Choice-based Demand Models for Emerging Applications in Retail and Online Platform Operations

by

Dmitry Mitrofanov

Submitted to the Stern School of Business
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Management

at the

NEW YORK UNIVERSITY

Thesis Committee:
Maxime Cohen, Anindya Ghose, Srikanth Jagabathula (chair), Gustavo Vulcano & Jiawei Zhang

Thesis Supervisor: Srikanth Jagabathula

May 2020

Choice-based Demand Models for Emerging Applications in Retail and Online Platform Operations

by

Dmitry Mitrofanov

Abstract

The digital technological revolution has transformed nearly every aspect of our life, disrupting industry after industry. This not only led to the explosive growth of online sharing platforms, but also created critical challenges as well as opportunities for traditional offline retailers to customize their operations. For instance, while personalized promotions have been around for several years now on online platforms, recent technological developments (e.g., electronic shopping carts), as well as increased availability of customer-level data, extended similar practices to brick-and-mortar settings. To ride this wave of change successfully, companies need to rethink traditional operational problems such as demand prediction or promotion planning. Demand forecasting models have always played a crucial role both in practice and in academia, where choice models are now ubiquitous in different areas of operations, marketing, and economics. However, these existing models face the challenge of staying relevant due to the dramatic changes mentioned above.

Demand calibration becomes increasingly important in the sharing economy era where online platforms need to make pricing decisions and match supply and demand. Unsurprisingly, demand forecasting in this context is very challenging because, at any given time, users on the online application face hundreds or even thousands of available alternatives, while suffering from limited attention. The failure to account for unobservable consideration sets in these settings will result in biased inferences and incorrect demand forecasts. Moreover, most existing literature regarding choice-based demand estimation and prediction makes the questionable assumption that we are always able to access accurate information about offer sets (which is a necessary input for choice models), while forecasting demand. For example, this assumption is clearly not valid in the case of online peer-to-peer car-sharing platforms where the availability of cars over time depends upon the decisions of many independent car owners and renters. Neither is it valid in the case of a retail store where product availability changes over time due to either stockouts or deliberate scarcity.

Overall, this thesis focuses on real-world implementations of choice models across a range of emerging applications both in retail as well as on online sharing economy platforms. The key contribution of this dissertation is in novel methodologies to effectively handle these aforementioned challenges, which have been created by the growing popularity of two-sided online platforms, recent advances in technology, and increased availability of customer-level data.

The first chapter proposes a back-to-back procedure for running personalized promotions in retail, from the construction of a nonparametric choice-based demand model where customer preferences are represented by directed acyclic graphs (DAGs), to the design of such promotions. We describe the process of obtaining the DAGs and explain how to mount a parametric, multinomial logit model (MNL) over them. We provide new bounds for the likelihood of a DAG and demonstrate how to conduct the MNL estimation. We test our model to predict purchases at the individual level using real-world retail data as compared to state-of-the-art benchmarks. Finally, we illustrate how to use the model to run personalized promotions. Our framework leads to significant revenue gains, which makes it an attractive candidate to be used in practice.

The second chapter investigates the problem of identifying consideration sets from sales transaction data in a data-driven manner. We assume that customers are boundedly rational and make their purchases in a two-stage process. First, they sample their consideration set. Next, they purchase the most preferred considered item. Theoretically, we address the problem of identifying consider-then-choose models from data. We use tools from Boolean function analysis to derive a closed-form expression for computing distribution over consideration sets from observed choice probabilities. Because calibrating this class of choice models is a hard problem, we propose a framework that effectively infers consideration sets. Our methodology of modeling the consideration set formation is founded on machine learning techniques (e.g., decision trees or random forests).

The last chapter is based on application of the proposed consider-then-choose framework in retail and in peer-to-peer car-sharing industries. We observe that accounting for consideration sets can boost the predictive performance in comparison with classical benchmarks. Our findings suggest that our models tend to be rather robust to the degree of ambiguity in the offer set definition. Their relative importance in prediction tasks increases with this noise. Moreover, we demonstrate that the proposed models can provide important managerial insights about the consideration set formation.

Thesis Supervisor: Srikanth Jagabathula
Title: Associate Professor of Tech, Ops, & Stats
Leonard N. Stern School of Business, NYU

Acknowledgments

I would like to express the deepest gratitude to my committee chair and advisor, Srikanth Jagabathula, for his motivation, immense knowledge, patience, and mentorship throughout my PhD. Srikanth was always supportive and assisted me not only by devoting an enormous amount of time to discussing our research, but also by supporting me academically and emotionally through the rough road to finish this thesis. I hope to continue to collaborating with and learning from Srikanth moving forward. I also greatly appreciate the continual encouragement and support of my co-advisor, Gustavo Vulcano. This thesis is a product of my close collaboration with Srikanth and Gustavo. Without their guidance and persistent help, this thesis would not have been possible. A special thanks goes to my co-author, Maxime Cohen, for his enthusiasm, insightful discussions, and for providing me with the opportunity to pursue various research projects. I am also very grateful to Arun Sundararajan for sharing data that significantly improved my dissertation.

I would like to thank the members of my thesis committee, Anindya Ghose and Jiawei Zhang, for their encouragement, feedback, and insightful comments on my work. During the course of my PhD program at NYU, I learned from a number of outstanding professors. I would especially like to thank Michael Pinedo for giving me important career and life advice and Josh Reed for being a wonderful PhD coordinator and for providing support. I would also like to thank Mor Armony, Ilan Lobel, and Wenqiang Xiao for all that they taught me. My PhD journey would not have been the same without interacting with many wonderful friends and colleges at NYU. In particular, I enjoyed working and spending time with Kevin Jiao, Sen Tian, Shixin Wang, Xinyi Zhao, Haotian Song, Ziran Liu, Yunchao Xu, Jiaoshuo Jiang, Sen Yang, Xuan Wang, Sandeep Chitla, Richard Bryant, and Ying Liu. A special thanks to Yuqian Xu, Tarek Abdallah, and Ashwin Venkataraman for their company and constant support. My appreciation also goes out to everyone who makes sure that the Stern doctoral program runs smoothly and that the students have everything they need. I want to especially thank Anya Takos, Donna Lashley, and Elisabeth Greenberg for all of their assistance over the years.

Last but not least, I want to thank my family for their understanding and love all throughout my life. I would not have made it this far without them. Specifically, I am extremely grateful

to my mother, Zoya, for her dedication, unconditional love, and many years of support during my graduate studies. My father, Valery, receives my deepest gratitude for encouraging and motivating me to work hard and for giving me thoughtful advise. *I am greatly indebted to my family for everything I am today and for all I have achieved. This thesis is dedicated to them.*

Preface

This dissertation describes joint research conducted with my advisors that is currently under review in journals. Parts I and II are based on the papers by [49] and [50], respectively. As a result, some of the content in this thesis was recorded verbatim from the above papers. The paper containing preliminary results of Part II was my job market paper, which received (1) 2nd prize in the 2019 MSOM student paper competition, and was selected as a finalist for the (2) 2019 INFORMS best service science student paper award.

During my PhD at Stern, I have also been conducting research related to ride-hailing operations with an empirical focus. In the first project within this field, we studied the impact of IPO on ride-hailing platforms' operational decisions. The manuscript is completed and is now under review in a journal [18]. In the second project, we analyzed loyalty and switching behaviors of price and delay sensitive riders using a comprehensive panel dataset, in which we observed choices for both Uber and Lyft. Even though this work is still ongoing [17], it was already presented in the 2019 INFORMS Annual Meeting. Both research projects are not included in this dissertation due to an NDA agreement that prevents us from making the work publicly available at this time. However, we will provide a short description of these projects below.

Lyft and Uber IPO: Before and After

The research question in [18] is motivated by the impact of business events on ride-hailing platforms' operational decisions. In 2019, there were two unicorn IPOs from ride-hailing platforms – Lyft filed its IPO on March 1 at a $24.3 billion valuation and Uber filed its IPO on April 11 at a $82.4 billion valuation. In [18], we investigate whether or not these platforms adjusted their operational decisions in anticipation of their IPOs. To answer this question, we use a comprehensive panel dataset provided by our industry partner with Uber and Lyft rides completed between January 2018 and July 2019. We treat each IPO filing day as a natural experiment and estimate several econometric models to quantify the IPO impact on platform decisions (e.g., promotion strategy), performance metrics (e.g., number of rides, market share, and number of users), and consumers (e.g., surplus and fairness). We find strong evidence that both platforms adjusted their operational decisions by issuing more promotions before their IPO filings. We also argue that the impact was different for Lyft versus Uber. Moreover, the richness of our dataset allows us to account for various sources of heterogeneity including market penetration, customers' past riding frequency, riders' deal-seeking behavior, and tip amount.

In [18], we find that Lyft and Uber offered more promotions to riders in cities with a lower market share, suggesting that both platforms tried to increase their presence in markets where they experienced low levels of penetration. We show that riders who received high promotions in the past were more strongly affected by promotional strategies, implying that the platforms prioritized short-term performance over the long-term impact of riders' deal-seeking behavior. We observe that the promotion amount was amplified for longer trips while the relative promotion was higher for shorter trips. This can be partially explained by the promotion structure in ride-hailing industry. Our analysis indicates that the IPO filings had a positive effect on rider surplus and fairness, in the sense that additional strategic promotions were fairly distributed among riders as opposed to being highly targeted.

Customer Loyalty in Ride-Hailing: Empirical Evidence

In [17], we analyze the loyalty and switching behaviors of price and delay sensitive customers in ride-hailing. Our dataset offers a unique opportunity to study this question as we observe choices for both Uber and Lyft. We first propose a way of overcoming the issue of missing data regarding waiting times and competitor prices. We assume that the competitor's price depends upon ride characteristics (e.g., trip length and duration) as well as a surge multiplier, which is controlled by the platform as a response to the demand-supply imbalance. We propose an approach to compute waiting times by calibrating an M/M/1 queuing model with a state-dependent service rate. We then estimate several reduced-form and choice models to examine customer loyalty and switching patterns.

From the reduced form analysis, we find that, even after controlling for prices and waiting times, customers are loyal to their most frequently used platform. In [17], by using choice models, we capture the dynamic interaction between customers and ride-hailing platforms. Specifically, we find that price reduction promotions have a positive short-term effect on riders' preferences but a rather negative long-term effect. This is because promotions induce deal-seeking behavior in users. Interestingly, our analysis shows that users' tipping decisions reveal their platform preferences. Moreover, the richness and granularity of our dataset allows us to take heterogeneity of riders into account to a great extent. Using revealed preferences of customers, we calibrate multilevel logit models and compute price and waiting time elasticities of each rider exploiting Bayesian approach. Herein, we analyze the differential responses of riders in regards to trip prices and delay; for instance, we investigate what riders tend to be more sensitive to trip prices and waiting times. Finally, we explicitly model riders' switching patterns by adopting an alternative interpretation of the multinomial logit. One emerging insight is that riders' deviations from their favorite platform are more likely to occur because of price reduction promotions issued by a competitor than because of a competitor offering lower waiting times.

Contents

Abstract 3

Acknowledgments 5

Preface 7

List of Figures 17

List of Tables 19

Introduction **21**

I Personalized Operations **31**

1 Customized Retail Promotions and Demand Estimation **33**

1.1 Introduction 33

1.1.1 Summary of results 35

1.1.2 Related literature 37

1.2 Choice model description 39

1.2.1 Modeling framework 40

1.2.2 Data model 42

1.2.3 DAG construction 43

1.2.4 Discussion of the model assumptions and the DAG construction procedure 47

1.2.5 Maximum likelihood estimation of the DAG-based choice model 49

1.3 Theoretical analysis of the DAG-based MNL Model 50

1.3.1 Tractable analytical bounds for the likelihood of a DAG 51

1.3.2 Tractable analytical bounds for the purchase probability prediction 54

1.4 Empirical study . . . 56

1.4.1 Data analysis . . . 56

1.4.2 Models compared . . . 58

1.4.3 Prediction performance measures . . . 62

1.4.4 Brand choice prediction results . . . 63

1.5 Robustness check for the prediction results . . . 66

1.5.1 Robustness with respect to different data aggregation strategies . . . 67

1.5.2 Robustness with respect to adding no-purchase observations . . . 70

1.5.3 Robustness with respect to the split between test and training datasets . . 72

1.5.4 Robustness with respect to the addition of implicit candidate edges in the DAGs . . . 73

1.5.5 Comparison with the DAG-based behavioral models studied by [52] 75

1.6 Optimization of personalized promotions . . . 78

1.6.1 Inferences from the DAG structures . . . 79

1.6.2 Promotion optimization: MILP formulation . . . 80

1.6.3 Customized promotions: performance evaluation . . . 85

1.6.4 Managerial insights: factors affecting improvements from personalization of promotions . . . 89

1.6.5 Additional insights on brand loyalty and the number of promoted items . . 90

1.6.6 Robustness check for the promotion optimization . . . 92

1.7 Conclusions and future work . . . 93

II Demand Models with Consideration Sets 95

2 Inferring Consideration Sets from Sales Transaction Data 97

2.1 Introduction . . . 97

2.1.1 Summary of the results and contribution . . . 100

2.1.2 Related literature . . . 101

2.2 General Consider-then-Choose model . . . 103

2.2.1 Identification of the GCC model . . . 106

2.3 Data model and estimation methodology 109

2.3.1 Single class ICC model 110

2.3.2 GCC model 111

2.3.3 Consider-then-choose models with features 113

2.3.4 Prediction scores and benchmark 115

2.4 Estimation methodologies of consider-then-choose models 116

2.4.1 MINLP formulation: Logistic-based consider-then-choose model 116

2.4.2 Estimation methodology: outer-approximation algorithm 118

2.4.3 GCC estimation methodology: EM algorithm 121

2.5 Conclusion 126

3 Robustness of Demand Prediction Models in Operations 127

3.1 Introduction 127

3.2 Study based on the synthetic data: robustness to the noise in offer sets 129

3.2.1 Synthetic data generation process 130

3.2.2 Results and discussion 131

3.3 Empirical study on the IRI academic dataset 134

3.3.1 Data analysis 135

3.3.2 Results and discussion 136

3.4 Case study on the car-sharing dataset: prediction analysis 141

3.4.1 Industry partner and data analysis 142

3.4.2 Modeling assumptions 143

3.4.3 Feature-based predictive accuracy results 145

3.5 Case study on the car-sharing dataset: explanatory analysis 148

3.5.1 Explanatory analysis 149

3.5.2 Discussion of model estimation assumptions 151

3.6 Conclusion 154

3.6.1 Model extension 155

3.6.2 GCGC model: estimation methodology 156

Conclusion and Future Directions 161

Appendices **165**

4 Proofs and Supplementary Materials for Chapter 1 **165**

4.1 Preliminaries on DAGs 165

4.2 Technical results 166

4.2.1 Propositions and proofs in Section 1.2 166

4.2.2 Propositions and proofs in Section 1.3 169

4.3 Heuristic for preference graph decycling 182

4.4 Benchmark models 184

4.4.1 LC-MNL model 184

4.4.2 RPL model 185

4.5 Evaluation of analytical bounds 185

4.5.1 Bounds on the probability of a DAG 186

4.5.2 Bounds on the probability of purchase 188

4.6 Optimization of personalized promotions 190

5 Proofs and Supplementary Materials for Chapter 2 **193**

5.1 Preliminaries on consider-then-choose models 193

5.2 Proofs of technical results 194

Bibliography **218**

List of Figures

1-1 Choice process example with three products. 42

1-2 Phases of DAG construction. The example with four products and three transactions. 45

1-3 Scatter plot of the average χ^2 scores. 64

1-4 Scatter plot of the average miss rate. 65

1-5 Brand choice χ^2 prediction results. In all the panels we estimate the benchmarks on the set of all individuals. 68

1-6 Brand choice miss rate prediction results. In all the panels we estimate the benchmarks on the set of all individuals. 69

1-7 Brand choice χ^2 prediction results with no-purchase option included. We assume that the number of no-purchases of every customer is equal to the number of times a customer buys her second most purchased product. 72

1-8 Brand choice miss rate prediction results with no-purchase option included. We assume that the number of no-purchases of every customer is equal to the number of times a customer buys her second most purchased product. 73

1-9 Brand choice prediction results with no-purchase option included. Every customer is assumed to choose no-purchase alternative on average 30% of times. 74

1-10 Brand choice prediction results with no-purchase option included. Every customer is assumed to choose no-purchase alternative on average 40% of times. 75

1-11 Brand choice prediction results with no-purchase option included. The training data consists of 21 weeks, and the test data consists of 31 weeks. 76

1-12 Brand choice prediction results with no-purchase option included. The training data consists of 31 weeks, and the test data consists of 21 weeks. 77

1-13 Brand choice prediction results with no-purchase option included. Robustness with respect to the addition of implicit candidate edges. 78

1-14 Scatter plot of the χ^2 scores. Comparison with the DAG-based behavioral models. 79

1-15 Scatter plot of the miss rate scores. Comparison with the DAG-based behavioral models. 80

1-16 Loyalty scores and improvements of PO-MNL Promotion over behavioral DAG-based models. 81

1-17 Revenue performance. 87

1-18 Revenue performance of promoting a single item. 88

1-19 Promotion optimization problem when we optimize the expected revenue for every purchasing transaction. 93

2-1 Results of applying the outer-approximation algorithm. 121

2-2 Results of applying the cutting plane algorithm. 122

3-1 Heatmap of the prediction scores under MNL model. 133

3-2 Heatmap of the prediction scores improvements under ICC model versus MNL. . . 135

3-3 Heatmap of the prediction scores improvements under ICC model versus MNL. Noise depth in the training dataset is 10% lower than in the test dataset. 137

3-4 Scatter plots of the prediction score improvements of GCC model over LC-MNL. . 139

3-5 The average prediction scores under GCC and LC-MNL choice models. 140

3-6 Scatter plots and liner regressions of the MAPE score improvement of ICC (vs. MNL model) over the noise intensity. 141

3-7 Scatter plots of the MAPE scores under the GCC vs. ICC models. 142

3-8 Prediction results under Consider-then-Choose (CC) and LC-MNL models with car features. 146

3-9 MAPE scores of car brands. 147

3-10 Scatter plot and linear regression of the percentage improvement of RF-CC versus LC-MNL over brand popularity and LC-MNL prediction accuracy. 148

3-11 Decision tree for consideration set formation of the renters based on the car-sharing dataset. 149

3-12 The average prediction scores under GCC and GCGC choice models. 156

3-13 Scatter plots of the prediction score improvements of GCC model over GCGC. . . . 157

4-1 Illustration for the proof in Lemma 4.2.1. The bottom part of a DAG is split in two independent terms. 170

4-2 Bottom parts of DAGs from Lemma 4.2.2. 176

4-3 Bottom parts of DAGs from Lemma 4.2.2. 176

4-4 Comparison of the performance of heuristic algorithm 3 vs. MIP (1.2). 183

4-5 Analysis of bounds for the probability of a DAG. 187

4-6 Comparison of approximations for purchase probabilities. 190

5-1 The case when $C_1 \not\subset C_2$ and $n - |C_1| - |C_5| > 0$. 204

5-2 The case when $C_1 \subset C_2$. 206

5-3 The case when $C_1 \not\subset C_2$ and $n - |C_1| - |C_5| = 0$. 206

List of Tables

1.1 Summary of the data. 58

1.2 Individual-level regressions with product category fixed effects. 90

1.3 Relevant summary statistics from the data. 92

3.1 Regression models where the dependent variable is the MAPE score under the MNL model. 134

3.2 Regression models where the dependent variable is the MAPE score improvement of the ICC model over the MNL model. 136

3.3 Summary statistics of the data used in IRI case study. 138

3.4 Descriptive statistics, the car-sharing dataset. 144

3.5 Additional descriptive statistics, the car-sharing dataset. 145

3.6 Logistics-based Consider-then-Choose (L-CC) and MNL model estimation results, the car-sharing dataset. 152

3.7 Statistical significance of three groups of car attributes. 153

Introduction

Digital technologies have dramatically reshaped industries in nearly every corner of the world. For instance, sharing economy platforms have become extremely popular in the last few years and have changed the ways in which we commute, travel, and borrow, among other activities. Even traditional large-scale brick-and-mortar retailers follow these trends in order to compete with newly popular online marketplaces. For example, technological developments have altered the manner in which offline retailers can customize their offers. Specifically, personalized promotions can now be implemented using electronic price tags, beacon-based technology, or electronic shopping carts embedded with a computer vision system, among other methods.

Technological advances, given their importance and widespread use, pose new challenges for companies to better solve classical problems within operational and marketing contexts, such as demand forecasting or running promotions. Demand estimation problems generally receive a lot of attention in operations because they are an integral part of revenue management as well as day-to-day operations in retail and online platform-enabled industries. Many models in practice assume each item in the product universe receives an independent stream of demand. Nevertheless, in the last two decades, we witnessed a paradigm shift from independent to choice-based demand models in order to capture the substitution within items in the product universe. Today choice-based demand modeling is a topic of great importance for both scholars and practitioners because it has significant revenue implications. However, demand calibration and prediction have always been very challenging for many reasons: (i) product availability that is hard to forecast (more so while making long-term demand predictions) because it varies over time as a result of either inventory stockouts or deliberate scarcity introduced by the firm; (ii) customer choices that are affected by promotion events (i.e., preferences may be altered by price or display promotions); (iii) the bounded rationality of individuals (i.e., customers evaluate only products in their consideration set which is unobservable; and (iv) scarcity of individual-level data (e.g., when only a few observations per customer are available for a particular category). The aforementioned technological revolution further exacerbates some of these challenges, which then require new solutions.

For example, technological developments have altered the manner in which offline retailers customize their offers. This creates a need to develop fine-grained demand models in order to evaluate individual customer responses to product promotions. The rise of the sharing economy creates an even greater need to account for users' bounded rationality as online platforms provide thousands of alternatives to choose from and the failure to account for consideration sets will result in biased demand estimates as well as incorrect predictions and inferences. Moreover, new online applications within the sharing economy exacerbate the problem of long-term demand forecasting because product assortments change rapidly over time. Think of an online peer-to-peer car-sharing platform where car availability over time depends upon the decisions of many independent car owners and renters that make car reservations in advance and at various times. To this end, success in making long-term demand predictions is largely dependent upon the robustness of choice-based demand models to the degree of ambiguity in offer set definitions. Note that to make demand predictions with choice models, we need to feed them accurate data regarding product availability, which is not always available. Most existing literature regarding choice-based demand estimation assumes that we are able to access accurate information on offer sets, which is a significant problem in practical usage. The previously mentioned emerging applications, both in retail and on online platform operations, reveal additional limitations within the existing work regarding choice-based demand modeling and operational decision:

- *Personalized promotions:* Most existing research focused on using aggregate sales transaction data to estimate demand models in the presence of stockouts and offer sets that change over time; then they use these estimates as inputs to solve assortment and pricing optimization problems. The paper by [6] provides an up-to-date overview of retail operations and revenue management literature regarding choice-based demand models. These models are used as inputs for key decisions, such as promotion optimization. The effect of sales promotions on retailers is studied extensively in marketing [8]. An overview of different promotion mechanisms used by retailers is provided by [34], where they summarize different objectives and effects of promotions. Only a few papers report a positive effect of sales promotions on long-run brand preferences [23, 32]. Most empirical studies conclude that mass sales promotions are good instruments to induce customers' substitution behaviors in the short run, but with neutral or negative effects on brand preference in the long run [24, 72, 86]. Most of this work focuses on an empirical understanding of the overall impact of promotions. We, on the other hand, seek to develop a methodology for carrying out personalized promotions.

- *Robust demand models accounting for consideration sets:* While calibrating a choice model from the sales transaction data, we encounter a major common challenge – the actual consideration set of a customer is unobserved. The offer set defines the set of products that the customer choose from, but the consideration set defines the set of products that the customer actually evaluates before making a final choice. In many emerging applications, the definition of the offer set is, itself, quite unclear. For example, in the context of online platform operations, a company needs to make short-term (or long-term) demand predictions in order to match with supply or to optimize strategic as well as marketing decisions. However, product availability varies over time and cannot always be known perfectly in advance. It is a common practical issue in retail when stockout events mask the true offer set information by adding noise to product availability data. Even if the offer set is well defined, the consideration set might still remain unclear, both on the online platform as well as in retail settings because of the physical and cognitive limitations of customers, that prevent them from evaluating all of the products offered. In all of these applications, it is important that the choice models remain robust amids the noisy definitions of the offer sets. Choice-based demand models accounting for consideration sets receive lots of attention both in marketing as well as operations fields [13, 35, 41, 99]. Our methodology is distinct because we provide a more flexible and data-driven framework for modeling consideration set formation of customers. As such, our model is able to subsume a variety of marketing consideration set heuristics and screening rules known in academia and practice, such as the *inertia in choice* or *short-term brand loyalty* [53]. This heuristic means that in terms of frequently purchased consumer goods, customers tend to stick to the same option rather than evaluate all available products in each store visit. In this thesis, we strive to showcase the application of a consider-then-choose framework in various real-world applications.

This thesis investigates choice-based demand estimation models in order to address the challenges stated above and explores emerging applications of the proposed methodologies in the operations field such as running personalized promotions or making robust demand predictions.

Our contributions

This dissertation consists of two parts. Part I investigates the problem of personalized operations in retail. In Part II, we study choice-based demand models, accounting for the consideration set formation of customers with applications both in retail and on online platforms. Within personalized operations, we focus on the problems of customized predictions and promotions in retail using consumer panel data. In the latter part, we specifically focus on two subproblems: (i) inferring consideration sets from sales transaction data and (ii) developing robust demand prediction models in retail and on online platform operations.

Personalized promotions

In the retail industry, it is crucial to build an efficient and profitable mechanism to run promotions, because a significant amount of money is spent on price reduction deals and an unprecedented number of goods are sold at discount prices. In the last several years, the rise of business analytics and the increased availability of data have motivated the retail industry to shift from using massive promotions to using customized offers in order to improve revenues and to better address the needs of consumers. These customized promotions rely on the heterogeneous product preferences of individuals within a given category as well as their different sensitivities to price discounts and/or to the location of products on the shelf. In Chapter 1, we tackle the problem of predicting individual customer responses to promotion decisions within a given product category and, based on these predictions, optimizing the portfolio of products to be promoted for a particular individual. The problem is challenging due to the limited number of observations available for each individual. To extract a signal from the limited data in the most efficient way, we model each individual through a directed acyclic graph (DAG) (i.e., partial order). In order to account for promotions in our model, each DAG consists of two copies of each product – under promotion and regularly priced. The DAG is constructed dynamically from a set of rules that account for the revealed preferences of each customer throughout their history of past purchases. In addition, for every item in the product category we add an edge to the DAG from its promoted copy to its non-promoted copy, assuming rational purchase behavior of the individuals. The DAG illustrates partial preferences of the form "promoted copy of product i is preferred to promoted copy of product j" through a directed edge from promoted copy of product i to promoted copy of product j, but typically does not provide a full preference list of all items in the product universe. The data sources required by our methodology include historical purchase transactions data tagged by customer ID, information about the assortment available for the product category of interest at the moment of purchase, and the identification of products that were on promotion at the time. Taking the collection of DAGs representing the customer basis as input, we calibrate an MNL model over the partial orders and quantify the prediction power on out-of-sample transactions. Then, we use this information to optimize personalized promotions. The illustrative DAG-based structure and intuitive properties of our systematic approach to running personalized promotions can appeal to both academia as well as industry, where many supermarkets still employ a rule-of-thumb approach in daily operations [20]. We summarize our contributions below.

- *We propose and analyze a nonparametric choice-based demand model that explicitly accounts for promotions.* We extend a nonparametric partial order-based choice model [52] to capture the promotion effect. The proposed model belongs to the family of Random Utility Maximization (RUM)-based choice models. In particular, we consider m customers making repeated purchases from a specific category of substitutable products over a finite period of time. In each store visit, customers sample a full list of preferences in accordance with their partial order, according to a distribution that corresponds to their market segment, and chooses the highest-ranking item within their consideration set.

- *We provide a preference graph decycling algorithm.* While processing the data source in the DAG construction phase, customers may exhibit an apparent inconsistency in their purchasing behavior, which may imply the creation of a cycle in the associated graph. Note that sales transaction data only provides us with a collection of revealed preferences for every individual – the set of available products at the time of purchase and the chosen product. Implying that the chosen brand is preferred over other brands that were available at time t but not chosen, we explicitly assume that the consideration set of a customer at time t consists of all the available items in the store. However, this might not be the case due to the limited attention of individuals. Therefore, ignoring consideration sets results in adding spurious edges (i.e., comparisons) to the DAG when building a customer's partial order. We address this challenge by running a decycling procedure based on a mixed integer linear programming (MILP) formulation in order to maintain a maximum number of arcs in describing customer preferences through a DAG. In other words, we want to keep as much information in the DAG as possible. To some extent, applying this decycling procedure is equivalent to accounting for consideration sets in a data-driven way, while building a customer's partial order.

- *We quantify predictive accuracy gains of the proposed choice-based demand model.* In our empirical analyses, we focus on real-world panel data regarding the sales transactions of 27 grocery categories across two large U.S. markets in 2007. These extensive empirical studies demonstrate that our approach to accounting for promotion effects results in more precise and fine-grained predictions of customer choice behavior. This is in comparison with state-of-the-art benchmarks that also incorporate promotion effects. In particular, we obtain up to 14% improvement in prediction accuracy, on average, across 27 product categories.

- *We derive bounds on estimation and prediction guarantees for the partial order-based choice model.* Obtaining an estimate of marginal distribution for partial preferences under the MNL model requires computing the likelihood of a DAG, which is known to be a #P-hard problem. The latter result follows immediately from the hardness of counting the number of rankings consistent with a partial order. Therefore, we use the approximation solution to efficiently compute the likelihood of the DAG and the probability that the customer chooses a specific product from an offer set conditioned on the partial order, which describes the customer's preferences. In the present paper, we derive lower and upper bounds relative to the exact likelihood of partial orders and to the exact distribution of partial preferences under the MNL model.

- *We propose a methodology for optimizing personalized promotions and test it on real-world panel data.* In the spirit of operations-related literature, the defined nonparametric choice-based demand estimation model is used as an input for a personalized promotion optimization framework in order to improve the retailer's revenue. Our approach to running personalized promotions allows for an intuitive and illustrative interpretation of the resulting customized promotions policy. This is a very appealing property for the retail industry, where many supermarkets still employ a manual process based on a rule-of-thumb approach as well as past experience, in order to decide price promotions. We observe that our methodology designed to optimize personalized promotions improves the retailers' revenue by more than 23% for certain product categories, based on the real-world panel sales transaction data.

Inferring consideration sets from sales transaction data

Chapter 2 investigates the problem of identifying consideration (or competition) sets from sales transaction data. This is a very important and challenging managerial decision for strategic planning as well as for managing day-to-day operations within the company. In particular, we study a general consider-then-choose (GCC) model. We fix the choice rule to be consistent with a single preference ordering but allow the distribution over consideration sets to be unrestricted. The key distinction of our work from the existing literature is that we allow the distribution over consideration sets to be general. At a high level, we contribute to the literature that infers consideration sets from the customer's perspective without imposing any prior belief on the consideration set formation. Our approach is rather general and completely data-driven. As such, our model subsumes the existing models (e.g., [68]). We outline our main contributions as follows.

- *Theoretical contribution.* We derive necessary and sufficient conditions in order for a collection of observed choice probabilities to be consistent with an underlying GCC model.

