_{\subseteq} {}_1(\ ) \cdot {}_2(\ ) = 1$ for all $\subseteq$ s.t. $\neq$:

$$\sum_{\subseteq} {}_1(\ ) \cdot {}_2(\ ) = \sum_{\subseteq} \mathbf{I}[\ \subseteq\ ] \cdot (-1)^{|\ 1|\ |} \mathbf{I}[\ \subseteq\ ]$$
$$= (-1)^{|\ 1|} \cdot \sum_{\subseteq} (-1)^{|\ |} \mathbf{I}[\ \subseteq\ \subseteq\ ]$$
$$= (-1)^{|\ 1|} \cdot (-1)^{|\ 1|} \cdot \sum^{|\ 2|-|\ 1|} (-1)\ \mathcal{C}^{|\ 2|-|\ 1|} \quad \text{where } \mathcal{C}\ = \frac{!}{!(\ -\ )!}$$

[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]

$$= (-1)^{|\ 1|} \cdot (1-1)^{|\ 2|-|\ 1|} = 0$$

Consequently, the probability to choose the no purchase option from the o er set $\{\ \setminus\ \}$ is given by

$$\mathbb{P}\ (\ \setminus\ ) = \sum_{\subseteq} (\ ) = \sum_{\subseteq} (\ ) \cdot (-1)^{|\ |} \cdot \mathbf{I}[\ \subseteq\ ] \tag{A3}$$
$$= \sum_{\subseteq} (\ ) \cdot (-1)^{|\ |} \cdot (\ )$$

Then it follows that

$$\begin{aligned}\sum_{\subseteq}(-1)^{|\ |-|\ |}\cdot\mathbb{P}(\quad\setminus\quad)&=\sum_{\subseteq}\mathbb{P}(\quad\setminus\quad)\cdot(-1)^{|\ |\ |\ |}\mathbf{I}[\quad\subseteq\quad] \\ &=(-1)^{|\ |}\cdot\sum_{\subseteq}\mathbb{P}(\quad\setminus\quad)\cdot\quad(\quad) \\ &=(-1)^{|\ |}\cdot\sum_{\subseteq}\sum_{_1\subseteq}\quad(\quad)\cdot(-1)^{|\ _1|}\cdot\quad{}_1(\quad)\cdot\quad(\quad) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\left[\text{by Equation (A3)}\right] \\ &=(-1)^{|\ |}\cdot\sum_{_1\subseteq}\quad(\quad)\cdot(-1)^{|\ _1|}\cdot\sum_{\subseteq}\quad{}_1(\quad)\cdot\quad(\quad) \\ &=(-1)^{|\ |}\cdot\quad(\quad)\cdot(-1)^{|\ |}\quad\left[\text{by Equation (A2)}\right] \\ &=\quad(\quad)\end{aligned}\tag{A4}$$

Now it remains to prove the uniqueness of probability distribution function obtained from purchasing transactions data under the CTC choice model. Note that Equation (A3) relates probability distribution over consideration sets to the choice frequencies $\mathbb{P}(\quad\setminus\quad)$ through the system of linear equations:

$$\mathbb{P}(\quad\setminus\quad)=\sum_{\subseteq}\quad(\quad)\cdot(-1)^{|\ |}\cdot\quad(\quad)\ \forall\quad\subseteq\quad\Longleftrightarrow\quad=\quad\cdot\tag{A5}$$

where $=(\quad)_{\subseteq}$ denotes the $|2\ |\times 1$ vector of choice fractions and $=(\quad)_{\subseteq}$ denotes the $|2\ |\times 1$ vector that represents the probability distribution function over consideration sets. is the $|2\ |\times|2\ |$ matrix such that s entry corresponding to the row and column is equal to $(-1)^{|\ |}\cdot\quad(\quad)$. Therefore, the relation between the choice frequencies and the underlying model can be represented in a compact form as $=\quad\cdot$. Then the proof of the uniqueness of reduces to showing that $\det(\quad)\neq 0$. From Equation (A4) we have

$$(\quad)=(-1)^{|\ |}\cdot\sum_{\subseteq}\Pr(\quad\setminus\quad)\cdot\quad(\quad)\ \forall\quad\subseteq\quad\Longleftrightarrow\quad=\quad\cdot$$

which establishes alternative linear relationship between choice frequencies $\Pr(\quad\setminus\quad)$ and the model parameters in a compact form as $=\quad\cdot$, where is the $|2\ |\times|2\ |$ matrix such that s entry corresponding to the row and column is equal to $(-1)^{|\ |}\cdot\quad(\quad)$. Therefore, we get

$$\begin{aligned}&=\quad\cdot\quad=\quad\cdot\quad\cdot\qquad\left[\text{by Equation (A5)}\right] \\ &\Longrightarrow\quad=\quad\cdot\quad\Longrightarrow\ \det(\quad)=\det(\quad)\cdot\det(\quad) \\ &\Longrightarrow 1=\det(\quad)\cdot\det(\quad)\ \Longrightarrow\ \det(\quad)\neq 0\end{aligned}$$

□

The next result will be invoked in the proof of upcoming Proposition 4.

LEMMA A2. *The combinatorial identity below is valid*

$$- \sum \mathcal{C} \cdot \left[\bar{\sum}_{-} (-1)^{-} \mathcal{C}^{-} \right] = \begin{cases} 1 & \text{if} \quad = \\ 0 & \text{if} \end{cases} \tag{A6}$$

where *when* = .

Proof: Let us consider two cases:

Case 1: = . In this case by invoking the assumptions of the lemma.

$$-\sum \mathcal{C} \cdot \left[\bar{\sum}_{-} (-1)^{-} \mathcal{C}^{-} \right] = -\sum \mathcal{C} \cdot \left[\bar{\sum}_{-} (-1)^{-} \mathcal{C}^{-} \right]$$

$$= -\sum \bar{\sum}_{-} (-1) \cdot \frac{!}{! \; !(- -)!} = 1$$

where the last equality is proved by induction on = − :

Base case: = 1.

$$-\sum \bar{\sum}_{-} (-1) \cdot \frac{!}{! \; !(- -)!} = -\bar{\sum} (-1)^{-} \mathcal{C} = (-1) \cdot \bar{\sum} (-1) \mathcal{C}$$

$$= (-1) \cdot ((1-1) - (-1)) = 1$$

Induction hypothesis: = .

Induction step: = + 1.

$$-\bar{\sum}_{- -} \bar{\sum}_{-} (-1) \cdot \frac{!}{! \; !(- -)!} \quad [\text{since} \quad = \quad - \quad - 1]$$

$$= -\bar{\sum}_{-} \bar{\sum}_{- -} (-1) \cdot \frac{!}{! \; !(- -)!} + \sum (-1) \cdot \frac{!}{!(-)!(-)!}$$

$$= -\bar{\sum}_{-} \bar{\sum}_{- -} (-1) \cdot \frac{!}{! \; !(- -)!} + \frac{!}{!(-)!} \cdot \sum (-1) \cdot \frac{!}{!(-)!}$$

$$= -\bar{\sum}_{-} \bar{\sum}_{- -} (-1) \cdot \frac{!}{! \; !(- -)!}$$

$$= -\bar{\sum}_{- -} \bar{\sum}_{-} (-1) \cdot \frac{!}{! \; !(- -)!} - \bar{\sum} (-1)^{- -} \cdot \frac{!}{(- -)! \; ! \; !}$$

$$= 1 - \bar{\sum} (-1)^{- -} \cdot \frac{!}{(- -)! \; ! \; !} \quad [\text{by induction hypothesis,} \quad = \quad - \;]$$

$$= 1 + (-1)^{- -} \cdot \frac{!}{(-)!} \cdot \bar{\sum} (-1) \cdot \frac{(-)!}{(- -)! \; !}$$

$$= 1$$

Case 2: . the last equality is proved by induction on $= - :$

Base case: $=1$. Then $= -1 \geq$, so that $\min( ) =$. And we have that

$$- \sum \mathcal{C} \cdot \left[\sum_{\_}^{\_} (-1) \mathcal{C}^{-} \right] = -\sum \mathcal{C} \cdot \left[\sum_{\_}^{\_} (-1) \mathcal{C}^{-} \right]$$

$$= (-1) \sum (-1) \mathcal{C} = 0$$

Induction hypothesis: $=$.

Induction step: $= +1$.

Condition 1: $-$. Then $\min(-) =$ and $\min(- -1) =$. We have that

$$-\sum \mathcal{C} \cdot \left[\sum_{\_}^{\_} (-1) \mathcal{C}^{-} \right] = -\sum \mathcal{C} \cdot \left[\sum_{\_\ \_}^{\_} (-1) \mathcal{C}^{-} \right] \text{ [since } = - -1]$$

$$= -\sum \mathcal{C} \cdot \left[\sum_{\_}^{\_} {}_{\_} (-1) \mathcal{C}^{-} \right] - \sum \mathcal{C} \cdot (-1)^{-} \cdot \mathcal{C}_{\_\ \_}^{\ \ \ -}$$

$$= -\sum \mathcal{C} \cdot (-1)^{-} \cdot \mathcal{C}_{\_\ \_}^{\ \ \ -} \quad [\text{ by induction hypothesis, } = -]$$

$$= (-1)^{-} \cdot \sum (-1) \cdot \mathcal{C} \cdot \mathcal{C}_{\_\ \_}^{\ \ \ -} = (-1)^{-} \cdot \frac{!}{!} \cdot \sum (-1) \cdot \frac{(-)!}{!(-)!(- -)!}$$

Now it is su cient to show that $\sum (-1) \cdot \frac{-}{\_\ \_\ \_} = 0$. We prove it by induction on . For $= 0$, it follows that $\sum (-1) \cdot \frac{-}{\_\ \_\ \_} = \sum (-1) \cdot \frac{}{\_} = -\sum (-1) \cdot \mathcal{C} = 0$. Assuming that the result holds for $=$, we prove it for $= +1$:

$$\sum (-1) \cdot \frac{(-)!}{!(-)!(- -1-)!} = \sum (-1) \cdot \frac{(-)!(- -)}{!(-)!(- -)!}$$

$$= (-) \cdot \sum (-1) \cdot \frac{(-)!}{!(-)!(- -)!} - \sum (-1) \cdot \frac{(-)!}{!(-)!(- -)!}$$

$$= -\sum (-1) \cdot \frac{(-)!}{!(-)!(- -)!} \quad [\text{by induction hypothesis, } =]$$

$$= -\sum (-1) \cdot \frac{(-)!}{!(-)!(- -)!}$$

$$= -\sum (-1) \cdot \frac{(-)!}{(-1)!(-)!(- -)!}$$

$$= \sum^{-} (-1) \cdot \frac{(-1-)!}{!(-1-)!(- -1-)!}$$

$$= \sum^{-} (-1) \cdot \frac{((-1)-)!}{!((-1)-)!((-1)- -)!}$$

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

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

$$
\begin{aligned}
&-\sum^{\bar{}}\ \bar{}\ \mathcal{C}\ \cdot\Big[\ \sum_{\bar{}\ \bar{}}^{\bar{}}(-1)^{\ }\ \mathcal{C}^{\ \bar{}}\ \Big]=-\sum^{\bar{}}\mathcal{C}\ \cdot\Big[\ \sum_{\bar{}\ \bar{}}^{\bar{}}(-1)^{\ }\ \mathcal{C}^{\ \bar{}}\ \Big]+\sum(-1)^{\ }\ \mathcal{C}\\
&=-\sum^{\bar{}}\mathcal{C}\ \cdot\Big[\ \sum_{\bar{}\ \bar{}}^{\bar{}}(-1)^{\ }\ \mathcal{C}^{\ \bar{}}\ \Big]\\
&=-\sum^{\bar{}}\mathcal{C}\ \cdot\Big[\ \sum_{\bar{}\ \ \ \bar{}}^{\bar{}}(-1)^{\ }\ \mathcal{C}^{\ \bar{}}\ \Big]-\sum^{\bar{}}\mathcal{C}\ \cdot(-1)^{\ \bar{}\ \bar{}}\cdot\mathcal{C}_{\ \bar{}\ \bar{}}\\
&=-\sum^{\bar{}}\mathcal{C}\ \cdot(-1)^{\ \bar{}\ \bar{}}\cdot\mathcal{C}_{\ \bar{}\ \bar{}}\quad\text{[by induction hypothesis,}\quad=\ -\ \text{]}\\
&=(-1)^{\ \bar{}}\quad\cdot\sum^{\bar{}}(-1)^{\ }\cdot\mathcal{C}^{\ }\cdot\mathcal{C}_{\ \bar{}\ \bar{}}=(-1)^{\ \ \bar{}}\cdot\frac{!}{!}\cdot\sum^{\bar{}}(-1)^{\ }\cdot\frac{(\ -\ )!}{!(\ -\ )!(\ -\ -\ )!}
\end{aligned}
$$

Now it is su cient to prove that $\sum^{\ \bar{}}(-1)^{\ }\cdot\frac{\bar{}}{\bar{}\ \bar{}}=0$. We prove it by induction on $-$. For $-\ =0$, it follows that $\sum^{\ \bar{}}(-1)^{\ }\cdot\frac{\bar{}}{\bar{}\ \bar{}}=\sum^{\ \bar{}}(-1)^{\ }\cdot\frac{}{\bar{}\ \bar{}}=\frac{}{\bar{}}\sum^{\ \bar{}}(-1)^{\ }\cdot\mathcal{C}^{\ \bar{}}=0$. Assuming that the result holds for $-\ =\ $, we prove it for $-\ =\ +1$:

$$
\begin{aligned}
&\sum^{\bar{}}(-1)^{\ }\cdot\frac{(\ -\ )!}{!(\ -\ -1-\ )!(\ -\ -\ )!}=\sum^{\bar{}}(-1)^{\ }\cdot\frac{(\ -\ )!(\ -\ -\ )}{!(\ -\ -\ )!(\ -\ -\ )!}\\
&=(\ -\ )\cdot\sum^{\bar{}}(-1)^{\ }\cdot\frac{(\ -\ )!}{!(\ -\ -\ )!(\ -\ -\ )!}-\sum^{\bar{}}(-1)^{\ }\cdot\frac{(\ -\ )!}{!(\ -\ -\ )!(\ -\ -\ )!}\\
&=-\sum^{\bar{}}(-1)^{\ }\cdot\frac{(\ -\ )!}{!(\ -\ -\ )!(\ -\ -\ )!}\quad\text{[by induction hypothesis,}\quad=\ -\ \text{]}\\
&=-\sum^{\bar{}}(-1)^{\ }\cdot\frac{(\ -\ )!}{!(\ -\ -\ )!(\ -\ -\ )!}\\
&=-\sum^{\bar{}}(-1)^{\ }\cdot\frac{(\ -\ )!}{(\ -1)!(\ -\ -\ )!(\ -\ -\ )!}\\
&=\sum^{\bar{}\ \bar{}}(-1)^{\ }\cdot\frac{(\ -1-\ )!}{!(\ -\ -1-\ )!(\ -\ -1-\ )!}\\
&=\sum^{\bar{}\ \ \bar{}}(-1)^{\ }\cdot\frac{((\ -1)-\ )!}{!((\ -1)-\ -\ )!((\ -1)-\ -\ )!}\\
&=0\quad\text{[by induction hypothesis]}
\end{aligned}
$$

□

Proof of Proposition 4: It follows from the proposition that

$$
\begin{aligned}
(\ ) &= \sum_{\subseteq} \sum_{\supseteq\ \cup} (-1)^{\ |\ |-|\ \ \ \ |} \cdot \mathbb{P}(\ ) \cdot \mathbf{I}[|\ \ \cup\ \ | \leq \ \ \ \ |\ \ |] \\
&= \sum_{\subseteq} \sum_{\supseteq\ \cup} \mathbb{P}(\ ) \cdot (-1)^{\ |\ \cap\ |} \cdot (-1)^{|\ |-|\ \cup\ |} \cdot \mathbf{I}[|\ \ \cup\ \ | \leq \ \ \ \ |\ \ |] \\
&= \sum_{\subseteq} \mathbb{P}(\ ) \cdot (-1)^{\ |\ \cap\ |} \cdot \mathbf{I}[|\ \ \cup\ \ | \leq\ ] \cdot \sum_{\supseteq\ \cup} (-1)^{|\ |-|\ \cup\ |} \cdot \mathbf{I}[|\ \ |\ \ \ ] \\
&\quad \Big[\text{since the expression depends only on the cardinality of sets} \quad \text{, the summation over the sets} \\
&\quad \text{is reduced to the summation over the cardinality of sets} \quad \Big] \\
&= \sum_{\subseteq} \mathbb{P}(\ ) \cdot (-1)^{\ |\ \cap\ |} \cdot \mathbf{I}[|\ \ \cup\ \ | \leq\ ] \cdot \sum_{-|\ \cup\ |}^{-|\ \cup\ |} (-1)^{\ } \mathcal{C}^{\ -|\ \cup\ |}_{\ } \\
&\quad \left[\text{where } \mathcal{C}^{\ }_{\ } = \frac{!}{!(\ \ -\ \ )!}\right]
\end{aligned}
$$

For every $\subseteq$ we de ne boolean functions $:2^{\ } \to \mathbb{R}$ and $:2^{\ } \to \mathbb{R}$ by

$$
\begin{aligned}
(\ ) &= \mathbf{I}[\ \ \subseteq\ \ |\ |\leq\ ] \\
(\ ) &= (-1)^{\ |\ \cap\ |} \cdot \mathbf{I}[|\ \ \cup\ \ | \leq\ ] \cdot \sum_{-|\ \cup\ |}^{-|\ \cup\ |} (-1)^{\ } \mathcal{C}^{\ -|\ \cup\ |}_{\ }
\end{aligned}
$$

Restricting consideration sets and o er sets by the size of up to (by the assumption of the proposition), we represent the probability to choose the no purchase option from the o er set through a linear combination of boolean functions $(\ )$ as follows:

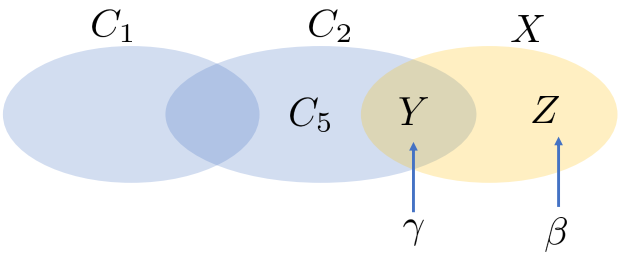
$$
\mathbb{P}(\ ) = \sum_{\subseteq} (\ ) \cdot \mathbf{I}[\ \ \subseteq\ \ |\ |\leq\ ] = \sum_{\subseteq} (\ ) \cdot (\ ) \tag{A7}
$$

Then for all $\subseteq$ such that $|\ \ |\ |\ \ |\leq$ we claim that

$$
\sum_{\subseteq} {}_1(\ ) \cdot {}_2(\ ) = \begin{cases} 1 & \text{if} \quad = \\ 0 & \text{otherwise} \end{cases} \tag{A8}
$$

Consequently, it follows from the claim that

$$
\begin{aligned}
&\sum_{\subseteq} \mathbb{P}(\ ) \cdot (-1)^{\ |\ \cap\ |} \cdot \mathbf{I}[|\ \ \cup\ \ | \leq\ ] \cdot \sum_{-|\ \cup\ |}^{-|\ \cup\ |} (-1)^{\ } \mathcal{C}^{\ -|\ \cup\ |}_{\ } \qquad (A9) \\
&= \sum_{\subseteq} \mathbb{P}(\ ) \cdot (\ \ )(\ ) = \sum_{\subseteq} \sum_{1\subseteq} (\ \ ) \cdot {}_1(\ ) \cdot (\ ) \\
&= \sum_{1\subseteq} (\ \ ) \cdot \sum_{\subseteq} {}_1(\ ) \cdot (\ ) = (\ ) \quad \Big[\text{by Equation (A8)}\Big]
\end{aligned}
$$



$C_1 \not\subset C_2$ $n - |C_1| - |C_5| >$ C_5 $C_2 \setminus \{C_1 \cap C_2\}$, Y $X \cap C_5$, Z
$X \setminus Y$ γ β Y Z

Now to complete the proof of the proposition, it is su cient to prove the claim and show the uniqueness of the solution. We prove the claim by considering two di erent cases.

Case 1: $\subseteq$.

$$\sum_{\subseteq} {}_1(\ )\cdot\ {}_2(\ ) = \sum_{\subseteq} (-1)^{\ }\cdot \mathbf{I}[|\ |\le\ ]\cdot \mathbf{I}[\ \cap\ =\varnothing]\cdot \mathbf{I}[|\ |\le\ -|\ |] \times \sum_{-|\ |-|\ 2|}^{-|\ |-|\ 2|} (-1)^{\ }\mathcal{C}^{\ -|\ |-|\ 2|}_{-|\ |-|\ 2|}$$

[in this case, $\cap\ =\varnothing$, $|\ \cap\ |=0$, $|\ \cap\ |=0$, and $|\ \cup\ |=|\ |+|\ |$]

$$= -\sum_{\subseteq} \mathbf{I}[|\ |\le\ ]\cdot \mathbf{I}[\ \cap\ =\varnothing]\cdot \mathbf{I}[|\ |\le\ -|\ |]\cdot \sum_{-|\ 2|}^{-|\ 2|-|\ |} (-1)^{\ }\mathcal{C}^{\ -|\ 2|-|\ |}_{-|\ 2|\ \ -|\ |}$$

[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]

$$= -\sum^{-|\ 2|\ \ -|\ 1|} \mathcal{C}^{\ -|\ 1|}_{\ } \cdot \left[\ \sum_{-|\ 2|\ \ -}^{-|\ 2|-} (-1)^{\ }\mathcal{C}^{\ -|\ 2|-}_{-|\ 2|\ \ -}\ \right] \left[\ \text{where}\ \ \text{- cardinality of set}\ \right]$$

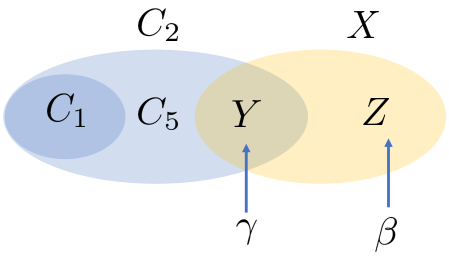
$$= \begin{cases} 1 & \text{if}\ \ = \\ 0 & \text{if}\ \ \subset \end{cases}$$

where the last equality follows by invoking Lemma A2, where $= -|\ |$, $= -|\ |$, and $= -|\ |$.

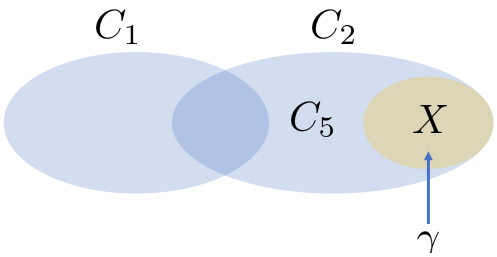
Case 2: $\not\subseteq$.

$$\sum_{\subseteq} {}_1(\ )\cdot\ {}_2(\ ) = \sum_{\subseteq} (-1)^{|\ \cap\ 2|}\cdot \mathbf{I}\left[\ \cap\ =\varnothing\ |\ |\le\ \right]\cdot \mathbf{I}\left[|\ \cup\ |\le\ \right] \times \sum_{-|\ \cup\ 2|}^{-|\ \cup\ 2|} (-1)^{\ }\mathcal{C}^{\ -|\ \cup\ 2|}_{-|\ \cup\ 2|}$$

[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]



$C_1 \subset C_2$ $k < n$ $n - |C_1| - |C_5| >$
C_5 $C_2 \setminus \{C_1 \cap C_2\}$, Y $X \cap C_5$, Z $X \setminus Y$ γ β
Y Z



$C_1 \not\subset C_2$ $n - |C_1| - |C_5|$ C_5 $C_2 \setminus \{C_1 \cap C_2\}, X \subseteq C_5$ γ
X

$$= \begin{cases} \sum^{|\ 5|}(-1) \quad \cdot \mathcal{C}^{|\ 5|} \cdot \left[\sum \quad {}^{-|\ 2|}\ {}^{-|\ 1|-|\ 5|}\ \mathcal{C}^{-|\ 1|-|\ 5|} \cdot \left[\sum {}^{-\ -|\ 2|}_{-\ -|\ 2|}(-1)\ \mathcal{C}^{-\ -|\ 2|} \right] \right] \\ \qquad\qquad \text{if} \quad \not\subset \quad \text{and} \quad -| \quad |-| \quad | \quad 0 \text{ see Figure A1} \\ \sum^{|\ 5|}(-1) \quad \cdot \mathcal{C}^{|\ 5|} \cdot \left[\sum \quad {}^{-|\ 2|}\ {}^{-|\ 2|}\ \mathcal{C}^{-|\ 2|} \cdot \left[\sum {}^{-\ -|\ 2|}_{-\ -|\ 2|}(-1)\ \mathcal{C}^{-\ -|\ 2|} \right] \right] \\ \qquad\qquad \text{if} \quad \subset \quad \text{see Figure A2} \\ \sum^{|\ 5|}(-1) \quad \cdot \mathcal{C}^{|\ 5|} \cdot \left[\sum {}^{-|\ 2|}_{-|\ 2|}(-1)\ \mathcal{C}^{-|\ 2|} \right] \\ \qquad\qquad \text{if} \quad \not\subset \quad \text{and} \quad -| \quad |-| \quad |=0 \text{ see Figure A3} \end{cases}$$

$$\left[\text{where} \quad = \quad \setminus \{ \quad \cap \quad \} \right]$$

$$= \begin{cases} \left[-\sum^{|\ 5|}(-1) \quad \cdot \mathcal{C}^{|\ 5|} \right] \cdot \left[\sum \quad {}^{-|\ 2|}\ {}^{-|\ 1|-|\ 5|}\ \mathcal{C}^{-|\ 1|-|\ 5|} \cdot \left[\sum {}^{-\ -|\ 2|}_{-\ -|\ 2|}(-1)\ \mathcal{C}^{-\ -|\ 2|} \right] \right] \\ \qquad\qquad \text{if} \quad \not\subset \quad \text{and} \quad -| \quad |-| \quad | \quad 0 \\ \left[-\sum^{|\ 5|}(-1) \quad \cdot \mathcal{C}^{|\ 5|} \right] \cdot \left[\sum \quad {}^{-|\ 2|}\ {}^{-|\ 2|}\ \mathcal{C}^{-|\ 2|} \cdot \left[\sum {}^{-\ -|\ 2|}_{-\ -|\ 2|}(-1)\ \mathcal{C}^{-\ -|\ 2|} \right] \right] \\ \qquad\qquad \text{if} \quad \subset \\ \left[-\sum^{|\ 5|}(-1) \quad \cdot \mathcal{C}^{|\ 5|} \right] \cdot \left[\sum {}^{-|\ 2|}_{-|\ 2|}(-1)\ \mathcal{C}^{-|\ 2|} \right] \quad \text{if} \quad \not\subset \quad \text{and} \quad -| \quad |-| \quad |=0 \end{cases}$$

$$= 0$$

where the last equality follows since | | 0, and $\sum^{|\ ^5|}(-1)\ \cdot \mathcal{C}^{|\ ^5|} = 0$.

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

$$\mathbb{P}\ (\ ) = \sum_{\subseteq}\ (\ )\cdot\ \ (\ )\ \ \forall\ \ \subseteq\ \ \Longleftrightarrow\ \ =\ \cdot \tag{A10}$$

where $=(\ )_{\subseteq}$ denotes the $|2\ |\times 1$ vector of choice fractions and $=(\ )_{\subseteq}$ denotes the $|2\ |\times 1$ vector that represents the probability distribution function over consideration sets. is the $|2\ |\times|2\ |$ matrix such that s entry corresponding to the row and column is equal to ( ). As a result, the relation between the choice frequencies and the underlying model can be represented in a compact form as $=\ \cdot$. Then the proof of the uniqueness of reduces to showing that $\det(\ ) \neq 0$. It follows from Equation (A9) that

$$(\ ) = \sum_{\subseteq}\mathbb{P}\ (\ )\cdot\ \ (\ )\ \ \forall\ \ \subseteq\ \ \Longleftrightarrow\ \ =\ \cdot$$

which provides another relationship between choice frequencies $\mathbb{P}\ (\ )$ and the model parameters in a linear form as $=\ \cdot$, where is the $|2\ |\times|2\ |$ matrix such that s entry corresponding to the row and column is equal to ( ). Therefore, we get

$$\begin{aligned}
&=\ \cdot\ =\ \cdot\ \cdot \quad \left[\text{by Equation (A10)}\right]\\
&\Longrightarrow\ \ =\ \cdot\ \ \Longrightarrow \det(\ ) = \det(\ )\cdot\det(\ )\\
&\Longrightarrow 1 = \det(\ )\cdot\det(\ )\ \Longrightarrow\ \det(\ ) \neq 0
\end{aligned}$$

□

We start this subsection with the proof of Proposition 6. We then prove Lemmas A3, A4, and A5 followed by the proof of Proposition 7 which invokes these lemmas.

Proof of Proposition 6: It follows directly from the proof of Proposition 1 that $\subseteq$. Then, it remains to show that the RUM model class is not a speci c case of the GCS model class. To this end, we provide a particular example of the RUM model class resulting in customers choice frequencies that are inconsistent with the GCS choice rule.

Let $=\{$ $\}$, and recall the presence of the no purchase option . Then let $:\mathscr{L}\rightarrow$ $[0\ 1]$ denote a speci cation of RUM class such that customers sample either preference list $=$ $\{$ $\}$ with probability $\in(0\ 1)$ or preference list $=\{$ $\}$ with probability $1-$.