We then demonstrate that the GCC model can be identified from sales transactions data alone. In particular, we provide a closed-form expression for computing distribution over consideration sets from observed choice probabilities. We also show that when the consideration sets are of size k at most, the consideration set distribution can be recovered from choice probabilities under offer sets of size k at most.

- *Methodological contribution.* We demonstrate how to estimate the parameters of the GCC model. We begin with the Maximum Likelihood Estimation (MLE) problem for the restricted version of the GCC model, where customers sample items in their consideration set independently. Formulating this problem as a mixed integer non-linear program (MINLP), we show that it can be calibrated by solving a sequence of MILPs using the outer-approximation algorithm. Then, in order to calibrate the GCC model, we provide the EM-based algorithm by dividing customers into segments. Every segment is characterized by specific attention parameters for sampling consideration sets. We also propose a methodology in order to model the consideration set formation of customers using machine learning techniques (e.g., decision trees or random forests) that can account for product attributes in a non-linear and tractable manner.

Robustness of demand prediction models in operational applications

In Chapter 3, we analyze various conditions and real-world scenarios in which choice models, based on the consider-then-choose framework, provide a better predictive performance than state-of-the-art benchmarks. First, we provide the results of an extensive simulation study that demonstrates robustness of consider-then-choose models to the noise in offer sets. We find that our model improves over the benchmark as we add more noise to offer sets in the synthetic data. Interestingly, our model's ability to outperform the benchmark is higher in scenarios when benchmark predictions deteriorate. In this simulation study, we assume that noise results in an estimate of the offer set, which is a superset of the true offer set. We also model real-world scenarios when we do not know the offer set exactly, but we can determine a superset of the true offer set. For example, this is true in retail settings where stockout events mask true offer set information. The summary of major contributions in Chapter 3 is outlined below.

- *We compare the predictive performance of choice models using the IRI academic dataset modeling several real-world scenarios when the retailer faces ambiguity in the offer sets.* Our predictive analysis across 20 product categories suggests that the relative performance of our model over the benchmark improves once we switch to scenarios with a higher noise level. Expectedly, we have only a moderate decrease in prediction accuracy when increasing noise in the offer sets for consider-then-choose models. Moreover, the improvements of our model vary significantly across product categories. We find a positive correlation between the improvements of GCC over the benchmark and noise intensity. The latter is measured as an average percentage of the items that are stocked out in a store, across product categories.

- *We also apply the proposed methodology to address the problem faced by online peer-to-peer car-sharing platforms.* To succeed in the long run, an online platform needs to make long-term or medium-term demand predictions for the listed cars. The major problem is that the company cannot rely on accurate data regarding car availability over time. Therefore, the robustness of consider-then-choose models to the noise in the offer set definition plays an important role in this context, as demonstrated in Chapter 3. Additionally, the proposed framework's flexibility enables us to estimate the choice model with non-linear in-product attributes formation of consideration sets (e.g., decision trees or random forests), which can significantly boost demand prediction performance. For example, using the industry partner dataset, we find that the random forest-based consider-then-choose model outperforms the benchmark by more than 50% in terms of the RMSE metrics. We also provide an explanatory analysis after calibrating our models using car feature information in order to gain insight about the consideration set formation of renters. We find that some car features are more important in explaining this consideration set formation than renter's preferences and vice versa. For example, our findings indicate that car age (as opposed to rental price) plays a relatively more important role in the consideration set formation than in the final choice. Our empirical analysis also suggests that the renters are more likely to build their consideration sets based on car brands rather than car features, even though customers are more likely to pay attention to car properties rather than brands while evaluating alternatives to their final choice.

Part I

Personalized Operations

Chapter 1

Customized Retail Promotions and Demand Estimation

1.1 Introduction

Recent advances in technology, the availability of individual-level transaction data, and analytics have resulted in an expansion of opportunities for companies to engage in personalized operations. Huge volumes of individual-level consumer information are collected from past purchases, through loyalty programs or third party data-brokers (e.g., Acxiom), which contain highly detailed digital profiles on many users. Today, we observe customized retail pricing in both online and offline marketplaces. Online retailers, such as Amazon, have been offering personalized pricing for several years now based on shoppers' demographic information, geographical location, purchase and search histories, and types of devices used for access [85].

Technological developments have altered the manner in which offline retailers can customize their offers. Personalized promotions can now be implemented using electronic price tags or beacon-based technology. These technologies have already been adopted. For instance, the B&Q retail chain [78] uses electronic price tags. Stores such as Macy's, Marsh supermarkets, and Gamestop as well as mall developers such as Simon Property Group and Macerich, have all tested the beacon-based technology. In fact, Simon Property Group installed about 4,800 beacons over 192 malls to target customers using the Simon app, and in 2015, the top 100 retailers saw approximately $4 billion of sales from the beacon-based technology. Another technological option for personalized in-store promotions is the use of computer vision. For instance, the Apricart's application [60] runs on a screen device attached to a shopping cart and provides customers with a "throw-it-in-the-cart" and "pay-on-the-cart" checkout experience, while eliminating the traditional checkout process and providing relevant content (e.g., real-time customized deals) by detecting what goes in and out of the shopper's cart.

Personalized promotions offer several benefits to retailers. They offer an effective tool for individual-level price discrimination. They reduce competition by making the price paid by customers opaque to other retailers, which is not always the same as the sticker price. They also induce stronger relationships with customers, driving up sales. According to an Accenture survey [97], more than 60% of customers want to participate in customized promotions and explore real-time deals. Along the same lines, a more recent study conducted on 1,250 global shoppers [11] reveals that 65% of customers appreciate personalized prices. It appears that consumers appreciate services accompanied with personalization more than they dislike sharing personal information about their purchasing habits [28].

Motivated by the significance of personalized promotions, we provide a full methodological roadmap to run personalized promotions in retail setting. The required input data consists of a history of sales transactions for a category of substitutable products (e.g., coffee in a grocery store) tagged by individual customer IDs. With each transaction, the data also supplies the set of products available for purchase (i.e., product availability) and the subset of items under promotion. Using this data, the retailer must first infer customer-level preferences for items within the category of analysis, which is used to predict *each customer's* purchases in response to the retailer's promotion decisions. This inference problem faces three main challenges: (i) data sparsity, because only a few observations per customer may be readily available for a particular category, (ii) variation in the availability of products (e.g., due to stockouts), and (iii) presence of promotions that may alter ex-ante customer choices. The first challenge is the most significant for any personalized prediction. The latter two challenges complicate preference inference because it is difficult to tell if a customer switched her purchase because of a change in preference, or because of a stockout or promotion. Once customer-level preferences are estimated, the retailer must decide an optimal subset of products to promote (if any) for each individual customer visit to the store or website, with the objective of maximizing revenue from each visit.

Our focus in this chapter is on the immediate-to-short-term brand switching effects of promotion decisions. Therefore, we consider a retailer who wants to maximize immediate revenue from a customer visit. Promotions also have medium-to-long-term effects, such as stockpiling, consumption stimulation (leading to a general increase in product consumption levels), new customer attraction, and customer retention (e.g., store, category, or brand loyalties). Our focus on brand switching effects of promotions allows us to develop methods that address implementation problems involving thousands of customers and millions of transactions. In the conclusion of this thesis, we briefly comment on the ways in which our decision-making system can be extended to account for some of the medium-to-long-term impacts of promotions. We also argue that our proposal may still be helpful in mitigating the stockpiling effect.

1.1.1 Summary of results

The building block of our proposal is a nonparametric choice-based demand model where each customer is characterized by a directed acyclic graph (DAG), representing a partial order among products in a particular category. In the DAG, each product is represented by two nodes: a full price version and a discounted counterpart. A directed edge from node a to node b indicates that the customer prefers the product corresponding to node a over the product corresponding to node b. The DAG captures the fact that customer preferences are acyclic. Unlike a full preference list, a DAG specifies pairwise preferences for only a subset of pairs of products; therefore, it represents a partial order. When visiting the store, the customer samples a full preference list *consistent* with her DAG,[1] according to a pre-specified distribution, and chooses the highest-ranking available product.

Inferring customer preferences from transaction data consists of two key elements. The first element is the construction of the DAGs. Starting from an empty graph (i.e., a collection of isolated nodes representing products), and using historical data as a source of revealed preferences for each individual, we start adding edges from the purchased product to the other products (i.e., nodes) offered, distinguishing between regular and discounted versions of the same product. This process may lead to a graph with cycles, reflecting the fact that a number of “incorrect edges”

[1] By *consistent* we mean that all the pairwise relationships between products represented in the DAG are also satisfied by the total order sampled by the customer in a store visit. This is formally stated in Section 1.2.1.

could have been added along the way. In order to associate each customer with a partial order, we run a decycling procedure with the objective of dropping spurious edges. The output of this first phase is a DAG for each customer.

The second key element involves fitting a choice model that specifies the distribution with which the customer samples full preference lists consistent with her DAG. We fit a multinomial logit (MNL) model as well as a multiclass version. This estimation requires computing the likelihood of the constructed DAGs, which is a computationally hard problem in general. In order to ease the estimation process and the posterior prediction, we provide lower and upper bounds for the likelihood of a DAG, which are easy to compute and which are then used as an approximation for the exact probability.

The predictive power of our method is illustrated through an extensive set of numerical experiments using real grocery panel data on purchases across two large U.S. markets in the year 2007. We split our dataset for each of 27 product categories into two parts. On the first half (i.e., the training data), we perform the aforementioned two stages: DAG construction and MNL estimation (both single and multi-class). Then, on the second half (i.e., the holdout sample), we predict what each customer would purchase under our model when confronted with historical offer sets and products on promotion. We compare our prediction with the reported purchase. Our study demonstrates that our approach results in more precise and more fine-grained predictions of customer choice behavior in comparison to state-of-the-art benchmarks that also incorporate promotion effects. Specifically, we obtain up to an average of 14% improvement in prediction accuracy, using standard measures, across the 27 product categories studied.

Confident in the predictive power of our model, we then use DAG construction outputs and estimation stages as inputs to run personalized promotions. We formulate a mixed integer linear program (MILP) that decides which products to promote when a particular customer, identified by her DAG, faces a given offer set. We analyze two types of scenarios. The first focuses on the setting in which the retailer runs personalized promotions in conjunction with mass promotions already in place (as reported in the dataset). Thus, the retailer can personalize the promotion of only those products not already under mass promotion. The second affords more flexibility to the retailer and assumes that the retailer can personalize the promotion of *any* product on offer. Our simulated results show average improvements of 16.7% and 23.9% respectively, for the two scenarios, across 27 categories, when compared to the existing promotion strategy in place.

The empirical validation of our model supports its use towards the implementation of customized promotions in a systematic, data-driven manner. Another key advantage of our method is that the DAG-based representation of preferences provides an intuitive and transparent interpretation of the personalized promotion decision. This is an appealing feature for the retail industry where several technically-sophisticated grocery chains still rely on manual processes to decide price promotions based on a rule-of-thumb approach as well as past experience (e.g., see [19]).

1.1.2 Related literature

This chapter touches upon two streams of literature: marketing and operations. While the use of panel data as a source to estimate choice models is still limited in the operations literature, it has been around for a while in the marketing field. A pioneering work in this regard is the seminal paper by [38], where the authors fit an MNL model to household panel data on regular ground coffee transactions, and which has led the way for choice modeling in marketing using scanner panel data. [16] and [101] provide a detailed overview of choice modeling using panel data in marketing. Much of this research stream focuses on understanding how various panel covariates affect the individual choice process.

This chapter is most closely related to the body of empirical research within marketing focused on developing a methodology for individual-level marketing policies. [103] provide a decision-support system to optimize the timing and the depth of promotions for a given brand. Their structural model accounts for three simultaneous components: interdependence in purchase incidence, brand choice, and purchase quantity, and assumes that preferences (even for a single customer) may vary over time. Like in our case, the building block for the model is the individual household level, but the likelihood function is based on a latent class market structure, which captures unobserved consumer heterogeneity. This function can be specified in closed form but lacks convexity properties that would ease the estimation process and make it hard to scale to a large number of alternatives in the category (in fact, in their experiments they report results based on two categories with only four options each). The promotion decision problem considered is also different. Whereas we focus on the optimal subset of brands to put on promotion, [103] assume that the retailer decides on a single brand or manufacturer to promote at a time. Given the brand, the decision focuses on how to set discounted prices (both the timing and the depth of promotions) for the next few store visits of a given individual. This price promotion problem is highly non-linear and lacks any structural property that also makes it hard to scale for a large number of variants in the category.

[56] develop a dynamic programming-based approach similar to the one in [103], but they use individual-level coefficients to evaluate the benefits of optimizing customized promotions at the level of each single customer. However, this twist makes their methodology even more computationally intractable, as for estimation, they need to use a Markov chain Monte Carlo procedure to simulate the posterior distribution of the model parameters and to compute household level estimates of preferences.

Our proposal here is rooted in a rank-based choice model of demand. This type of nonparametric choice model specifies customer classes defined by their rank orderings of all alternatives within the product category. When visiting the store, a customer is assumed to purchase the available product with the highest ranking in her preference list, or to leave without making any purchase. This model, which provides the full flexibility of random utility models, has been gaining increasing attention in the OM-related literature [29, 66, 83, 95]. However, these references still assume a market-level choice-based demand model.

In [52], a first step is taken towards the specialization of the rank-based choice model to capture and estimate individual preferences. In that paper, the authors propose to model individual preferences through DAGs, but their construction is guided by heuristic definitions of the consideration sets (e.g., see [39]). Therein, a customer samples a full preference list of items in the product universe (along with the no purchase alternative) in accordance with her partial order, forms a consideration set, and then buys the available product among the considered ones with the highest ranking. Three models based on different consideration set definitions were studied: i) standard, where the consideration set is equal to the offer set; ii) inertial, where the consideration set is a subset of the offer set given by the previous purchase and the current products on promotion; and iii) censored, which is a slight generalization of the inertial model. Both (ii) and (iii) were designed to capture the inertia in choice [53], which is a principle claiming that customers tend to stick to the same option when facing frequently purchased consumer goods.

Motivated by the promising predictive results of that model (which was also successfully applied recently to model preferences for virtual items in video games; [57]), in this chapter we leverage the performance of the DAG-based approach with the objective of designing customized promotions. Our contribution with respect to [52] spans several dimensions. First, our consideration set formation is purely data-driven, providing greater flexibility without imposing any prior beliefs on bounded rationality of individuals, such as the stickiness principle for the aforementioned inertial model. This approach allows us to extend the coverage of the number of individuals whose behavior our model can explain with non-empty DAG structures. Second, in our new proposal we explicitly account for promotions as part of the DAG definition (and not indirectly through the heuristic formation of the consideration set). Our method of incorporating promotions forms the fundamental backbone of the proposal. It provides clean managerial insights about customer preferences (as explained later) and allows running promotion optimization in a transparent way. Third, from a theoretical perspective, we develop tractable analytical lower and upper bounds for the likelihood of DAGs under the MNL model. To this end it is known that computing the exact likelihood is a #P-hard problem. The lower bound is indeed the exact probability of a DAG when it is a forest of directed trees, as shown in [52]. Here, we demonstrate that under some technical conditions, the bounds are asymptotically tight. In addition, we derive tractable analytical lower and upper bounds for the MNL probability of a customer choosing a specific product conditioning on her DAG and the available offer set. Finally, we address the promotion optimization problem as a key distinguishing feature of our work, whereas in [52] the focus was limited to establishing the predictive power of the behavioral-based DAG model.

1.2 Choice model description

This section formally introduces our general modeling framework, starting from some basic notation and explaining the choice process derived from the customers' DAGs. We continue with the description of the data model that serves as input for our choice model, followed by the presentation of the different phases involved in the DAG construction procedure. Next, we discuss the underlying assumptions for our model, and close the section with the formulation of the associated maximum likelihood estimation problem.

1.2.1 Modeling framework

Consider a category of n substitutable products on which a set of m individuals make purchases over a finite horizon. Both the set of customers and the set of products remain constant over time. Each product has two different versions: the regularly priced version and its promoted counterpart. The promotion could be a price or display promotion or any form of product presentation that highlights its presence on the shelf. We denote by $\mathcal{N}$ the set $\{a_1, a_2, \ldots, a_n\}$ of regularly priced versions of the products. For any $j \in [n]$ (i.e., $1 \leq j \leq n$), we let a_{j+n} denote the promoted version of product a_j. Furthermore, we let $\mathcal{N}' = \mathcal{N} \bigcup \{a_{n+1}, a_{n+2}, \ldots, a_{2n}\}$ denote the expanded product universe with the corresponding promoted counterparts.

The preferences of each customer over the product universe $\mathcal{N}'$ are described through a partial order, which could be visualized as a directed acyclic graph (DAG). A DAG D consists of $2n$ nodes, with two copies for each product (one for the non-promoted version, and one for its promoted counterpart), and a collection of directed edges (or pairwise preference relations) denoted by $E_D \subset \{(a_k, a_j) : 1 \leq k, j \leq 2n, k \neq j\}$, so that for any $(a_k, a_j) \in E_D$ we have that item a_k is preferred to item a_j. With the assumption that a customer always prefers the promoted version of a product over its regularly priced counterpart, the DAG has n arcs of the form (a_{j+n}, a_n).

The DAG captures the strong preferences the customer has over the products. These preferences remain constant from one purchase instance to the next one. For instance, suppose that a customer *always* prefers caffeinated (regular) coffee over decaffeinated (decaf) coffee. Such a customer will be captured by a DAG with preference edges from every regular coffee brand (promoted or not) to every decaf coffee brand (promoted or not). The customer's brand preferences may change from one purchase instance to another, but she will always purchase regular coffee over decaf coffee. Note that a customer may have no strong preferences, in which case her DAG would be rather sparse (or even empty). At the other extreme, a customer may have very strong preferences over all the products, in which case her DAG would be a total ordering over the $2n$ products. The DAGs provide us with a flexible tool to capture customers between these two extremes.

We will describe the complete process for constructing the DAG from the observed transaction data below. But for now, given these DAGs, we describe the choice process. In general, DAGs can only specify what a customer *will not* purchase – rather than what she will purchase – in each store visit. For instance, in the example above, the DAG specifies that the customer will not purchase decaf coffee in the presence of regular coffee, but it remains silent on which of the regular coffee brands she will purchase. Because the preferences not present in the DAG may change between purchase instances, we capture them through a probabilistic model. We let λ denote a distribution over all possible total orderings of the $2n$ products. A total ordering (unlike a partial order) specifies the pairwise preferences for *all* possible $\binom{2n}{2}$ pairs. Equivalently, a total ordering is a ranking (i.e., permutation or preference list) of the $2n$ products. In each interaction with the retailer, the customer samples a ranking that is *consistent* with her DAG D according to distribution λ (to be estimated from data as explained below). She then chooses the most preferred product according to the sampled ranking from a subset of products she considers from among the offered products.

More formally, if σ denotes a preference list, $\sigma(a_j)$ indicates the preference rank of product a_j. A lower ranking indicates a higher preference order; in other words, we have that a_k is preferred to a_j according to σ, written as $a_k \succ_\sigma a_j$, if and only if $\sigma(a_k) < \sigma(a_j)$. We say that a preference list σ is consistent with partial order D if and only if $\sigma(a_k) < \sigma(a_j)$ for each $(a_k, a_j) \in E_D$. Upon arrival to the store a customer samples a full ranking σ consistent with her DAG D according to a distribution λ. In other words, we can interpret any partial order D as a censored representation of the underlying full rankings σ that a customer could sample. Given the probability mass function $\lambda(\sigma)$ for all full rankings σ, we define S_D as the set of rankings compatible with D, i.e., $S_D = \{\sigma : \sigma(a_k) < \sigma(a_\ell) \text{ whenever } (a_k, a_\ell) \in D\}$. As a result, we can compute the likelihood of DAG D as follows:

$$\lambda(D) = \sum_{\sigma \in S_D} \lambda(\sigma). \tag{1.1}$$

In the store, the customer is offered a subset of products $S \subset \mathcal{N}'$. Naturally, at most one element between a_j and a_{j+n} is included in S. Let $C \subseteq S$ denote the subset of products the customer considers during this visit. Then, she purchases the most preferred product a_k within the set of considered products, i.e., $a_k = \arg\min_{a_i \in C} \sigma(a_i)$. The customer could sample a different

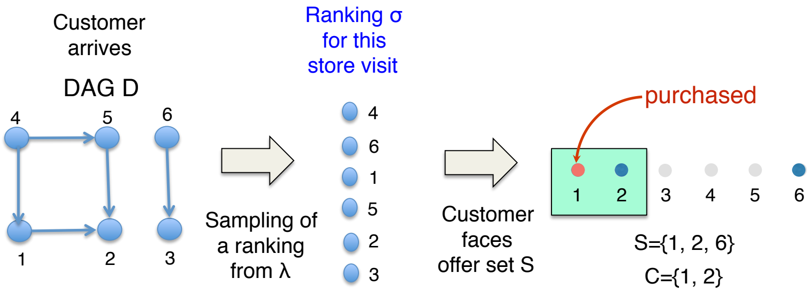

Figure 1-1: Choice process example with three products.

ranking, independently, in each store visit, but the ranking is always consistent with her DAG D. Figure 1-1 illustrates the choice process for a particular store visit given a DAG D, a distribution over full rankings λ, and an offer set S. We do not impose any structural assumptions on how the consideration set is formed by the customer. In the absence of any additional information, in principle we assume that $C = S$, i.e., the customer considers everything on offer (this is indeed our approach in the numerics in Section 1.4).

We note that different customers may have different DAGs, but they all use the *same* distribution λ to sample the rankings. In other words, the distribution λ is a population attribute whereas the DAG is an individual attribute. Even though the same distribution λ is being used by all the customers, our model easily captures preference heterogeneity. For instance, the distribution λ could be a latent-class multinomial logit (LC-MNL) model, which assumes that the population is comprised of K latent classes and the preferences of each class of customers is described by a different MNL model, thereby allowing for preference heterogeneity. In addition, the rankings sampled by customers must be consistent with their respective DAGs, so the effective distribution used for each customer is the conditional distribution λ given her DAG. Because DAGs differ across customers, these conditional distributions will also differ.

1.2.2 Data model

Our data model is the same as the one used in [52]. We consider a dataset with transactions tagged by the IDs from m customers. For a given customer i, we consider a training horizon of T_i transactions of the form $(a_{j_{it}}, S_{it})$, for $t = 1, 2, \ldots, T_i$, that we use to infer her partial order of preferences. The offer set is $S_{it} \subset \mathcal{N}'$, and $a_{j_{it}} \in S_{it}$ denotes the product she purchased in period t.

The subset $P_{it} \subset S_{it}$ denotes the set of promoted products in period t. In our dataset the promotion could be either *display* or *price*. In our numerics we restrict to price promotions, and in particular we consider the promotion feature as a binary attribute of a product. That is, we do not distinguish between different levels of price promotions although our model could be easily extended to account for a finite number of price discount points by simply adding a product copy (i.e., node) for each discount level in the discrete set.

In order to partially mitigate the data sparsity issue, in our implementation we aggregate products within a category by brand, so as to have at least a few observations of offerings and purchases for each of the items.

1.2.3 DAG construction

We now describe how we build the DAG for each customer using her historical purchase transactions within the category. We process customer transactions one-at-a-time to dynamically build the DAG. The whole process involves four steps, but at the core, it relies on a set of preference inferences made from each transaction. In order to illustrate the process and its challenges, consider a transaction in which the customer was offered products a and b and she purchased product a. Given this transaction, we can reason that there are three different possibilities: (i) the preference $a \succ b$ is strong and so the edge (a, b) must be part of the customer's DAG; (ii) the preference $a \succ b$ is *not* a strong preference and the customer simply sampled a ranking σ with $\sigma(a) < \sigma(b)$ for this purchase instance, so neither edge (a, b) nor (b, a) should be part of the customer's DAG (if the edge (b, a) were part of the DAG, then the preference list σ with $\sigma(a) < \sigma(b)$ would never be sampled because it would be inconsistent with the DAG); and (iii) product b was not considered by the customer and therefore, we cannot make any inferences about the preference relation between a and b in the DAG (in fact, it is perfectly possible for $b \succ a$ in the DAG, but it would not matter because as far as the customer is concerned, product b was never under consideration). All three inferences are consistent with our model, and therefore, more than one DAG is consistent with the given data. Our challenge lies in identifying the set of DAGs that are consistent with the given data, and then using a reasonable criterion to pick one from this set.

At a high level, we deal with the above challenge in the DAG construction by first building a directed graph G including what we call *candidate* edges. A candidate edge is an edge that we are unsure of, with the understanding that it may be removed at a later stage in the DAG construction process. As seen below, the graph G allows us to keep track of the set of DAGs that are consistent with the given transaction data. We then make an identification assumption to pick a DAG that is a subgraph of G. To keep our presentation clean, we first describe all the steps involved in constructing a DAG. We then discuss in Section 1.2.4 all the assumptions implicit in our DAG construction process. Figure 1-2 illustrates the DAG construction process for a small running example with four products ($n = 4$) and three transactions ($T = 3$), following the sequence of four phases below.

Phase 0: Initializing the preference graph with edges from promoted versions to corresponding non-promoted versions of products. We start from an empty graph G, and add $2n$ isolated nodes from the product universe $\mathcal{N}'$, where each node represents either a non-promoted or a promoted version of a product. Let E_G denote the set of edges in the graph G. Starting from the empty set E_G, we add n edges (a_{j+n}, a_j) for each non-promoted item a_j, $j \in [n]$. These edges capture the fact that a promoted copy a_{j+n} of every product $a_j \in \{a_1, \ldots, a_n\}$ is preferred to its regularly priced copy a_j since both products have the same attributes except the promotion feature. Note that these are not candidate edges because we are certain of their presence in the final DAG.

Phase 1: Adding *candidate* edges from sales transactions. We incrementally add *candidate* edges to the preference graph G by processing the customer's transactions one-at-a-time. For each transaction $(a_{j_{it}}, S_{it})$ of individual i, we draw edges from $a_{j_{it}}$ to the other items in the offer set, i.e., $E_G \leftarrow E_G \cup \left\{(a_{j_{it}}, a_\ell) : \forall a_\ell \in S_{it} \setminus \{a_{j_{it}}\}\right\}$. These edges signify that potentially all the offered products were considered by the individual, and all preference edges were indeed "strong" preferences (i.e., they are all part of the DAG and not just sampled preferences). We keep track of the purchase events where each edge (a_j, a_ℓ) is added through the weight $w_{j\ell}$, defined as the number of times the customer chose product a_j when a_ℓ was also offered.

Phase 2: Adding *implicit* candidate edges. In order to make the DAG denser, we enrich it with *implicit* candidate edges, based on the assumption that if a customer has a strong preference

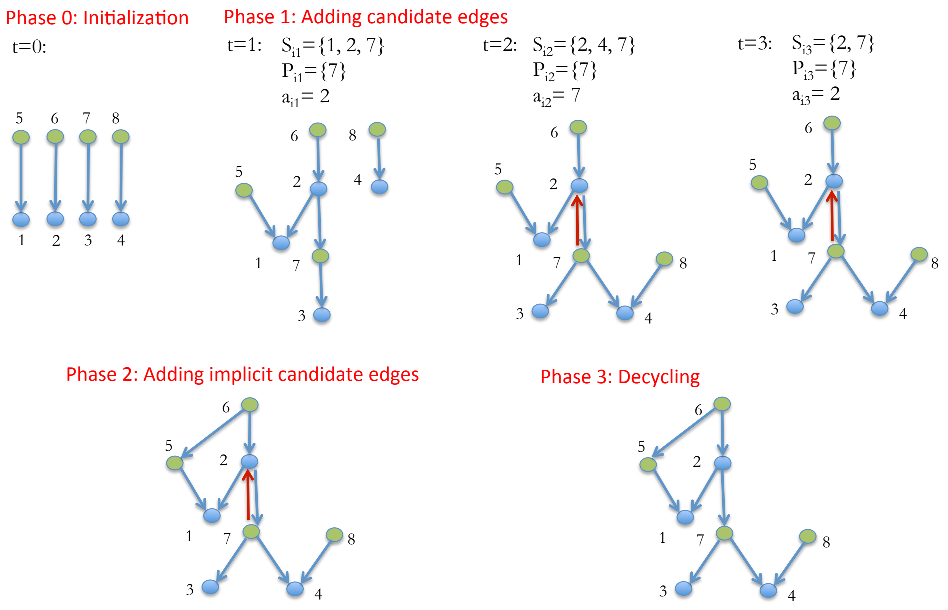

Figure 1-2: Phases of DAG construction. The example with four products and three transactions.

between the non-promoted (promoted) copies of two products, then the preference extends also to the corresponding promoted (non-promoted) copies. More precisely, for any $1 \leq j, \ell \leq n$, if $(a_j, a_\ell) \in E_G$, then we add the edge $(a_{j+n}, a_{\ell+n})$ to G, i.e., $E_G \leftarrow E_G \cup \{(a_{j+n}, a_{\ell+n})\}$. Similarly, if $(a_{j+n}, a_{\ell+n}) \in E_G$, then we add the edge (a_j, a_ℓ) to G, resulting in $E_G \leftarrow E_G \cup \{(a_j, a_\ell)\}$. To deemphasize the implicit counterparts of these edges, we assign the weight $w_{j+n,\ell+n} \leftarrow w_{j,\ell}/(T_i n^2)$ when $(a_{j+n}, a_{\ell+n})$ is the implicit edge, and $w_{j,\ell} \leftarrow w_{j+n,\ell+n}/(T_i n^2)$ when (a_j, a_ℓ) is the implicit edge. In other words, the weights of the implicit counterparts are scaled down by a factor of $T_i n^2$. Intuitively, by scaling down the weights of the implicit edges, we prioritize candidate edges over implicit candidate edges, since candidate edges inferred directly from the revealed preferences of customers are likely to be more informative. The reason for this precise choice of scaling factors will become clear below.

Phase 3: Graph decycling. This is a critical step in the DAG construction process, in which we attempt to eliminate the spurious edges added in G so far to arrive at the final DAG, where by *spurious* we mean that the edge contradicts the interpretation of other edges in the graph. The first indication that there are spurious edges in G is the presence of directed cycles. As discussed above, the data does not identify the DAG and therefore, we need to make an identification assumption to arrive at the final DAG from G. We assume that the underlying DAGs of customers are large, so we find the largest weight DAG that is supported by the choice observations. In other words, we assume that all candidate and implicit candidate edges are part of the underlying DAG, unless contradicted by data. This assumption translates to deleting cycles from G while maximizing the aggregate weight of the edges retained (or similarly, minimizing the total weight of the edges deleted). This problem is known in the graph theory literature as the *minimum weight feedback arc set* problem and is known to be NP-hard even when all weights are equal to 1 ([55] provides a reduction from the minimum vertex cover problem).