Consequently, probability distribution function over preference lists results in the following choice frequencies:

$$\mathbb{P}(\{\quad\}) = \quad \mathbb{P}(\{\quad\}) = 1 \Rightarrow \quad \text{is preferred to} \quad \text{(by Proposition 5)},$$
$$\mathbb{P}(\{\quad\}) = 1 - \quad \mathbb{P}(\{\quad\}) = 1 \Rightarrow \quad \text{is preferred to} \quad \text{(by Proposition 5)}.$$

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

LEMMA A3. *Assume that for all consideration sets* $\subseteq$ *we have that*

$$\sum_{\subseteq} (-1)^{|\ |-|\ |} \mathbb{P}(\quad \setminus \quad) \geq 0$$

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

$$\sum_{\subseteq} (-1)^{|\ |-|\ |} \mathbb{P}(\quad \setminus \quad) \geq 0$$

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

Proof: Suppose that $\subseteq$ and $\subseteq$. Let denote $\setminus$. We can now establish the following chain of equalities:

$$\begin{aligned}
\sum_{\subseteq} (-1)^{|\ |-|\ |} \mathbb{P}(\quad \setminus \quad) &= \sum_{\subseteq} (-1)^{|\ |-|\ |} \cdot \mathbb{P}(\{\quad \setminus \quad\} \setminus \quad) \\
&= \sum_{\subseteq} (-1)^{|\ |-|\ |} \cdot \mathbb{P}(\{\quad \setminus \quad\} \setminus \quad) \\
&= \sum_{\subseteq} \sum_{\subseteq} \sum_{\subseteq} (-1)^{|\ |-|\ |} \cdot (-1)^{|\ |-|\ |} \cdot \mathbb{P}(\{\quad \setminus \quad\} \setminus \quad) \\
&\quad \Big[\text{by invoking Lemma A1 for every} \quad \subseteq \quad \text{, where} \quad = \quad \setminus \quad = \\
&\quad = \quad \text{and} \quad (\quad \setminus \quad) = (-1)^{|\ |-|\ |} \cdot \mathbb{P}(\{\quad \setminus \quad\} \setminus \quad)\Big] \\
&= \sum_{\subseteq} \sum_{\subseteq} \sum_{\subseteq} (-1)^{|\ |\ |\ |-|\ |-|\ |} \cdot \mathbb{P}(\{\quad \setminus \quad\} \setminus \quad) \\
&= \sum_{\subseteq} \sum_{\in\langle\ \cup\ \rangle} (-1)^{|\ |\ |\ |-|\ |} \cdot \mathbb{P}(\quad \setminus \quad) \quad \Big[\text{where} \quad = \quad \cup \quad \text{, since} \quad \cap \quad = \varnothing\Big] \\
&= \sum_{\subseteq} \sum_{\in\langle\ \cup\ \rangle} (-1)^{|\ \cup\ |-|\ |} \cdot \mathbb{P}(\quad \setminus \quad) \quad \Big[\text{since} \quad \cap \quad = \varnothing\Big]
\end{aligned}$$

$$= \sum_{\subseteq} \sum_{\subseteq'} (-1)^{|\ '|-|\ |} \cdot \mathbb{P}(\quad \setminus \quad) \left[\text{where} \quad ' = \quad \cup \quad\right]$$

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

$$\text{since} \sum_{\subseteq'} (-1)^{|\ '|-|\ |} \cdot \mathbb{P}(\quad \setminus \quad) \geq 0 \text{ with strict inequality when } |\quad'| \leq 3\Big]$$

□

The following three lemmas will be used in the proof of upcoming Proposition 7.

LEMMA A4. *If a sample of sales transaction data satisfies Conditions 1, 2, and 3, then for all* $\in \quad \cap \quad$ *where* $\quad \subseteq \quad$ *and* $\quad \subseteq \quad$ *we have that* $\mathbb{P}(\quad) \geq \mathbb{P}(\quad)$. .

Proof: Prove the result by induction on the n=$|\quad|-|\quad|$. We consider $\quad \in \quad \cap \quad$ and $\quad \subseteq \quad$. For the base case $\quad = 0$ we have that $\quad = \quad$ and $\mathbb{P}(\quad) = \mathbb{P}(\quad)$. Assume that the result holds for $\quad = \quad$, i.e., $\quad = \quad$ and $|\quad|-|\quad| = \quad$. Then we prove it for $\quad = \quad + 1$. Let us suppose w.l.o.g. that $\quad = \quad \cup \{\quad\}$ and $\quad \notin \quad$. Next, assume, by contradiction, that $\mathbb{P}(\quad \setminus \{\quad\}) \quad \mathbb{P}(\quad)$. Consequently, by Condition 2 it follows that $\mathbb{P}(\{\quad\}) \quad \mathbb{P}(\{\qquad\})$. Then by Condition 1 we have that

$$\mathbb{P}(\{\quad\}) = \mathbb{P}(\{\qquad\}) \tag{A11}$$

It now follows that

$$\begin{aligned}
&\mathbb{P}(\{\quad\}) - \mathbb{P}(\{\qquad\}) \\
&= \Big(1 - \mathbb{P}(\{\quad\})\Big) - \Big(1 - \mathbb{P}(\{\qquad\}) - \mathbb{P}(\{\qquad\})\Big) \quad [\text{ by standard probability property}] \\
&= \Big(1 - \mathbb{P}(\{\quad\})\Big) - \Big(1 - \mathbb{P}(\{\qquad\}) - \mathbb{P}(\{\quad\})\Big) \quad \Big[\text{by Equation (A11)}\Big] \\
&= \Big(1 - \mathbb{P}(\{\quad\})\Big) - \Big(\mathbb{P}(\{\quad\}) - \mathbb{P}(\{\qquad\})\Big) \quad [\text{by standard probability property}] \\
&= 1 - \mathbb{P}(\{\quad\}) - \mathbb{P}(\{\quad\}) + \mathbb{P}(\{\qquad\}) \quad 0 \quad \Big[\text{by Condition 3 and Lemma A3, when} \quad = \quad = \{\qquad\}\Big]
\end{aligned}$$

which contradicts to $\mathbb{P}(\{\quad\}) \quad \mathbb{P}(\{\qquad\})$. Then we have

$$\begin{aligned}
\mathbb{P}(\quad) &\leq \mathbb{P}(\quad \setminus \{\quad\}) = \mathbb{P}(\quad) \quad [\text{note that } |\quad|-|\quad| = \ ] \\
&\leq \mathbb{P}(\quad) \quad [\text{by induction hypothesis}]
\end{aligned}$$

Therefore, the result now follows by induction. □

LEMMA A5. *Consider* $\quad \in \quad \subseteq \quad \neq \quad$. *Then GCS choice model, with strict preference list* $\quad$ *and distribution over consideration sets* $\quad$ *where* $(\quad) \quad 0$ *if* $|\quad| \leq 3$, *implies the following list of implications:*

) $\mathbb{P}(\ \setminus\{\ \}) \quad \mathbb{P}(\ ) \implies \quad \succ \quad$ *and* $\forall\ ' \subseteq \qquad \in\ ' : \mathbb{P}(\ '\setminus\{\ \}) \quad \mathbb{P}(\ ')$,

) $\mathbb{P}(\ \setminus\{\ \}) = \mathbb{P}(\ ) \implies \quad \succ \quad$ *and* $\forall\ ' \subseteq \qquad \in\ ' : \mathbb{P}(\ '\setminus\{\ \}) = \mathbb{P}(\ ')$

) $\mathbb{P}(\ \setminus\{\ \}) \neq \mathbb{P}(\ ) \implies \mathbb{P}(\ \setminus\{\ \}) = \mathbb{P}(\ )$.

Proof: a) Suppose that $\mathbb{P}(\ \setminus\{\ \}) \quad \mathbb{P}(\ )$. Assume, by contradiction, that $\quad \succ \quad$. Then it can be inferred from purchase probability de nition under the GCS, see Equation (1), that $\mathbb{P}(\ \setminus\{\ \}) = \mathbb{P}(\ )$, which leads to a contradiction. As a result, we have that $\quad \succ \quad$ since preferences are strict and asymmetric. Then $\forall\ ' \subseteq \qquad \in\ '$ we establish that

$$\mathbb{P}(\ '\setminus\{\ \}) - \mathbb{P}(\ ') \geq (\{\qquad\}) \quad \Big[\text{by Equation (1)}\Big]$$

$$0 \quad \Big[\text{by Assumption that} \quad (\ ) \quad 0 \text{ if } |\ | \leq 3\Big]$$

b) Suppose that $\mathbb{P}(\ \setminus\{\ \}) = \mathbb{P}(\ )$. Assume, by contradiction, that $\quad \succ \quad$. Then it follows that

$$\mathbb{P}(\ \setminus\{\ \}) - \mathbb{P}(\ ) \geq (\{\qquad\}) \quad \Big[\text{by Equation (1)}\Big]$$

$$0 \quad \Big[\text{by Assumption that} \quad (\ ) \quad 0 \text{ if } |\ | \leq 3\Big]$$

which contradicts to the assumption above. As a result, we have that $\quad \succ \quad$, since preferences are strict and asymmetric. Then by Equation (1) we have that $\forall\ ' \subseteq \qquad \in\ ' : \mathbb{P}(\ '\setminus\{\ \}) = \mathbb{P}(\ ')$.

c) Suppose that $\mathbb{P}(\ \setminus\{\ \}) \neq \mathbb{P}(\ )$. Then it is straightforward to verify that $\mathbb{P}(\ \setminus\{\ \}) \quad \mathbb{P}(\ )$, since the following inequality holds from the Lemma A4: $\mathbb{P}(\ \setminus\{\ \}) \geq \mathbb{P}(\ )$. Consequently, invoking the implication from part), we have $\quad \succ \quad$, and by Equation (1) we obtain that $\mathbb{P}(\ \setminus\{\ \}) = \mathbb{P}(\ )$ $\square$

Proof of Proposition 7: *Necessity:* if purchasing transactions data is consistent with the GCS choice model with strict preference list and distribution over consideration sets where $(\ ) \quad 0$ if $|\ | \leq 3$, then we claim that three axioms Condition 1, Condition 2, and Condition 3 are satis ed. First, it follows from Proposition 3 that Condition 3 is satis ed. Then Condition 1 and Condition 2 are satis ed by Lemma A5.

Sufficiency: we claim that the choice rule that satis es Condition 1, Condition 2, and Condition 3 is a GCS choice model with the strict preference list where no purchase option is the least preferred item, and probability distribution function over consideration sets such that $(\ ) \quad 0$ if $|\ | \leq 3$.

De ne a binary relation between products $\quad \subseteq \quad \neq \quad$, where $\quad = 1$ if $\mathbb{P}(\ \setminus\{\ \}) \quad \mathbb{P}(\ )$ for some $\quad \subseteq \quad$ s.t. $\quad \in \quad$ (note, by Condition 2 it implies that $\mathbb{P}(\ \setminus\{\ \}) \quad \mathbb{P}(\ )$

for all $\subseteq$ s.t. $\in$), and zero otherwise. We claim that is complete, asymmetric, and transitive binary relation.

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

$$
\begin{aligned}
&\mathbb{P}(\{\}) - \mathbb{P}(\{\}) \\
&= \Big(1 - \mathbb{P}(\{\})\Big) - \Big(1 - \mathbb{P}(\{\}) - \mathbb{P}(\{\})\Big) \quad \text{[by standard probability property]} \\
&= \Big(1 - \mathbb{P}(\{\})\Big) - \Big(1 - \mathbb{P}(\{\}) - \mathbb{P}(\{\})\Big) \quad \Big[\text{by Condition 2, see above}\Big] \\
&= \Big(1 - \mathbb{P}(\{\})\Big) - \Big(\mathbb{P}(\{\}) - \mathbb{P}(\{\})\Big) \quad \text{[by standard probability property]} \\
&= 1 - \mathbb{P}(\{\}) - \mathbb{P}(\{\}) + \mathbb{P}(\{\}) \quad 0 \quad \Big[\text{by Condition 3 and Lemma A3, where} \quad = \quad = \{\}\Big]
\end{aligned}
$$

which concludes that $=1$. Therefore, completeness of binary relation now follows.

Second, we establish that the de ned binary relation is asymmetric, i.e., if $=1$ then $=0$. Suppose that $\mathbb{P}(\setminus\{\}) \quad \mathbb{P}(\,)$ for some $\subseteq$, i.e., $=1$. Then by Condition 1 we have that $\mathbb{P}(\setminus\{\}) = \mathbb{P}(\,)$ (note, by Condition 2 we have that for all $' \subseteq$ s.t. $\in$ $' : \mathbb{P}(\,'\setminus\{\}) = \mathbb{P}(\,')$), which further implies that $=0$. As a result, asymmetry of binary relation now follows.

Third, we show the transitivity of binary relation , i.e., if $=1$ and $=1$ then $=1$ for all $\in$. Assume by contradiction that binary relation is not transitive. To this end, there exist $\in$ such that $=1$, $=1$, $=0$ with the following list of implications:

$$
\begin{aligned}
=1 &\Rightarrow \mathbb{P}(\setminus\{\}) \quad \mathbb{P}(\,) \quad [\text{ for some } \subseteq\ ] \\
&\Rightarrow \mathbb{P}(\{\}) \quad \mathbb{P}(\{\}) \quad \Big[\text{by Condition 2}\Big] \\
&\Rightarrow \mathbb{P}(\{\}) = \mathbb{P}(\{\}) \quad \Big[\text{by Condition 1}\Big] \qquad \text{(A12)} \\
&\Rightarrow \mathbb{P}(\{\}) = \mathbb{P}(\{\}) \quad \Big[\text{by Condition 2}\Big] \qquad \text{(A13)} \\
=1 &\Rightarrow \mathbb{P}(\setminus\{\}) \quad \mathbb{P}(\,) \quad [\text{ for some } \subseteq\ ] \\
&\Rightarrow \mathbb{P}(\{\}) \quad \mathbb{P}(\{\}) \quad \Big[\text{by Condition 2}\Big] \\
&\Rightarrow \mathbb{P}(\{\}) = \mathbb{P}(\{\}) \quad \Big[\text{by Condition 1}\Big] \qquad \text{(A14)} \\
&\Rightarrow \mathbb{P}(\{\}) = \mathbb{P}(\{\}) \quad \Big[\text{by Condition 2}\Big] \qquad \text{(A15)} \\
=0 &\Rightarrow \mathbb{P}(\setminus\{\}) \leq \mathbb{P}(\,) \quad [\text{ for some } \subseteq\ ]
\end{aligned}
$$

$$\Rightarrow \mathbb{P}\left(\quad \setminus \{\quad\}\right) = \mathbb{P}\left(\quad\right) \quad \Big[\text{by Lemma A4}\Big] \tag{A16}$$

$$\Rightarrow \mathbb{P}\left(\{\quad\}\right) = \mathbb{P}\left(\{\qquad\}\right) \quad \Big[\text{by Condition 2}\Big] \tag{A17}$$

Using the property of the choice rule, i.e., $\forall \quad \subseteq \quad : \sum_{r\in} {}_{+}\mathbb{P}\left(\quad\right) = 1$, for o er sets $\quad = \{\qquad\}$, $\quad = \{\qquad\}$, $\quad = \{\qquad\}$, and $\quad = \{\qquad\}$ we further establish the following list of implications:

For $\quad = \{\qquad\}$: $\mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) = 1$

$$\Rightarrow \mathbb{P}\left(\{\quad\}\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) = 1 \quad \Big[\text{by Equation (A13)}\Big]$$

$$\Rightarrow \mathbb{P}\left(\{\quad\}\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) = 1 \quad \Big[\text{by Equation (A14)}\Big]$$

$$\Rightarrow \mathbb{P}\left(\quad\right) = \mathbb{P}\left(\{\quad\}\right) - \mathbb{P}\left(\quad\right) \quad [\text{by standard probability property}] \tag{A18}$$