We formulate the above decycling procedure as a mixed integer linear program (MILP). For a given graph G, and for every edge $(a_k, a_\ell) \in E_G$, define the binary variable $x_{k\ell}$ that takes the value 1 if edge (a_k, a_ℓ) is finally retained in the induced acyclic subgraph $D \subset G$, and takes value zero otherwise. To ensure that the DAG defined by the variables $\{x_{k,\ell} \colon (a_k, a_\ell) \in E_G\}$ does not contain cycles, we introduce auxiliary binary variables $y_{k,\ell}$, for all $1 \leq k, \ell \leq 2n$ and $k \neq \ell$. These variables represent a total order over all the products in $\mathcal{N}'$ with $y_{k,\ell} = 1$ if a_k is preferred over a_ℓ, and $y_{k,\ell} = 0$ otherwise. The following MILP enforces that the final DAG is a subset of some total order defined by the $\boldsymbol{y}$ variables:

$$\max_{\boldsymbol{x},\boldsymbol{y}} \sum_{(a_k,a_\ell)\in E_G \setminus \{(a_{j+n},a_j)\colon 1\leq j\leq n\}} w_{k\ell}\, x_{k\ell} \tag{1.2}$$

$$\text{s.t.:}\ x_{k+n,k} = 1, \quad \forall\ 1 \leq k \leq n \tag{C1}$$

$$x_{k,\ell} = y_{k,\ell}, \quad \forall (a_k, a_\ell) \in E_G, \tag{C2}$$

$$y_{k\ell} + y_{\ell k} = 1, \quad \forall a_k, a_\ell \in \mathcal{N}', \quad k \leq \ell, \tag{C3}$$

$$y_{k\ell} + y_{\ell p} + y_{pk} \leq 2, \quad \forall a_k, a_\ell, a_p \in \mathcal{N}', \quad k \neq \ell \neq p, \tag{C4}$$

$$y_{k,\ell} \in \{0, 1\} \ \ \forall a_k, a_\ell \in \mathcal{N}', \quad k \leq \ell.$$

The constraints guarantee that the induced subgraph D defined by those edges $(a_k, a_\ell) \in E_G$ for which $x_{k\ell} = 1$, is a DAG. The first set of equalities (C1) ensures that all the edges added in Phase 0 are retained in the final DAG. The second set of equalities (C2) ensure that the DAG D defined by the variables $\boldsymbol{x}$ is a subset of the graph defined by the total order corresponding to the variables $\boldsymbol{y}$. The third (C3) and fourth (C4) constraints together ensure that $\boldsymbol{y}$ indeed defines a total order. Specifically, the third set of constraints ensures that either a_k is preferred over a_ℓ or a_ℓ is preferred over a_k but not both, and the fourth set of constraints imposes the total ordering among any three products. The correctness of the MILP is shown in Proposition 1 in Section 4.2.1 in the Appendix. This proposition shows that the MILP maximizes the weight of the candidate edges in the resulting DAG, and it never deletes a candidate edge if an implicit candidate edge can be deleted to break a directed cycle.

The size of the MILP (1.2) scales quadratically with n in the number of variables and cubically in the number of constraints. In Section 4.3 in the Appendix we propose a greedy heuristic to approximately solve the preference graph decycling in polynomial time. We compare the output DAGs of the heuristic and MILP (1.2) on our dataset and observe that under the heuristic we obtain only 0.4% sparser DAGs (in aggregate), which indicates its promising applicability in other real size problems.

1.2.4 Discussion of the model assumptions and the DAG construction procedure

We now discuss the most relevant assumptions we make for developing the model and constructing the DAGs.

Model assumptions. The assumption that the product universe $\mathcal{N}'$ and the set of customers remains constant over a finite horizon is needed to infer the customer DAGs. Echoing [52], our approach can be run periodically to update the DAGs and incorporate new customers. In between updates, new products in the category can be considered as part of a family of products (say, products of the same brand represented by only two elements: a_ℓ and $a_{n+\ell}$), which is the minimal level of data aggregation that we consider.

Note that in our presentation, without loss of generality, we do not provide a special treatment to the always available, no-purchase option. This option can be handled in the same way as other items in the product universe except that only one copy of this alternative would be part of the DAG, i.e., the no-purchase option can be represented by a particular node in the DAG, say a_0.

DAG construction assumptions. There are two key distinctions between the DAG construction process here and the one in [52]. First and foremost, our proposal here is purely data-driven and in Phase 1 mirrors the *standard* consideration set definition therein (under which the consumer chooses among all the products in the offer set), though accounting explicitly for promoted products, here represented by node entities. The other two models presented in [52], inertial and censored, which actually showed the best predictive performance, are based on behavioral rules to build the consideration sets.

The second key distinction is the way we address apparent *inconsistencies* in the purchasing behavior of a customer. According to [52], during the DAG construction process, as soon as the addition of arcs into a customer DAG D_i implies the creation of a cycle or the customer's transaction can not be explained by the pre-specified behavioral assumptions, the process stops and all the arcs are deleted, keeping D_i as the empty DAG (i.e., a collection of isolated nodes). In such case, no structure is superimposed and the customer could be described by a standard choice-based demand model (e.g., a typical, single-class MNL). In our new proposal, Phase 2 could end with a graph with cycles, which are then deleted in Phase 3. In the context of our model, cycles could originate because of spurious edges introduced for reasons (ii) and (iii) laid out in Section 1.2.3. Under this interpretation, consumers are fully rational and the modeler incorrectly added edge (a, b) either because $\sigma(a) < \sigma(b)$ in the particular ranking σ sampled in this particular store visit (although (a, b) is not a strong preference, case (ii)), or because the modeler incorrectly assumed that b was part of the consideration set (case (iii)). As discussed in the description of Phase 3, our decycling procedure deletes a minimum number of spurious edges added along the way.

Another possible way to rationalize the decycling process is model misspecification. Customers exhibit a bounded rational behavior, including possible inconsistencies in their purchases. In this case, the addition of candidate edges in Phase 1 assuming that the consideration set of the customer is indeed the entire offer set is *correct* for that purchase instance, but the customer is inconsistent over time. The decycling step in Phase 3 provides the largest DAG that is sustained by the customer's inconsistent purchase behavior although that behavior over time cannot be explained by a DAG.

The *identification assumption* that the underlying customer DAGs are large is driven by our desire to retain a rich representation of the customers' strong preferences. This assumption is reasonable for product categories in which customers make repeated purchases, which increases their familiarity of the product category, allowing them to develop strong preferences. Grocery categories are a good example of that, as also evidenced by our empirical study. One can imagine other assumptions that are appropriate in particular settings. Such assumptions will result in a different decycling step and adjustments in the corresponding MILP, while keeping the rest of our framework intact.

Another point that deserves discussion is the addition of implicit candidate edges. These edges are not directly revealed in the customer's choices. So the customer may have revealed that a_k is preferred over a_ℓ, but she has not revealed if the same preference extends to their respective promoted copies a_{k+n} and $a_{\ell+n}$. Yet, we make the assumption that the customer is likely to prefer a_{k+n} over $a_{\ell+n}$. The reason is that strong preferences of the customer (those that are part of the DAG and do not change from one purchase instance to next) are likely to be driven by characteristics other than promotion activity, which does vary from one visit to another. As an example, a customer may prefer regular coffee to decaf coffee because of taste. Such a customer would continue to prefer regular coffee over decaf coffee even if both of them are on promotion. There is still the possibility that our assumption is wrong in specific cases because of which we scale down the weights of implicit edges by $T_i^2 n$. Proposition 1 in Section 4.2.1 in the Appendix shows that with this scaling the MILP strictly prioritizes implicit candidate edges over the candidate edges for deletion. Overall, since we do not observe whether the pairwise comparisons in the DAG are correct or spurious, we test empirically whether it is effective to add those implicit edges in the preference DAG (see Section 1.5 in the Appendix). We notice that by adding implicit edges in the DAG construction process and obtaining denser DAGs, the improvements in the prediction performance are significant.

1.2.5 Maximum likelihood estimation of the DAG-based choice model

Once we infer the customers' DAGs, we use maximum likelihood estimation (MLE) to calibrate a probability distribution over the full rankings consistent with these DAGs. In order to compute the panel data log-likelihood function, we consider only revealed preferences that are consistent with the inferred DAGs. That is, if during the DAG construction process no cycle was formed, then every transaction pair $(a_{j_{it}}, S_{it})$ (which is a star graph with head $a_{j_{it}}$ and set of leaves $S_{it} \setminus$

$\{a_{j_{it}}\}$), is a subgraph of the corresponding DAG D_i. In case a cycle was formed, say for customer i, consider a transaction $(a_{j_{it}}, S_{it})$ such that one of the edges $(a_{j_{it}}, a_k)$, with $a_k \in S_{it} \setminus \{a_{j_{it}}\}$, was deleted in the decycling procedure. Since $(a_{j_{it}}, a_k)$ was part of a cycle, it follows that there is a directed path from a_k to $a_{j_{it}}$ in the final DAG D_i. This implies that conditioned on the customer having DAG D_i, she did not consider product a_k when choosing $a_{j_{it}}$ even though a_k was on offer. Therefore, product a_k can be ignored for computing data log-likelihood.

Once we filter out these inconsistent preferences, the likelihood function that we maximize to calibrate the model is just the sum of likelihoods of customers' partial orders, i.e.,

$$\log \mathcal{L}(\text{Panel Data}) = \sum_{i=1}^{m} \log \lambda(D_i) = \sum_{i=1}^{m} \log \left(\sum_{\sigma \in S_{D_i}} \lambda(\sigma) \right).$$

See Proposition 2 in Section 4.2.1 in the Appendix for a formal justification of this log-likelihood expression.

Finally, note that the tractability of the MLE problem depends on the distribution λ over preference lists, e.g., the log-likelihood function is concave under the MNL/Plackett-Luce distribution.

1.3 Theoretical analysis of the DAG-based MNL Model

We now focus on two computational problems that arise in using our model with data: computing (a) the probability of a DAG D and (b) the choice probability given an offer set S conditioned on a DAG D. The first computation is needed to solve the estimation problem discussed above, and the second one is needed to predict the purchase of a customer with DAG D. Both computations are difficult for a general DAG D. In fact, even the problem of counting the number of total orders that are consistent with a given DAG D (i.e., the size of set S_D) is a #P-hard problem [7]. For that reason, we limit our attention to the standard Plackett-Luce (PL) [69] model for the underlying distribution λ over rankings, for which at least there is a closed form expression for the likelihood of a ranking. In the PL model, each product $a \in \mathcal{N}$ is associated with parameter (i.e., weight) $v_a > 0$. The probability of sampling ranking σ is given by

$$\lambda(\sigma) = \prod_{r=1}^{n} \frac{v_{\sigma_r}}{\sum_{j=r}^{n} v_{\sigma_j}}.$$

For brevity of notation, we also use v_j to refer to v_{a_j}, for a given indexing of the products. As shown by [51], the choice probabilities under the PL model are consistent with those under a standard MNL model with the same parameters $(v_a)_{a \in \mathcal{N}}$. In other words, we have

$$\Pr(a_i|S) = \frac{v_i}{\sum_{a_j \in S} v_j}$$

under both the PL and MNL models. Under the PL model, the choice probability $\Pr(a_i|S)$ is equal to the probability of a star DAG with edges from product a_i to all the products in the set $S \setminus \{a_i\}$ because this DAG always results in the choice of a_i from S. The choice probability under the MNL model, on the other hand, can be derived from its random utility specification [5]. Because of this equivalence of both models on the choice probabilities, we use the terms PL and MNL interchangeably.

1.3.1 Tractable analytical bounds for the likelihood of a DAG

We first focus on the problem of computing the likelihood $\lambda(D) = \sum_{\sigma \in S_D} \lambda(\sigma)$ of a DAG D under the PL model. [52] derive a closed form expression for $\lambda(D)$ when D satisfies a special structure. To state the result, we introduce the concept of *reachability*. The reachability function Ψ_D of DAG D maps each node a to the set of nodes that can be reached from a through the edges in D. More precisely, $\Psi_D(a) = \{b\colon$ there is a directed path from a to b in $D\}$. We assume that a node is reachable from itself, so $a \in \Psi_D(a)$ for all a, and Ψ_D is always non-empty. The DAG D is equivalently described by the reachability function $\Psi_D(\cdot)$ of its nodes. Without loss of generality, we represent the DAG D by its unique transitive reduction, which is the unique graph with the fewest number of edges possible and the same reachability function as D. We start from DAGs that are forests of *directed trees*[2] with unique roots, where *root* is any node with no incoming

[2] A *directed tree* is a connected and directed graph that would still remain acyclic if the directions are ignored.

edges. It is then shown in [52, Proposition 3.2] that

$$\lambda(D) = \prod_{a \in \mathcal{N}} \frac{v_a}{\sum_{a' \in \Psi_D(a)} v_{a'}}, \tag{1.3}$$

whenever D is a forest of directed trees, each with a unique root.

[52] propose to use the equation (1.3) to approximate the probability of a general DAG, even if it is not a directed tree. For the general case, they do not provide any guarantees for this approximation, suggesting that computing the probability of a DAG is difficult in the presence of v-nodes, defined as the nodes with at least two incoming edges. We now show that equation (1.3) provides a lower bound approximation for the probability of a general DAG. In particular, we establish the following result.

Proposition 1.3.1. *Under the PL model, we have that for any DAG D,*

$$\tilde{\lambda}(D) \leq \lambda(D), \ \ \textit{where } \tilde{\lambda}(D) := \prod_{a \in \mathcal{N}} \frac{v_a}{\sum_{a' \in \Psi_D(a)} v_{a'}}.$$

The inequality above is strict if D has at least one v*-node.*

The proof is rather involved and the details are provided in Section 4.2.1 in the Appendix. Here, we provide a sketch. The proof uses induction on v-degree, k, of D, defined as the sum of the degrees of the v-nodes in D minus the number of v-nodes. The base case of $k = 0$ follows from the equation (1.3) because the v-degree of D is zero if and only if D is a forest of directed trees, each with a unique root. To establish the induction step, we consider a DAG with v-degree of $k+1$ and carry out the following "splitting" operations to create a DAG with v-degree of at most k, in order to apply the induction hypothesis. We pick a v-node a_y with the property that the subgraph $D[a_y]$ "hanging" from node a_y – which is induced by D on the set of nodes $\Psi_D(a_y)$ – is a directed tree. Such a v-node always exists (at the minimum, it is a leaf in the DAG). Then, we split D into DAG $D[a_y]$ and the remaining DAG $\bar{D}[a_y]$, which is induced by D on the set of nodes $(\mathcal{N} \setminus \Psi_D(a_y)) \cup \{a_y\}$. We then split the node a_y in DAG $\bar{D}[a_y]$ to create a new copy a'_y such that one of the incoming edges into a_y moves to the node a'_y while the other incoming edges remain with node a_y, resulting in the DAG D_y^{split}. This splitting operation results in new nodes for which the PL parameters values must be appropriately defined. With these parameter values, we show that our splitting operation can only reduce the probability of the resulting collection of DAGs. We then establish the result by invoking the induction hypothesis on D_y^{split}, which by construction has a v-degree of at most k.

An upper bound for the likelihood of a DAG D can be readily obtained by deleting some edges in D. Deleting an edge strictly increases the set of permutations that are consistent with the DAG, so for any $\bar{D} \subset D$, we have that $S_{\bar{D}} \supset S_D$, where recall that S_D is the set of all rankings that are consistent with D. It thus follows that $\lambda(\bar{D}) \geq \lambda(D)$. We state this result formally in the following proposition and prove it in Section 4.2.1 in the Appendix.

Proposition 1.3.2. *For any two DAGs D and $\bar{D}$ such that $\bar{D} \subset D$, we must have that $\lambda(D) \leq \lambda(\bar{D})$, with strict inequality under the PL model if all the parameter values are strictly positive.*

Note that the above result is true for any distribution λ and not just for the PL model. To obtain a tractable upper bound under the PL model, we choose a DAG $\bar{D}$ that is a forest of directed trees, each with a unique root. Multiple such DAGs may exist and we can pick the one that provides the tightest upper bound. Finding the optimal DAG $\bar{D}$ is a hard problem, so we propose a greedy heuristic that recursively deletes all, except one, of the incoming edges to each of the v-nodes in the DAG. See Section 4.5 in the Appendix for details of the algorithm.

Next we explore the tightness of the developed lower and upper bounds of a DAG's likelihood. Let $R(D, \bar{D}) = \lambda(\bar{D})/\tilde{\lambda}(D)$ denote the ratio between them for any DAG $\bar{D} \subset D$. It is clear that $R(D, \bar{D}) \geq 1$ for all $\bar{D} \subset D$, so we express our tightness guarantee by deriving a parametric upper bound for $R(D, \bar{D})$. For that, let ℓ denote the size of the largest reachability set in DAG D, i.e., $\ell = \max_{a \in \mathcal{N}} |\Psi_D(a)|$, and let p denote the number of nodes with v-nodes in their reachability sets, i.e., $p = |\{a \in \mathcal{N} \colon \exists \, \mathsf{v}\text{-node } b \in \Psi_D(a)\}|$. Further, let $\Delta := \max_{a \in \mathcal{N}} \max_{b \in \Psi_D(a) \setminus \{a\}} v_b / v_a$ be the maximum ratio between the weights of nodes within the same directed path in the DAG. We can derive the following guarantee:

Proposition 1.3.3. *Consider DAGs D and $\bar{D}$ such that $\bar{D} \subset D$ is obtained by deleting all, except one, of the incoming edges into each of the v-nodes. Then, we have that*

$$0 \leq \log R(D, \bar{D}) \leq p \cdot \log(1 + \ell \cdot \Delta).$$

Then, if $\Delta \in o(n^{-2})$, where n is the number of nodes, then we have that $\lim_{n \to \infty} \tilde{\lambda}(D) = \lambda(D)$.

The proof is given in Section 4.2.1 in the Appendix. Note that the bound above applies to *any* DAG $\bar{D}$ that satisfies the conditions stated in the proposition; in particular, it applies to the DAG $\bar{D}$ that is constructed using the heuristic described in Section 4.5 in the Appendix. The above result shows that our approximation guarantee depends on the number of v-nodes in the DAG (more precisely, on the number of nodes with v-nodes in their reachability sets) and the ratio Δ of PL parameters. The bound is derived in the most general setting, and in this generality, it is tight. For instance, when there are no v-nodes, then $p = 0$ and we obtain the guarantee $R(D, \bar{D}) = 1$, as expected.[3] In several other cases, however, the bound can be weak. In fact, we show on the actual sales data that the approximation ratio $R(D, \bar{D})$ can be much smaller than what is suggested by our theoretical bound.

1.3.2 Tractable analytical bounds for the purchase probability prediction

We now turn to the prediction problem, that of predicting the probability that a customer with DAG D purchases product a_j from offer set S. Recall that a customer with DAG D always samples a preference list σ that is consistent with D, i.e., $\sigma \in S_D$. The probability that such a customer will purchase product a_j from offer set S is then equal to the conditional probability that the sampled permutation is consistent with the star DAG $C(a_j, S)$, in which there are edges only from a_j to all the products in $S \setminus \{a_j\}$. Then, the probability $f(a_j, S, D)$ that the customer will purchase a_j from S is given by

$$f(a_j, S, D) = \Pr\big(S_{C(a_j,S)} | S_D\big) = \frac{\Pr\big(S_{C(a_j,S)} \cap S_D\big)}{\Pr(S_D)},$$

where the second equality follows from the Bayes rule. Now, given any offer set S, let $h_D(S) \subset S$ denote the subset of "heads" (i.e., the subset of nodes without parents) in the subgraph of the transitive closure of D restricted to set S. Since it follows by definition that every node in $S \setminus h_D(S)$ has at least one incoming edge from a node in $h_D(S)$, the customer with DAG D will never purchase the products in $S \setminus h_D(S)$. Therefore, we obtain that $f(a_j, S, D) = 0$ for all $a_j \in S \setminus h_D(S)$. For the products in $h_D(S)$, the probability of choosing a_j from S depends

[3] Note that in the absence of v-nodes, then any connected component must have a single root.

on the probability of the DAG representing the collection of permutations $S_{C(a_j,S)} \cap S_D$, which corresponds to the merged DAG $D \uplus C(a_j, S)$ obtained by taking the union of the graphs D and $C(a_j, S)$. We thus obtain

$$f(a_j, S, D) = \begin{cases} \frac{\lambda(D \uplus C(a_j,S))}{\lambda(D)}, & \text{if } a_j \in h_D(S), \\ 0, & \text{otherwise,} \end{cases}$$

In computing the choice probabilities for the products in $h_D(S)$, we run into similar #P-hardness issues as mentioned above. To deal with this challenge, [52] focus on the special case when D is a forest of directed trees, each with a unique root, and all the nodes in $h_D(S)$ are roots in D. With these assumptions, [52, Proposition 3.3] shows that

$$f(a_j, S, D) = \frac{\tilde{\lambda}(D \uplus C(a_j, S))}{\tilde{\lambda}(D)} = \frac{v_{\Psi_D(a_j)}}{\sum_{a_\ell \in h_D(S)} v_{\Psi(a_\ell)}},$$

where we define $v_{\Psi_D(a)} = \sum_{b \in \Psi_D(a)} v_b$. More generally, they propose to use the above expression as an approximation, but do not provide any performance guarantee. We can now use the results of Propositions 1.3.1 and 1.3.2 to obtain bounds for the choice probability prediction. For that, we define

$$\underline{f}(a_j, S, D) := \frac{\tilde{\lambda}(D \uplus C(a_j, S))}{\lambda(\overline{D})} \text{ and } \overline{f}(a_j, S, D) = \frac{\lambda(\overline{D \uplus C(a_j, S)})}{\tilde{\lambda}(D)}, \tag{1.4}$$

where for any DAG D, we let $\overline{D}$ denote the DAG with the properties described in Proposition 1.3.3. We can now establish the following:

Corollary 1.3.1. *For a given DAG D, under the Plackett-Luce model, the following tractable bounds of purchase probabilities apply:*

$$\underline{f}(a_j, S, D) \le f(a_j, S, D) \le \overline{f}(a_j, S, D), \text{ and}$$
$$\underline{f}(a_j, S, D) \le \hat{f}(a_j, S, D) \le \overline{f}(a_j, S, D),$$

where

$$\hat{f}(a_j, S, D) = \frac{\tilde{\lambda}(D \uplus C(a_j, S))}{\tilde{\lambda}(D)} = \frac{v_{\Psi_D(a_j)}}{\sum_{a_\ell \in h_D(S)} v_{\Psi_D(a_\ell)}}. \tag{1.5}$$

This corollary follows immediately from our definitions and the results of Propositions 1.3.1

and 1.3.2. We are also able to provide a parametric approximation guarantee similar to the one in Proposition 1.3.3. We define the parameters ℓ and p as above, but now for the merged DAG $D \uplus C(a_j, S)$. That is, $\ell = \max_{a \in \mathcal{N}} \left|\Psi_{D \uplus C(a_j,S)}\right|$ and $p = \left|\{a \in \mathcal{N} \colon \exists \, \textsf{v-node} \ b \in \Psi_{D \uplus C(a_j,S)}(a)\}\right|$. We also define $\Delta = \max_{a \in \mathcal{N}} \max_{b \in \Psi_{D \uplus C(a_j,S)} \setminus \{a\}} v_b / v_a$. We can then establish the following result:

Proposition 1.3.4. *Given DAG D, offer set S, and product $a_j \in h_D(S)$, we have that*

$$0 \leq \log \frac{\overline{f}(a_j, S, D)}{\underline{f}(a_j, S, D)} \leq 2p \cdot \log(1 + \ell \cdot \Delta).$$

Further, if $\Delta \in o(n^{-2})$, where n is the number of nodes, then we have that $\lim_{n \to \infty} \hat{f}(a_j, S, D) = f(a_j, S, D)$.

The tightness of the bound above again follows from the case when $p = 0$. For other cases, the approximation ratio can be much better than that suggested by the bound above, as demonstrated on real-world data in Section 4.5 in the Appendix.

1.4 Empirical study

We now test our proposal on the IRI Academic dataset [12], which consists of real-world purchase transactions from grocery and drug stores. We compare the predictive power of our method against standard benchmarks, such as the latent-class MNL (LC-MNL) and the random parameters logit (RPL) models. We show that our method significantly outperforms the benchmarks on holdout data on standard performance metrics for measuring predictive accuracy. In the next section, we show how our method can be used to customize product promotions.

1.4.1 Data analysis

We analyze consumer packaged goods (CPG) purchase transaction data for year 2007 over a chain of grocery stores in two large Behavior Scan markets in USA. For every purchase instance in the data set, we have the week and the store id of the purchase, the universal product code (UPC) of the purchased item, the panel ID of the purchasing customer, quantity purchased, price paid, and an indicator of whether the purchased item is on promotion or not. Overall we considered 27 categories (see Table 1.1) of products out of the available 31 categories, skipping four because of data sparsity.

The data consists of 1.2M records of weekly purchase transactions from 84K customers over 52 weeks.[4] The transaction data is split into the training set, consisting of the first 26 weeks of purchase observations, and the test set, consisting of the last 26 weeks. We considered only customers with two or more transactions over the training period. After filtering out customers with less than two observations over the training data within each category, we were left with a total of 64K customers and 1.1M purchase transactions. To alleviate data sparsity, we aggregated all the items with the same vendor code (comprising digits 3 through 7 in 13-digit-long UPC code) into a unique "product". For each transaction, we know the purchased product, say a_j, but we do not have explicit knowledge of the offer set. As a result, we approximately constructed the offer set S by taking the union of all the products that were purchased in the same category as a_j, in the same week, and in the same store. The transaction also contains a promotion indicator, which is set to 1 if product a_j was on display or price promotion at the time of purchase. Using this information, we also approximately constructed the set of promoted products consisting of all the products in S that were on promotion at least once during the week.

Using the purchase transactions within training data, we constructed a DAG for each of the customers according to our model, which we label partial order MNL (PO-MNL) Promotion model from here onwards. The construction resulted in the set woCyc of customers with preference graphs *without* cycles and set Cyc of customers with preference graphs *with* cycles. For the customers in Cyc, we decycled their preference graphs using MILP (1.2) formulated in Section 1.2.3 to obtain DAGs. We implemented this decycling procedure in Python (version 2.7.2) using Gurobi (version 7.0) as the optimization engine, and ran it on a 3.0Ghz processor with 16GB of RAM. We set the time limit to 30 seconds. The mean running time was 8 seconds, with most instances solved to optimality.

On average, we deleted only 3.4% of edges during this process. As shown in Table 1.1, the preference graphs of 40% of customers had no cycles. Customers with cycles in their preference graphs generally had more purchases and, hence, denser graphs than customers without cycles,

[4] The number of unique customers/panelists across the 27 product categories is far less than 84K. But we analyze categories separately, so we treat each "customer-category" combination as a separate customer.

for all categories of products. In particular, the DAGs of customers in Cyc, on average, had 61% more edges and 74% larger height (defined as the length of the longest directed path in the graph after decycling) than the customers in woCyc. On average, each category had 38 vendors and about 45.5% of vendors were offered in each store and week combination.

Category			**Individuals**			**PO-MNL Promotion Model**					
Expanded name	**Vend**	**AvOS**	**Total**	**$\geq$2 sales**	**AvTr**	$\lvert$woCyc$\rvert$	**% Del**	**Dens$_1$**	**Dens$_2$**	**H$_1$**	**H$_2$**
Beer	67	43.87	1796	1154	7.11	466	1.10	208.11	319.51	2.35	4.20
Carbonated beverages	46	15.36	4677	4387	17.63	666	3.38	92.45	148.11	2.22	4.43
Cigarettes	13	7.14	452	307	10.39	233	2.95	35.37	47.53	2.45	3.53
Coffee	59	19.80	3101	2255	5.59	1098	1.93	130.37	186.81	2.85	4.66
Cold cereal	39	17.66	4438	3998	10.94	687	3.69	88.41	158.14	2.47	4.81
Deodorant	32	14.55	1345	653	3.47	347	2.63	69.56	98.48	2.67	4.06
Facial tissue	10	4.17	2967	2063	4.96	1148	7.64	22.48	28.36	2.52	3.80
Frozen dinners/Entrees	77	33.14	3707	3288	13.46	942	2.45	179.24	349.34	2.65	5.51
Frozen pizza	38	15.50	3460	2946	7.83	1213	3.38	84.41	132.93	2.62	4.68
Household cleaners	68	31.42	2725	1699	4.14	388	1.58	170.89	269.30	2.94	4.72
Hot dogs	41	16.81	3318	2187	3.82	1190	1.89	89.95	127.86	2.84	4.37
Laundry detergent	18	10.08	3196	2181	4.04	1367	3.62	49.28	69.71	2.60	4.19
Margarine/Butter	16	10.35	3474	2750	5.65	1405	4.38	47.00	71.07	2.85	4.45
Mayonnaise	14	6.86	3761	2386	3.28	1853	3.41	35.10	44.24	2.46	3.59
Milk	33	11.69	4851	4652	14.90	1674	3.12	78.84	118.78	2.88	4.70
Mustard	52	17.07	3728	2515	3.66	895	1.83	107.38	142.27	2.77	4.06
Paper towels	11	6.94	3072	2051	5.20	977	7.20	29.82	42.74	2.79	4.44
Peanut butter	19	7.99	3153	1923	3.89	1232	3.13	43.38	56.88	2.55	3.70
Salt snacks	95	26.79	4727	4446	15.09	629	2.58	179.08	320.94	2.38	5.26
Shampoo	41	18.74	1466	738	3.73	357	1.67	127.54	172.51	2.62	4.16
Soup	90	32.87	4636	4322	12.02	988	1.71	207.27	353.24	2.67	5.07
Spaghetti/Italian sauce	52	17.85	3473	2698	5.46	1363	1.98	120.92	178.83	2.76	4.62
Sugar substitutes	10	5.05	750	308	3.30	258	3.76	28.45	33.74	2.41	3.62
Toilet tissue	11	7.66	3760	2817	5.10	1552	6.86	32.25	47.75	2.71	4.52
Toothbrushes	36	15.86	1115	499	3.06	260	1.96	90.46	124.48	2.70	4.15
Toothpaste	25	12.05	2110	1186	3.58	708	2.01	64.15	86.54	2.46	3.72
Yogurt	26	9.84	3766	3491	19.81	1349	5.41	55.55	86.79	2.57	4.79

Table 1.1: Summary of the data.

1.4.2 Models compared

We fitted our PO-MNL Promotion model to the data described above and compared its predictive performance against two widely used benchmarks: the LC-MNL and the RPL models. Both these models belong to the general class of random utility models (RUMs), so that in each purchase instance, a customer samples product utilities and then chooses the available product with the highest value.

LC PO-MNL Promotion model.

First, we fitted a single class PO-MNL Promotion model to the DAGs. Recall that to deal with promoted products we expanded our product universe to consist of two copies, a promoted one and a non-promoted one, of each product. Following the notation introduced in Section 1.2.1, products $a_1, a_2, \ldots, a_n$ are the non-promoted copies and $a_{n+1}, a_{n+2}, \ldots, a_{2n}$ are the promoted copies. For any $j \in [n]$, product a_{j+n} is the promoted copy corresponding to product a_j. We let τ_j denote the MNL parameter of product a_j, so that $v_j = \exp(\tau_j)$. We parameterize the model as follows: for any $a_j \in \mathcal{N}'$:

$$\tau_j = \begin{cases} \beta_j^0, & \text{if } 1 \leq j \leq n \\ \beta_{j-n}^0 + \beta_{j-n}, & \text{if } n+1 \leq j \leq 2n, \end{cases}$$

where β_j^0 is the utility derived from the non-promoted copy of product $j \in [n]$, and β_j is the additional utility from the promotion feature. We estimate the parameters by solving the following approximated regularized likelihood problem:

$$\max_{\boldsymbol{\beta},\boldsymbol{\beta}_0} \sum_{i=1}^{m} \sum_{j=1}^{2n} \left[\tau_j - \log\left(\sum_{a_\ell \in \Psi_{D_i}(a_j)} \exp(\tau_l) \right) \right] - \alpha(\left\|\boldsymbol{\beta}^0\right\|_1 + \|\boldsymbol{\beta}\|_1), \tag{1.6}$$

where $\Psi_{D_i}(a_j)$ is the set of nodes that are reachable from a_j in DAG D_i of customer i. To arrive at the above approximation, we used the lower bound $\tilde{\lambda}$ for computing the likelihood of a DAG, as discussed in Section 1.3. When the value of α is fixed, it can be shown that the optimization problem in (1.6) is globally concave and therefore can be solved efficiently [92]. We tuned the value of α by 5-fold cross-validation. The above likelihood problem is exact only if every DAG D_i is a forest of directed trees, each with a unique root. Otherwise, as shown in Proposition 1.3.1, it provides a lower bound. Once we estimated the parameters, we predicted purchase probabilities on holdout data using the following approximation:

$$\tilde{f}(a_j, S, D) = \begin{cases} \frac{v_{\Psi_D(a_j)}}{\sum_{a_\ell \in h_D(S)} v_{\Psi_D(a_\ell)}}, & \text{if } a_j \in h_D(S), \\ 0, & \text{otherwise.} \end{cases} \tag{1.7}$$

In Section 4.5 in the Appendix we provide empirical evidence that this approximation is a good and easy-to-compute proxy for the exact $f(a_j, S, D)$.