For $\quad = \{\qquad\}$: $\mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) = 1$

$$\Rightarrow \mathbb{P}\left(\{\quad\}\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) = 1 \quad \Big[\text{by Equation (A15)}\Big]$$

$$\Rightarrow \mathbb{P}\left(\{\quad\}\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) = 1 \quad \Big[\text{by Equation (A16)}\Big]$$

$$\Rightarrow \mathbb{P}\left(\quad\right) = \mathbb{P}\left(\{\quad\}\right) - \mathbb{P}\left(\quad\right) \quad [\text{by standard probability property}] \tag{A19}$$

For $\quad = \{\qquad\}$: $\mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) = 1$

$$\Rightarrow \mathbb{P}\left(\{\quad\}\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) = 1 \quad \Big[\text{by Equation (A17)}\Big]$$

$$\Rightarrow \mathbb{P}\left(\{\quad\}\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) = 1 \quad \Big[\text{by Equation (A12)}\Big]$$

$$\Rightarrow \mathbb{P}\left(\quad\right) = \mathbb{P}\left(\{\quad\}\right) - \mathbb{P}\left(\quad\right) \quad [\text{by standard probability property}] \tag{A20}$$

For $\quad = \{\qquad\}$: $\mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right) = 1$

$$\Rightarrow 0 = \mathbb{P}\left(\varnothing\right) - \mathbb{P}\left(\{\quad\}\right) - \mathbb{P}\left(\{\quad\}\right) - \mathbb{P}\left(\{\quad\}\right) + \mathbb{P}\left(\quad\right) + \mathbb{P}\left(\quad\right)$$

$$+ \mathbb{P}\left(\quad\right) - \mathbb{P}\left(\quad\right) \quad \Big[\text{since } \mathbb{P}\left(\varnothing\right) = 1, \text{ and by Equations (A18)-(A20)}\Big]$$

$$0 \quad \Big[\text{by Condition 3 and Lemma A3, where} \quad = \quad = \quad\Big]$$

which leads to a contradiction. Therefore, the preference relation $\quad$ is transitive. Since we proved that binary relation $\quad$ is complete, asymmetric, and transitive, it speci es strict preference list $\succ$ over products in $\quad$, s.t. $\quad \succ \quad$ i $\quad = 1$. In addition, it immediately follows from the axioms that

is the least preferred item in the product universe according to the preference list $\succ$, i.e., for all $\in$ we have that $= 0$:

$$\mathbb{P}(\varnothing) - \mathbb{P}(\{\ \}) \quad 0 \left[\text{by Condition 3 and Lemma A3, where} \quad = \quad = \{\ \}\right]$$

which implies that $= 0$ by de nition.

Next, we prove that

$$\mathbb{P}(\ ) = \mathbb{P}(\ '\setminus\{\ \}) - \mathbb{P}(\ ') \ \ \forall \quad \in \quad \text{s.t.} \quad \subseteq$$

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

Case 1: product is not the least preferred item in . In other words there exists $\in$ s.t. $\succ$. Then by de nition of the binary relation we have that $\mathbb{P}(\ \setminus\{\ \}) \quad \mathbb{P}(\ )$, and the result now follows:

$$\begin{aligned}\mathbb{P}(\ ) &= \mathbb{P}(\ \setminus\{\ \}) \quad [\text{by Condition 1 }]\\ &= \mathbb{P}(\ '\setminus\{\ \}) - \mathbb{P}(\ ') \quad [\text{by induction hypothesis,}\\ &\text{and note that} \quad \in\ ' \text{ since } \quad \succ \quad]\end{aligned}$$

Case 2: product is the least preferred item in . Consider o er set $= \{ \qquad _{-} \}$ such that w.l.o.g. $_{-} \succ \ \succ \ \succ \ \succ$. Assuming $\in$, we can now establish the following chain of equalities:

$$\begin{aligned}\mathbb{P}(\ ) &= 1 - \mathbb{P}(\ ) - \sum^{-} \mathbb{P}(\ )\\ &= -\mathbb{P}(\ ) + \mathbb{P}(\varnothing) - \sum^{-} \mathbb{P}(\{ \qquad _{-} \})\\ &= -\mathbb{P}(\ ) + \mathbb{P}(\varnothing) - \sum^{-} \mathbb{P}(\{ \qquad _{-} \}) \quad \left[\text{by Condition 1}\right]\\ &= -\mathbb{P}(\ ) + \mathbb{P}(\varnothing) - \sum^{-} \Big(\mathbb{P}(\{ \qquad _{-} \}) - \mathbb{P}(\{ \qquad _{-} \})\Big)\\ &\qquad [\text{by induction hypothesis}]\\ &= -\mathbb{P}(\ ) + \mathbb{P}(\{ \qquad _{-} \})\\ &= \mathbb{P}(\{ \qquad _{-} \}) - \mathbb{P}(\{ \qquad _{-} \}) = \mathbb{P}(\ '\setminus\{\ \}) - \mathbb{P}(\ ')\end{aligned}$$

Let us denote two particular sets $\quad$ and $\quad'$ as follows: $\quad = \quad \setminus\{\ '\setminus\{\quad\}\}$, $\quad' = \quad \setminus\ '$. We can now establish the following chain of equalities:

$$
\begin{aligned}
\mathbb{P}(\quad) &= \mathbb{P}(\ '\setminus\{\quad\}) - \mathbb{P}(\ ') \\
&= \mathbb{P}(\ '\setminus\{\quad\}) + \Bigg(\sum_{\subseteq}\sum_{\subseteq}(-1)^{|\ |-|\ |}\cdot\mathbb{P}(\quad\setminus\quad) - \mathbb{P}(\quad\setminus\quad)\Bigg) - \mathbb{P}(\ ') \\
&\qquad \Big[\text{by invoking Lemma A1, where} \quad = \quad = \quad = \quad \text{and} \quad (\quad\setminus\quad) = \mathbb{P}(\quad\setminus\quad)\Big] \\
&= \sum_{\subseteq}\sum_{\subseteq}(-1)^{|\ |-|\ |}\cdot\mathbb{P}(\quad\setminus\quad) - \mathbb{P}(\ ') \Big[\text{since} \quad \setminus \quad = \ '\setminus\{\quad\}\Big] \\
&= \sum_{\subseteq}\sum_{\subseteq}(-1)^{|\ |-|\ |}\cdot\mathbb{P}(\quad\setminus\quad) - \Bigg(\sum_{\subseteq\ '}\sum_{\subseteq}(-1)^{|\ |-|\ |}\cdot\mathbb{P}(\quad\setminus\quad) \\
&\qquad\qquad - \mathbb{P}(\quad\setminus\ ')\Bigg) - \mathbb{P}(\ ') \\
&\qquad \Big[\text{by invoking Lemma A1, where} \quad = \quad = \ ' \quad = \quad \text{and} \quad (\quad\setminus\quad) = \mathbb{P}(\quad\setminus\ ')\Big] \\
&= \sum_{\subseteq}\sum_{\subseteq}(-1)^{|\ |-|\ |}\cdot\mathbb{P}(\quad\setminus\quad) - \sum_{\subseteq\ '}\sum_{\subseteq}(-1)^{|\ |-|\ |}\cdot\mathbb{P}(\quad\setminus\quad) \Big[\text{since} \quad \setminus\ ' = \ '\Big] \\
&= \sum_{\in\langle\ '\cup\{\ r\}\rangle}\sum_{\subseteq}(-1)^{|\ |-|\ |}\cdot\mathbb{P}(\quad\setminus\quad) - \sum_{\subseteq\ '}\sum_{\subseteq}(-1)^{|\ |-|\ |}\cdot\mathbb{P}(\quad\setminus\quad) \Big[\text{since} \quad = \ '\cup\{\quad\}\Big] \\
&= \sum_{\in\langle\ '\rangle\uplus\{\ r\}}\sum_{\subseteq}(-1)^{|\ |-|\ |}\mathbb{P}(\quad\setminus\quad) \\
&= \sum_{\in\langle\ '\rangle\uplus\{\ r\}}(\quad) \quad \text{where} \quad (\quad) = \sum_{\subseteq}(-1)^{|\ |-|\ |}\mathbb{P}(\quad\setminus\quad) \\
&= \sum_{\subseteq}(\quad)\cdot\mathbf{I}[\quad\in\quad]\cdot\mathbf{I}[\quad\in\langle\ '\rangle\uplus\{\quad\}] \\
&= \sum_{\subseteq}(\quad)\cdot\mathbf{I}[\quad\in\quad]\cdot\mathbf{I}[\quad\succ\quad\forall\quad\in\quad\cap\quad\neq\quad] \\
&= \sum_{\subseteq}(\quad)\cdot\mathbf{I}[\quad\in\quad\cap\quad]\cdot\mathbf{I}[\quad\succ\quad\forall\quad\in\quad\cap\quad\neq\quad] \Big[\text{since we assume that} \quad\in\quad, \\
&\qquad\qquad \text{otherwise the choice probability is } 0\Big]
\end{aligned}
$$

which is exactly the equation to compute the probability to purchase $\quad\in\quad$ for the o er set $\quad\subseteq\quad$ under the GCS choice model. As a result, we also have $\mathbb{P}(\quad) = \sum_{\quad\subseteq\quad}(\quad)\cdot\mathbf{I}[\quad\cap\quad = \varnothing]$ becasue of the standard probability law, i.e., $\mathbb{P}(\quad) = 1 - \sum_{\ r\in\quad}\mathbb{P}(\quad)$. Note that the above chain of equations speci es probability distribution function $\quad$ over consideration sets. Moreover, it follows from Proposition 3 that $\quad$ is de ned uniquely. In order to complete the proof, we show that the preference relation $\succ$ is also de ned uniquely. Suppose, by contradiction, there is another strict

preference order $\succ'$ such that $\succ' \neq \succ$ and $\mathbb{P}_{\cdot}(\cdot)_{\succ'} = \mathbb{P}_{\cdot}(\cdot)_{\succ}$. Therefore there exist items $\in$ s.t. $\succ$ and $\succ'$. By de nition of the GCS choice rule, we have

$$\mathbb{P}(\{\quad\})_{\succ} = \sum_{\subseteq} \mathbf{I}[\quad \in \quad] \cdot (\quad)$$

$$\mathbb{P}(\{\quad\})_{\succ'} = \sum_{\subseteq} \mathbf{I}[\quad \in \quad] \cdot \mathbf{I}[\quad \in \quad] \cdot (\quad)$$

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

$$\mathbb{P}(\{\quad\})_{\succ} - \mathbb{P}(\{\quad\})_{\succ'} \geq (\{\quad\}) \quad 0 \quad \Big[\text{by Condition 3}\Big]$$

which contradicts to $\mathbb{P}_{\cdot}(\cdot)_{\succ'} = \mathbb{P}_{\cdot}(\cdot)_{\succ}$. $\square$

We start this section by showing how we can capture the impact of product features on consideration set formation and by providing the MINLP formulation for the estimation of the single class, logistic-based ICS (L-ICS) model. Then we describe the outer-approximation algorithm which is used to calibrate di erent variants of consider-then-choose models followed by the empirical validation of this algorithm. We nish this section by describing the EM algorithm to calibrate the GCS and CTC models.

To capture the impact of product features on consideration set formation, we use the following three ways to make propensity parameter a function of the product features, where is the observed th feature of product :

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

$$= + \sum +$$

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

$$\in \quad \text{i} \quad = \quad + \sum \quad + \quad \geq 0$$

Then the propensity of product is given by

$$= \Pr[\quad \in \quad] = \frac{\exp(\quad + \sum \quad)}{1 + \exp(\quad + \sum \quad)}$$

- *Decision tree-based ICS model (DT-ICS).* Here it is assumed that individuals decide which items to consider based on a tree with leaves $\in \{1\ 2\ \ \}$ to which we can associate a mean probability $\ $ of whether the item is going to be considered or not (see Murphy (2012)). Then we can write the probability to consider the item $\ $ in the following way:

$$= \Pr[\ \ \in\ \ ] = \sum\ \ \mathbb{I}[\ \ \in\ \ ] = \sum\ \ (\ \ )$$

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

- *Random forest-based ICS model (RF-ICS).* In this case, we assume that individuals $\ $rst 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 over $\ $tting problem of decision trees by adding more trees instead of building one big tree. We can write the probability of considering the item $\ $ as follows:

$$= \Pr[\ \ \in\ \ ] = \sum \frac{1}{\ \ } \ \ (\ \ )$$

  where $(\ \ )$ is the probability of considering item $\ $ according to the $\ $th decision tree.

Estimation methodology. In a similar spirit to the preliminaries in Section 4.1, we can formulate the maximum likelihood estimation problem for the single class, logistic-based ICS (L-ICS) model with product features in such a way so that we can apply the outer-approximation algorithm in Appendix A3.2 in order to calibrate it.

On the other hand, the calibration of the DT-ICS and RF-ICS models is more challenging. To this end, we need to estimate both the ranking $\ $ and the decision tree (or random forest). Intuitively, both decision trees and random forests are not limited to the generalized linear model class. Instead, they can model the relationship between the features in a non-linear way by bisecting the space into smaller and smaller regions. Note that if the ranking $\ $ is known, then the log-likelihood optimization is equivalent to calibrating a classi $\ $cation 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 $\ $nd $\ $. We could then heuristically iterate between these two steps of $\ $nding $\ $ and $\ $ until $\ $nding a $\ $xed-point or a time limit is reached. As a practical matter in Section 7, we assume that the ranking $\ $ is known a priori (see Section A5.2 for the details).