To account for heterogeneity among the customers, we also fitted a K latent class PO-MNL Promotion model, which assumes that each customer belongs to one of the $h \in \{1, .., K\}$ latent classes. A customer from class h samples her DAGs according to the PO-MNL Promotion model with parameters τ_{jh}, defined as

$$\tau_{jh} = \begin{cases} \beta_{jh}, & \text{if } 0 \leq j \leq n \\ \beta^0_{j-n,h} + \beta_{j-n,h}, & \text{if } n+1 \leq j \leq 2n. \end{cases}$$

We let the prior probability that a customer belongs to class h by $\gamma_h \geq 0$, so that $\sum_{h=1}^{K} \gamma_h = 1$. Then, similar to the PO-MNL model, we estimate the parameters by solving the following approximated regularized likelihood problem:

$$\max_{\boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}} \sum_{i=1}^{m} \log \left[\sum_{h=1}^{K} \gamma_h \prod_{j=1}^{2n} \frac{\exp(\tau_{jh})}{\sum_{a_\ell \in \Psi_{D_i}(a_j)} \exp(\tau_{\ell h})} \right] - \alpha \sum_{h=1}^{K} (\left\|\boldsymbol{\beta}^0_h\right\|_1 + \|\boldsymbol{\beta}_h\|_1),$$

The above optimization problem is nonconcave for $K > 1$, even with the value of α fixed. Therefore, we use the standard expectation-maximization (EM) based algorithm described in [92] to obtain a stationary point.[5] Specifically, we initialize the EM with a random allocation of customers to one of the K classes, resulting in an initial allocation $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$, which form a partition of the collection of all the customers. Then we set $\gamma_h^{(0)} = |\mathcal{D}_h| / \big(\sum_{d=1}^{K} |\mathcal{D}_d| \big)$. In order to get a parameter vector $\boldsymbol{\tau}_h^{(0)}$, we fit a PO-MNL Promotion model, described above, to each subset of customers. After calibrating the model, we make predictions in the following way. For each individual i with DAG D_i, we estimate the posterior membership probabilities $\hat{\gamma}_{ih}$ for each class $h \in [1, .., K]$:

$$\hat{\gamma}_{ih} = \frac{\gamma_h \prod_{j=1}^{2n} \left[v_{jh} / \sum_{a_\ell \in \Psi_{D_i}(a_j)} v_{\ell h} \right]}{\sum_{d=1}^{K} \gamma_d \prod_{j=1}^{2n} \left[v_{jd} / \sum_{a_\ell \in \Psi_{D_i}(a_j)} v_{\ell d} \right]},$$

[5] See the details in Appendix A2.1.2 in [52].

where $v_{jh} = \exp(\tau_{jh})$, and then make the prediction:

$$\tilde{f}(a_j, S, D_i) = \sum_{h=1}^{K} \hat{\gamma_{ih}} \tilde{f}_h(a_j, S, D),$$

where $\tilde{f}_h(a_j, S, D)$ are the approximated probabilities in the equation (1.7). We estimated the model for $K = 1, 2, \ldots, 10$, and report the best performance measure from these 10 variants, for every performance metric that we introduce below.

Benchmark models

We compare our models with two benchmark models succinctly described here (see Section 4.4 in the Appendix for details). The first benchmark is the LC-MNL choice model with K latent classes. In this model, each customer belongs to one unobservable class, and customers from class $h \in \{1, ..., K\}$ make purchases according to the MNL model associated with that class. The model is described by the parameters of the MNL characterizing each class and by the prior probabilities of customers belonging to each of the classes. Once the model parameters are estimated, we make customer-level predictions by averaging the predictions from K single-class models, weighted by the posterior probability of class-membership. Similarly to the LC PO-MNL Promotion model, we estimated the model for $K = 1, 2, \ldots, 10$, and report the best performance measure from these 10 variants, for every performance metric that we introduce in the upcoming subsection.

The second benchmark model, which also captures heterogeneity in customer preferences, is the RPL model, which assumes that in each purchase instance, a customer samples the $\boldsymbol{\beta}$ parameters of the product utilities according to some distribution and then makes the choice according to a single-class MNL model with parameter vector $\boldsymbol{\beta}$. In comparison with LC-MNL benchmark, RPL model allows the parameter vectors $\boldsymbol{\beta}$ to take a continuum of values. Particularly, we assume that parameter vector $\boldsymbol{\beta}$ is sampled according to multivariate normal distribution with mean $\boldsymbol{\mu}$ and diagonal variance-covariance matrix Σ, i.e., $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \Sigma)$. Calibration of the RPL choice model is based on the sample average approximation approach, which is computationally intensive.

In both benchmark models, we account for product promotion status by introducing a product-specific parameter in order to capture the additional utility from this feature.

1.4.3 Prediction performance measures

Broadly, we want to predict the product purchased by customer i in period time $t+1$ given the set of offered as well as promoted items at time $t+1$. For that, we compare the models based on a one-step-ahead prediction experiment for every category under two different metrics: χ^2 and miss-rate. Recall that in our case, a *period* corresponds to a week. For each category of products, we separately fit the benchmark models and the PO-MNL Promotion and LC PO-MNL Promotion models to the following three subsets of individuals: customers without cycles in their preference graph, customers with cycles in their preference graph, and the combination of all the customers. Then, for each category and each subset of customers, we report the comparisons of the different fitted models.

The "chi-square" score is computed as follows:

$$\chi^2 \text{ score} = \frac{1}{|\mathcal{N}||U|} \sum_{i \in U, a_j \in \mathcal{N}} \frac{(n_{ij} - \hat{n}_{ij})^2}{0.5 + \hat{n}_{ij}}, \text{ where } \hat{n}_{ij} = \sum_{t=1}^{T_i} f_i(j_{it}, t),$$

where U is the set of all individuals, and n_{ij} is the observed number of times individual i purchased product j during the time horizon of length T_i. The indicator function $f_i(j_{it}, t)$ takes value 1 if the product indexed j has the highest choice probability for individual i at time t, and 0 otherwise. This score measures the ability of the model combinations to predict the aggregate market shares of the products purchased by every individual, where lower scores indicate better prediction accuracy. The 0.5, added in the denominator, allows to deal with undefined instances.

The miss-rate is computed as follows:

$$\text{miss rate} = \frac{1}{|U|} \sum_{i \in U} \frac{1}{|T_i|} \sum_{t=1}^{T_i} \mathbb{I}[f_i(j_{it}, t) = 0],$$

where $\mathbb{I}[A]$ is the indicator function that takes value 1 if A is true and 0 otherwise, and j_{it} is the index of the item purchased at time t by individual i. Miss rate is a more stringent predictive measure than "chi-square" score, because it rewards or penalizes a method on every individual transaction assessment, as opposed to the long-term aggregate prediction of the chi-square score. Both scores are designed to reflect the types of prediction problems that would be relevant in practice.

1.4.4 Brand choice prediction results

Figure 1-3 presents scatterplots of the "chi-square" scores of LC-MNL and RPL versus "chi-square" scores of PO-MNL Promotion (single class) and LC PO-MNL Promotion (multi-class), across the 27 product categories, for three subsets of customers. We conclude that the PO-MNL Promotion models outperforms both LC-MNL and RPL benchmarks to a big extent (i.e., most of the points lie above the 45-degree line). Note that both benchmark models also account for the promotion status of the products. First, consider the left two panels in Figure 1-3. Here, we calibrate the models on the subset of individuals who do not have cycles in their preference graph (i.e., up to Phase 2 in the DAG construction process). The "chi-square" score of PO-MNL Promotion model exhibits an average improvement of 10.25% over LC-MNL and 4.55% over RPL. This improvement in prediction performance can be explained by the effectiveness of the DAGs in capturing partial preferences of the customers. Table 1.3 reports the distribution of the number of unique brands purchased by customers across the training data. On average customers purchase no more than 4 unique brands in the training data, indicating that customers have strong preferences and their purchases do not change very much from week to week. This brand loyal behavior of customers also explains the significant gains in performance that our method obtained over the benchmark methods. These improvements are significant, especially considering the fact that both benchmark models have more parameters to estimate and require around 300× more time than it takes to estimate the PO-MNL Promotion model. The key attribute of the PO-MNL Promotion model making it superior to the benchmarks is that it accounts for heterogeneous customer preferences through their partial orders, so that it makes more efficient use of the limited purchase transaction data. Using the LC PO-MNL Promotion model, we can further boost performance, resulting in average improvement of 14.72% over LC-MNL and 8.98% over RPL.

Second, consider the middle column in Figure 1-3, where we calibrated the models on the subset of individuals that have cycles in their preference graph. The formation of cycles in the

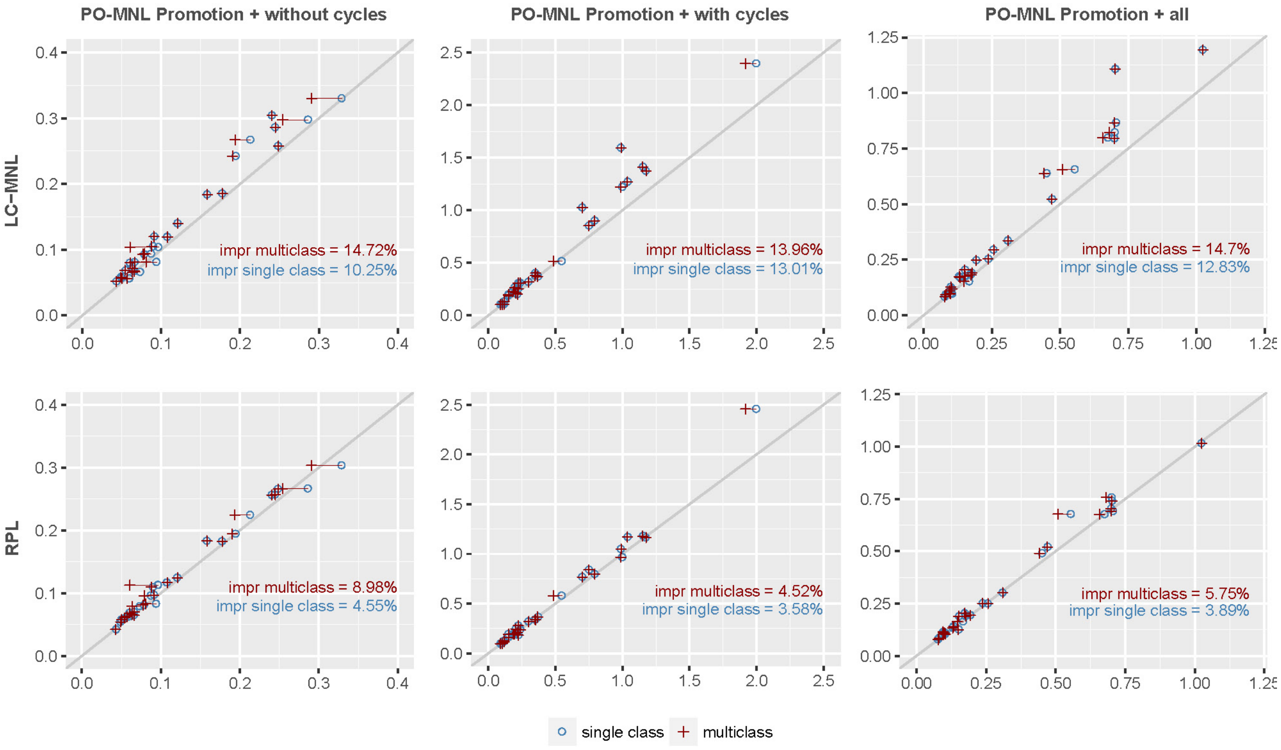

Figure 1-3: Scatter plot of the average χ^2 scores.

preference graph is symptomatic of a less consistent choice behavior of the customers in the first place. In fact, by checking the y-scale we can observe that the performance of all the models deteriorate in this case as compared to their performance in the left panels. Yet, we see that PO-MNL Promotion model exhibits an average improvement of 13.01% over LC-MNL and 3.58% over RPL, whereas LC PO-MNL Promotion model, capturing heterogeneity of customers to a greater extent, has an average improvement of 13.96% over LC-MNL and 4.52% over RPL.

Third, consider the right panels in Figure 1-3, where we use the previous separate calibrations but report the joint prediction over all the individuals for each category of products. The performance here is a weighted average between the two types of customers: with and without cycles in the preference graphs, achieving significant improvements overall: PO-MNL Promotion model exhibits an average improvement of 12.83% over LC-MNL and 3.89% over RPL, while LC PO-MNL Promotion model shows an average improvement of 14.7% over LC-MNL and 5.75% over RPL.

Figure 1-4 presents scatterplots of the miss-rates, using a display format similar to that of

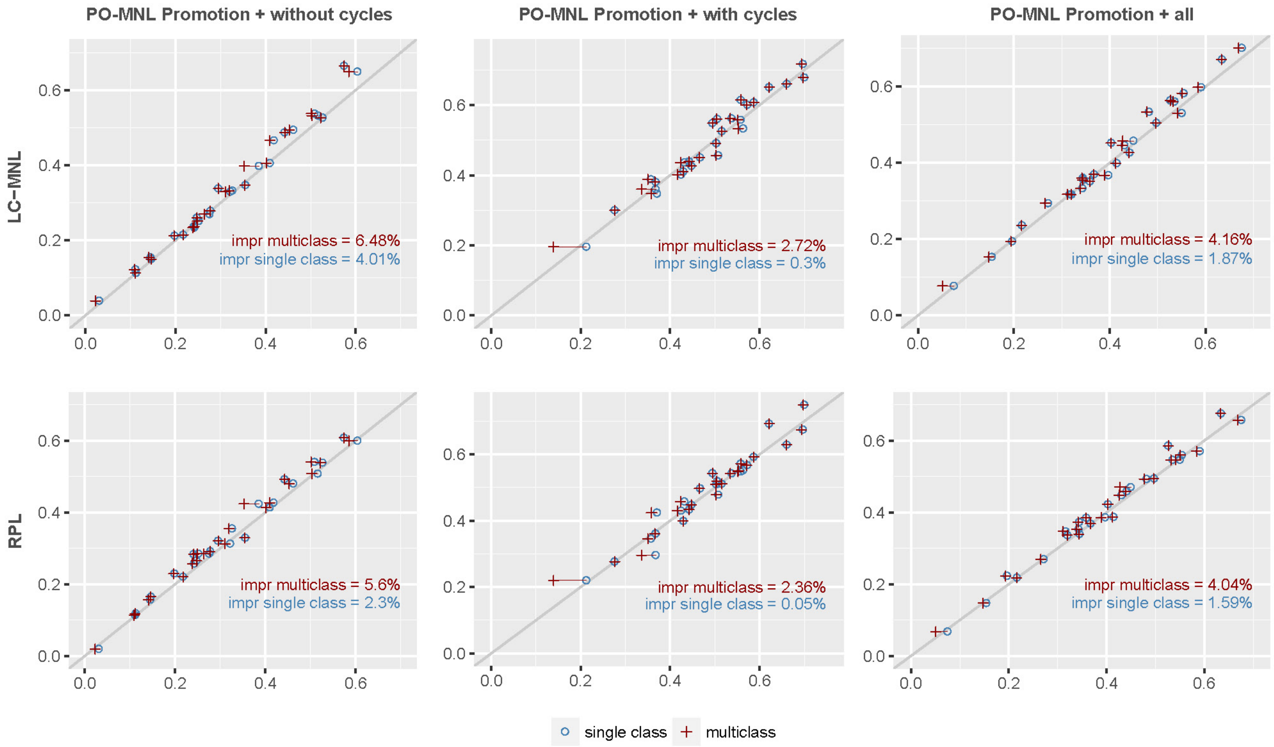

Figure 1-4: Scatter plot of the average miss rate.

Figure 1-3. From it, we observe that our model combinations obtain improvements of between 0.05% and 4.01% under PO-MNL Promotion, and further improvements of between 2.36% and 6.48% under LC PO-MNL Promotion over the benchmarks in all six panels. Even though these numbers appear to be low, we emphasize here that this metric is a very stringent one and therefore it is expected that our PO-MNL Promotion models obtain moderate (but still significant) improvements over state-of-the-art alternatives.

We make the following observations from the results. First, recall that the decycling in Phase 3 of the DAG construction process allows us to calibrate the PO-MNL Promotion model also for the subset of individuals that have cycles in their preference graph. As a result, it further boosts the improvement of PO-based models over the classical benchmarks by increasing the coverage of individuals to the maximum level of 100%; in other words, we can calibrate PO-MNL Promotion and make predictions for both subsets of the customers, those with and without cycles in their preference graph. Second, from all the panels it can be concluded that the RPL model outperforms the LC MNL model on average across 27 categories of products.

Third, for all the panels we have that LC PO-MNL model boosts the performance of PO-MNL model by accounting for additional heterogeneity of customers. Fourth, we observe that PO-MNL Promotion model (or LC PO-MNL model) outperforms in most of the categories both LC MNL and RPL benchmarks, which incorporate the same information on promotions. Therefore, this model can be used to measure customer response to product promotions even when we have very few observations for each customer by capturing partial preferences of the customers by DAGs.

In Section 1.5, we perform several robustness checks with respect to some of the assumptions that we made here, including (i) the way we aggregate data from customers to estimate the benchmark models, (ii) accounting explicitly for the no-purchase option, and (iii) the way we split data between training and holdout samples. The key insights remain the same. We also tested (iv) the impact of not adding the implicit candidate edges in Phase 2 of the DAG construction process, and noticed a poorer performance of around 1.85% on average with respect to both χ^2 and miss rates compared to including them. Finally, in Section 1.5.5 we report comparative statistics on the predictive performance of the behavioral models studied by [52]. We find that for categories with high loyalty index, and within them, for customers having non-empty behavioral DAGs, practitioners may prefer to use the PO-MNL Inertial and Censored models. Other than these (category, individual) combinations, the use of the PO-MNL Promotion model proposed in this chapter leads to more accurate predictions. Yet, in this chapter, we apply the PO-MNL Promotion model to run customized promotions for all categories and individuals because it relies on a completely data-driven approach to build the DAGs, and these DAG structures with both promoted and non-promoted nodes serve as basis to design and run the personalized promotions discussed in the next section.

1.5 Robustness check for the prediction results

In this section we summarize the major empirical experiments conducted in order to check the robustness of the prediction results reported in Section 1.4.4. We start checking the robustness with respect to different data aggregation strategies, followed by the effect of accounting for no-purchase observations, the effect of different cutoff points between training and holdout sample data, and the effect of adding implicit candidate edges in Phase 2 of the DAG construction process.

1.5.1 Robustness with respect to different data aggregation strategies

We start by demonstrating the robustness of the results to changes in how we calibrate the benchmarks. In particular, in Section 1.4.4, we calibrated all the models separately on (i) customers who do not have cycles in their preference graph under the PO-MNL Promotion, and (ii) customers who have cycles in their preference graph under the PO-MNL Promotion model. Then we used previous separate calibrations but presented the joint prediction over all the individuals for each category of products, i.e., the weighted average prediction performance between both types of customers: with and without cycles in their preference graph. Here we show the prediction performance of our model versus the benchmarks using a display format similar to that in Section 1.4.4, but when calibrating both benchmarks on the set of all individuals. Note that there is a tension about the benchmarks here since one side they are estimated on a larger volume of data, but at the same time this extra volume comes at the expense of higher customer heterogeneity (i.e., individuals without and with cycles pool together for the estimation process).

Similarly to Figure 1-3, Figure 1-5 presents scatterplots of the "chi-square" scores of LC-MNL and RPL versus "chi-square" scores of PO-MNL Promotion (single class) and LC PO-MNL Promotion (multi-class), across the 27 product categories. Note that in all the panels we calibrate the benchmarks on the set of all individuals and then separately make predictions for three subsets of customers: without and with cycles, and the entire population.

First, consider the left two panels in Figure 1-5. Here, we calibrate the PO-MNL Promotion and make predictions with all the models on the subset of individuals who do not have cycles in their preference graph. The "chi-square" score of PO-MNL Promotion model exhibits an average improvement of 9.77% over LC-MNL and 5.79% over RPL. Using the LC PO-MNL Promotion model, we can further boost performance, resulting in average improvement of 14.31% over LC-MNL and 10.28% over RPL.

Second, consider the middle column in Figure 1-5, where we calibrate the PO-MNL Promotion and make predictions with all the models on the subset of individuals that have cycles in their preference graph. We see that PO-MNL Promotion model exhibits an average improvement

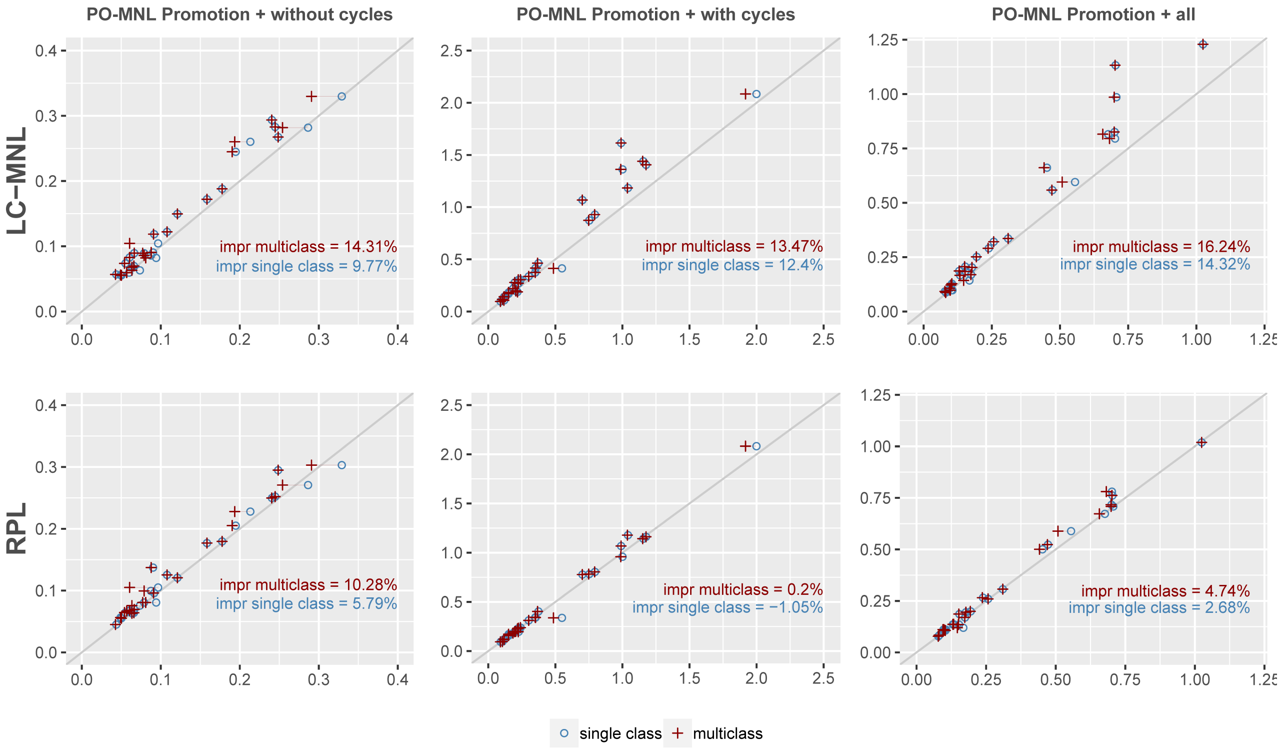

Figure 1-5: Brand choice χ^2 prediction results. In all the panels we estimate the benchmarks on the set of all individuals.

of 12.4% over LC-MNL and deterioration of 1.05% over RPL, while LC PO-MNL Promotion model leverages the performance to an average improvement of 13.47% over LC-MNL and 0.2% over RPL.

Third, consider the right panels in Figure 1-5, where we use the previous separate calibrations for all the models but report the joint prediction over all the individuals for each category of products. The performance here is a weighted average between the two types of customers: with and without cycles in the preference graphs, achieving significant improvements overall: PO-MNL Promotion model exhibits an average improvement of 14.32% over LC-MNL and 2.68% over RPL, while LC PO-MNL Promotion model shows an average improvement of 16.24% over LC-MNL and 4.74% over RPL.

Figure 1-6 presents scatterplots of the miss-rates, using a display format similar to that of Figure 1-4. From it, we observe that our model combinations obtain improvements of up to 8.92% under PO-MNL Promotion, and further improvements of up to 11.3% under LC PO-MNL

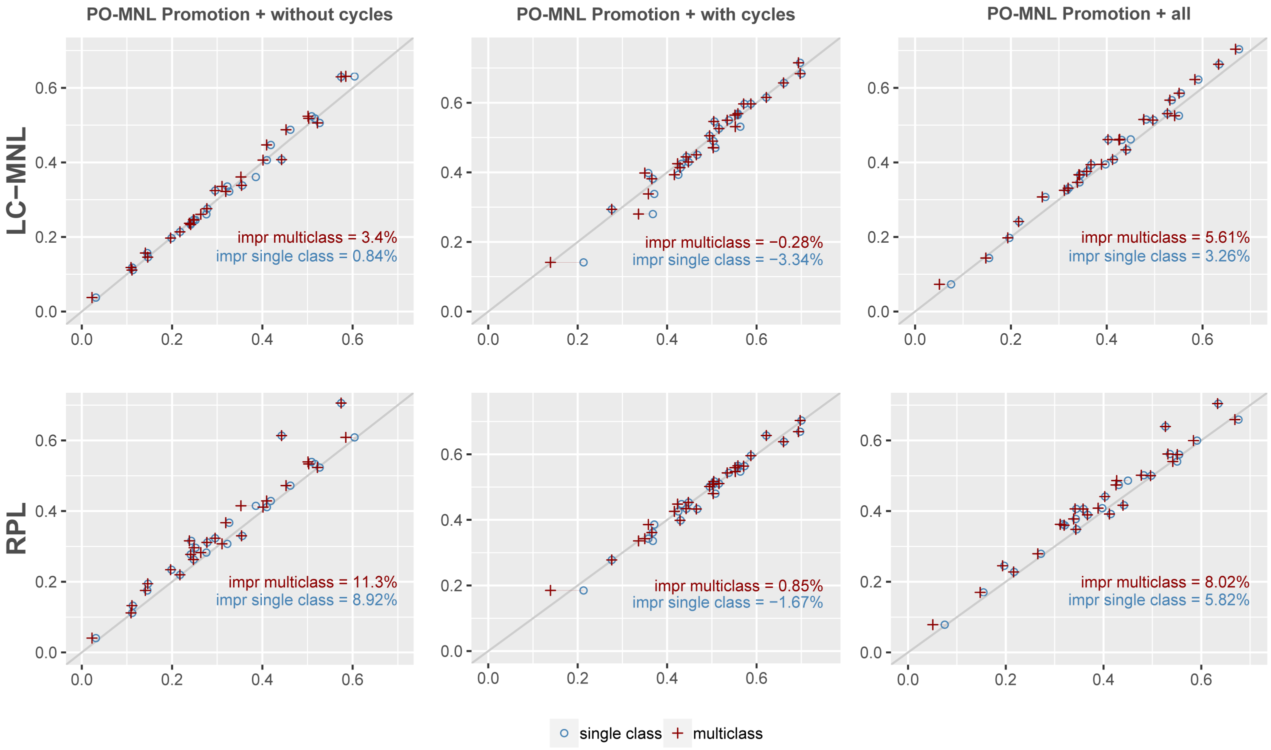

Figure 1-6: Brand choice miss rate prediction results. In all the panels we estimate the benchmarks on the set of all individuals.

Promotion over the benchmarks within the six panels.

The key observations can be summarized as follows: (i) the PO-MNL Promotion model, especially the multiclass version of it, outperforms the state-of-the-art competitive benchmarks even when the latter ones are allowed to be estimated on a larger population of customers; (ii) the best results for our model are observed on the individuals without cycles in the first place, who are the ones displaying the most consistent behavior; and (iii) LC-MNL is more competitive than RPL with respect to χ^2 scores, but is dominated by RPL in terms of miss rates. The last observation is different from what we report in Section 1.4.4, where RPL dominated LC-MNL with respect to both χ^2 and miss rates. The larger volume of data to train the models seems to favor more LC-MNL than RPL.

1.5.2 Robustness with respect to adding no-purchase observations

In order to streamline the comparison of the models in the empirical case study in Section 1.4.4 we did not include the no-purchase observations in the prediction tasks since we did not have explicit data on the no-purchase alternatives in our dataset. Here, we demonstrate that the brand choice prediction results remain qualitatively the same when we include the no-purchase option in our calibration and prediction tasks.

To this end, we approximately build the no-purchase observations from our data. In particular, we can approximately infer from the data the times the customer visits to the store to make at least one category purchase. Therefore, we can easily obtain instances when the customer visited the store but ended up not making a specific category purchase. However, these observations in the data cannot be considered as the no-purchase instances since we do not know if the customer had the intent to make a category purchase and ended up choosing the no-purchase option. In fact, the number of store visits is around ten times higher than the number of purchases for some categories. As a result, in order to minimize the number of "spurious" no-purchase observations inferred from the data, we first assume that the number of times a customer chooses the outside option is comparable to the number of times a customer makes a category purchase. In particular, we say that the number of no-purchases of every customer is equal to the number of times a customer buys her second most purchased product. It implies that the customers chose a no-purchase alternative on average αT_c times, where T_C is the total number of times a customer made a category purchase and $\alpha = 20.5\%$. We also show below that the obtained prediction results are robust to other values of α. As a result, we randomly sample the fixed portion of the no-purchase observations from the data on the store visits of the customers when they decided not to make a category purchase, and include these additional transactions into our dataset for every category. Then we use the same approach described above to calibrate the models and test their predictive performance.

Analogously to Figure 1-3, Figure 1-7 presents scatterplots of the "chi-square" scores of LC-MNL and RPL versus "chi-square" scores of PO-MNL Promotion (single class) and LC PO-MNL Promotion (multi-class), across the 27 product categories, for three subsets of customers. First, consider the left two panels. Here, we calibrate the models on the subset of individuals that do not have cycles in their preference graph. The "chi-square" score of PO-MNL Promotion model exhibits a moderate average deterioration of 2.79% over LC-MNL and an average improvement of 7.43% over RPL. Second, consider the middle column in Figure 1-7, where we calibrated the models on the subset of individuals that have cycles in their preference graph. We see that PO-MNL Promotion model exhibits an average improvement of 12.88% over LC-MNL and 4.17% over RPL.