In this part of the section we formulate the maximum likelihood estimation problem for the logistic-based ICS (L-ICS), and then simplify it in such a way so that we can apply the outer-approximation algorithm in Section A3.3 in order to calibrate it. Recall that , $\forall$ $\neq$ is binary linear ordering variable such that $=1$ if product goes before product in the preference list $\succ$ (or, equivalently,), and $=0$ otherwise. The data log-likelihood function under this model is given by

$$\mathscr{L}(\quad) = \sum \left[\log \frac{\boldsymbol{\beta} \boldsymbol{X}_{j_t}}{1 + \boldsymbol{\beta} \boldsymbol{X}_{j_t}} + \sum_{k \in \ / \ t} [\quad _t \log \frac{1}{1 + \boldsymbol{\beta} \boldsymbol{X}_{k_t}}] \right]$$

and the ML problem can be represented in the following way

$$\begin{aligned} \max_{\boldsymbol{\beta}\ \boldsymbol{\delta}} \quad & \mathscr{L}(\quad) \\ \textbf{s.t.:} \quad & + \quad = 1 \quad \forall \quad \leq \\ & + \quad + \quad \leq 2 \quad \forall \quad \neq \quad \neq \\ & \in \{0 \ 1\} \quad \forall \end{aligned} \tag{A21}$$

where the constraints ensure that indeed represents a total order. In particular, the rst set of constraints ensures that either is preferred over 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 de ned as $=$, $\forall$ and rewrite the likelihood function in the following way

$$\mathscr{L}(\quad) = \sum \left[\log \frac{\boldsymbol{\beta} \boldsymbol{X}_{j_t}}{1 + \boldsymbol{\beta} \boldsymbol{X}_{j_t}} + \sum_{k \in \ / \ t} (\quad _t - 1) \log\left(\frac{1}{2}\right) + \sum_{k \in \ / \ t} [\log \frac{1}{1 + \sum_i \ _{ikj_t} \boldsymbol{X}_{ik_t}}] \right]$$

since if $_t = 1$ we have that $_t =$ $\forall$, and

$$_t \log \frac{1}{1 + \boldsymbol{\beta} \boldsymbol{X}_{k_t}} = \log \frac{1}{1 + \boldsymbol{\beta} \boldsymbol{X}_{k_t}} = \log \frac{1}{1 + \sum_i \ _{ikj_t} \boldsymbol{X}_{ik_t}} = (\quad _t - 1) \log\left(\frac{1}{2}\right) + \log \frac{1}{1 + \sum_i \ _{ikj_t} \boldsymbol{X}_{ik_t}}$$

if $_t = 0$ we have that $_t = 0 \ \ \forall$, and

$$_t \log \frac{1}{1 + \boldsymbol{\beta} \boldsymbol{X}_{k_t}} = 0 = -\log\left(\frac{1}{2}\right) + \log \frac{1}{1 + } = (\quad _t - 1) \log\left(\frac{1}{2}\right) + \log \frac{1}{1 + \sum_i \ _{ikj_t} \boldsymbol{X}_{ik_t}}$$

Let be the value of the largest component in vector , i.e., $=$. And we also de ne $\mathscr{L}(\quad)$ in the following way

$$\mathscr{L}(\quad) = \sum \left[\log \frac{\boldsymbol{\beta} \boldsymbol{X}_{j_t}}{1 + \boldsymbol{\beta} \boldsymbol{X}_{j_t}} + \sum_{k \in \ / \ t} [\log \frac{1}{1 + \sum_i \ _{ikj_t} \boldsymbol{X}_{ik_t}}] \right]$$

We can then formulate the MLE problem in terms of the variables ( ):

$$\max_{\boldsymbol{\beta}\,\boldsymbol{\tau}\,\boldsymbol{\delta}} \mathscr{L}(\quad) + \sum_{k \in \; / \; t} (\quad_t - 1)\log\left(\frac{1}{2}\right) \tag{A22}$$

$$\begin{aligned}
\textbf{s.t.:}\quad & \leq \quad \forall \\
& \leq \quad \forall \\
& \geq \; + \quad - \quad \forall \\
& \geq - \quad \forall \\
& + \quad = 1 \quad \forall \quad \leq \\
& + \quad + \quad \leq 2 \quad \forall \quad \neq \; \neq \\
& \in \{0\ 1\} \quad \forall
\end{aligned}$$

where the rst four sets of linear constraints ensure that = , ∀ , given that is a binary variable.

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) de ned below.

$$\min_{\boldsymbol{\theta}\,\boldsymbol{\tau}\,\boldsymbol{\delta}} \quad_1 \tag{P}$$

$$\begin{aligned}
\textbf{s.t.:}\quad & \leq \mathscr{L}(\quad) \\
& + \quad + \quad \leq 0 \\
& = \quad + \\
& \leq \quad \leq \\
& \in \{0\ 1\} \quad \forall
\end{aligned}$$

where $\mathscr{L}(\quad)$ is a concave function; , , , and are continuous decision variables; is a binary variable; , , , and are constant vectors; and and are lower and upper bounds of ,

respectively. Note that optimization problems (10) 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 de ne the linear constraint ( ) that is added to the optimization problem at the th iteration for given ( ), i.e., we linearize the convex constraint at ( ):

$$( ) = \{ : \mathscr{L}( ) + \frac{\mathscr{L}( )}{ } + \frac{\mathscr{L}( )}{ } - \geq 0 \in \mathcal{R} \}$$

The broad idea of the outer-approximation algorithm is that at the th iteration, we substitute convex constraint $\leq \mathscr{L}( )$ with a set of linear constraints $= \{ ( )\}|$.

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

$$\boldsymbol{\theta}\boldsymbol{\tau}_{1} \qquad (( ))$$

$$\begin{aligned} \textbf{s.t.:} \quad & \leq \mathscr{L}( ) \\ & + \quad + \quad \leq 0 \\ & = \quad + \end{aligned}$$

Note that solving MINLP (P) reduces to solving concave subproblem above if is known. Second, let us de ne the MILP subproblem for given and as follows

$$\boldsymbol{\theta}\boldsymbol{\tau}\boldsymbol{\delta}_{1} \qquad ( )$$

$$\begin{aligned} \textbf{s.t.:} \ & ( ) \in \\ & + \quad + \quad \leq 0 \\ & = \quad + \\ & \leq \quad \leq \\ & \in \{0 \ 1\} \quad \forall \end{aligned}$$

Note that solving MINLP (P) reduces to solving MILP subproblem if we approximate convex constraint $\leq \mathscr{L}( )$ with a set of linear constraints .

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 e ectively 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 the intersection of its collection of supporting half-spaces. To this end, the outer approximation de nes the optimization problem ( ) 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 nite sequence of convex problems (( )) and relaxed versions of a MILP ( ).

Algorithm 1 Outer-Approximation algorithm for optimization problem (P)

procedure OUTER-APPROXIMATION(P)

$= \mathcal{R} \times \mathcal{R}$, $= -\infty$, $= \infty$, $= 1$

Select arbitrary , i.e., it can be arbitrary full ranking

while | − | **do**

Solve concave subproblem ( ) such that $=$ * (i.e., the optimal objective function of ( )), and ( ) = (* *) (i.e., the optimal solution of ( ))

Set = ⁻ ∩ ( )

Solve MILP subproblem such that $=$ * (i.e., the optimal objective function of), and ( ) = (* * *) (i.e., the optimal solution of))

$= +1$

return ( ).

A3.3.1. Outer-approximation algorithm vs. cutting plane algorithm Note that Algorithm 1 to solve the optimization problem (P) requires the solution of both convex optimization problem (( )) and MILP ( ). The solution of the convex optimization problem (( )) in each iteration might be extremely computationally demanding, while in solving the MILP ( ) the computational work, on the other hand, might be more moderate, because for every iteration we need to solve the MILP problem ( ) which is the previous MILP problem (⁻) 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 nite sequence of MILP problem ( ), see Algorithm 2. Note that Algorithm 2 is identical to the Algorithm 1 except that we skip Step 5 in the cutting plane Algorithm 2. Even though the main iteration loop of Algorithm 1 is, generally, more e cient, we have global convergence for both Algorithms 1 and 2.

Algorithm 2 Cutting plane algorithm for optimization problem (P)

procedure CUTTING PLANE(P)

$= \mathcal{R} \times \mathcal{R}$, $= -\infty$, $= \infty$, $= 1$

Select arbitrary , i.e., it can be arbitrary full ranking

Select arbitrary , i.e., it can be arbitrary distribution over consideration sets

Set $=$ $\cdot$, $= -\infty$, $= \infty$

while $|$ $-$ $^{-}|$ **do**

Set $=$ $^{-} \cap$ ( )

Solve MILP subproblem such that $=$ *, $=$ * (i.e., the optimal objective function of), and ( ) = (* * *) (i.e., the optimal solution of))

$=$ $+1$

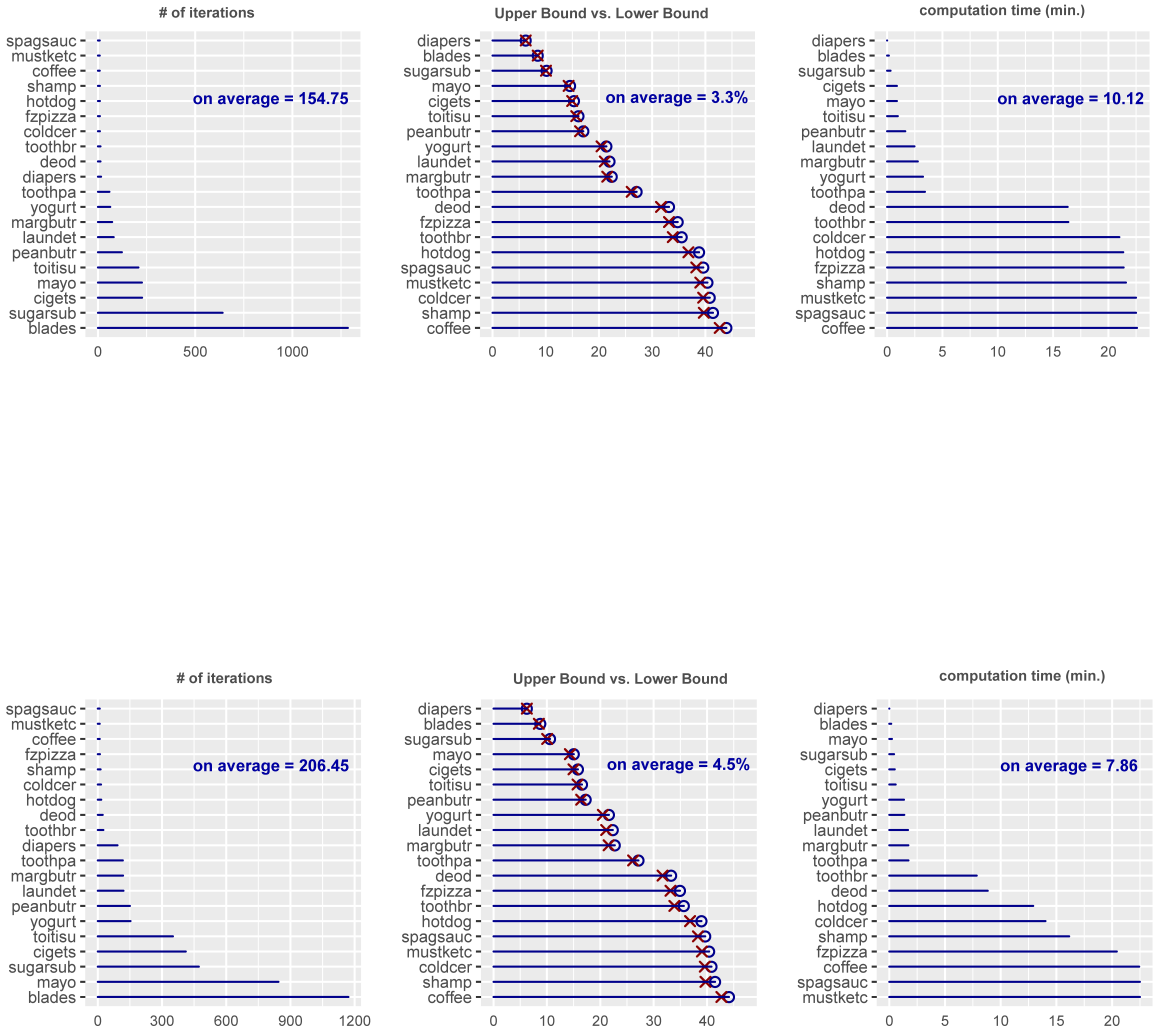
return ( ).

A3.3.2. Empirical validation of the algorithms. In this section, we analyze the performance of outer-approximation algorithm (1) and cutting plane algorithm (2) to estimate ICS model with IRI Academic dataset (see data descriptive statistics in Table A1). We limited the running time of the algorithms by 3 hours, and the precision was set to 1e-6. It follows from Figure A4 that the optimality gap of the outer approximation algorithm (1) to calibrate the ICS model is 3.3% on average over 20 product categories. On the other hand, it is shown in Figure A5 that the optimality gap of the cutting plane algorithm (2) to calibrate the ICS model is 4.5% on average over 20 product categories. Following these ndings, we apply outer-approximation algorithm (1) to calibrate the ICS model in our analysis as it provides signi cantly faster convergence to the optimal solution, which is consistent with previous studies.

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

A3.4.1. Estimation with aggregate level data. The log-likelihood function to calibrate the GCS model, after we reparametrize it by dividing all the transactions into segments, is given by

$$\log \mathscr{L}(\succ) = \sum \log \Big(\sum_{t} \prod_{\substack{j \in t \\ j \succ_{\sigma} j_t}} (1 -) \Big) \tag{A23}$$



where ≥ 0 is the weight of the class a priory, s.t. $\sum = 1$; denotes the set of o ered items at time ; $_t$ denotes the product purchased at time ; denotes the time horizon.

Unsurprisingly, 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 $\mathbf{x}$. Then, we compute the conditional expected value of the log-likelihood function $\mathrm{E}[\log \mathscr{L}(\mathbf{x})|\mathbf{x}]$ (the E , expectation, step). Next, the resulting expected log-likelihood function is maximized to compute new estimates $\mathbf{x}$ (the M , maximization, step), and we repeat the

algorithm until convergence to get a sequence of estimates $\{\mathbf{x} \quad = 1\ 2 \quad\}$. 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 classes, resulting in an initial allocation $\mathcal{D}\ \ \mathcal{D} \quad \mathcal{D}$, which form a partition of the collection of all the transactions. Then we set $= |\mathcal{D}\ |\ (\sum \quad |\mathcal{D}\ |)$. Then $\quad$ (i.e., $\succ \quad$) and $\quad$ for all $\in \{1 \quad\}, \quad \in \quad$ are obtained by solving the following optimization problem:

$$\max_{\boldsymbol{\theta}_h} \sum_{\in \mathcal{D}_h} \left(\log \quad _t + \sum_{\substack{j \in \ t \\ j \succ_\sigma \ j_t}} \log(1 - \quad) \right)$$

which is solved by using the outer approximation algorithm in Section A3.3.

E-step: we compute $\quad$, which is the membership probability of every transaction at time $\ $ to belong to the segment $\ $ (i.e., $\in \quad$, where $\ $ is the set of transactions in class $\ $) based on the parameter estimates $\{ \quad ^- \quad ^- \quad ^- \}$ and the purchasing transactions data $(\quad _t \quad)|\ :$

$$= \Pr\left(\in \ \middle|\ ^- \quad ^- \quad ^- \ (\quad _t \quad)| \ \right)$$

$$= \Pr\left(\in \ \middle|\ ^- \quad ^- \quad ^- \ (\quad _t \quad) \right) \ [\text{ independence of purchases}]$$

$$= \frac{\Pr\left(\left(\quad _t \quad \right) \middle| \quad - \quad - \quad - \quad \in \quad \right) \cdot \Pr\left(\quad \in \quad \middle| \quad - \quad - \quad - \right)}{\Pr\left(\left(\quad _t \quad \right) \middle| \quad - \quad - \quad - \right)} \quad [\text{ Bayes theorem }]$$

$$= \frac{\Pr\left(\left(\quad _t \quad \right) \middle| \quad - \quad - \quad - \quad \in \quad \right) \cdot \Pr\left(\quad \in \quad \middle| \quad - \quad - \quad - \right)}{\sum \quad \Pr\left(\left(\quad _t \quad \right) \middle| \quad - \quad - \quad - \quad \in \quad \right) \cdot \Pr\left(\quad \in \quad \middle| \quad - \quad - \quad - \right)}$$

[Law of total probability]

$$= \frac{^{-} \left[\quad \frac{-}{t} \prod_{j \succ^{(q-1)}} \quad ^{j \in \ t} \quad _{jt} (1 - \quad ^{-}) \right]}{\sum \quad \left[\quad ^{-} \left(\quad \frac{-}{t} \prod_{j \succ^{(q-1)}} \quad ^{j \in \ t} \quad _{jt} (1 - \quad ^{-}) \right) \right]}$$

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

$$\sum\sum \quad \log\left(\quad _t \prod_{j \succ_\sigma} ^{j \in \ t} \quad _{jt} (1 - \quad) \right)$$

M-step: rst, we update class membership probabilities for every segment $\in \{1 \ 2 \quad \}$:

$$= \frac{\sum \quad}{\quad}$$

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

$$_{\boldsymbol{\theta}} \sum\sum \quad \log\left(\quad _t \prod_{j \succ_\sigma} ^{j \in \ t} \quad _{jt} (1 - \quad) \right)$$

which is solved using outer-approximation algorithm in Section A3.3.