In the left two panels in Figure 1-7, it is demonstrated that using LC PO-MNL Promotion model, we can further boost performance of the proposed methodology, resulting in average improvement of 14.64% over LC-MNL and 22.8% over RPL. Similarly, as illustrated in the middle column of Figure 1-7, LC PO-MNL Promotion model, capturing heterogeneity of customers to a greater extent, has an average improvement of 16.07% over LC-MNL and 7.43% over RPL.

Third, consider the right panels in Figure 1-7, where we use the previous separate calibrations but report the joint prediction over all the individuals for each category of products. The performance here is a weighted average between the two types of customers: with and without cycles in the preference graphs, achieving significant improvements overall: PO-MNL Promotion model exhibits an average improvement of 12.59% over LC-MNL and 3.48% over RPL, while LC PO-MNL Promotion model shows an average improvement of 16.18% over LC-MNL and 7.27% over RPL.

Analogously to Figure 1-4, Figure 1-8 presents scatterplots of the miss-rates, using a display format similar to that of Figure 1-7. From it, we observe that we obtain improvements of between 1.28% and 10.04% under LC PO-MNL Promotion over the benchmarks in all six panels.

Like in the base case in Section 1.4.4 without the no-purchase option, we observe that the PO-MNL Promotion model continues to outperform both benchmarks in most of the categories, with respect to χ^2 and miss rate, especially for its multiclass version.

Next, we showcase the robustness of the predictive results for other values of α. In particular, we test the predictive performance of the LC PO-MNL Promotion model versus the LC-MNL benchmark for $\alpha = 30\%$ and $\alpha = 40\%$ in Figures 1-9 and 1-10, respectively. Here we focus only on the LC-MNL benchmark provided its competitive performance and the heavy computational burden of RPL. We observe that the improvements of PO-MNL Promotion over the LC-MNL benchmark vary with α between 12.36% and 14.95% in terms of the χ^2 score, and between 1.24%

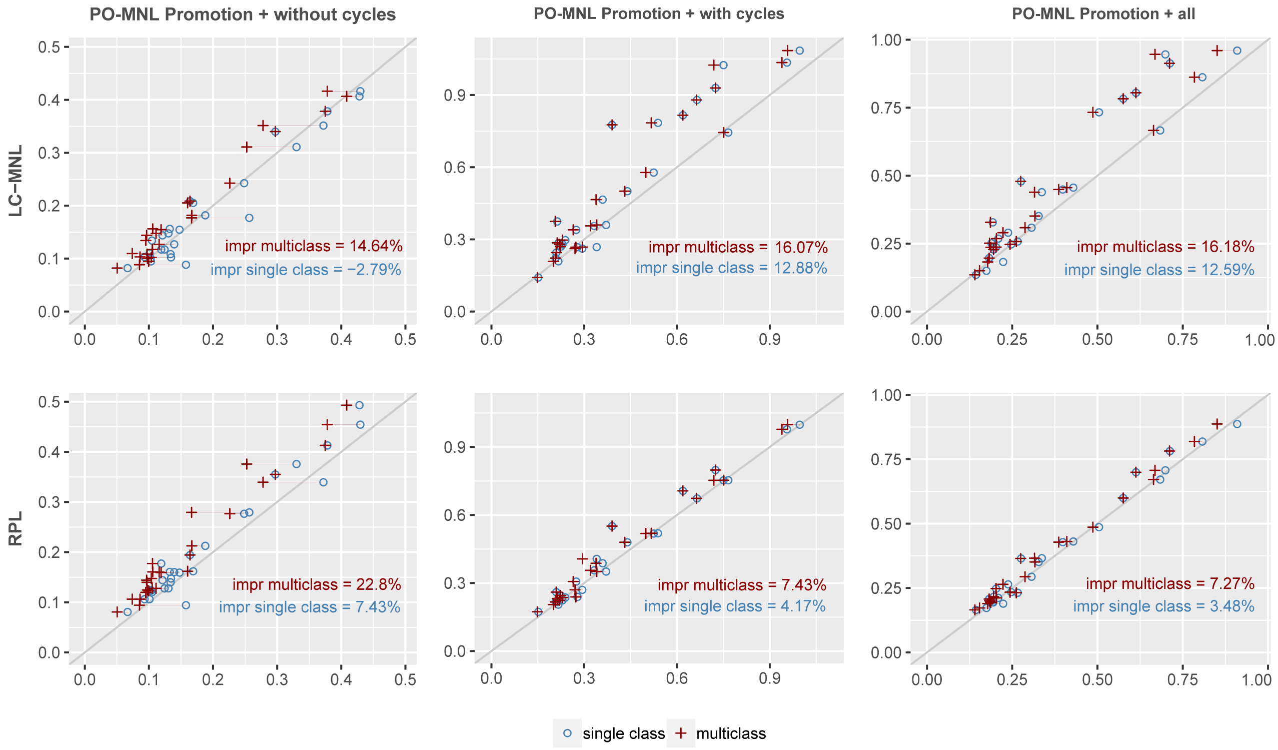

Figure 1-7: Brand choice χ^2 prediction results with no-purchase option included. We assume that the number of no-purchases of every customer is equal to the number of times a customer buys her second most purchased product.

and 1.76% in terms of the miss rate.

Figures 1-9 and 1-10 confirm that the superior performance of PO-MNL Promotion is robust to different fractions of no-purchases in the dataset.

1.5.3 Robustness with respect to the split between test and training datasets

In all our experiments so far, the training set consists of the first 26 weeks and the test set consists of the last 26 weeks of transactions. Here we show the robustness of the results to changes in how we split the data into the training and test sets. We also include the no-purchase observations (with $\alpha = 20.5\%$).

When we reduce the volume of the training set to the first 21 weeks and enlarge the test set to the last 31 weeks of transactions, Figure 1-11 shows that the LC PO-MNL Promotion model

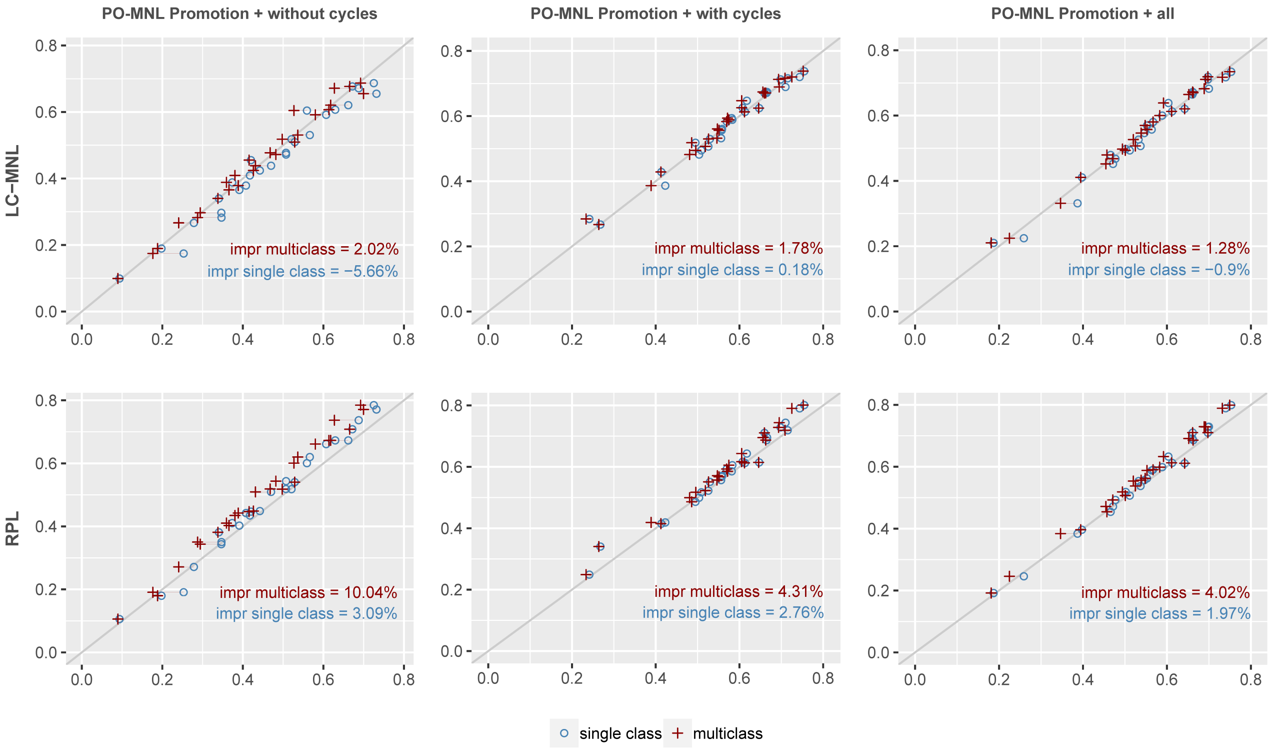

Figure 1-8: Brand choice miss rate prediction results with no-purchase option included. We assume that the number of no-purchases of every customer is equal to the number of times a customer buys her second most purchased product.

outperforms the LC-MNL benchmark by 16.14% and 1.02% based on the miss rate and χ^2 scores, respectively.

When increasing the training set to the first 31 weeks and reducing the test set to the last 21 weeks of transactions, Figure 1-12 shows that the LC PO-MNL Promotion model outperforms the LC-MNL benchmark by 14.66% and 1.12% based on the miss rate and χ^2 scores, respectively. We observe a sustained relative performance of the PO-MNL Promotion model over the LC-MNL model as we reduce or increase the training dataset.

1.5.4 Robustness with respect to the addition of implicit candidate edges in the DAGs

Recall that the Phase 2 in the DAG construction process is about the inclusion of *implicit candidate edges* in the DAG that identifies each individual. This is a heuristic step that assumes

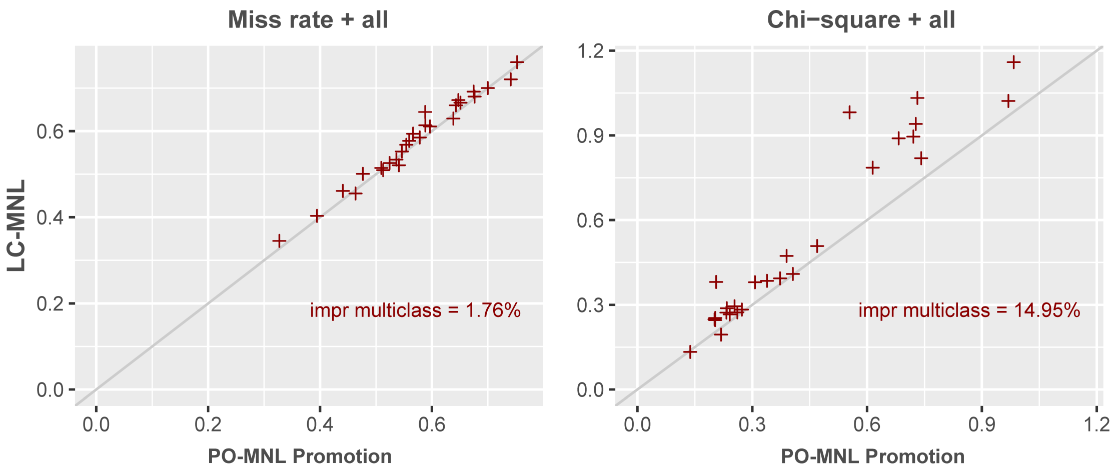

Figure 1-9: Brand choice prediction results with no-purchase option included. Every customer is assumed to choose no-purchase alternative on average 30% of times.

that the relative preference between two products is preserved regardless the promotion status of the products. That is, if from Phase 1 a full price version of a_j is preferred over the full price version of a_k, then the promoted version a_{j+n} is also preferred to the promoted version a_{k+n}. Similarly, if the Phase 1 preference is stated on the promoted versions, then the relative preference is extended to the corresponding full price versions. These Phase 2 edges have a low weight and some of them are the ones to be deleted in Phase 3 in case a cycle arises in the DAG. Of course, the heuristic could add spurious implicit candidate edges, and the final justification for the existence of the edges is their empirical performance.

To this end, Figure 1-13 illustrates that the LC PO-MNL Promotion model with implicit candidate edges outperforms the LC PO-MNL Promotion without implicit edges by 1.82% and 1.9% based on miss rate and χ^2 scores, respectively. This observation provides enough support for their inclusion in the DAG. It allows us to conclude that adding implicit edges in the DAG construction process boosts the benefit in prediction performance because of the denser DAG which outweighs the biases from adding few spurious edges along the way.

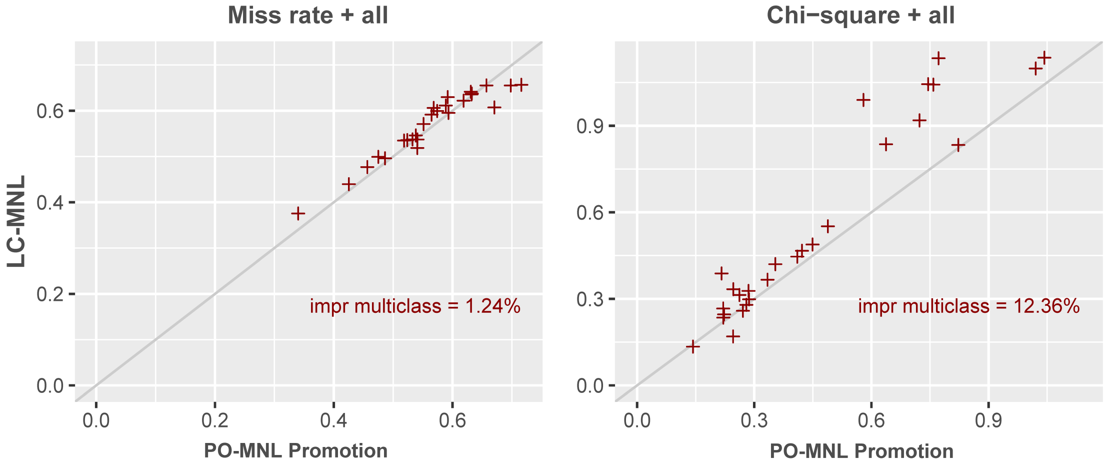

Figure 1-10: Brand choice prediction results with no-purchase option included. Every customer is assumed to choose no-purchase alternative on average 40% of times.

1.5.5 Comparison with the DAG-based behavioral models studied by [52]

In this subsection, we compare the predictive performance of the PO-MNL Promotion model with the PO-MNL Inertial and PO-MNL Censored models studied by [52]. Note that both PO-MNL Inertial and PO-MNL Censored models take into account the information about product promotions implicitly through modeling the consideration sets of customers via behavioral rules. Relying on the pre-specified behavioral assumptions this approach cannot consistently explain the purchasing behavior of all the customers. As a result, the DAGs of customers whose purchasing transactions are inconsistent with these assumptions are assumed to be empty (i.e., without edges), which reduce the representation of customer preferences to any standard random utility model such as the MNL or the LC-MNL. In particular, to run the prediction performance of PO-MNL Inertial and PO-MNL Censored for customers that have empty DAGs, we use the best of up to 10 LC-MNL [LC-MNL].

Recall that the approach taken in this chapter for the DAG construction is different, since it is completely data-driven and accounts explicitly for promotion effects. The approach could still lead to cycles in the preference graph, which are then deleted in Phase 3 such that all customers are characterized by non-empty DAGs. Figure 1-14 illustrates the scatter plot of the χ^2 scores of all 29 product categories under the LC PO-MNL Promotion vs. PO-MNL Inertial with clustering

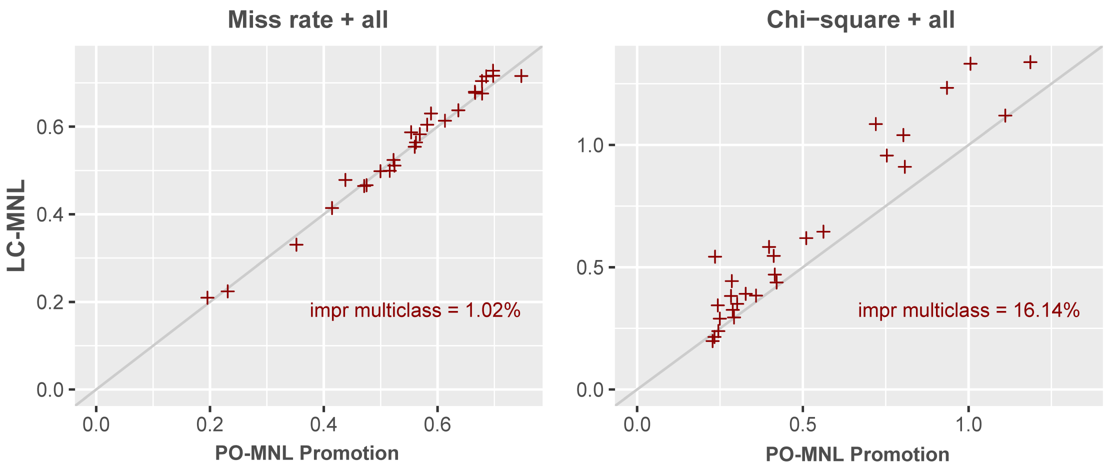

Figure 1-11: Brand choice prediction results with no-purchase option included. The training data consists of 21 weeks, and the test data consists of 31 weeks.

and PO-MNL Censored with clustering (see [52, Section 5]). In all the panels we calibrate the LC-MNL and LC PO-MNL Promotion models on the set of all individuals. First, consider the left two panels in Figure 1-14. Here, we calibrate the PO-MNL Inertial [top left panel] and PO-MNL Censored [bottom left panel] and represent the prediction performance of all the models over the subset of individuals that can be explained by behavioral assumptions. The χ^2 score of LC PO-MNL Promotion model exhibits an average deterioration of 15.52% over PO-MNL Inertial with clustering and 18.68% over PO-MNL Censored with clustering. Second, consider the right column in Figure 1-14, where we calibrate the PO-MNL Inertial [top right panel] and PO-MNL Censored [bottom right panel] and represent the prediction performance of all the models over the subset of individuals that can not be explained by behavioral assumptions. In this case, both PO-MNL Inertial and PO-MNL Censored models are reduced to the LC-MNL model. We see that LC PO-MNL Promotion model exhibits an average improvement of 11.18% over PO-MNL Inertial and 2.14% over PO-MNL Censored.

Figure 1-15 presents scatterplots of the miss rates, using a display format similar to that of Figure 1-14. The insights are the same as in Figure 1-14. In the left column, we calibrate the PO-MNL Inertial [top left panel] and PO-MNL Censored [bottom left panel] models and represent the prediction performance of all the models over the subset of individuals who can be explained by behavioral assumptions. We observe that LC PO-MNL Promotion obtains an

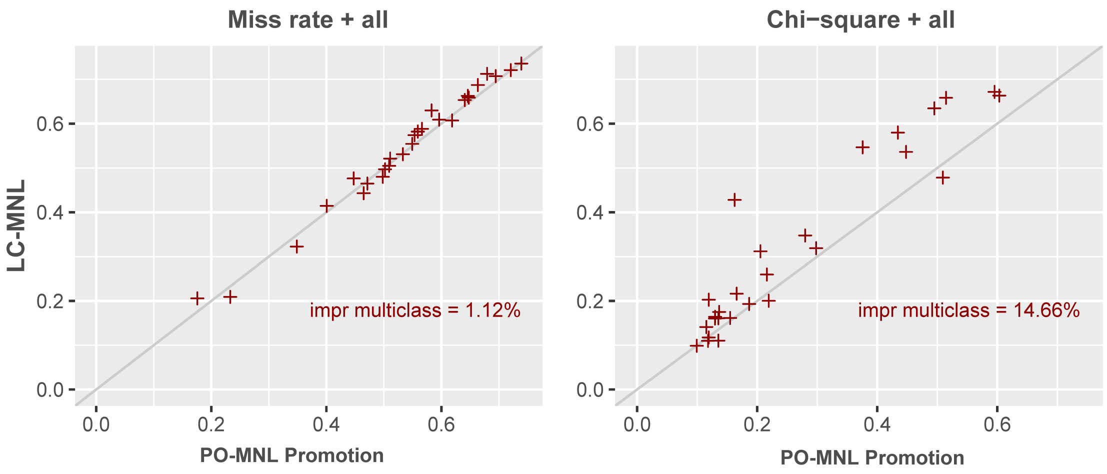

Figure 1-12: Brand choice prediction results with no-purchase option included. The training data consists of 31 weeks, and the test data consists of 21 weeks.

average deterioration of 1.36% over PO-MNL Inertial with clustering and of 3.66% over PO-MNL Censored with clustering. Then, in the right column in Figure 1-15, we calibrate the PO-MNL Inertial [top right panel] and PO-MNL Censored [bottom right panel] and represent the prediction performance of all the models over the subset of individuals that can not be explained by behavioral assumptions. In this case, both PO-MNL Inertial and PO-MNL Censored models are reduced to LC-MNL model. We notice that LC PO-MNL Promotion model exhibits an average improvement of 2.13% over PO-MNL Inertial and 2.53% over PO-MNL Censored.

Even though the results in Figures 1-14 and 1-15 show an average dominance of the behavioral models over the PO-MNL Promotion optimization model with respect to both χ^2 and miss rates, the presence of points above the diagonal indicates that for some categories PO-MNL Promotion still dominates. In order to characterize those categories, in Figure 1-16, we report the loyalty score of each category computed on the training data (left panel).[6] Then, in the middle and right panels we explore possible correlations between the percentage of χ^2 improvement of the behavioral models with respect to PO-MNL Promotion (vertical axis) vs. loyalty score (horizontal axis). We note a negative correlation for PO-MNL Promotion improvements with respect to both PO-MNL Inertial and Censored models, meaning that the behavior of customers for most of the

[6] To compute the loyalty score of a category (c.f. [52, Section 5.4]), we calculate the fraction of the total purchases coming from the most frequently purchased product (i.e., vendor) of each customer buying from that category and take the average of those fractions across customers purchasing from the category.

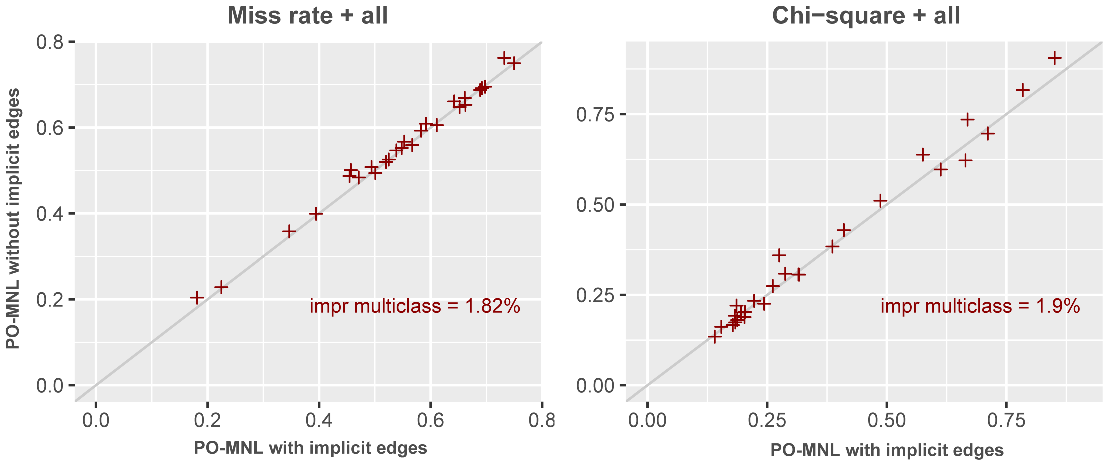

Figure 1-13: Brand choice prediction results with no-purchase option included. Robustness with respect to the addition of implicit candidate edges.

categories with low loyalty index (which exhibit the least stickiness in customers' preferences) are better explained by the PO-MNL Promotion model, as it is the case for customers represented by empty DAGs in the Jagabathula and Vulcano's approach.

These findings suggest that the practitioners might use the PO-MNL Inertial and Censored models for categories with high loyalty index, and within them, for customers having non-empty DAGs. Other than this, the use of the PO-MNL Promotion model proposed in this chapter leads to more effective predictions.

1.6 Optimization of personalized promotions

Having established that our model provides a more faithful representation of customer choice behavior than existing standard benchmarks, we now turn to the problem of customizing promotions. We take the standpoint of a retailer who wants to decide which products to put on promotion for each customer visit to maximize the expected revenue. In our study, the offer set is already decided and the retailer can only change the promotion activity from one customer to another. As discussed in Section 1.1, this setup reflects the practical situation faced by brick-and-mortar retailers, who cannot customize the shelf display to each visiting customer, but can adjust the promotion activity by launching personalized coupons to different customers.

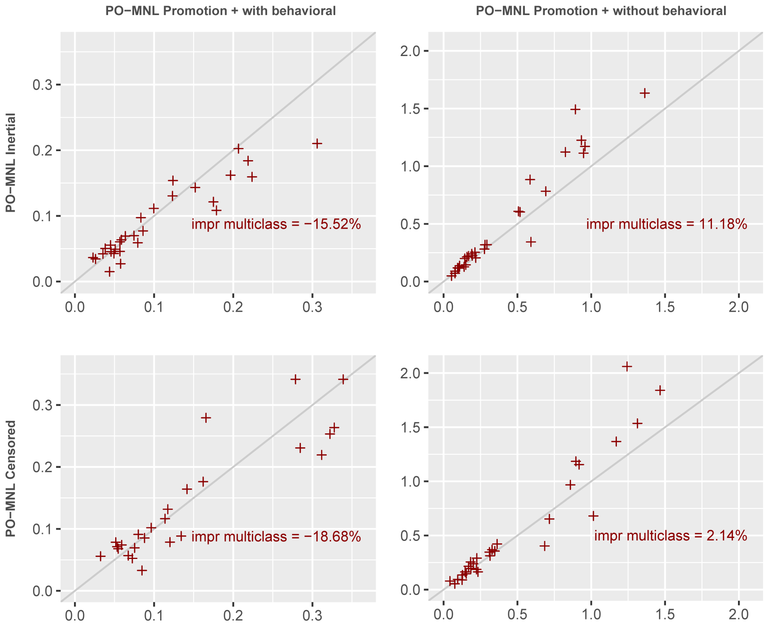

Figure 1-14: Scatter plot of the χ^2 scores. Comparison with the DAG-based behavioral models.

We start by illustrating some basic facts that the retailer can infer about the preferences of each customer from the structure of the corresponding DAGs. Then, we formulate the retailer's decision problem under the PO-MNL Promotion model as an MILP, followed by a test of our proposed methodology using the DAGs trained as described in Section 1.4 on the IRI Academic Dataset. For each purchase instance of the customer in the holdout sample, we use the MILP to determine the optimal promotion set. We then use the PO-MNL Promotion model to predict the purchase decisions of the customer under the optimal and the existing (i.e., those that are part of the holdout data) promotion sets in order to assess potential revenue improvements.

1.6.1 Inferences from the DAG structures

Our DAG-based representation of the customers' preferences has inherent value to a retailer reasoning about his promotion strategy; namely, the retailer can come to some key conclusions about the promotion decision purely from the nonparametric structure DAG.

To illustrate this, consider a customer whose preferences are described by DAG D in Figure 1-2 (after Phase 3) facing the full offer set including products 1 through 4, each of them in either its promoted or non-promoted version. From this DAG alone, the retailer can make the following

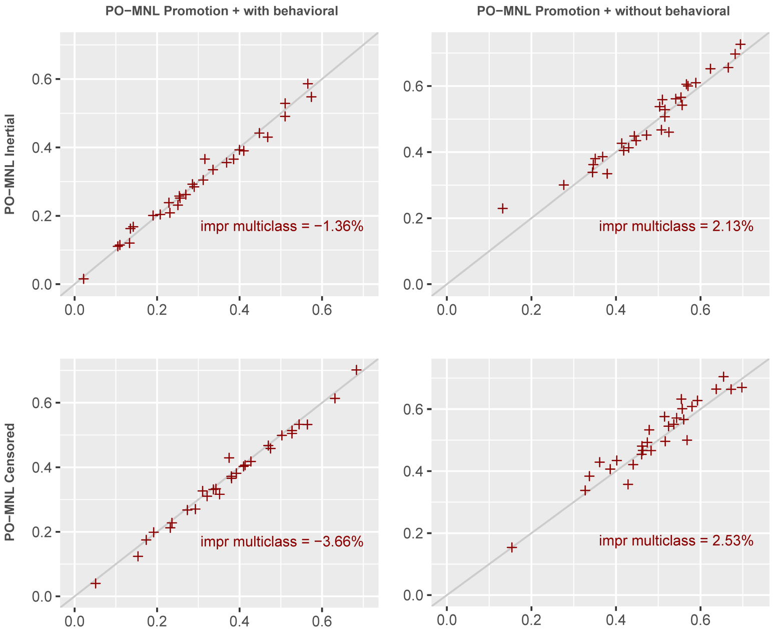

Figure 1-15: Scatter plot of the miss rate scores. Comparison with the DAG-based behavioral models.

inferences about what his promotion strategy should be for this customer: (a) product 4 will be purchased only on promotion since it is dominated by both versions of product 2; (b) product 3 will *not* be purchased whether it is put on promotion or not since it is also dominated by both versions of product 2; (c) the promotion strategy for product 1 depends on what is done for product 2 – if product 2 is on promotion, product 1 will not be purchased whether it is on promotion or not because there is a directed path from promoted 2 to promoted 1 (i.e., node 5 therein), and hence to non-promoted 1; but if product 2 is not promoted, then product 1 *could* be purchased if it is put on promotion –note there is no directed path between nodes 2 and 5.

Similar reasoning can be applied in other cases. In this way, our proposed DAG structures provide a visual, intuitive, and systematic way for retailers to reason about their promotion strategy on a per customer basis.

1.6.2 Promotion optimization: MILP formulation

We now systematize the intuitive reasoning above through an MILP to formulate the retailer's promotion optimization problem. The retailer must solve this problem each time a customer

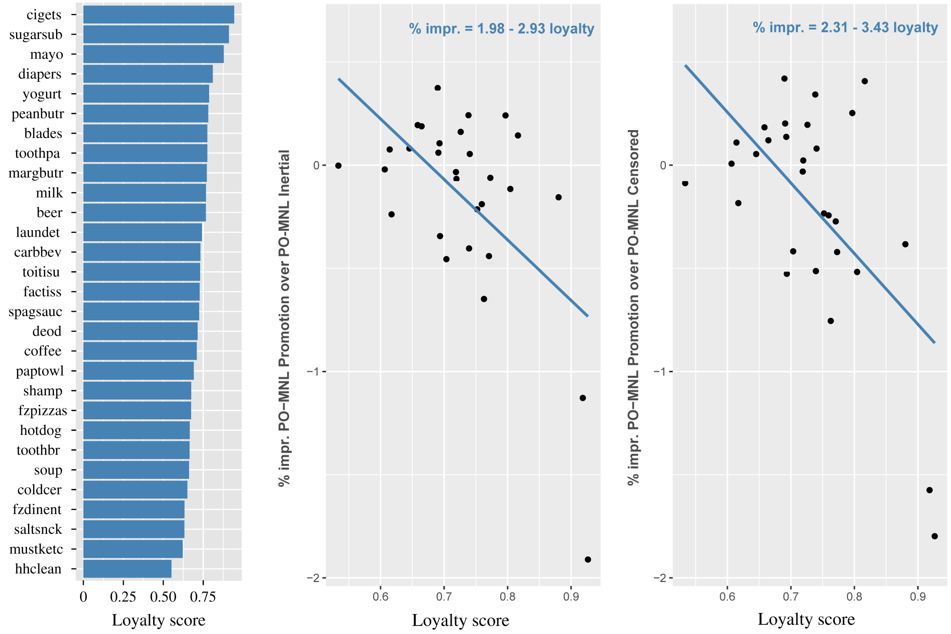

Figure 1-16: Loyalty scores and improvements of PO-MNL Promotion over behavioral DAG-based models.

visits the store. The formal setup is as follows. Recall that the universe $\mathcal{N}'$ consists of $2n$ products, where for all $j \in [n]$, the products a_j and a_{j+n} are the non-promoted and promoted copies, respectively, of the same product. For each $j \in [n]$, we let r_j denote the revenue from the non-promoted copy a_j and d_j the discount offered for the promoted copy a_{j+n}, for a total revenue of $r_j - d_j$. Also for each $j \in [n]$, let q_j and q_{j+n} denote the expected purchase quantities when the customer purchases the non-promoted copy a_j and the promoted copy a_{j+n}, respectively. We assume that the no-purchase option a_0 is always available and $r_0 = d_0 = 0$. Note that throughout this chapter so far, the no-purchase option was included implicitly in our analysis because as far as our methodology is concerned, there is no distinction between the no-purchase option and any other product (except that there is no promoted version of the no-purchase option). We now make it explicit because the promotion decision of the retailer not only impacts brand switching but also affects the purchase propensity of the customer.

The retailer must decide which products to offer on promotion. For any $j \in [n]$, if the retailer decides to offer product a_j on promotion, then we say that the retailer has decided to offer the promoted copy a_{j+n}, whereas if the retailer decides *not* to promote product a_j, then we say that the retailer has decided to offer the non-promoted copy a_j. As a result, the promotion decision of the retailer reduces to an assortment decision. To capture this, we let $S_{\mathcal{A}} \subseteq \mathcal{N}' \cup \{a_0\}$, with $a_0 \in S_{\mathcal{A}}$ denote the subset of available products from which the retailer must select his offer set. To be consistent with our set up, $S_{\mathcal{A}}$ has the property that if product $a_j \in S_{\mathcal{A}}$ for $j \in [n]$, then product $a_{j+n} \in S_{\mathcal{A}}$. Then, the goal of the retailer is to decide the subset of products in $S_{\mathcal{A}}$ to offer to a customer, with the constraint that exactly one of the promoted or non-promoted copies of each product in $S_{\mathcal{A}}$ is offered, as discussed in Section 1.2.1.

Our MILP model includes three sets of decision variables: $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{z}$. We start from defining binary variables $\boldsymbol{y}$, used to determine which product version (promoted or non-promoted) is offered within the available set $S_{\mathcal{A}}$, i.e., $(y_j \in \{0,1\} \colon a_j \in S_{\mathcal{A}})$, where $y_j = 1$ means that product copy a_j is offered. This can be captured by the constraint:

$$y_j \in \{0,1\} \text{ and } y_j + y_{j+n} = 1 \;\forall\; a_j \in S_{\mathcal{A}},\; j \in [n].$$

Since the no-purchase alternative is always available, we set $y_0 = 1$.

The binary variables $\boldsymbol{x}$ are used to indicate the product that will be purchased. Let $(x_j \in \{0,1\} \colon a_j \in S_{\mathcal{A}})$, with $x_j = 1$ if and only if the customer purchases product a_j. Of course, only available products could be purchased, and the set of binary variables $\boldsymbol{y}$ enforces this connection. Let $S(\boldsymbol{y}) := \{a_j \in S_{\mathcal{A}} \colon y_j = 1\}$ denote the specific assortment offered to the customer under the offer decision $\boldsymbol{y}$. Further, let $(z_j \in \{0,1\} \colon a_j \in S_{\mathcal{A}})$ denote auxiliary variables with $z_j = 1$ for all products a_j in the set $h_D(S(\boldsymbol{y}))$ of heads (i.e., the nodes without parents) in the subgraph of the transitive closure of D restricted to the set $S(\boldsymbol{y})$.

Customers only purchase the head products (see Section 1.3.2); therefore, we have that $x_j = 1$ only if $z_j = 1$. To determine which of the head products the customer purchases, we use the approximate posterior probabilities $\tilde{f}(a_j, S(\boldsymbol{y}), D)$ from (1.7) and assume that the customer purchases the product with the highest posterior probability. That is, we assume that the customer purchases the product $a_j \in h_D(\boldsymbol{y})$ such that

$$\frac{v_{\Psi_D(a_j)}}{\sum_{a_\ell \in h_D(\boldsymbol{y})} v_{\Psi_D(a_\ell)}} \geq \frac{v_{\Psi_D(a_k)}}{\sum_{a_\ell \in h_D(\boldsymbol{y})} v_{\Psi_D(a_\ell)}} \quad \forall\; a_k \in h_D(\boldsymbol{y}) \setminus \{a_j\},$$

where we define $v_j = \exp\big(\beta_j^0\big)$ and $v_{j+n} = \exp\big(\beta_j^0 + \beta_j\big)$ for all $j \in [n]$. Since the denominators on both sides of the inequality are equal, the customer purchases product a_j only if $v_{\Psi_D(a_j)} \geq v_{\Psi_D(a_k)}$ for all $a_k \in h_D(S(\boldsymbol{y})) \setminus \{a_j\}$. These constraints can together be expressed as

$$a_j \notin \operatorname*{arg\,max}_{a_k \in S_{\mathcal{A}}:\, z_k = 1} v_{\Psi_D(a_k)} \implies x_j = 0, \tag{1.8}$$

$$x_j \leq z_j \quad \forall\, a_j \in S_{\mathcal{A}}, \tag{1.9}$$

$$\sum_{j:\, a_j \in S_{\mathcal{A}}} x_j = 1, \ x_j \in \{0, 1\} \quad \forall\, a_j \in S_{\mathcal{A}}, \tag{1.10}$$

where the first inequality ensures that product a_j will not be purchased if it does not have the maximum attraction value (i.e., probability of being purchased) and the second inequality ensures that only heads are purchased. The normalization constraint (1.10) ensures that exactly one product is purchased.

To relate the head variables $\boldsymbol{z}$ to the offer variables $\boldsymbol{y}$, let $B \in \{0,1\}^{(2n+1)\times(2n+1)}$ denote the adjacency matrix of the transitive closure of D, so that $B_{kj} = 1$ if and only if there is a path from node a_k to node a_j in D, for any $k, j \in \{0, 1, 2, ..., 2n\}$. Now, product a_j becomes a head if and only if it is offered and there is no other product preferred over a_j that is also offered. We can express this condition as the following set of linear constraints:

$$z_j \leq y_j, \quad \forall\, a_j \in S_{\mathcal{A}}, \tag{1.11}$$

$$z_j \leq 1 - B_{kj} y_k, \quad \forall\, a_k, a_j \in S_{\mathcal{A}}, k \neq j, \tag{1.12}$$

$$z_j \geq y_j - \sum_{a_k \in S_{\mathcal{A}} \setminus \{a_j\}} B_{kj} y_k, \quad \forall\, a_j \in S_{\mathcal{A}}, \tag{1.13}$$

$$z_j, y_j \in \{0, 1\}, \quad \forall\, a_j \in S_{\mathcal{A}}, \tag{1.14}$$

where the first constraint ensures that only offered products can become heads, the second constraint ensures that a_j is not a head if an offered product $a_k \in S_{\mathcal{A}}$ is preferred over a_j in DAG D, and the third constraint ensures that a_j becomes a head if it is offered and there is no other offered product $a_k \in S_{\mathcal{A}}$ that is preferred over a_j in D.

Finally, it remains to express the objective function of the retailer in terms of the decision variables. The objective of the retailer is to maximize the expected revenue $\sum_{j:\, a_j \in S_{\mathcal{A}}} R_j q_j x_j$ from the customer, where $R_j = r_j$ and $R_{j+n} = r_j - d_j$ for any $j \in [n]$, and $R_0 = 0$.

Combining the above, we can express the retailer's optimization problem as follows:

$$\begin{aligned} \max_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{y}} \quad & \sum_{j:\, a_j \in S_{\mathcal{A}}} R_j q_j x_j \\ \text{subject to} \quad & \boldsymbol{x}, \boldsymbol{z} \text{ satisfy } (1.8) - (1.10), \\ & \boldsymbol{z}, \boldsymbol{y} \text{ satisfy } (1.11) - (1.14), \\ & y_j + y_{j+n} = 1 \ \ \forall\ a_j \in S_{\mathcal{A}}, j \in [n], \\ & y_0 = 1. \end{aligned}$$

To convert the above optimization problem into an MILP, we need to formulate constraint (1.8) as a linear constraint. For that, we introduce continuous variables $\{0 \leq p_j \leq 1 \colon a_j \in S_{\mathcal{A}}\}$ such that p_js are the attraction values of the head products, but normalized to sum to less than 1. Also, $p_j = 0$ for a non-head product. Given such $\boldsymbol{p}$'s, constraint (1.8) can be expressed as

$$x_j \leq 1 + p_j - p_k, \quad \forall\ a_k, a_j \in S_{\mathcal{A}}, k \neq j, \tag{1.15}$$

where it is clear that $x_j = 0$ whenever $p_j < p_k$ for some $a_k \in S_{\mathcal{A}}$. We show in Lemma 4.6.1 (see Section 4.6 in the Appendix) that the following set of constraints ensure that $\boldsymbol{p}$'s are the normalized attraction values:

$$p_j \leq z_j \ \ \forall\ a_j \in S_{\mathcal{A}}, \tag{1.16}$$

$$p_0 + \sum_{j:\, a_j \in S_{\mathcal{A}}} p_j = 1, \tag{1.17}$$

$$0 \leq p_j \leq v_{\Psi_D(a_j)} p_0, \quad \forall\ a_j \in S_{\mathcal{A}} \tag{1.18}$$

$$p_0 + z_j - 1 \leq p_j / v_{\Psi_D(a_j)} \ \ \forall\ a_j \in S_{\mathcal{A}}. \tag{1.19}$$

Putting everything together, we obtain the following MILP:

$$\begin{aligned} \max_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{y},\boldsymbol{p}} \quad & \sum_{j:\, a_j \in S_{\mathcal{A}}} R_j q_j x_j \qquad (1.20) \\ \text{subject to} \quad & \boldsymbol{x}, \boldsymbol{z} \text{ satisfy } (1.9) - (1.10), \\ & \boldsymbol{z}, \boldsymbol{y} \text{ satisfy } (1.11) - (1.14), \\ & \boldsymbol{x}, \boldsymbol{p}, \boldsymbol{z} \text{ satisfy } (1.15) - (1.19), \\ & y_j + y_{j+n} = 1 \ \ \forall\, a_j \in S_{\mathcal{A}}, j \in [n], \\ & y_0 = 1. \end{aligned}$$

Conceptually, the formulation is determining, through its variables $\boldsymbol{y}$, which ones of the $O(2^n)$ subsets of products should be put on promotion. Variables $\boldsymbol{z}$ determine the set of heads of the intersection between the customer DAG and the offer set that are candidates to be purchased, variables $\boldsymbol{p}$ are normalized attraction values for those heads, and variables $\boldsymbol{x}$ indicate the product to be purchased (by identifying the one with highest purchasing likelihood). The size of MILP (1.20) scales linearly in the number of variables and quadratically in the number of constraints with respect to the number of products n. In our experience, the implementation of the promotion optimization ran very fast, taking just 0.17 seconds on average and always solving to optimality.

1.6.3 Customized promotions: performance evaluation

We now evaluate the performance of the MILP proposed above to personalize promotions. We carry out our analysis using the DAGs that were trained as described in Section 1.4, but with the no-purchase option added. As noted above, the data do not consist of no-purchase observations. Therefore, as described in Section 1.5, we combine the purchases of a panelist across all categories to approximately infer customer visits to the store and then use a simple heuristic to infer which of these visits ended with no-purchase. Our robustness checks (also reported in Section 1.5) show that our results are persistent to this specific heuristic.

Because of our aggregation, each product was purchased at different prices in the training data. Therefore, to arrive at the price r_j for each product a_j, and the discounted price $r_j - d_j$

its promotion counterpart a_{j+n}, we averaged the full price and the discounted price, respectively, across different customers and stores in the training data. Similarly, to find the expected full price purchase quantity q_j and the discounted price purchase quantity q_{j+n} for each product, we averaged the full price purchase quantity and discount price purchase quantity, respectively, in the training data.

In order to assess the potential gains from our promotion strategy, we need a way to determine the purchases of the customers under different promotion strategies. We demonstrated in Section 1.4 that our DAG model provides the best accuracy for predicting individual customer purchases, when compared to existing benchmarks. Therefore, the DAG-based model is a promising candidate to anticipate purchases. As an extra check, we also first verified its accuracy in predicting revenues from each customer. To this end, for each customer and each hold-out time period, we compared the revenue from the purchase predicted by the PO-MNL Promotion model to the revenue from the actual purchase. The left panel in Figure 1-17 illustrates a scatterplot of the 'predicted revenue' vs. the realized revenue from actual purchases, where by 'predicted revenue' we mean the predicted revenue when customers choose according to the PO-MNL Promotion model for a given set of promoted and non-promoted items. Each point on the plot represents the revenues from one of the 27 product categories, averaged over all the customers and all the hold-out time periods. We found that the absolute revenue prediction errors are relatively small across all 27 product categories with a mean absolute error (MAE) of only 6.34%. This observation builds confidence on the predictive power of our model in terms of revenue assessment on top of the already verified purchase instance predictive power.

The revenue gains from customizing promotions are depicted in the middle panel in Figure 1-17. First, we consider the impact of the personalized promotions while ignoring the existing mass in-store promotions already offered in the store. Therefore, the retailer can set *any* subset of available products on promotion for each customer. We find that the retailer can increase the overall revenue by an estimated 23.93% on average across the 27 categories, when compared to the existing promotion strategy.

From the crosses depicted in the middle panel, we notice that the revenue gains from personalizing promotions vary significantly from category to category. To better explain this variation, we regressed the percentage improvement in revenue from personalization for each category against

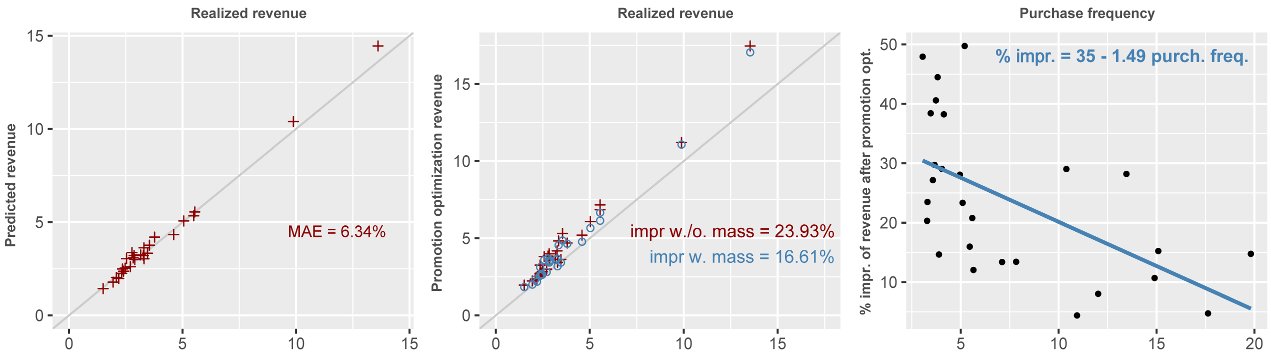

Figure 1-17: Revenue performance.

the average purchase frequency for items in the category. We measure the purchase frequency as the average number of times a customer makes a category purchase. The right panel of Figure 1-17 illustrates the regression. We see a clear negative correlation between the percentage revenue improvement and purchase frequency, suggesting the personalization could have the biggest impact for less frequently bought categories of products. Section 1.6.4 provides further analysis on the factors that explain the variation in the gains from personalization at the individual customer level. The main takeaway is that personalization is more beneficial for customers who are sensitive to promotions and who purchase frequently and is less beneficial to customers who are brand loyal; see the Section 1.6.4 for precise definitions of these terms.

In reality, personalized promotions need to coexist with the mass promotions already in place in the store (as reported in the dataset). To capture this, we impose the additional constraint that a retailer can personalize the promotions of only those products that are not already on mass promotions. The middle panel of Figure 1-17 illustrates with small circles that if the personalized promotions are mounted on top of existing in-store mass promotions the retailer can still increase the overall revenues by an estimated 16.61%, on average, across the 27 categories, when compared to the existing promotion strategy. Thus, under PO-MNL Promotion model, personalization boosts the flexibility of the promotion implementation, providing extra flexibility and enhancing the strategic promotion space.

Sometimes, a particular brand will impose a constraint to the retailer about not being promoted jointly with a competitive brand. In what follows, we empirically study the case where at most one item could be put on promotion at the personalized level. This could be implemented

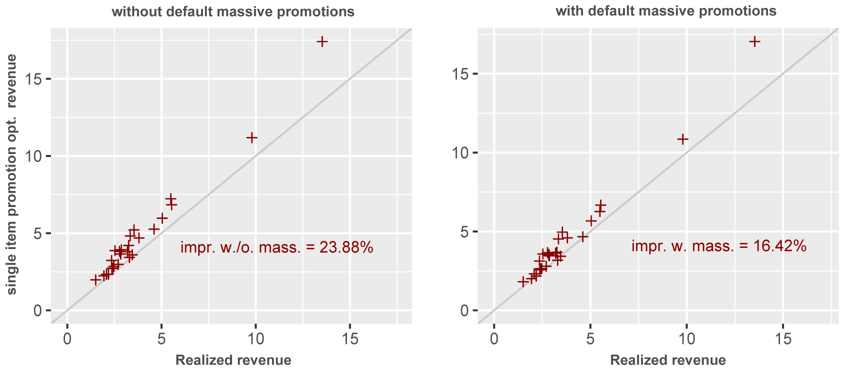

Figure 1-18: Revenue performance of promoting a single item.

by taking the MILP (1.20) and adding the constraint: $\sum_{j=n+1}^{2n} y_j \leq 1$. However, since this constraint reduces the search space from $O(2^n)$ to $O(n)$, it could be executed via a simple search algorithm that effectively sets $y_{j+n} = 1$ and $y_{k+n} = 0$ for all $k \in [n] \setminus \{j\}$, for each $j \in [n]$, and finally retains the value assignment that leads to the highest objective function. Analogous to our previous analysis, Figure 1-18 illustrates this limited promotion situation under two cases: no mass promotions simultaneously present (left panel), and the case where personalized promotions are run on top of the mass ones (right panel). If we promote at most one item for every customer arriving to the store, in the absence of mass promotions (left panel), the retailer can increase the overall revenues by an estimated 23.88% on average across the 27 categories, when compared to the existing promotion strategy. The right panel illustrates that if the personalized single-item promotions are mounted on top of the existing in-store mass promotions, then the retailer can increase the overall revenues by an estimated 16.42% on average across the 27 categories, when compared to the existing promotion strategy. These results indicate that by promoting just one item for every customer arriving to the store the retailer can get close to all the additional revenue extractable through personalization; see Section 1.6.5 for a partial explanation of why a small number of promotions is sufficient to extract most of the benefit from personalization. Consequently, the strategy of customized promotions where we promote at most one item for every customer visit, might help the retailer to mitigate the negative effects of running mass promotions and still lead to near optimal revenues.

1.6.4 Managerial insights: factors affecting improvements from personalization of promotions

The revenue improvements from personalization vary across different the customers in each category. To explain this variation, for each category, we consider three different customer-level characteristics:

1. *Brand loyalty*, measured as the percentage of times a customer buys her most frequently purchased brand from the category;
2. *Purchase frequency*, measured as the number of purchases a customer makes from the category; and
3. *Promotion sensitivity*, measured as the percentage of times a customer buys a promoted product from the category.

We regressed the revenue improvement for each customer and category combination against the brand loyalty, purchase frequency, and promotion sensitivity variables. Table 1.2 reports the results from fitting four different models:

$$\begin{aligned}
&\text{RevImpr}_{i,c} = \beta_{01} \cdot \text{Cat}_c + \beta_{11} \cdot \text{Bloyalty}_{i,c} + \varepsilon_{i,c} && \text{(Model 1)} \\
&\text{RevImpr}_{i,c} = \beta_{02} \cdot \text{Cat}_c + \beta_{22} \cdot \text{PurFreq}_{i,c} + \varepsilon_{i,c} && \text{(Model 2)} \\
&\text{RevImpr}_{i,c} = \beta_{03} \cdot \text{Cat}_c + \beta_{33} \cdot \text{PromSens}_{i,c} + \varepsilon_{i,c} && \text{(Model 3)} \\
&\text{RevImpr}_{i,c} = \beta_{04} \cdot \text{Cat}_c + \beta_{14} \cdot \text{Bloyalty}_{i,c} + \beta_{24} \cdot \text{PurFreq} + \beta_{34} \cdot \text{PromSens}_{i,c} + \varepsilon_{i,c} && \text{(Model 4)},
\end{aligned}$$

where the variables $\text{Bloyalty}_{i,c}$, $\text{PurFreq}_{i,c}$, and $\text{PromSens}_{i,c}$ respectively denote the brand loyalty, purchase frequency, and promotion sensitivity computed for customer i under category c. The brand loyalty and purchase frequency variables were computed using the training data. To ensure exogeneity, the promotion sensitivity variable was computed using the data from the previous year (2006). The variable Cat_c is an indicator variable denoting category c to capture category fixed effects. Finally, $\text{RevImpr}_{i,c}$ is the average revenue improvement from personalization for customer i under category c, computed over the holdout sample.

The results from the regressions are consistent and intuitive. The benefits from personalization are negatively correlated with brand loyalty (Model 1) and purchase frequency (Model 2) but positively correlated with promotion sensitivity (Model 3). In other words, customers who purchase infrequently and concentrate their purchases only on a few brands are harder to pursuade to switch to more profitable brands through personalized promotions. On the other hand, customers who frequently purchase promoted items are easier to be influenced by personalizing promotions. These findings are consistent in a multiple regression of the revenue improvement against all three variables (Model 4). The coefficients in the multiple regression are all statistically significant, indicating that all three factors together influence the brand switching behavior of customers in response to promotions.

	Dependent variable:			
	Revenue Improvement (%)			
	Model 1	Model 2	Model 3	Model 4
Brand loyalty	-14.415***			-20.514***
	(-2.742)			(-3.521)
Purch. frequency		-1.108***		-1.218***
		(-3.133)		(-3.224)
Prom. sensitivity			14.575***	13.891***
			(3.666)	(3.610)
Category FE	Yes	Yes	Yes	Yes
No. Observations:	57,059	57,059	57,059	57,059
R-squared:	0.001	0.003	0.001	0.006

t statistics in parentheses

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

Table 1.2: Individual-level regressions with product category fixed effects.

1.6.5 Additional insights on brand loyalty and the number of promoted items

In this section, we provide additional descriptive statistics to gain insights into the extent of brand loyalty of the customers and the number of promoted items in our dataset. Table 1.3 presents these statistics.

The columns second to the sixth in the table report the statistics describing the distribution of the number of unique brands purchased by the customers in each category; specifically, the columns report the mean, the standard deviation, and the first, second, and the third quartiles of the distribution, respectively. We note that on average customers purchase no more than 4 unique brands in the training data, inidicating that customers have strong preferences and their purchases don't change very much from week to week. This brand loyal behavior of customers also explains the significant gains in performance that our method obtained over the benchmark methods.

Columns seven and eight report the average number of products our method offers on promotion across all the transactions in the holdout sample. Column seven reports this number when existing mass promotions are ignored and column eight reports the number of promoted products on top of the products that are already on mass promotion. We note that most of these numbers are less than 1, indicating that at optimality, our method offers only a small number of products on promotion. This observation explains why our method is able to extract most of the revenues even with the constraint of offering at most one product on promotion; see Section 1.6.3.

In order to provide a partial explanation for the small number of products that are put on promotion, the last two columns of the table report the average offer set size and the average number of products that can be potentially promoted (NonDom) across all the transactions. To calculate the number of products that can be potentially promoted, consider a transaction in which the offer set is S and the customer has DAG D. We note that product a_j will *not* be promoted if there is another product $a_i \in S$ such that the non-promoted copy of a_i (and consequently, the promoted copy of a_i) is preferred over the promoted copy of a_j in DAG D. The reason is that product a_j will not be purchased whether promoted or not because either the promoted or the non-promoted copy of a_i will be offered to the customer. We call such a product a_j a dominated product and any product that is not dominated, the non-dominated prodcut. Given this, NonDom reports the average number of non-dominated products across all the transactions in the holdout sample.

We observe from the table that customer DAGs are such that the average number of non-dominated products is far smaller than the average offer set size. Because the number of non-dominated products is an upper bound on the number of products that will promoted, this table provides a partial explanation as to why at optimality, only a small number of products are promoted.

Category	# Unique brands purchased					# prom. items		AvOS	NonDom
	mean	stdev	1st q.	2nd q.	3rd q.	w./o. mass	w. mass		
beer	2.09	1.29	1	2	3	1.11	0.26	43.87	23.72
carbbev	2.91	1.31	2	3	4	0.76	0.14	15.36	4.36
cigets	1.29	0.52	1	1	2	0.78	0.68	7.14	5.11
coffee	2.17	1.19	1	2	3	0.61	0.28	19.80	6.72
coldcer	3.03	1.37	2	3	4	0.76	0.16	17.66	5.02
deod	2.03	1.04	1	2	2	0.88	0.29	14.55	9.86
factiss	1.90	0.87	1	2	2	0.71	0.49	4.17	2.53
fzdinent	3.36	2.10	2	3	4	0.63	0.32	33.14	7.97
fzpizza	2.52	1.33	2	2	3	0.62	0.15	15.50	5.55
hhclean	2.90	1.40	2	3	4	0.88	0.61	31.42	11.01
hotdog	2.09	1.04	1	2	3	0.54	0.39	16.81	7.71
laundet	1.96	1.07	1	2	2	0.66	0.26	10.08	4.89
margbutr	1.93	1.02	1	2	2	0.62	0.41	10.35	3.61
mayo	1.36	0.55	1	1	2	0.24	0.20	6.86	4.53
milk	2.36	1.17	1	2	3	0.36	0.23	11.69	2.84
mustketc	2.29	0.92	2	2	3	0.59	0.33	17.07	7.79
paptowl	2.15	1.08	1	2	3	0.94	0.49	6.94	2.62
peanbutr	1.65	0.75	1	2	2	0.28	0.23	7.99	4.39
saltsnck	3.67	1.89	2	3	5	0.77	0.18	26.79	6.87
shamp	2.15	1.14	1	2	3	0.93	0.29	18.74	8.33
soup	2.94	1.46	2	3	4	0.92	0.24	32.87	15.99
spagsauc	2.19	1.25	1	2	3	0.77	0.39	17.85	7.12
sugarsub	1.23	0.47	1	1	1	0.91	0.70	5.05	3.60
toitisu	2.01	1.06	1	2	3	0.99	0.49	7.66	2.98
toothbr	2.06	1.00	1	2	3	0.80	0.44	15.86	7.84
toothpa	1.65	0.74	1	2	2	0.95	0.23	12.05	7.80
yogurt	2.41	1.30	1	2	3	0.46	0.18	9.84	3.67

Table 1.3: Relevant summary statistics from the data.

1.6.6 Robustness check for the promotion optimization

In the revenue results in Section 1.6, we assumed that, when arriving at the store, every customer would buy the product with highest probability of being purchased under PO-MNL Promotion model. An alternative objective would be to compute expected revenues accounting for the probabilities of purchase for every single product on offer. Figure 1-19 illustrates the results of running personalized promotions under the modified objective function of the optimization

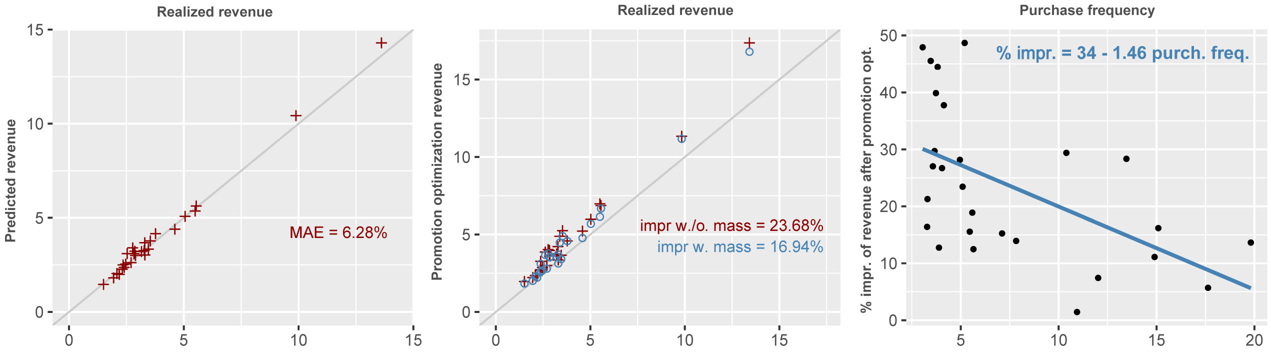

Figure 1-19: Promotion optimization problem when we optimize the expected revenue for every purchasing transaction.

problem. Since the new promotion optimization results, illustrated in Figure 1-19, almost exactly resemble the ones in Figure 1-17, we conclude that all the insights remain qualitatively the same under the updated formulation of the promotion optimization problem.

1.7 Conclusions and future work

Sales promotions planning is an important part of day-to-day operations in the retail industry, where a large proportion of products is sold under discounted prices. For many years, grocery retailers have run massive promotions, offering the same deal to all the customers. This approach is appealing due to its simple practical implementation, however it may lead to a neutral or a negative impact on revenue in the long run. Because different customers are affected by promotions differently, the retailer benefits from offering personalized deals. Nowadays, this is more feasible given the unprecedented volume of panel data regarding sales transactions that businesses are able to collect. New technology also assists in personalizing the customer experience. As a result, customization can mitigate the negative effects of promotions and be used as an appealing means for price discrimination.

In this chapter, we considered a back-to-back methodology of running personalized promotions with the objective of increasing retailers' revenue by inducing the brand switching effect. Naturally, an important step in personalized promotion planning is to understand individual preferences for different products within a category. The building blocks of our proposal identify each customer with a nonparametric DAG that explicitly accounts for promotions by creating two copies for every item in the product category: promoted and non-promoted versions. Edges in the DAG of an individual reflect the relative preference between two products or, more precisely, between the two versions of each product. We described how to build each customer's DAG for a given category in a purely data-driven way, and explained how to calibrate a parametric (multiclass) MNL model over the collection of customer DAGs. We demonstrated its ability to make more precise and finely grained predictions of customers' responses to price promotions on real retail data compared to state-of-the-art benchmarks. Theoretically, we derived tractable lower and upper bounds relative to the exact likelihood of partial orders and to the likelihood of purchasing a particular product from a given offer set.

The successful performance of these purchasing prediction results served as the basis for the next phase: the implementation of customized promotions. We formulated a compact MILP to solve the personalized promotion optimization problem. On the same dataset, we verified via simulation studies that our personalized promotions provide revenue gains across 27 categories of the order of 16% if run on top of the current mass promotions already in place and of the order of 23% if instead *any* subset of available products can be promoted. Similar revenue gains were observed even after constraining the retailer to promote, at most, a single item. Overall, based on the results obtained from the real retail data, we believe that our methodology constitutes an interesting framework to be further tested in the retail operations practice.