A3.4.2. 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 customers:

Initialization: we initialize the EM with a random allocation of individuals to one of the classes, resulting in an initial allocation $\mathcal{D} \quad \mathcal{D} \quad \mathcal{D}$, which form a partition of the collection

of all the individuals. Then we set = |$\mathcal{D}$ | (∑ |$\mathcal{D}$ |). Then (i.e., ≻) and for all ∈{1 }, ∈ are obtained by solving the following optimization problem:

$$\boldsymbol{\theta}_h \sum_{\in \mathcal{D}_h} \Big(\log \; _{it} + \sum_{\substack{j \in \; it \\ j \succ \sigma \; j_{it}}} \log(1 - \;) \Big)$$

which is solved by using the outer approximation algorithm in Section A3.3.

E-step: we compute , which is the membership probability of every individual to belong to the segment based on the parameter estimates { − − − } and the purchasing transactions data ($_{it}$)| ^{i} :

$$= \frac{^{-} \prod^{i} \left[\; _{it}^{-} \prod_{\substack{j \in \; it \\ j \succ^{(q-1)}}} (1 - \; ^{-}_{j_{it}}) \right]}{\sum \left[\; ^{-} \prod^{i} \left(\; _{it}^{-} \prod_{\substack{j \in \; it \\ j \succ^{(q-1)}}} (1 - \; ^{-}_{j_{it}}) \right) \right]}$$

M-step: rst, we update class membership probabilities for every segment ∈{1 2 }:

$$= \frac{\sum}{}$$

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

$$_{\boldsymbol{\theta}} \sum\sum \quad \sum^{i} \log \Big(\; _{it} \prod_{\substack{j \in \; it \\ j \succ \sigma \; j_{it}}} (1 - \;) \Big)$$

which is solved using the outer-approximation algorithm in Section A3.3. In order to simplify the optimization of the aforementioned conditional expected value of the log-likelihood function in Section 6, we assume that the ranking is known a priori, i.e., the products are assumed to be ranked according to their sales.

A3.4.3. EM algorithm heuristics. Note that the proposed EM algorithm might become computationally challenging for large-scale problems as we need to run an outer-approximation algorithm for every th iteration. Alternatively, we might further assume that the preference order (i.e., ≻) over items in the product universe is known, e.g., we can rank the products according to their popularity in the sales transaction data or we can estimate the ranking from calibrating single class ICS model (see Section 4.1). In this case, the M step for 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 an outer-approximation algorithm) given by:

$$= \frac{\sum \quad \mathbb{I}[\ _t = \ ]}{\sum \quad \mathbb{I}[\ _t = \ ] + \sum \quad \mathbb{I}[\ \in \quad \succ \ _t]}$$

which can be applied with aggregate level data (see Section A3.4.1), and

$$= \frac{\sum \quad \sum^{\ i} \quad \mathbb{I}[\ _{it} = \ ]}{\sum \quad \sum^{\ i} \quad \mathbb{I}[\ _{it} = \ ] + \sum \quad \sum^{\ i} \quad \mathbb{I}[\ \in \quad \succ \ _{it}]}$$

which can be applied with panel data (see Section A3.4.2).

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

A3.5.1. Estimation with aggregate level data. Similarly to the Section A3.4, we calibrate the CTC model by dividing transactions into segments such that customers in segment sample their consideration sets based on the attention parameters and have their preferences characterized by the ranking . Then the log-likelihood function can be represented in the following way

$$\log \mathscr{L}(\quad) = \sum \log \Big(\sum \quad _t \prod_{j \in \ \succ_h \ _t} (1 - \quad _{j_t}) \Big) \tag{A24}$$

where ≥ 0 is the weight of the class , s.t. $\sum \quad = 1$; denotes the set of o ered items at time ; $_t$ denotes the product purchased at time ; denotes the time horizon. Conceptually, we can obtain all the parameters of the CTC model (i.e., distributions over the preference lists and considerations sets) by maximizing the log-likelihood function above for a su ciently large .

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

Initialization: we randomly allocate sales transaction to one of the classes, resulting in an initial allocation $\mathcal{D} \quad \mathcal{D} \quad \mathcal{D}$, which form a partition of the collection of all the transactions. Consequently, we set $= |\mathcal{D} \ | \ (\sum \quad |\mathcal{D} \ |)$. Then (i.e., $\succ$) and for all $\in \{1 \quad \}$, $\in$ are obtained by solving the following optimization problem:

$$_{h \ \boldsymbol{\theta}_h} \sum_{\in \mathcal{D}_h} \Big(\log \quad _t + \sum_{j \in \ \succ_h \ _t} \log(1 - \quad _{j_t}) \Big)$$

which is solved by using the outer approximation algorithm for the ICS model in Section A3.3.

E-step: we compute , which is the membership probability of every transaction at time to belong to the segment based on the parameter estimates { – – – } and the purchasing transactions data (t)| :

$$= \frac{^{-}\left[\begin{array}{c} ^{-} \\ t \end{array} \prod_{\substack{j \in t \\ j \succ_h^{(q-1)} j_t}} (1 - \quad ^{-}) \right]}{\sum \left[^{-}\left(\begin{array}{c} ^{-} \\ t \end{array} \prod_{\substack{j \in t \\ j \succ_h^{(q-1)} j_t}} (1 - \quad ^{-}) \right) \right]}$$

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

$$= \frac{\sum}{}$$

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

$$_h \boldsymbol{\theta}_h \sum \quad \log\Big(\quad _t \prod_{\substack{j \in t \\ j \succ_h j_t}} (1 - \quad) \Big)$$

which is solved by using the outer approximation algorithm for the ICS model in Section A3.3.

Note that in 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-scale problems we might solve the optimization problem at M -step by ranking the products according to their popularity for each segment . This way we can obtain the preference order for every segment for $-$th iteration. In this case, the M step in the EM algorithm reduces to solving a globally concave maximization problem with a unique and closed form solution given by:

$$= \frac{\sum \quad \mathbb{I}[\ _t = \ ]}{\sum \quad \mathbb{I}[\ _t = \ ] + \sum \quad \mathbb{I}[\ \in \ \succ \ _t]}$$

A3.5.2. Estimation with panel data. We update the EM algorithm above in the following way:

Initialization: we randomly allocate individuals to one of the classes, resulting in an initial allocation $\mathcal{D}\ \ \mathcal{D}\ \ \mathcal{D}$. Consequently, we set $= |\mathcal{D}\ |\ (\sum |\mathcal{D}\ |)$. Then (i.e., $\succ$) and for all $\in \{1\ \ \}$, $\in$ are obtained by solving the following optimization problem:

$$_h \boldsymbol{\theta}_h \sum_{\in \mathcal{D}_h} \Big(\log \quad _{it} + \sum_{\substack{j \in it \\ j \succ_h j_{it}}} \log(1 - \quad) \Big)$$

which is solved by using the outer approximation algorithm for the ICS model in Section A3.3.

E-step: we compute , which is the membership probability of every individual to belong to segment based on the parameter estimates { } and the purchasing transactions data ( )| :

$$= \frac{\quad \prod \left[\quad \prod_{\substack{j \in \\ j \succ_h^{(q-1)} \quad _{jit}}} (1 - \quad) \right]}{\sum \quad \prod \left[\quad \left(\quad \prod_{\substack{j \in \\ j \succ_h^{(q-1)} \quad _{jit}}} (1 - \quad) \right) \right]}$$

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

$$= \frac{\sum}{\quad}$$

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

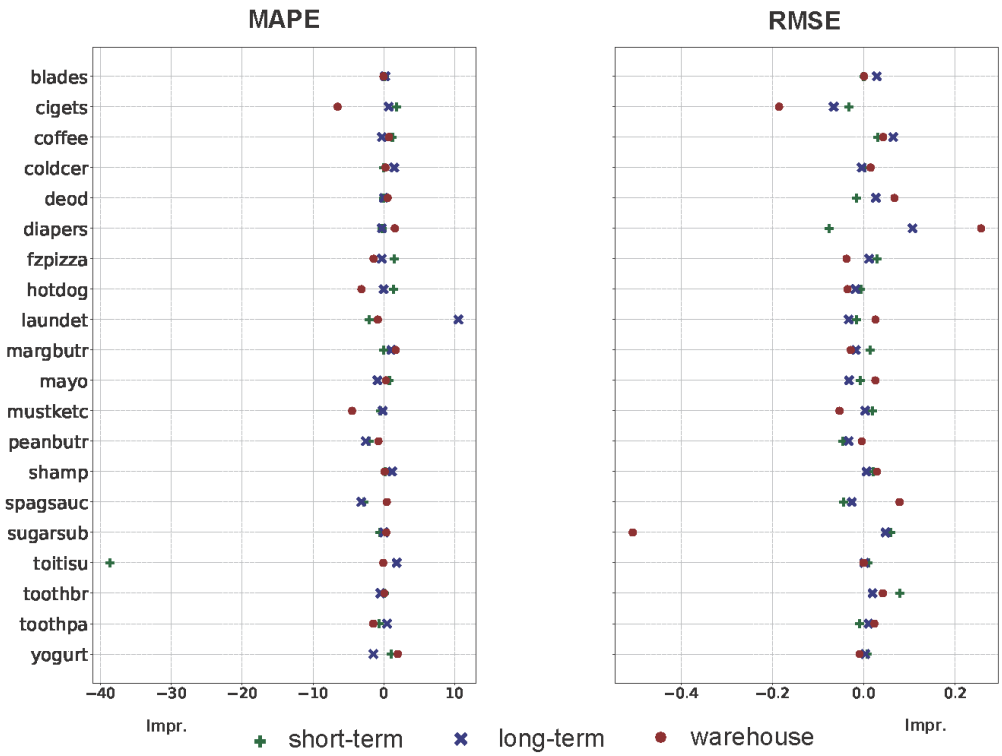
$$_h \boldsymbol{\theta}_h \sum \quad \sum^{i} \log \left(\quad _{it} \prod_{\substack{j \in \\ j \succ_h \quad _{jit}}} (1 - \quad) \right)$$

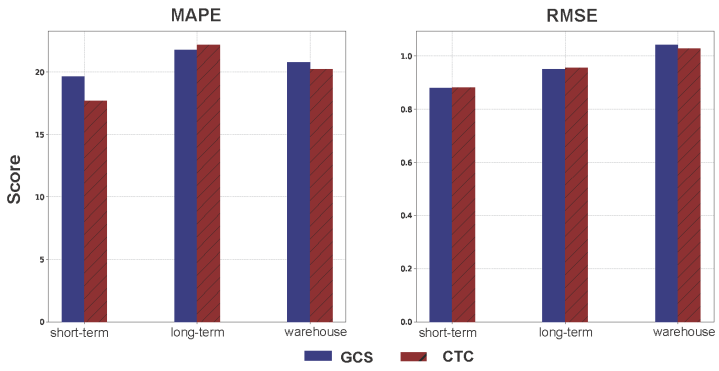
which is solved by using the outer approximation algorithm for the ICS model in Section A3.3.

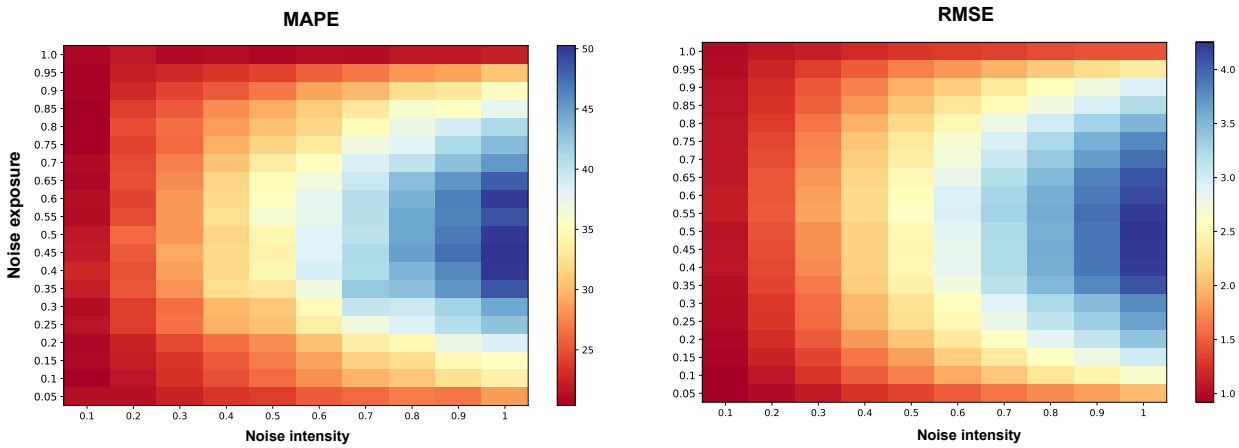
In this section, we compare the prediction performance of the GCS with the CTC based on IRI dataset. It follows from Figures A6 and A7 that the GCS provides higher prediction accuracy than the CTC under the long-term forecast scenario whereas the CTC outperforms the GCS under the warehouse forecast scenario. As a result, we can not claim dominance of the GCS or the CTC.

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

The setup of the experiments in this section is identical to the one in Section 5. In the new set of experiments we simulate sales transaction data according to the rank-based model with fteen







customer types where type customers are making purchases according to the ranking , 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 fteen 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 fteen of them are the rankings from the ground truth and the remaining fteen rankings are sampled randomly. Moreover, we start the EM algorithm with an equal probability of sampling each preference list.

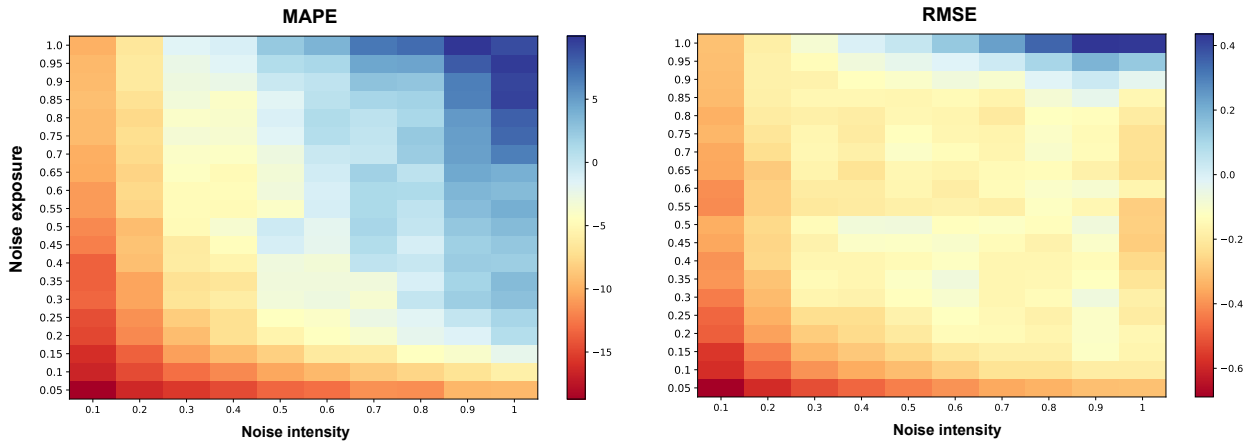
Similarly to Section 5, in Figures A8 and A9 we present the heatmaps of the prediction scores under the rank-based model, and the prediction scores improvements of the ICS model versus the rank-based model, respectively. And in Tables A2 and A3, we report the results for the regression (16), where the dependent variables are the prediction scores and the improvement scores, respectively. The main insights remain the same the results of this extensive simulation study demonstrate that choice models based on the consider-then-choose framework are more robust to noise in o er sets than their classical counterparts, i.e., the ICS model outperforms the ground truth rank-based model under noisier regimes.