An industry implementation of our proposal would need to fine-tune a few details. For instance, there are two MILPs that need to be solved: the decycling and the promotion optimization. The decycling procedure is run periodically for each customer (e.g., once every six months) and could be solved as an overnight batch process. However, the promotion optimization must be solved in real time during each store visit since it depends on both the particular DAG of the customer as well as the subset of offered products. Even though, in our experience, the problem is solved to optimality within a fraction of a second for up to 100 products, as the size of the product category scales, the computational performance could suffer. As such, developing valid inequalities and designing a branch-and-cut procedure, or testing polynomial running time heuristics, could be fundamental for real applications.

Part II

Demand Models with Consideration Sets

Chapter 2

Inferring Consideration Sets from Sales Transaction Data

2.1 Introduction

Modeling consumer preferences is a fundamental task in many operations applications. It provides the necessary inputs for optimizing assortments and prices in retail settings as well as when matching demand and supply on online platforms. Over the past several years, discrete choice models have become the predominant method to modeling customer demand. At the core, these models are designed to predict the choice a customer makes in response to an offered set of products. These models are generally calibrated using sales transaction data, which consists of a collection of observations of the form (a, S), where a is the chosen product from offer set S.

The estimation of a choice model from the transaction data faces a major common challenge: the actual consideration set, or the choice set of the customer, is unobserved. The offer set consists of the set of items a customer could possibly purchase, but the consideration set is the set of items that the customer actually evaluates before purchasing. Knowing the consideration set is critical for calibrating choice models. At its core, calibration relies on the inference that the customer prefers the chosen product to all other products the she considered. When the consideration set is unknown, we may make erroneous inferences, which result in biased model estimates. For instance, we may infer that a product that was not chosen is preferred less than the chosen one when, in fact, the customer never even considered that product. This issue has been recognized in existing literature. Methods have been proposed to infer consideration sets when the sales transaction data is complemented by tagging individual transactions through customer IDs [49, 52] and by running surveys [88].

In the absence of any fine-grained data to complement sales transaction data, the common assumption within operations has been to assume that customers consider *everything* on offer. However, in many emerging applications, the definition of the offer set is, itself, quite unclear. Choice models are increasingly being applied to model demand on online platforms [62]. Think, for instance, of an online platform for peer-to-peer car sharing, and consider a renter about to make a booking request. Which cars should be considered as being "offered" by the platform at the moment of booking? Does the offer set comprise of all the available cars in a radius of 0.2 miles from her location? Or are all of the available cars parked in a garage in a radius of 0.1 miles? Or maybe only cars of a particular make parked on the street nearby? The precise mechanism by which the renter makes a choice is opaque to the platform,[1] which only observes car availability at the moment of the booking and the realized transaction.

In addition, the availability of the products is not always perfectly known. Both on the online platform as well as in retail settings, we only have access to the sales transaction data. From this data, offer sets are commonly constructed by assuming that a product is available at a particular time if it appeared in a transaction close to this time; the offer set is formed by taking a combination of all available products at that time [96]. Such an assumption is reasonable for frequently purchased products. However, it generally results in a noisy estimate of the true offer set.

Finally, even in cases when the availability information is known, the consideration set still remains unclear. Think of a customer facing a category of products in a grocery store. We could argue that the offer set is the collection of all SKUs available on the shelf, which has been the default assumption in retail applications of operations management (OM)-related literature (e.g., [75, 80]). This approach is justified from classical parametric discrete choice models of the random utility maximization (RUM) class [5, 59, 64, 92], in which a customer, assumed to be fully rational, evaluates all feasible product options before making a final decision. The immediate

[1] Browsing data, if available, may provide greater insight, but it is still imprecise because clicks and appearance in search results are not perfect signals of consideration.

criticism of this assumption is that customers are boundedly rational, and the associated cognitive burden prevents them from evaluating all offered products. Therefore, they consider only a subset of offered products.

To address these challenges, we model customer choices through a consider-then-choose model. Under this framework, during the first stage customers form a consideration set by eliminating a few products, and then choose from the remaining options. Products that are ruled out during the first step are clearly not going to be purchased. The customer then evaluates products from the consideration set before making a final choice. In cases when the customer is fully rational and there is noise in the offer set, the consideration set may be viewed as the true (latent) offer set. In cases when the customer is boundedly rational, the consideration set may be viewed as a subset that is formed using simple screening rules to limit the cognitive burden of evaluating all offered products. We assume that the first stage is described by a distribution over consideration sets, and the second stage is described by a choice rule. In order to be able to identify our model from sales transaction data, we make a simplifying assumption about the choice rules. Following recent approaches in the literature, we assume that the choice rule is fully specified by a unique ranking over the products in the product category [3, 68, 70]. As a result, the heterogeneity of customers and stochasticity of choice is captured in the model through the consideration set formation of individuals.

The consideration sets, we estimate, also address a related challenge of identifying a firm's competitors, which provides necessary input for the firm's business strategy development, marketing decisions (e.g., advertising, positioning, segmentation strategy, and promotions), pricing strategy, and service operations (e.g., distribution strategy) of every firm. Our methodology allows firms to determine competition based on demand side considerations. Specifically, if two brands appear in the consideration set of the same customer, then the firms owning those brands will be competing with each other. While every other firm is a potential competitor, the firm can prioritize its competitors by analyzing which are most likely to appear alongside their own firm in customers' consideration sets.

In summary, understanding the set of products considered by customers at the moment of making a choice in a systematic way (e.g., considering either the availabilities for services declared on an online platform, the assortment provided by a retailer, or competitors' offers), may provide an advantage, which ultimately affects the bottom line. This chapter contributes to the existing literature along these lines. In order to provide a consistent presentation in our work, we focus on the consideration set version of the problem.

2.1.1 Summary of the results and contribution

The main contribution of this chapter is a methodology designed to estimate consideration sets of customers from sales transaction data. Our methodology relies on fitting a general consider-then-choose model to the data. Specifically, we make the following contributions:

- *A general consider-then-choose model.* We propose the general consider-then-choose (GCC) choice model to infer consideration sets from sales transactions data. We assume that all customers have the same preference order for items in the product universe, and the stochasticity of choice comes from the bounded rationality of customers who make every purchase in a two-stage process. First, they sample their consideration set, which is any subset of the product universe. Secondly, they purchase the most preferred item in the sampled consideration set.

- *Identification conditions for the GCC model.* The primary question we address is whether the GCC model is identifiable from purchasing transaction data. We provide necessary and sufficient conditions that can be used to discover whether or not sales data was generated by the model, and provide arguments about how to infer the preference order and the probability distribution function over consideration sets from observed choice frequencies.

- *Methodology to estimate the parameters of the GCC model.* We start with the MINLP formulation of the MLE problem to infer model parameters from sales transaction data of the restricted version of the GCC model where customers include items in the consideration set independently. We show that, in this case, the MINLP can be solved by solving a sequence of MILPs by iteratively linearizing the optimizing problem. It follows from existing results that this procedure converges to the global maximum. The procedure also provides a bound on the optimality gap at every step. Then, we propose the EM-based algorithm in order to calibrate the GCC model.

2.1.2 Related literature

The consideration set literature originates in the marketing and psychology field and dates back to the papers by [14], [47] and [102]. It has long been recognized that consumers usually make choices in a two-stage process [65, 82, 89]. 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 very hard to observe whether a product is included or not in a consumer's consideration set, and it might depend on a number of factors not necessarily related to the consumer's preferences. Nevertheless, there is ample empirical evidence in the literature of the consider-then-choose behavior of customers. In his seminal paper, [40] shows that a model based on consideration set phenomenon accounts for as much as 78% of the explainable uncertainty in purchase transaction data. [41] reports that the average size of the consideration set of consumer packaged goods in US is 1/10th of the total number of brands in the product category. In a previous study, customers consider on average only 3 brands of deodorants, 4 brands of shampoos, 4 brands of laundry detergents, and 4 brands of coffee [43].

The notion of consideration sets might arise from the limited information gathering ability of consumers, because they incur a search cost to learn a detailed information about the products [13, 79]. The search cost might be both cognitive and explicit, when a customer reads reviews about the product and tries to find the available information in the Internet. In this stream of literature it is assumed that consumers keep searching for products until the expected gain from search is less than the searching cost. In particular, [81] calibrated and tested a choice model of brand set composition that incorporated the search cost for consideration set formation and obtained significant improvement in predictive performance. Another stream of literature assumes that customers apply simple screening rules to alleviate the cognitive burden of overwhelmingly large variety of products [35, 42, 94], instead of sequentially adding products to the consideration set based on benefit-vs.-cost trade-off. As a result, the product evaluation process within the consideration set is assumed to be much more exhaustive than the heuristic screening to find which products will be included in the consideration set. [41] reports consideration set heuristics that are the most popular in marketing and psychology literature while being of great importance for managerial decisions in advertising, product development, and strategic planning (e.g., conjunctive, disjunctive, compensatory, elimination by aspects) [45, 73, 93]. In particular, customers might only consider the products that they can afford based on their budget constraints [51], or consider only the products under promotion and the ones purchased previously [52].

The paper by [68] is similar to our work in terms of some modeling assumptions. They study a choice model of boundedly rational agents having limited attention. In particular, consideration set formation is stochastic, where every alternative is considered with a given unobservable probability, the attention parameter. After forming a consideration set, a consumer purchases the product that maximizes a preference relation within considered products. As the main result, they demonstrate 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. In contrast to their model, we assume that there is a distribution function over all possible subsets in the product universe, and every time customers make a purchase, they sample a consideration set according to this distribution, and then choose the most preferred item. Our model generalizes the one by [68], relaxing the assumption of the lack of correlation between attention parameters, i.e., independent formation of consideration sets. [70] introduce the concept of attention filter that relates to consideration set formation. The authors showed that their consideration set model of choice can be characterized by relaxing the weak axiom of the revealed preferences (WARP), and provided a choice theoretical foundation to retrieve consumer's attention and preferences from choice data. There are some other papers in the economics literature that study rational inattention choice models [15, 27, 71].

The ideas of identifying consideration sets from sales transaction data can be extended to finding competition sets. For the latter, a recent alternative approach was developed by [63], who present a network-driven methodology to find competition sets and applied it to the hotel industry using a combination of search and clickstream data. [61] present an approach to characterize competition sets in the hotel travel industry. They estimate demand with a random-coefficient multinomial logit model and identify customer segments to build competition sets. Theoretical models of spatial competitions dates back to the seminal paper by [46], where the authors assume that companies compete only with nearest neighbors. Several other recent empirical papers consider the models of market competition through location-based measures in different industries:

bike-sharing [36], movie theaters [22], fast food industry [2, 91], car dealer networks [1, 76], and gasoline market [77]. Alternative empirical studies focus on estimation of competition through cross-price elasticities [9, 21].

The approach that we use to model consideration sets is part of the current trend of choice-based demand estimation in the OM-related literature [67, 90, 98]. [3] consider the problem of assortment optimization under the general consider-then-choose model. In particular, they proposed a dynamic programming formulation by introducing a bipartite graph representation of the assortment problem which is proved to be NP-hard [4]. They demonstrate that the dynamic program can be solved in polynomial time for the special cases of consideration set formation. [33] showed that assortment optimization problem under the restricted version of consideration set choice model [68] runs in polynomial time even with capacity constraints. The general consider-then-choose (GCC) choice model that we study in this chapter is a special case of a general nonparametric discrete choice model, well studied in the literature [29, 83, 95], where all the rankings are drawn from the same permutation of items in the product category and where the stochasticity is coming from the consideration set formation. Motivated by the notion of consideration sets, there are several other recent articles in the operations field that incorporate cognitive limitations of consumers in their models [30, 31, 99].

2.2 General Consider-then-Choose model

In this section, we formally describe the setup and the model. We also present theoretical results on the conditions under which our model can be identified from aggregated sales transaction data.

We consider a universe N of n products $\{a_1, a_2, \ldots, a_n\}$. We let a_0 denote the 'no-purchase' or the 'outside' option. Customers arrive to the store sequentially, and in each choice instance, a customer is presented with a subset $S \subseteq N$ of products and the customer chooses either one of the products in S or the outside option a_0. We let $\mathbb{P}_j(S)$ denote the probability that a customer chooses product $a_j \in S$ and $\mathbb{P}_0(S)$ – the probability that the customer chooses the outside option. Our goal is to model this choice process through a probabilistic model that specifies all the choice probabilities $\{\mathbb{P}_j(S) \colon a_j \in S^+, S \subseteq N\}$, where we use S^+ to denote the set $S \cup \{a_0\}$. We assume that the choice probabilities satisfy the standard probability laws: $\mathbb{P}_j(S) \geq 0$ for all $a_j \in S^+$ and $\sum_{a_j \in S^+} \mathbb{P}_j(S) = 1$ for all $S \subseteq N$.

In order to explicitly account for the fact that customers may not consider all the offered products before making a choice, we assume that customer choices follow a two-stage consider-then-choose model. In the first stage, the customer forms a consideration set $C \subseteq N$ and in the second stage, the customer chooses either a product from the set $S \cap C$ of products or the outside option a_0, when the offered set of products is S. In this model, for a product to be purchased, it must be both offered *and* considered. The seller restricts customers' choices by deciding which products to offer. But the customer further restricts her choices to just the ones in her consideration set because either she has strong unobserved preferences (which prevent her from ever buying certain products), or cognitive overload prevents her from evaluating all the products on offer before choosing.

The model is specified by two mathematical objects: a distribution $\lambda \colon 2^N \to [0, 1]$ over consideration sets such that $\sum_{C \subseteq N} \lambda(C) = 1$ and a choice rule that specifies which product is chosen from the subset $S \cap C$. In each choice instance, a randomly drawn customer from the population samples a consideration set C according to λ and then chooses a product from the set $(S \cap C)^+$ according to the choice rule. The most general model would accommodate any distribution λ and choice rule, but such a general model cannot be identified from transaction data alone. Therefore, we restrict the degrees of freedom, by assuming that customers' choice rule in the second stage is described by a single preference ordering over the products. The preference ordering or ranking of the products in N is described by a bijective ranking function $\sigma \colon N \to \{1, \ldots, n\}$ specifying a preference rank $\sigma(a_j)$ for each product a_j. Assuming that lower-valued ranks are preferred over higher-valued ranks, ranking σ indicates that product a is preferred over product b if and only if $\sigma(a) < \sigma(b)$. The preference ordering σ induces an antireflexive, antisymmetric, and transitive preference relation $\succ_\sigma$, defined as $a \succ_\sigma b$ if and only if $\sigma(a) < \sigma(b)$. Under this choice rule, the customer chooses the most preferred product according to σ from the subset $S \cap C$; that is, the customer chooses the product $\arg\min \{\sigma(a_j) \colon a_j \in S \cap C\}$ if $S \cap C \neq \varnothing$ and the outside option a_0 otherwise. Note that we are implicitly assuming that the outside option is the least preferred product in the ranking σ, so the customer always makes a purchase whenever $S \cap C$ is non-empty. This assumption is without loss of generality (WLOG) because otherwise the customer will never purchase a product a_j that is less preferred than a_0, in which case product a_j can be eliminated from the universe N. With this restriction on the choice rule, we show below that the model can be identified from the transaction data alone, even without any further assumptions on λ.

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

$$\mathbb{P}_j(S) = \begin{cases} \sum_{C \subseteq N} \lambda(C) \cdot \mathbf{I}[a_j \in S \cap C] \cdot \mathbf{I}[a_j \succ_\sigma a_k \; \forall a_k \in S \cap C, a_k \neq a_j], \text{ if } a_j \in S \\ \sum_{C \subseteq N} \lambda(C) \cdot \mathbf{I}[S \cap C = \varnothing], \text{ if } a_j = a_0, \end{cases} \tag{2.1}$$

where $\mathbf{I}[A]$ is the standard indicator function taking the value 1 if condition A is satisfied and the value 0 otherwise. We further assume that the empty condition, that is, $A = \varnothing$, is always satisfied.

In order to understand how our model is related to existing models, we first ask if the GCC model belongs to the general random utility maximization (RUM) class [10]. The RUM class is the most studied choice model class in the literature and includes popular models, such as the MNL, nested logit (NL), and mixture of MNLs (MMNL) models. At the core, the model assumes that in each choice instance, customers sample utility values for the products according to some joint distribution and chooses the offered product with the highest sampled utility value. Equivalently, the RUM class of models is described by a distribution over preference orderings of products [29, 87], so that a customer samples a preference ordering according to the distribution and chooses the most preferred offered product according to the sampled preference list. We show that the GCC model belongs to the class of RUM models, but the RUM models is a strict superset of the GCC class. Specifically, we establish the following result.

Proposition 2.2.1. *The GCC choice model is a special case of the RUM choice rule, that is,* $GCC \subset RUM$*, but* $GCC \neq RUM$.

The proof of the proposition is given in Section 5.2. It exhibits an example choice model that belongs to the RUM class but not to the GCC class.

While prior literature has studied consider-then-choose models, almost all of them have imposed restrictions on the structure of λ, perhaps with less restrictive assumptions on the choice rule. We differ from the literature, by allowing λ to be very general, but with the restriction that the choice rule is fully described by a single preference list. As such, our model subsumes the models of [68] and [3]. Proposition 2.2.1 is the first to study the relationship between the GCC and the RUM model classes.

In the rest of this section, we study the theoretical properties of the GCC model—identification conditions and robustness to noise in offer set information. Of course, when estimating such a general distribution λ from data, we run into computational challenges in addition to identification issues. To deal with those, we need to make additional assumptions on λ. But we defer these considerations to Sections 2.3 and beyond.

2.2.1 Identification of the GCC model

We now study the case when the offer sets are perfectly observed. In this case, we show that the GCC model is fully identfied only from the observed choice probabilities. In contrast, the RUM model is not fully identified when $n \geq 4$ [84].

To establish our result, suppose that purchases are generated according to an underlying GCC model, and we observe the choice probabilities $\mathbb{P}_j(S)$ for all products $a_j \in S^+$ and all offer sets $S \subseteq N$; here, we are ignoring any finite sample issues and assuming that the choice probabilities are exactly known. Then, we can recover the underlying parameters of the GCC model from the collection of choice probabilities $\{\mathbb{P}_j(S) \colon a_j \in S^+, S \subseteq N\}$. In particular, we have the following result:

Proposition 2.2.2. *Suppose that the collection of choice probabilities $\{\mathbb{P}_0(S) \colon S \subseteq N\}$ are consistent with an underlying GCC model. Then, we have that*

$$\lambda(C) = \sum_{X \subseteq C} (-1)^{|C|-|X|} \mathbb{P}_0(N \setminus X). \tag{2.2}$$

The result of the proposition follows immediately from a particular form of the inclusion-exclusion principle stated in [37]. For any finite set Z, if $f \colon 2^Z \to \mathbb{R}$ and $g \colon 2^Z \to \mathbb{R}$ are two real-valued set functions defined on the subsets of Z such that $g(X) = \sum_{Y \subseteq X} f(Y)$, then the inclusion-exclusion principle states that $f(Y) = \sum_{X \subseteq Y} (-1)^{|Y|-|X|} g(X)$. Our result then follows from noting that $\mathbb{P}_0(N \setminus X) = \sum_{C \subseteq X} \lambda(C)$. For completeness, we provide an alternative proof of this result in Section 5.2 from first principles.

The result of Proposition 2.2.2 shows that to recover λ, we only need the choice probabilities of the outside option under all the offer sets. On the other hand, the underlying preference ordering σ can be recovered independently from λ using the choice probabilities under all offer sets of size at most two; that is, $\{\mathbb{P}_j(S) \colon a_j \in S, |S| \leq 2\}$. Specifically, we have the following result:

Proposition 2.2.3. *Suppose that the collection of choice probabilities* $\{\mathbb{P}_j(S) \colon a_j \in S, |S| \leq 2\}$ *are consistent with an underlying GCC model. Then,*

$$\sigma(a_j) < \sigma(a_i) \quad \textit{if} \;\; \mathbb{P}_i(\{a_i\}) > \mathbb{P}_i(\{a_i, a_j\}), \;\; \textit{for all}\; 1 \leq i, j \leq n, i \neq j.$$

The proof of Proposition 2.2.3 follows directly from the definition of the GCC model, and is presented in Section 5.2.

Empirical evidence in the marketing literature suggests that the size of the consideration sets for most of the customers in different categories is relatively small, e.g., [48] concludes that the median number of laundry detergents that a consumer considers before making a purchase is one. When the size of consideration sets in the GCC model is bounded above by $k < n$, it follows immediately from Proposition 2.2.2 that to recover λ, we need choice probabilities under offer sets of size $n - k$ and larger.

Corollary 2.2.1. *In a GCC model, suppose that customers sample consideration sets of size at most k for some $1 \leq k \leq n$; that is, $\lambda(C) = 0$ whenever $|C| > k$. Then, the distribution over consideration sets λ can be identified using choice probabilities under offer sets of size $n - k$ or larger, that is, from the collection* $\{\mathbb{P}_0(S) \colon |S| \geq n - k\}$.

When the consideration sets are small, Corollary 2.2.1 shows that it is sufficient to collect choice probabilities for large offer sets. In many applications, however, firms cannot offer very large offer sets to its customers because of space constraint either in a physical store or on the website. The next proposition shows that when consideration sets are small, firms can identify λ by offering only small offer sets:

Proposition 2.2.4. *In a GCC model, suppose that customers sample consideration sets of size at most k for some $1 \leq k \leq n$; that is, $\lambda(C) = 0$ whenever $|C| > k$. Let $\{\mathbb{P}_0(S) \colon S \subseteq N, |S| \leq k\}$ be a collection of choice probabilities that are consistent with such a GCC model. Then, we have*

$$\lambda(C) = \sum_{X \subseteq N} \sum_{Y \supseteq X \cup C} (-1)^{1+|Y|-|X \Delta C|} \cdot \mathbf{I}[|X \cup C| \leq k < |Y|] \cdot \mathbb{P}_0(X),$$

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

The proof of the proposition is quite involved. It requires establishing several combinatorial identities. We present the proof in Section 5.2. The proposition shows that when the consideration sets are of size at most k, then the consideration set distribution can be recovered using choice probabilities of offer sets of size at most k.

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

Proposition 2.2.5. *The collection of choice probabilities $\{\mathbb{P}_j(S) \colon a_j \in S^+, S \subseteq N\}$ is consistent with a GCC model with unique parameters σ and consideration distribution λ such that $\lambda(C) > 0$ for all $|C| \leq 3$ if and only if it satisfies the following conditions:*

Condition 1. For all offer sets $S \subseteq N$ and $a_1, a_2 \in S$ such that $a_1 \neq a_2$: if $\mathbb{P}_1(S \setminus \{a_2\}) \neq \mathbb{P}_1(S)$, then it must be that $\mathbb{P}_2(S \setminus \{a_1\}) = \mathbb{P}_2(S)$.

Condition 2. For all offer sets $S, S' \subseteq N$ and $a_1, a_2 \in S \cap S'$ such that $a_1 \neq a_2$: if $\mathbb{P}_1(S \setminus \{a_2\}) \overset{(=)}{>} \mathbb{P}_1(S)$, then it must be that $\mathbb{P}_1(S' \setminus \{a_2\}) \overset{(=)}{>} \mathbb{P}_1(S')$.

Condition 3. For all offer sets $S \subseteq N$, we have that $\sum_{X \subseteq S} (-1)^{|S|-|X|} \mathbb{P}_0(N \setminus X) \geq 0$ with a strict inequality when $|S| \leq 3$.

Propostion 2.2.5 is similar to the set of conditions established in [68, Theorem 1] for the case when the consideration set distribution λ has the product form. Our result extends their result to a general consideration set distribution λ. Condition 1 is similar to the I-Asymmetry assumption in [68], which states that either product a_2 influences (note that the influence may either be an increase or decrease) the sales of product a_1 or vice versa but not both. In other words, influence is one directional and the products cannot influence the sales of each other. Condition 2 states that if product a_2 cannibalizes the sales of product a_1 is one offer set, then it must continue to do that in all the offer sets. That is, the direction of infleunce is consistent across all the offer sets. Condition 3 is a technical restriction to ensure the existence of a valid probability distribution function λ over the consideration sets. The strict inequality in Condition 3 is needed to ensure the preference list over products in N satisfies the transitivity requirement. The proof of Proposition 2.2.5 is presented in Section 5.2 in the Appendix. Establishing necessity is straightforward. But establishing sufficiency is challenging.

2.3 Data model and estimation methodology

We begin this section by focusing on a special case of GCC model, in which products enter consideration sets independently. We provide mixed integer non-linear programming (MINLP) formulation of the maximum likelihood estimation (MLE) problem to calibrate this model. Then, we show that solving MINLP can be reduced to solving a sequence of mixed integer linear programs (MILPs) and propose an outer-approximation and cutting plane algorithms (see Section 2.4) in order to implement it. We continue with a MINLP formulation of the MLE problem to infer GCC model parameters from sales transaction data. We propose the EM-based algorithm to estimate GCC model. Next, we demonstrate how to model consideration set formation with covariates, such as product features and price. In particular, we focus on three widely used methods in machine learning to describe consideration sets of customers: logistic-based, decision tree-based, and random forest-based consideration set models. We show that logistic-based consider-then-choose model can be calibrated using outer-approximation algorithm (see Section 2.4). We conclude the section by presenting two score metrics, which are used in this chapter to assess the prediction performance of the models, followed by the description of the benchmark.

2.3.1 Single class ICC model

We first consider a special case of the GCC model in which the consideration set distribution has a product form. Specifically, we assume that customers build a consideration set by tossing a coin for each product a_j and deciding to independently include it with probability $\theta_j \in [0, 1]$. It then follows that $\lambda(C) = \prod_{a_j \in C} \theta_j \prod_{a_j \notin C}(1 - \theta_j)$. We call this model the independent consider-then-chooses (ICC) model. This was also the model studied in [68]. It can be shown that under the ICC model, the probability of choosing product a_j from offer set S is given by

$$\mathbb{P}_j(S) = \theta_j \prod_{a_i \in S:\ a_i \succ_\sigma a_j} (1 - \theta_i).$$

We then formulate the maximum likelihood estimation problem for the ICC model and simplify it in such a way so that we can apply the outer-approximation algorithm in Section 2.4 in order to calibrate it. The data log-likelihood function under the single class ICC model is given by

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

and the MLE problem can be represented in the following way:

$$\max_{\boldsymbol{\theta}, \boldsymbol{\delta}} \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\delta}) \tag{2.3}$$

$$\text{subject to } \boldsymbol{\delta} \text{ satisfy } (2.15) - (2.17),$$

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

To simplify the likelihood function, we introduce a new variable $\boldsymbol{\tau}$, defined as $\tau_{jk} = \delta_{jk}\theta_k, \forall\ j, k,$ and rewrite the likelihood function as follows:

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\tau}) = \sum_{t=1}^{T} \Bigg[\log \theta_{j_t} + \sum_{\substack{a_k \in S_t: \\ k \neq j_t}} \log(1 - \tau_{j_t k}) \Bigg].$$

We can then formulate the MLE problem in terms of the variables $(\boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\theta})$:

$$
\begin{aligned}
& \max_{\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{\delta}} \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\tau}) && (2.4)\\
\textbf{s.t.:}\ & \tau_{jk} \leq \theta_k, \quad \forall\, j, k, && (2.5)\\
& \tau_{jk} \leq \delta_{jk}, \quad \forall\, j, k, && (2.6)\\
& \tau_{jk} \geq \theta_k + \delta_{jk} - 1, \quad \forall\, j, k, && (2.7)\\
& \tau_{jk} \geq 0, \quad \forall\, j, k, && (2.8)\\
& \delta_{jk} + \delta_{kj} = 1, \quad \forall\, j, k,\ j \leq k, && (2.9)\\
& \delta_{jk} + \delta_{kp} + \delta_{pj} \leq 2, \quad \forall\, j, k, p\ j \neq k \neq p, && (2.10)\\
& 0 \leq \theta_j \leq 1, \quad \forall\, j, && (2.11)\\
& \delta_{jk} \in \{0, 1\}, \quad \forall\, j, k. && (2.12)
\end{aligned}
$$

Estimation methodology

Note that the optimization problem (2.4) can be represented as the optimization problem (P) in Section 2.4.2 without loss of generality. As a result, we can formally apply the outer-approximation method [26] to solve the optimization problem (P), see Algorithm 1 in Section 2.4. The proposed algorithm effectively exploits its structure, where we have a linearity of the binary variables and convexity of the non-linear constraint, which only depend on continuous variables. 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. To this end, the outer-approximation defines the optimization subproblem as an MILP. Because of the potentially many continuous points required for outer-approximation, we solve a sequence of MILPs to build up increasingly tight relaxations of the original MINLP. Overall, the proposed Algorithm 1 consists of solving a finite sequence of convex problems and relaxed versions of MILPs.

2.3.2 GCC model

We start by formulating the associated maximum likelihood estimation problem for the GCC model. We assume access to sales data, which consists of the purchasing transactions over T

periods. Every purchasing instance is represented by a tuple (a_{j_t}, S_t) for $t \in \{1, ..., T\}$, where S_t denotes the subset of products offered in period t, and a_{j_t} denotes the product purchased by then. Next, we define binary linear ordering variable δ_{kj} for $a_k, a_j \in N,\ k \neq j$, where $\delta_{kj} = 1$ if product a_k goes before product a_j in the preference list $\succ$ (or equivalently, σ), and $\delta_{kj} = 0$ otherwise. The associated log-likelihood function is given by

$$\mathscr{L}(\boldsymbol{\delta}, \boldsymbol{\lambda}) = \sum_{t=1}^{T} \log\left(\sum_{C \subseteq N} \lambda(C) \cdot \mathbf{I}[a_{j_t} \in S_t \cap C] \cdot \mathbf{I}[\delta_{j_t k} = 1,\ \forall a_k \in S_t \cap C, k \neq j_t] \right). \tag{2.13}$$

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

$$\max_{\boldsymbol{\delta}, \lambda} \ \mathscr{L}(\boldsymbol{\delta}, \boldsymbol{\lambda}) \tag{2.14}$$

$$\textbf{s.t.:}\ \delta_{jk} + \delta_{kj} = 1, \quad \forall\ j, k,\ j < k, \tag{2.15}$$

$$\delta_{jk} + \delta_{kp} + \delta_{pj} \leq 2, \quad \forall\ j, k, p,\ j \neq k \neq p, \tag{2.16}$$

$$\lambda(C) \geq 0, \quad \forall\ C \subseteq N,$$

$$\sum_{C \subseteq N} \lambda(C) = 1,$$

$$\delta_{jk} \in \{0, 1\}, \quad \forall\ j, k. \tag{2.17}$$

The first set of equalities guarantees that either product a_j is preferred to product a_k in the rank list or product a_k is preferred to product a_j. The second set of constraints ensures a linear ordering of products. The third and fourth sets of inequalities ensure the validity of the probability distribution function λ over consideration sets.