With the objective of benchmarking the ICS model with a more competitive variant of the MNL, we resort to the LC-MNL. As it was mentioned above, we estimate the LC-MNL model for $= 1\ 2\ \ 5$, classes and report the best performance measure from these ve variants. Figure A11 illustrates the heatmaps of the MAPE and RMSE prediction score improvements of the ICS model

	Model (1) Score	Model (2) Score	Model (3) Score	Model (4) Score	Model (5) Score
	38.341***				48.526***
	(23.312)				(15.904)
		32.639***			-9.259***
		(20.358)			(-3.425)
			30.882***		42.516***
			(5.470)		(29.661)
				4.827*	13.450***
				(1.657)	(19.378)
	7.072***	15.593***	17.891***	26.428***	-13.927***
	(6.930)	(19.324)	(8.679)	(19.091)	(-14.501)
No. Observations:	200	200	200	200	200
R-squared:	0.733	0.677	0.131	0.014	0.955
Adj. R-squared:	0.732	0.675	0.127	0.009	0.954

t

* $p <$. ** $p <$. *** $p <$.

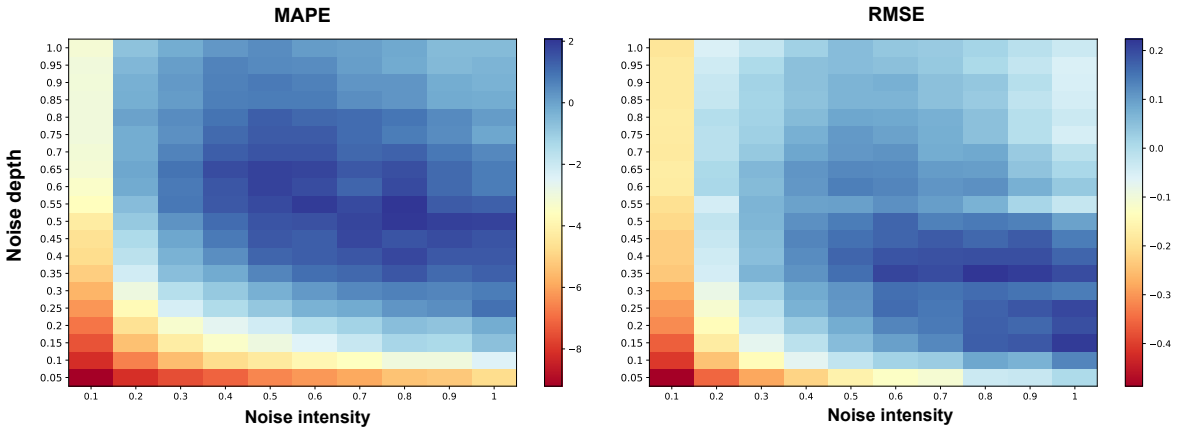


versus the LC-MNL when the ground truth is an MNL model as in Section 5. As expected, Figure A11 con rms that the LC-MNL is a more competitive benchmark. However, our qualitative results remain the same: the ICS model outperforms the LC-MNL model, which subsumes the ground truth MNL model, under su ciently noisy regimes.

	Model (1) Impr.	Model (2) Impr.	Model (3) Impr.	Model (4) Impr.	Model (5) Impr.
	16.211***				23.908***
	(16.814)				(13.640)
		13.646***			-6.997***
		(14.968)			(-4.506)
			1.991		12.733***
			(0.697)		(15.464)
				9.837***	12.419***
				(8.231)	(31.147)
	-12.304***	-8.641***	-4.050***	-6.917***	-22.532***
	(-20.567)	(-18.832)	(-3.884)	(-12.178)	(-40.841)
No. Observations:	200	200	200	200	200
R-squared:	0.588	0.531	0.002	0.255	0.933
Adj. R-squared:	0.586	0.528	-0.003	0.251	0.931

t

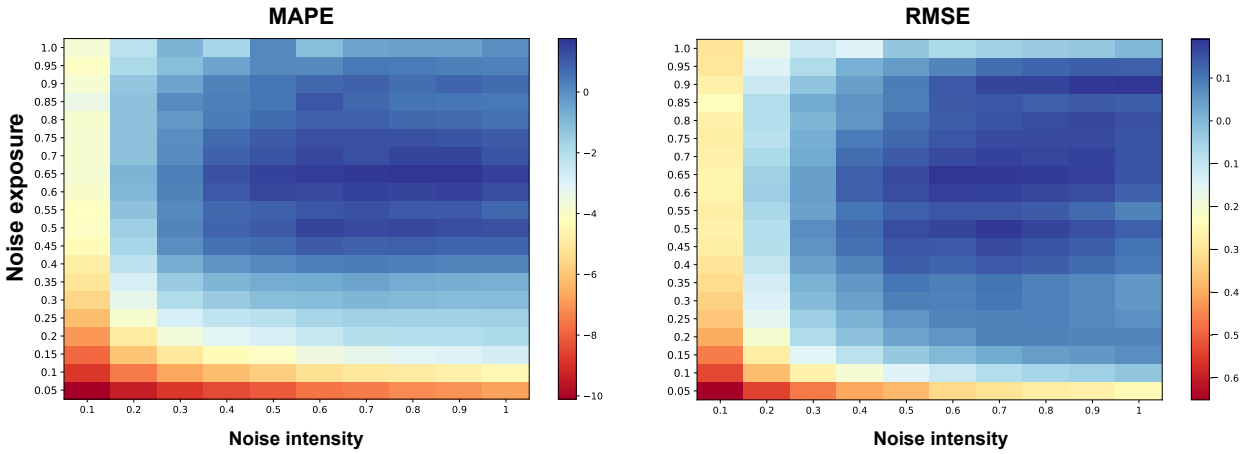
* $p <$. ** $p <$. *** $p <$.



η

γ

In this section, we provide more details on how the one-directional cannibalization property of the ICS model helps it to outperform the ground-truth MNL model in the simulation study in Section 5.



γ η

This is just one potential mechanism that can explain the superior performance of the ICS model in noisy regimes. To streamline our analysis, we focus on the speci c case of the simulation study in Section 5 where $= 0\,5$ and $= 1$, which corresponds to the case of maximum asymmetry between the training and testing datasets (i.e., the probability that a product is only o ered in either the training or the testing dataset is 0.5, and every product in the exposure set will be added to either the training or the testing datasets). As it was mentioned in Section 5, the one-directional cannibalization property of the ICS model provides robustness to the noise in the o er sets because it alleviates its impact on the demand prediction for the higher ranked items (i.e., the presence or absence of lower-ranked products in the o er set does not a ect the demand prediction for higher ranked products). We con rm this intuition by estimating the following regression speci cation:

$$= + \cdot + \tag{A25}$$

where corresponds to item $\in$ and corresponds to a speci c instance, i.e., to each of the 1,000 generated instances. is the improvement of item demand forecasting error in instance obtained by the ICS model over MNL, i.e., $= -$, where we rely on two ways to measure the item s demand forecasting accuracy: (1) absolute percentage error $( ) = \frac{\cdot\left| _j - _j\right|}{_j}$, where denotes the observed sales for in the test dataset and denotes our prediction and (2) demand prediction error $( ) = \frac{\cdot\left(_j - _j\right)}{\sum_{a_j \in N} _j}$. Note that the former and the latter prediction errors are factored into the computations of the MAPE and RMSE scores, respectively, and that they are expressed in percentage points so that the score improvement is also expressed in percentage points. is the position of item in the ranking inferred

from the ICS model calibration on instance . is a categorical variable to control for the instance-level xed e ects. Table A4 presents the results from the regression speci cation (A25) where the improvement by the ICS model over MNL is measured using score in columns (1) and (3), while the improvement by the ICS model over MNL using the score is presented in columns (2) and (4). In columns (1) and (2) we do not control for the instance-level xed e ects, while in columns (3) and (4) we do. It follows from Table A4 that the coe cient of the variable is negative and statistically signi cant, with a magnitude that implies a relevant decrease in for the low-ranked products, which con rms the intuition that the ICS model is especially competitive when predicting the market shares of the top-ranked items because those items are immune to the presence or absence of lower-ranked products in the o er sets. Note that this nding is consistent across all four columns in Table A4 (with and without instance-level xed e ects, under both prediction scores). Note that the in all the columns in Table A4 is relatively small which is not unexpected given that we only have a single covariate in the regression models and there might be other factors (in addition to the variable) which could explain the variation in the outcome variable .

	Model (1) Impr.	Model (2) Impr.	Model (3) Impr.	Model (4) Impr.
	-13.385*** (-3.274) 331.279*** (12.408)	-1.428*** (-5.987) 27.625*** (14.699)	-13.385*** (-3.177)	-1.428*** (-5.815)
Instance FE:	No	No	Yes	Yes
No. Observations:	15,000	15,000	15,000	15,000
R-squared:	0.001	0.003	0.001	0.003

t
* $p<$. ** $p<$. *** $p<$.

In this section, we show the robustness of our prediction results in Section 6 to di erent speci - cations of the prediction metrics. To this end, we focus on the variations of RMSE and MAPE metrics when aggregating predictions either over a one week intervals (i.e., RMSE and MAPE)

or over every o er set (i.e., RMSE and MAPE). More formally, RMSE and MAPE metrics when aggregating predictions over one-week intervals can be computed in the following way

$$\text{MAPE} = \frac{100}{\sum | |} \sum \sum_{j \in t} \frac{| - |}{10+} \tag{A26}$$

$$\text{RMSE} = 100 \sqrt{\frac{1}{\sum | |} \sum \frac{\sum_{j \in t} (-)}{(\sum_{j \in t})}} \tag{A27}$$

where is the set of items that were o ered to customers in week , is the total number of weeks in the test dataset, is the observed number of times product was purchased in week , and is the predicted number of times product to be purchased in week . In the same spirit, we can compute RMSE and MAPE metrics when aggregating predictions over each o er set as follows

$$\text{MAPE} = \frac{100}{\sum \left| \right|} \sum \sum_{j \in m} \frac{| - |}{10+} \tag{A28}$$

$$\text{RMSE} = 100 \sqrt{\frac{1}{\sum \left| \right|} \sum \frac{\sum_{j \in m} (-)}{(\sum_{j \in m})}} \tag{A29}$$

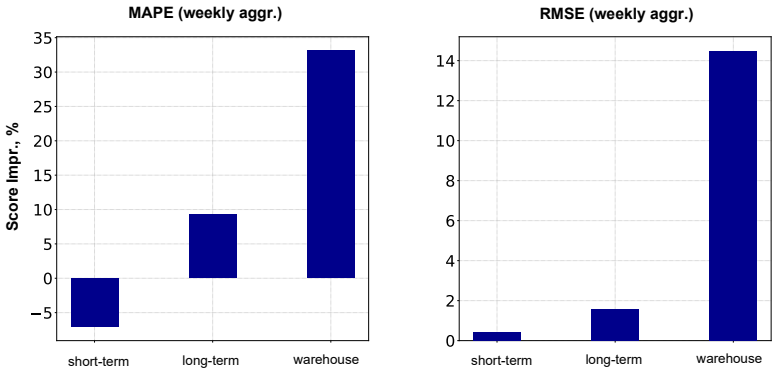
where is the th o er set, is the total number of di erent o er sets in the test dataset, is the observed number of times product was purchased under the o er set , and is the predicted number of times product to be purchased under the o er set . Alternatively, we could aggregate predictions based on the true o er sets in the following way

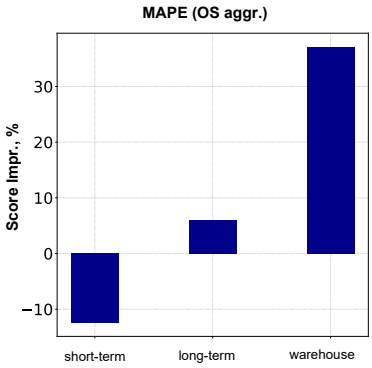
$$\text{MAPE} = \frac{100}{\sum ' | |} \sum' \sum_{j \in} \frac{| - |}{10+} \tag{A30}$$

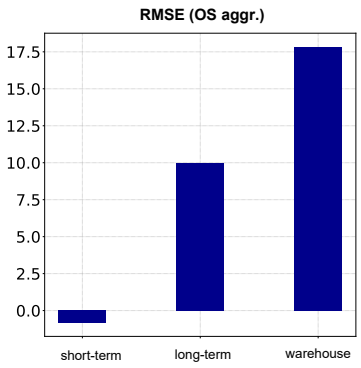
$$\text{RMSE} = 100 \sqrt{\frac{1}{\sum ' | |} \sum' \frac{\sum_{j \in} (-)}{(\sum_{j \in})}} \tag{A31}$$

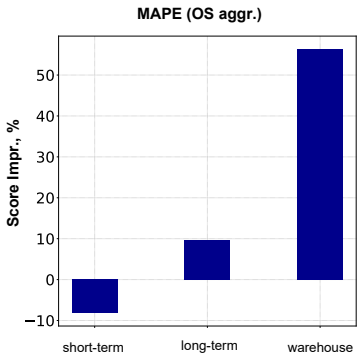
where is the th true o er set, ′ is the total number of di erent true o er sets in the test dataset.

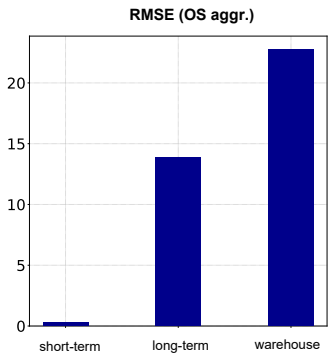
In Figure A12, we exhibit improvements of the GCS over the LC-MNL under MAPE (left panel) and RMSE (right panel), averaging across 20 product categories, for the three di erent scenarios. From these panels, we observe that the GCS signi cantly improves over the LC-MNL once we shift from short to long-term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

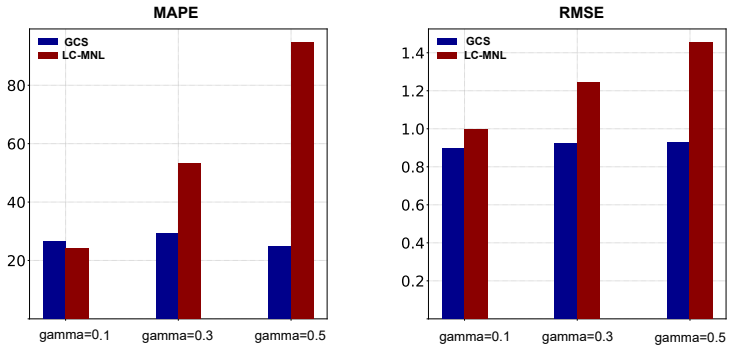










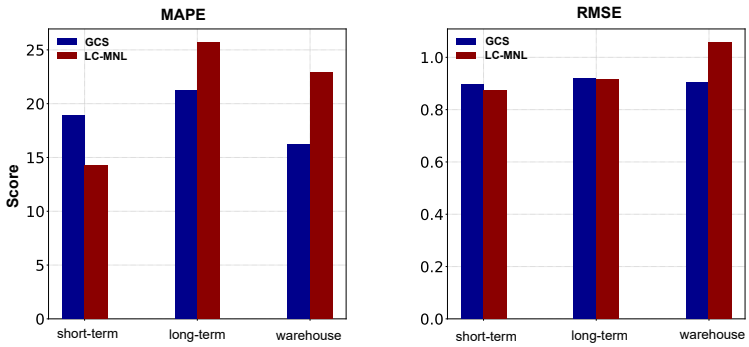


η γ

In this section, we show the robustness of our prediction results in Section 6 to an alternative noise generation process. To this end, we add the noise only to the test dataset and we do it in a similar way to the simulation study presented in Section 5. Figure A15 illustrates the predictive performance of the GCS and the LC-MNL models when is equal to 0.5 and takes three di erent values (0.1, 0.3, and 0.5). It follows from Figure A15 that our results are qualitatively the same under this alternative noise regime the relative predictive performance of the GCS over the LC-MNL improves with the level of noise (measured by) in the o er sets.

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

In this section, we calibrate the Logistic-based ICS (L-ICS) and MNL models accounting for car features and discuss the modeling assumptions based on the car sharing dataset (see data descriptive statistics in Table A5). We also provide explanatory analysis of choice models in order to gain insights about the consideration set formation of renters using the car feature information. In addition, we address the problem of a potential price endogeneity in our empirical explanatory



analysis. We argue that in our setting we are unlikely to have any price endogeneity problems while calibrating the models. We conclude this section by presenting Figure A17 which illustrates an instance of the decision tree obtained after tting the DT-ICS model.

We start this section by calibrating the single-class L-ICS model with features to examine the extent to which various variables impact the consideration set structure. Assuming that the cars are ranked according to their popularity among renters (see Section 7.2), the problem of tting the L-ICS model is the one of estimating the coe cients . 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 are logistically distributed, we estimate the vector using logistic regression analysis.

The results for the L-ICS model appear in the rst column of Table A6. In the middle column, the table lists the average marginal e ects (AME) of the L-ICS model when all the covariates are at their mean. Then we also calibrate the usual linear-in-parameters MNL model where the utility from reserving the car alternative is represented with linear in parameters function , i.e., = + . On the right, Table A6 presents the estimates of the MNL model parameters. However, the interpretation of the vector for the L-ICS and MNL models is di erent. The parameters of

	Mean	Std.	Min	Max
Brands				
Acura	2.52%	15.68%	0%	100%
Audi	4.54%	20.82%	0%	100%
BMW	11.73%	32.18%	0%	100%
Buick	0.21%	4.61%	0%	100%
Chevrolet	0.79%	8.84%	0%	100%
Chrysler	0.41%	6.37%	0%	100%
Dodge	0.82%	9%	0%	100%
Fiat	0.9%	9.42%	0%	100%
Ford	2.63%	16.01%	0%	100%
Honda	16.77%	37.36%	0%	100%
Hyundai	3.42%	18.16%	0%	100%
Infiniti	0.21%	4.61%	0%	100%
Jeep	0.41%	6.39%	0%	100%
Kia	0.49%	6.95%	0%	100%
Land Rover	0.17%	4.14%	0%	100%
Lexus	1.24%	11.06%	0%	100%
Mazda	3.44%	18.22%	0%	100%
Mercedes Benz	3.73%	18.95%	0%	100%
Mercury	0.07%	2.59%	0%	100%
Mini	7.66%	26.59%	0%	100%
Mitsubishi	0.68%	8.21%	0%	100%
Nissan	4.05%	19.71%	0%	100%
Pontiac	0.21%	4.57%	0%	100%
Porsche	1.5%	12.14%	0%	100%
Saab	0.03%	1.73%	0%	100%
Saturn	0.28%	5.25%	0%	100%
Scion	0.62%	7.85%	0%	100%
Subaru	3.71%	18.89%	0%	100%
Suzuki	0.53%	7.29%	0%	100%
Smart	5.44%	22.69%	0%	100%
Tesla	1.42%	11.83%	0%	100%
Toyota	10.33%	30.43%	0%	100%
Volkswagen	7.54%	26.41%	0%	100%
Volvo	1.53%	12.28%	0%	100%
Car types				
Economy	14.83%	35.54%	0%	100%
Standard	48.83%	49.99%	0%	100%
Fullsize	19.56%	39.67%	0%	100%
SUV	9.41%	29.2%	0%	100%
Trucks	3.31%	17.88%	0%	100%
Luxury	4.06%	19.74%	0%	100%
Car location type and accessibility				
Car access [open]	81.89%	38.51%	0%	100%
Car access hours [all hours]	93.52%	24.61%	0%	100%
Car location type [garage]	28.50%	45.14%	0%	100%
Car location type [street]	23.86%	42.62%	0%	100%
Car location type [surface lot]	43.85%	49.62%	0%	100%
Car location type [valet]	0.24%	4.84%	0%	100%
Car features				
Price (per hour)	8.63	4.61	2.0	300.0
Car age	5.32	3.18	-0.3	18.3
Transmission [automatic]	95.21%	21.35%	0%	100%
Premium wheels	29.38%	45.55%	0%	100%
Power seats	46.88%	49.90%	0%	100%
Bluetooth/wireless	33.74%	47.28%	0%	100%
Leather interior	53.56%	49.87%	0%	100%
Sunroof/moonroof	53.48%	49.88%	0%	100%
Premium sound	46.25%	49.86%	0%	100%
Power windows	92.90%	25.68%	0%	100%
GPS navigation system	23.05%	42.11%	0%	100%
Roof rack	6.98%	25.48%	0%	100%
Tinted windows	13.24%	33.89%	0%	100%

the L-ICS model listed in the left column of Table A6 show the estimated impact of exogenously imposed changes in car features on consideration set formation. Rather, the parameters of the MNL model in the right column of Table A6 show the in uence of car features on the customer s choices, i.e., revealed preferences. Notably, quite a few coe cients (17 out of 53) estimated based on L-ICS and MNL models are not aligned, i.e., the covariates that increase (or decrease) the likelihood of considering the car under the L-ICS model might not necessarily increase (or decrease) the likelihood of booking the car under the MNL model, e.g., the utility of the renter from considering the car brand Jeep is higher by 0.98 (= 4.9, 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 (= -6.2, 0.01) than the utility from reserving baseline brands. Also, some of the covariates (7 out of 53) that are statistically signi cant in explaining the choice of renters under the MNL model might be statistically insigni cant under the L-ICS model, e.g., the utility from choosing a car parked in the street is lower (= -8.33, 0.01) than the utility from choosing the car located in the valet parking area while the discrepancy between these two parking location types is insigni cant (= 1.36, 0.10) under the L-ICS model. The price and car age coe cients are statistically signi cant and negative for both the L-ICS and MNL models. However, the impact of an 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 an 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 ndings, car age plays a relatively more important role during the formation of the consideration set in the L-ICS model compared to its role in the choice process under the MNL model.

Next, we consider three types of car attributes (i.e., car brand, car location type and accessibility, and car type and features), with the objective of empirically verifying their impact on consideration set formation under the L-ICS model and on the choice probabilities under the MNL model. Models 1,2, and 3 (both under the L-ICS and MNL) incorporate all the covariates except car types and features, car location type and accessibility, and brands, respectively, e.g., Model 1 excludes car types and features while including all the other covariates. According to Table A7, the car type and features attributes are less statistically signi cant than car brand attributes under the L-ICS model, whereas the opposite e ect takes place under the MNL model. These ndings are robust to the various measures of statistical signi cance and goodness-of- t presented in Table A7 such as LL, AIC, BIC, Likelihood Ratio (LR) statistics, and Wald statistics. Overall, it is implied that car location type and accessibility play the least important role both for consideration set formation and for the nal 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.

	L-ICS		AME (L-ICS)		MNL	
	Coeff.	Std.err.	Coeff.	Std. err.	Coeff.	Std. err.
Brands						
Acura	0.29**	0.10	0.067**	0.024	-0.33***	0.094
Audi	-0.11	0.097	-0.025	0.023	0.22*	0.088
BMW	0.0061	0.089	0.0014	0.021	0.14	0.083
Chrysler	-0.19	0.15	-0.044	0.035	-1.24***	0.13
Dodge	0.21	0.12	0.048	0.029	-0.0095	0.11
Fiat	1.01***	0.15	0.24***	0.035	-0.030	0.11
Ford	-0.50***	0.095	-0.12***	0.022	0.055	0.087
Honda	0.13	0.086	0.029	0.020	0.0011	0.079
Hyundai	0.010	0.094	0.0024	0.022	-0.064	0.085
Infiniti	0.21	0.25	0.049	0.059	0.30	0.24
Jeep	0.98***	0.20	0.23***	0.047	-0.93***	0.15
Kia	0.36*	0.15	0.085*	0.035	-0.47***	0.13
Land Rover	1.01***	0.25	0.24***	0.058	0.74***	0.20
Lexus	0.043	0.11	0.0100	0.027	0.20	0.11
Mazda	0.27**	0.098	0.063**	0.023	0.070	0.089
Mercedes Benz	-0.32**	0.100	-0.074**	0.023	0.20*	0.090
Mercury	2.53***	0.75	0.59***	0.18	-0.064	0.29
Mini	0.37***	0.094	0.086***	0.022	0.23**	0.086
Mitsubishi	-0.37**	0.13	-0.087**	0.030	-0.51***	0.12
Nissan	0.39***	0.094	0.090***	0.022	0.11	0.085
Pontiac	0.87***	0.23	0.20***	0.054	0.54**	0.18
Porsche	-0.24*	0.11	-0.057*	0.027	0.45***	0.11
Scion	0.74***	0.17	0.17***	0.039	-0.67***	0.14
Subaru	0.32***	0.096	0.075***	0.023	0.43***	0.084
Suzuki	-0.071	0.15	-0.017	0.036	0.47**	0.16
Smart	-0.15	0.096	-0.036	0.022	0.090	0.084
Tesla	1.29***	0.16	0.30***	0.038	3.24***	0.15
Toyota	0.27**	0.089	0.063**	0.021	0.13	0.080
Volkswagen	0.26**	0.091	0.061**	0.021	0.46***	0.084
Volvo	1.26***	0.13	0.29***	0.029	0.26*	0.11
Baseline brands	Baseline		Baseline		Baseline	
Car types						
Economy	0.36***	0.068	0.084***	0.016	-0.28***	0.055
Standard	0.29***	0.060	0.068***	0.014	-0.13**	0.050
Fullsize	0.34***	0.066	0.079***	0.015	-0.36***	0.054
SUV	0.055	0.070	0.013	0.016	-0.43***	0.057
Luxury	0.37***	0.083	0.087***	0.019	-0.19**	0.070
Trucks	Baseline		Baseline		Baseline	
Car location type and accessibility						
Car access [open]	-0.36***	0.029	-0.084***	0.0067	0.029	0.026
Car access hours [all hours]	-0.089*	0.045	-0.021*	0.010	-0.097*	0.038
Car location type [garage]	-0.24***	0.061	-0.057***	0.014	0.13*	0.052
Car location type [street]	0.080	0.059	0.019	0.014	-0.40***	0.048
Car location type [surface lot]	-0.19***	0.057	-0.044***	0.013	0.27***	0.048
Car location type [valet]	Baseline		Baseline		Baseline	
Car features						
Price (per hour)	-0.022***	0.0033	-0.0051***	0.00077	-0.12***	0.0042
Car age	-0.042***	0.0039	-0.0099***	0.00091	-0.032***	0.0033
Transmission [automatic]	0.34***	0.046	0.080***	0.011	0.45***	0.037
Premium wheels	-0.0025	0.025	-0.00059	0.0058	-0.18***	0.021
Power seats	-0.21***	0.024	-0.048***	0.0055	0.043*	0.021
Bluetooth/wireless	-0.13***	0.025	-0.031***	0.0059	-0.29***	0.021
Leather interior	0.087**	0.030	0.020**	0.0070	0.12***	0.025
Sunroof/moonroof	0.0011	0.027	0.00026	0.0064	0.14***	0.024
Premium sound	0.25***	0.027	0.059***	0.0063	-0.14***	0.022
Power windows	-0.0058	0.042	-0.0013	0.0099	0.40***	0.036
GPS navigation system	-0.085**	0.029	-0.020**	0.0067	0.18***	0.023
Roof rack	0.16***	0.046	0.036***	0.011	-0.26***	0.037
Tinted windows	-0.087**	0.030	-0.020**	0.0070	-0.30***	0.026
Constant	-0.13	0.15				
No. of obs.	26791		26791		26791	
AIC	76980.3				69788.7	
BIC	77464.8				70309.2	
Log likelihood	-38436.1				-34841.4	
Pseudo R^2 square	0.024					

*$p <$. , **$p <$. , ***$p <$.

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

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

A5.2.1. Semiparametric approach. Using the semiparametric approach in order to calibrate the two-stage CTC model, we assume that renters form their consideration set taking into account car features. Then we assume that during the second stage renters choose the most preferred car among the considered ones according to the preference order over the universe of car alternatives, which remains the same over time, i.e., the ranking is xed over time. Modeling the second stage choice process this way, we do not parameterize the ranking which implies that the cars are assumed to have the same attributes over time. In this subsection we justify this assumption according to our dataset.

We start by analyzing the variation of the hourly price parameter over car alternatives. In Table A8, we report that the average coe cient of variation (CV) of the hourly price across all the car alternatives is around 5% while owners of cars listed on average around two di erent 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, which 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 e ciently match demand and supply as opposed to many ride-sharing platforms (e.g., Uber, Lyft) which optimize the price of the ride to match riders with drivers on-demand. Then in the same Table A8 we can also observe that more than 98% of car owners did not alter their car access (i.e., open or closed), access hours (i.e., 24 hours or restricted), and location type (i.e., garage, street, surface lot, or valet).

Additional descriptive statistics:	
Number of rentals	26791
Number of car owners	514
Number of available alternatives (within 0.3 mile)	5.7
Rental duration (days)	0.62
Rental request in advance (days)	1.24
Price CV (averaging over car owners)	0.053
Average number of price modes	2.33
The most frequent price (percentage)	0.78
The second most frequent price (percentage)	0.16
Average number of car access modes	1.07
The most frequent car access (percentage)	0.99
Average number of car access hours modes	1.03
The most frequent car access hours (percentage)	0.99
Average number of car location type modes	1.10
The most frequent car location type (percentage)	0.98

A5.2.2. Price endogeneity problem. Next, we want to address the concerns of potential price endogeneity in our empirical analysis. First of all, estimating the demand with personalized data signi cantly alleviates the price endogeneity problem since each renter has only a trivial in uence 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 very severe 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 on 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 insigni cant as it was discussed above, i.e., the price does not react to any unobservable shocks (see Table A8).