Estimation methodology

We divide all the transactions into K segments such that for every segment $h \in \{1, ..., K\}$ a customer considers an arbitrary subset of items $C \subseteq N$ with likelihood $\lambda(C) = \prod_{a_j \in C} \theta_{hj} \prod_{a_j \notin C} (1 - \theta_{hj})$, where θ_{hj} is the probability to include item a_j in the consideration set in the segment h. Consequently, every segment of customers forms their consideration set probabilistically, and knowing the attention parameters $\boldsymbol{\theta}_h$, we can easily infer the “central” consideration set (i.e., the subset of items in the product universe that has the highest probability to be considered) for customers from the segment h, i.e., the “central” consideration set consists of all the items a_j in the product universe such that $\theta_{hj} \geq 0.5$. Recall that preferences of individuals are homogeneous, i.e., they can be characterized by a unique preference order σ.

A key observation is that, for sufficiently large K, we can calibrate GCC model by estimating the following three components: (1) segment probabilities, (2) attention parameters for each segment, and (3) the ranking. Therefore, this parametrization of the GCC model is a natural approach when we have sparsity in customer segments. In Section 2.4.3, we provide the detailed analysis of how to calibrate the GCC model with the EM algorithm. Note that if we have access to the panel data, then for each individual i we can estimate the posterior membership probabilities for each segment based on individuals' history of sales transactions. In the subsequent sections, we estimate the model for $K = 1, 2, ..., 5$ and report the best performance measure from these 5 variants, for every prediction metric that we introduce below. Note that we use the similar procedure to estimate the GCC model with features, with the only difference that customers from each segment h sample consideration sets according to the Logistic-based consider-then-choose (L-CC) model, which takes into account the feature representation of products (see Section 2.3.3).

2.3.3 Consider-then-choose models with features

Recall our general framework – the two-stage choice process, where consumers consider a subset of the offered products in the first stage and then, in the second stage, choose the most preferred product from the set considered. In this section, we revisit the first stage, i.e., the consideration set formation stage. We describe three types of consideration set formation models with features:

- *Logistic-based consider-then-choose model.* We assume that customers have linear-in-parameters utility U_j from considering product $a_j \in \mathcal{N}$, given by

$$U_j = \beta_j^0 + \sum_k \beta_k x_{jk} + \varepsilon_j,$$

where x_{jk} is the observed kth feature of product a_j; and ε_j is a random variable distributed as a standard logistics, i.e., $\varepsilon_j \sim \mathit{Logistic}(1)$. Therefore, product a_j is considered by an individual if and only if the utility from paying attention on it is non-negative, i.e.,

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

Then the attention probability of product a_j is given by

$$\Pr[a_j \in C] = \frac{\exp\left(\beta_j^0 + \sum_k \beta_k x_{jk}\right)}{1 + \exp\left(\beta_j^0 + \sum_k \beta_k x_{jk}\right)}.$$

- *Decision tree-based consider-then-choose model.* Here, it is assumed that individuals decide which items to consider based on a tree with leaves $m \in \{1, 2..., M\}$, to which we can associate a mean probability w_m of whether the item is going to be considered or not (see [74]). Then, we can write the probability to consider the item a_j in the following way:

$$\Pr[a_j \in C] = \sum_{m=1}^{M} w_m \mathbb{I}[\boldsymbol{x}_j \in R_m] = \sum_{m=1}^{M} w_m \phi(\boldsymbol{x}_j, \boldsymbol{v}_m),$$

where R_m is the mth region, i.e., the mth leaf; $\boldsymbol{v}_m$ encodes the choice of features to split on and the threshold value, on the path from the root to the mth leaf; and $\phi(\boldsymbol{x}_j, \boldsymbol{v}_m)$ is equal to 1 if $\boldsymbol{x}_j$ belongs to the mth leaf, and equal to 0 otherwise.

- *Random forest-based consider-then-choose model.* In this case, we assume that individuals, first, randomly sample a tree and then decide which items to consider based on the sampled tree (see [74]). Note that a random forest avoids the overfitting problem of decision trees by adding more trees instead of building one big tree. We can write the probability to consider the item a_j as follows:

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

where $f_k(\boldsymbol{x})$ is the probability to consider the item a_j according to the k'th decision tree.

Estimation methodology

In a similar spirit to the Section 2.3.1, we can formulate the maximum likelihood estimation problem for the logistic-based consider-then-choose model with product features in a such a way so that we can apply the outer-approximation algorithm in Section 2.4 in order to calibrate it (see Section 2.4.1). On the other hand, the calibration of DT-CC and RF-CC models is more challenging. To this end, we need to estimate both the ranking σ and a decision tree (or a random forest). Intuitively, both a decision tree and a random forest map product features into the binary outcome variable of whether the product is going to be considered or not, in non-linear way. Note that if the ranking σ is known, then the log-likelihood optimization is equivalent to calibrating a classification decision tree or a random forest, with the splitting criteria based on the entropy function. Then, given a decision tree and random forest, the log-likelihood optimization problem reduces to solving MILP to find σ. Therefore, the heuristic would be to repeat the following steps: (1) optimize log-likelihood function condition on $\hat{\sigma}$ by calibrating the decision tree or random forest, (2) optimize log-likelihood function condition on the calibrated decision tree or random forest and obtain the ranking σ, until either convergence (note that the convergence is not guaranteed) or hitting the maximum number of iterations.

2.3.4 Prediction scores and benchmark

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

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

where $f(a_j, S_t)$ is the probability to choose item j_t under the offer set S_t for the transaction at time t; and n_j is the observed number of times product a_j was purchased during the time period of length T. Note that we add 10 in the denominator to deal with undefined instances. The second score, RMSE, is given by

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{N}|}\sum_{a_j \in \mathcal{N}} \left(\frac{n_j - \hat{n}_j}{T}\right)^2}.$$

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

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

2.4 Estimation methodologies of consider-then-choose models

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

2.4.1 MINLP formulation: Logistic-based consider-then-choose model

In this subsection, we formulate the maximum likelihood estimation problem for the logistic-based consider-then-choose model, and then simplify it in such a way so that we can apply the outer-approximation algorithm in Section 2.4 in order to calibrate it. The data log-likelihood function under this model is given by

$$\mathscr{L}(\boldsymbol{\beta},\boldsymbol{\delta}) = \sum_{t=1}^{T}\left[\log\frac{e^{\boldsymbol{\beta}\boldsymbol{X}_{j_t}}}{1+e^{\boldsymbol{\beta}\boldsymbol{X}_{j_t}}} + \sum_{\substack{a_k\in S_t:\\ k\neq j_t}}\left[\delta_{j_tk}\log\frac{1}{1+e^{\boldsymbol{\beta}\boldsymbol{X}_{k_t}}}\right]\right],$$

and the ML problem can be represented in the following way:

$$\begin{aligned}
&\max_{\boldsymbol{\beta},\boldsymbol{\delta}}\ \mathscr{L}(\boldsymbol{\beta},\boldsymbol{\delta}) \qquad\qquad (2.18)\\
\textbf{s.t.:}\ \ &\delta_{jk}+\delta_{kj}=1, \quad \forall\ j,k,\ \ j\leq k,\\
&\delta_{jk}+\delta_{kp}+\delta_{pj}\leq 2, \quad \forall\ j,k,p\ j\neq k\neq p,\\
&0\leq\theta_j\leq 1, \quad \forall\ j,\\
&\delta_{jk}\in\{0,1\}, \quad \forall\ j,k.
\end{aligned}$$

To simplify the likelihood function, we introduce a new variable $\boldsymbol{\tau}$, defined as $\tau_{ijk}=\delta_{jk}\beta_i$, $\forall\ i,j,k$, and rewrite the likelihood function in the following way:

$$\mathscr{L}(\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{\tau}) = \sum_{t=1}^{T}\left[\log\frac{e^{\boldsymbol{\beta}\boldsymbol{X}_{j_t}}}{1+e^{\boldsymbol{\beta}\boldsymbol{X}_{j_t}}} + \sum_{\substack{a_k\in S_t:\\ k\neq j_t}}(\delta_{j_tk}-1)\log\left(\frac{1}{2}\right) + \sum_{\substack{a_k\in S_t:\\ k\neq j_t}}\left[\log\frac{1}{1+e^{\sum_i\tau_{ij_tk}\boldsymbol{X}_{ik_t}}}\right]\right],$$

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

$$(\delta_{j_tk}-1)\log\left(\frac{1}{2}\right)+\log\frac{1}{1+e^{\sum_i\tau_{ij_tk}\boldsymbol{X}_{ik_t}}} = \log\frac{1}{1+e^{\sum_i\tau_{ij_tk}\boldsymbol{X}_{ik_t}}} = \log\frac{1}{1+e^{\sum_i\boldsymbol{\beta}\boldsymbol{X}_{k_t}}},$$

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

$$(\delta_{j_tk}-1)\log\left(\frac{1}{2}\right)+\log\frac{1}{1+e^{\sum_i\tau_{ij_tk}\boldsymbol{X}_{ik_t}}} = -\log\left(\frac{1}{2}\right)+\log\frac{1}{1+e^{0}} = 0.$$

Let M denote the value of a very large constant. We can then formulate the MLE problem in terms of the variables (δ, β, θ):

$$\max_{\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{\delta}} \mathscr{L}(\boldsymbol{\beta},\boldsymbol{\tau}) + \sum_{\substack{a_k \in S_t: \\ k \neq j_t}} (\delta_{j_tk} - 1) \log\left(\frac{1}{2}\right) \tag{2.19}$$

$$\begin{aligned} \textbf{s.t.:}\ & \tau_{ijk} \leq \beta_i, \quad \forall\ i, j, k, \\ & \tau_{ijk} \leq M\delta_{jk}, \quad \forall\ i, j, k, \\ & \tau_{ijk} \geq \beta_i + M\delta_{jk} - M, \quad \forall\ i, j, k, \\ & \tau_{jk} \geq -M\delta_{jk}, \quad \forall\ j, k, \\ & \delta_{jk} + \delta_{kj} = 1, \quad \forall\ j, k,\ j \leq k, \\ & \delta_{jk} + \delta_{kp} + \delta_{pj} \leq 2, \quad \forall\ j, k, p\ j \neq k \neq p, \\ & \delta_{jk} \in \{0, 1\}, \quad \forall\ j, k, \end{aligned}$$

where

$$\mathscr{L}(\boldsymbol{\beta},\boldsymbol{\tau}) = \sum_{t=1}^{T} \Big[\log \frac{e^{\boldsymbol{\beta X}_{j_t}}}{1 + e^{\boldsymbol{\beta X}_{j_t}}} + \sum_{\substack{a_k \in S_t: \\ k \neq j_t}} \big[\log \frac{1}{1 + e^{\sum_i \tau_{ij_tk} \boldsymbol{X}_{ik_t}}} \big] \Big].$$

2.4.2 Estimation methodology: outer-approximation algorithm

The optimization problems (2.4) and (2.19) have a similar structure and can be represented as the following optimization problem without loss of generality:

$$\begin{aligned} \max_{\boldsymbol{\theta},\boldsymbol{\tau},\boldsymbol{\delta},\mu}\ & \mu \\ \textbf{s.t.:}\ & \mu_1 \leq \mathscr{L}(\boldsymbol{\theta},\boldsymbol{\tau}), \\ & A\boldsymbol{\theta} + B\boldsymbol{\tau} + C\boldsymbol{\delta} \leq 0, \\ & \mu = \mu_1 + E\boldsymbol{\delta}, \\ & \mu_L \leq \mu \leq \mu_U, \\ & \delta_{jk} \in \{0, 1\}, \quad \forall\ j, k, \end{aligned} \tag{P}$$

where $\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\tau})$ is a concave function.
Define, for given $(\boldsymbol{\theta}^i, \boldsymbol{\tau}^i)$:

$$D(\boldsymbol{\theta}^i, \boldsymbol{\tau}^i) = \{\boldsymbol{\theta}, \boldsymbol{\tau} : \mathscr{L}(\boldsymbol{\theta}^i, \boldsymbol{\tau}^i) + \frac{\partial \mathscr{L}(\boldsymbol{\theta}^i, \boldsymbol{\tau}^i)}{\partial \boldsymbol{\theta}} + \frac{\partial \mathscr{L}(\boldsymbol{\theta}^i, \boldsymbol{\tau}^i)}{\partial \boldsymbol{\tau}} - \mu_1 \geq 0,\ \mu \in \mathcal{R}^1\}$$

Define for given $\boldsymbol{\delta}^i$ the concave subproblem $S(\boldsymbol{\delta}^i)$:

$$\begin{aligned} \max_{\boldsymbol{\theta}, \boldsymbol{\tau}, \mu} \quad & \mu \\ \textbf{s.t.:} \quad & \mu \leq \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\tau}), \\ & A\boldsymbol{\theta} + B\boldsymbol{\tau} + C\boldsymbol{\delta}^i \leq 0. \\ & \mu = \mu_1 + E\boldsymbol{\delta}^i, \end{aligned} \tag{$S(\boldsymbol{\delta}^i)$}$$

Define, for given Ω^i, μ_L^i, and μ_U^i, the MILP subproblem M^i:

$$\begin{aligned} \max_{\boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\delta}, \mu} \quad & \mu \\ \textbf{s.t.:} \quad & (\boldsymbol{\theta}, \boldsymbol{\tau}) \in \Omega^i, \\ & A\boldsymbol{\theta} + B\boldsymbol{\tau} + C\boldsymbol{\delta} \leq 0, \\ & \mu = \mu_1 + E\boldsymbol{\delta}, \\ & \mu_L \leq \mu \leq \mu_U, \\ & \delta_{jk} \in \{0, 1\}, \quad \forall\ j, k. \end{aligned} \tag{M^i}$$

We can now formally apply the outer-approximation method [26] to solve the optimization problem (P), see Algorithm 1. The proposed algorithm effectively exploits the structure of the optimization problem (P) where we have a linearity of the binary variables and convexity of the non-linear constraint, which only depends on continuous variables. In order to linearize the optimization problem, we use the outer-approximation of a convex set by intersection of its collection of supporting half-spaces. To this end, the outer approximation defines the optimization problem (M^i) as MILP. Because of the potentially many continuous points required for outer-approximation, we solve a sequence of MILPs to build up increasingly tight relaxation of the original MINLP. Overall, the proposed Algorithm 1 consists of solving a finite sequence of convex problems $(S(\boldsymbol{\delta}^i))$ and relaxed versions of a MILP (M^i).

Note that Algorithm 1 to solve optimization problem (P) requires the solution of both convex optimization problem $(S(\boldsymbol{\delta}^i))$ and MILP (M^i). The solution of the convex optimization problem $(S(\boldsymbol{\delta}^i))$ in each iteration might be computationally intensive; while in solving the MILP (M^i), the computational work, on the other hand, might be more moderate, because for every iteration i we need to solve the MILP problem (M^i), which is the previous MILP problem (M^{i-1}) with only one additional linear constraint added. Therefore, we propose to use the cutting plane algorithm to solve the MINLP in this case [100], which would require the solution of only the finite sequence of MILP problem (M^i), see Algorithm 2. Even though the main iteration loop of Algorithm 1 is, generally, more efficient, we have global convergence for both Algorithms 1 and 2.

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

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

    return $(\boldsymbol{\theta}^i, \boldsymbol{\tau}^i, \boldsymbol{\delta}^i)$.
```

Empirical validation of the algorithms

In this subsection, we analyze the performance of the outer-approximation algorithm 1 and cutting plane algorithm 2 to estimate ICC model with IRI Academic dataset. We limited the running time of the algorithms by 3 hours, and the precision was set to 1e-6. It follows from Figure 2-1 that the optimality gap of the outer-approximation algorithm 1 to calibrate the ICC

Algorithm 2 Cutting plane algorithm for optimization problem (P)

```
procedure CUTTING PLANE(P)
$\Omega^0 = \mathcal{R}^n \times \mathcal{R}^m$, $\mu_L = -\infty$, $\mu_U = \infty$, $i = 1$
Select arbitrary $\boldsymbol{\delta}^1$, i.e., it can be arbitrary full ranking
Select arbitrary $\boldsymbol{\lambda}^1$, i.e., it can be arbitrary distribution over consideration sets
Set $\boldsymbol{\tau}^1 = \boldsymbol{\delta}^1 \cdot \boldsymbol{\lambda}^1$, $\mu_U^0 = -\infty$, $\mu_U^1 = \infty$
    while $\left|\mu_U^i - \mu_U^{i-1}\right| > \varepsilon$ do
        Set $\Omega^i = \Omega^{i-1} \cap D(\boldsymbol{\theta}^i, \boldsymbol{\tau}^i)$
        Solve MILP subproblem $M^i$ such that $\mu_U = \mu^*$, $\mu_U^i = \mu^*$ (i.e., the optimal objective
    function of $M^i$), and $(\boldsymbol{\theta}^i, \boldsymbol{\tau}^i, \boldsymbol{\delta}^i) = (\boldsymbol{\theta}^*, \boldsymbol{\tau}^*, \boldsymbol{\delta}^*)$ (i.e., the optimal solution of $M^i$))
        $i = i + 1$

    return $(\boldsymbol{\theta}^i, \boldsymbol{\tau}^i, \boldsymbol{\delta}^i)$.
```

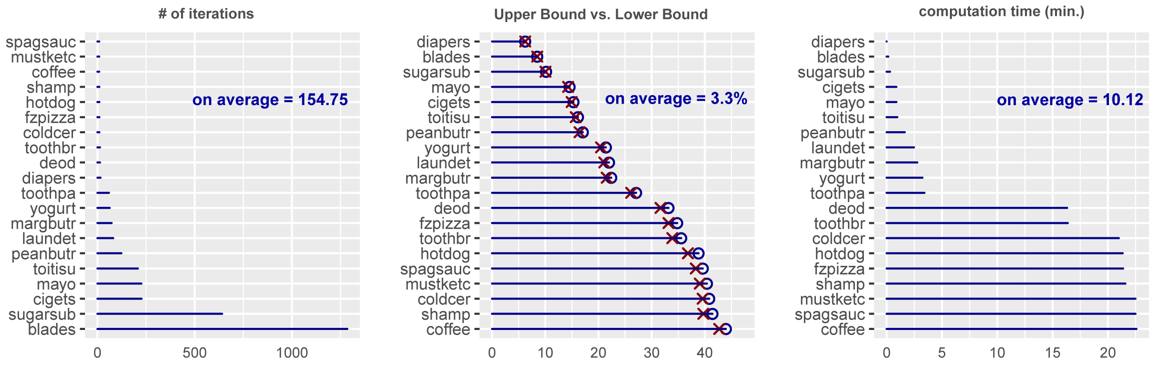

Figure 2-1: Results of applying the outer-approximation algorithm.

model is 3.3%, on average, over 20 product categories. On the other hand, it is shown in Figure 2-2 that the optimality gap of the cutting plane algorithm 2 to calibrate the ICC model is 4.5%, on average, over 20 product categories. Following these findings, we apply outer-approximation algorithm 1 to calibrate ICC model in this chapter, as it provides significantly faster convergence to the optimal solution, which is consistent with previous studies.

2.4.3 GCC estimation methodology: EM algorithm

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

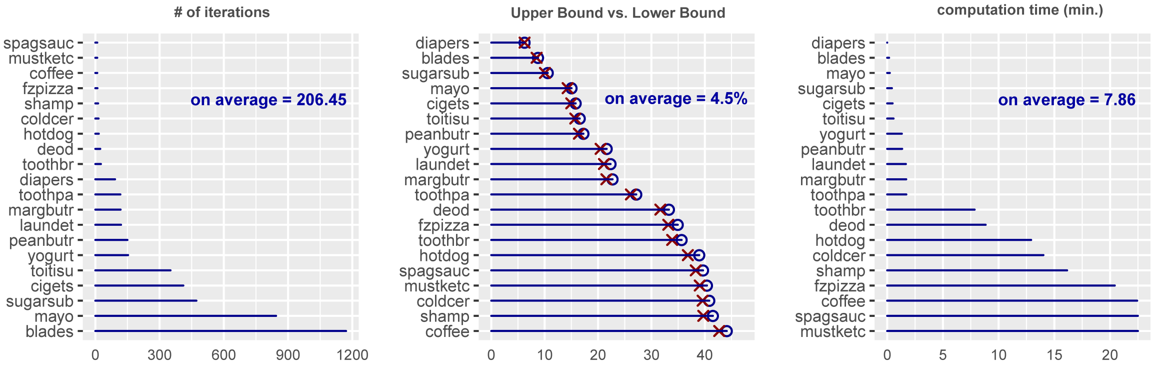

Figure 2-2: Results of applying the cutting plane algorithm.

Estimation with aggregate level data

The log-likelihood function to calibrate the GCC model, after we reparametrize it by dividing all the transactions into K segments, is given by

$$\log \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\gamma}, \succ_{\boldsymbol{\sigma}}) = \sum_{t=1}^{T} \log \Bigg(\sum_{h=1}^{K} \gamma_h \theta_{h,j_t} \prod_{\substack{a_j \in S_t: \\ a_j \succ a_{j_t}}} (1-\theta_{hj}) \Bigg), \tag{2.20}$$

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

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

Initialization: we initialize the EM with a random allocation of observations to one of the K classes, resulting in an initial allocation $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$, which form a partition of the collection of all the transactions. Then, we set $\gamma_h^{(0)} = |\mathcal{D}_h|/(\sum_{d=1}^{K} |\mathcal{D}_d|)$. Next, $\succ$ (i.e., σ) and $\theta_{hj}^{(0)}$, for all

$h \in \{1, ..., K\}$, $a_j \in N^+$ are obtained by solving the following optimization problem:

$$\max_{\succ, \boldsymbol{\theta}_h} \sum_{t \in \mathcal{D}_h} \Big(\log \theta_{h,j_t} + \sum_{\substack{a_j \in S_t: \\ a_j \succ a_{j_t}}} \log(1 - \theta_{hj}) \Big),$$

which is solved by using the outer-approximation algorithm in Section 2.4.2.

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

$$\begin{aligned}
P_{ht}^{(q)} &= \Pr\Big(t \to h \Big| \succ^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}, (a_{j_t}, S_t)|_{t=1}^{T}\Big) \\
&= \Pr\Big(t \to h \Big| \succ^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}, (a_{j_t}, S_t)\Big) \quad [\text{ independence of purchases}] \\
&= \frac{\Pr\Big((a_{j_t}, S_t)\Big| \succ^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}, t \to h\Big) \cdot \Pr\Big(t \to h \Big| \succ^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}\Big)}{\Pr\Big((a_{j_t}, S_t)|\Big| \succ^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}\Big)} \\
&= \frac{\Pr\Big((a_{j_t}, S_t)\Big| \succ^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}, t \to h\Big) \cdot \Pr\Big(t \to h \Big| \succ^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}\Big)}{\sum_{r=1}^{K} \Pr\Big((a_{j_t}, S_t)\Big| \succ^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}, t \to r\Big) \cdot \Pr\Big(t \to r \Big| \succ^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}\Big)} \\
&= \frac{\gamma_h^{(q-1)} \Big[\theta_{h,j_t}^{(q-1)} \prod_{\substack{a_j \in S_t: \\ a_j \succ^{(q-1)} a_{j_t}}} (1 - \theta_{hj}^{(q-1)})\Big]}{\sum_{r=1}^{K} \Big[\gamma_r^{(q-1)} \Big(\theta_{r,j_t}^{(q-1)} \prod_{\substack{a_j \in S_t: \\ a_j \succ^{(q-1)} a_{j_t}}} (1 - \theta_{rj}^{(q-1)})\Big)\Big]},
\end{aligned}$$

where "$t \to h$" denotes " transaction at time t belongs to the segment h". As a result, conditional expected value of the log-likelihood function is given by

$$\sum_{h=1}^{K} \sum_{t=1}^{T} P_{ht}^{(q)} \log \Big(\theta_{h,j_t} \prod_{\substack{a_j \in S_t: \\ a_j \succ a_{j_t}}} (1 - \theta_{hj})\Big).$$

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

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

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

$$\max_{\succ, \boldsymbol{\theta}} \sum_{h=1}^{K} \sum_{t=1}^{T} P_{ht}^{(q)} \log \Big(\theta_{h,j_t} \prod_{\substack{a_j \in S_t: \\ a_j \succ a_{j_t}}} (1 - \theta_{hj}) \Big),$$

which is solved using outer-approximation algorithm in Section 2.4.2.

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 m customers:

Initialization: we initialize the EM with a random allocation of individuals to one of the K classes, resulting in an initial allocation $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$, which form a partition of the collection of all the individuals. Then, we set $\gamma_h^{(0)} = |\mathcal{D}_h| / (\sum_{d=1}^{K} |\mathcal{D}_d|)$. Next, $\succ$ (i.e., σ) and $\theta_{hj}^{(0)}$, for all $h \in \{1, ..., K\}$, $a_j \in N^+$ are obtained by solving the following optimization problem:

$$\max_{\succ, \boldsymbol{\theta}_h} \sum_{i \in \mathcal{D}_h} \Big(\log \theta_{h,j_{it}} + \sum_{\substack{a_j \in S_{it}: \\ a_j \succ a_{j_{it}}}} \log(1 - \theta_{hj}) \Big),$$

which is solved by using the outer-approximation algorithm in Section 2.4.2.
E-step: we compute $P_{hi}^{(q)}$, which is the membership probability of every individual i to belong to the segment h based on the parameter estimates $\{\succ^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}\}$ and the purchasing

transactions data $(a_{j_{it}}, S_{it})|_{t=1}^{T_i}$:

$$P_{hi}^{(q)} = \frac{\gamma_h^{(q-1)} \prod_{t=1}^{T_i} \left[\theta_{h,j_{it}}^{(q-1)} \prod_{\substack{a_j \in S_{it}: \\ a_j \succ^{(q-1)} a_{j_{it}}}} (1-\theta_{hj}^{(q-1)})\right]}{\sum_{r=1}^{K} \left[\gamma_r^{(q-1)} \prod_{t=1}^{T_i} \left(\theta_{r,j_{it}}^{(q-1)} \prod_{\substack{a_j \in S_{it}: \\ a_j \succ^{(q-1)} a_{j_{it}}}} (1-\theta_{rj}^{(q-1)})\right)\right]}.$$

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

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

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

$$\max_{\succ, \boldsymbol{\theta}} \sum_{i=1}^{m} \sum_{h=1}^{K} P_{hi}^{(q)} \sum_{t=1}^{T_i} \log \Big(\theta_{h,j_{it}} \prod_{\substack{a_j \in S_{it}: \\ a_j \succ a_{j_{it}}}} (1-\theta_{hj})\Big),$$

which is solved using outer-approximation algorithm in Section 2.4.2.

EM algorithm heuristics

The EM algorithm, proposed above, might become computationally challenging for the large-scale problems, as we need to run an outer-approximation algorithm for every qth iteration. Alternatively, we might further assume that the preference order $\succ_\sigma$ over items in the product universe is known, e.g., we can rank the products according to their popularity in the sales transaction data or we can estimate the ranking from calibrating single class ICC model (see Section 2.3.1). In this case the "M" step for qth iteration in the EM algorithm reduces to solving a globally concave maximization problem with a unique, closed form solution (i.e., we don't need to apply outer-approximation algorithm), given by

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

which can be applied using aggregate level data (see Section 2.4.3), and

$$\boldsymbol{\theta}_{hj}^{(q)} = \frac{\sum_{i=1}^{m} \sum_{t=1}^{T_i} P_{hi}^{(q)} \mathbb{I}[a_{j_{it}} = a_j]}{\sum_{i=1}^{m} \sum_{t=1}^{T_i} P_{hi}^{(q)} \mathbb{I}[a_{j_{it}} = a_j] + \sum_{i=1}^{m} \sum_{t=1}^{T_i} P_{hi}^{(q)} \mathbb{I}[a_j \in S_{it}, a_j \succ a_{j_{it}}]},$$

which can be applied if we have access to the panel data (see Section 2.4.3).

2.5 Conclusion

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

In the spirit of the consider-then-choose framework, we assume that customers make a purchasing decision in two stages. First, a boundedly rational consumer, who suffers from limited attention, forms her consideration/competition set, which is usually a small subset of substitutable items in the product category, due to cognitive limitations. Secondly, a consumer evaluates all products in her consideration set and purchases the one that is the most preferred. Focusing on the fist stage of the choice process, we propose an effective means of modeling the consideration set formation using existing machine learning methods. Although consideration sets are unobservable, our modeling approach allows us to infer the most likely subset of items considered by each individual, depending on past purchasing transactions.

Chapter 3

Robustness of Demand Prediction Models in Operations

3.1 Introduction

Demand prediction is critical when optimizing prices and planning retail operations as well as when matching supply and demand on online platforms. The fundamental unit of analysis under this choice-based demand paradigm is the customer. In the context of retail, customers are subjects creating product demand, and understanding the driving forces behind their decision-making process allows us to build better demand models, and, hence, to make better operational decisions. In the context of the sharing economy, where services are exchanged between private individuals in online platforms, understanding customer preferences leads to more effective matchings between parties. In pursuing such objective, accounting for the consideration set of the consumers (i.e., the set of products really evaluated by consumers prior to making a choice) is indeed a fundamental input. One of the intuitive properties of consideration set identification is the fact that the exclusion of non-considered items from the offer set does not impact the choice of customers. For this reason, choice-based demand models accounting for consideration sets of customers might be more robust to the errors in offer set definitions.

Note that the classical discrete choice models [5, 59, 92] from Random Utility Maximization (RUM) class, including the pioneering work of [64], imply that customers consider all products available from the offer set. Without this simplifying assumption, models from the RUM class that attract major attention among scholars and practitioners (e.g., the multinomial logit (MNL), nested logit (NL), and latent class logit (LC) models) become difficult to estimate. However, this common assumption is acknowledged to be an overestimation of the cognitive burden exercised by a customer when facing an assortment. Evidence suggests that customers' limited attention leads to the two-stage consider-then-choose process to purchase an item given limited physical and cognitive abilities in considering the full offer set, no matter how noisy its definition is. Under this framework, during the first stage they eliminate a few alternatives through a simple screening rule, and then choose from the remaining options. Items that are ruled out during the first step are clearly not going to be purchased. Instead, items from the consideration sets are extensively evaluated by a customer based on their attributes.

The ability of consider-then-choose models to account for both customer preference and cognitive limitation is essential in practice. First of all, identifying whether sales volume depends mainly on the customer's evaluation of the product or on the customer's attention is important for practitioners and might result into developing different business strategies to improve sales. Secondly, explicitly accounting for the consideration set formation of customers, the consider-then-choose types of models proposed in Chapter 2 are likely to be robust to the noise in sales transactions data. It is a common practical issue that noise corrupts product availability data (e.g., in retail settings, stockout events mask the true offer set information). Moreover, companies (e.g., retail companies and online platforms) need to make long-term demand predictions in order to optimize strategic marketing decisions and to be successful in the long run. To this end, the company can not rely on the accurate data regarding product availability over time in the distant future. In this case, it is very important that choice models are robust to errors in the offer set.

In this chapter, we evaluate the prediction performance of various choice-based demand models. Specifically, we showcase the robustness of consider-then-choose models to the noise in offer set definitions under different real-world scenarios and noise regimes, using the modeling framework proposed in Chapter 2. Overall, we make the following contributions:

- *Numerical experiments: robustness to noise in offer sets.* Because the consideration set formation of customers is explicitly modeled in a consider-then-choose type of framework, it is highly likely that its predictive performance is robust to the noise in the definition of the offer sets in comparison with their classical counterparts (e.g., MNL, LC-MNL). We verify this proposal by explicitly incorporating the noise factor in synthetic sales transaction datasets and by testing the predictive performance of choice models under different noise scenarios.

- *Empirical analysis: better demand predictions for the retail industry.* We compare choice models under several real-world scenarios in retail when we are likely to face significant noise in offer set definitions. Our findings suggest that the relative performance of our model as opposed to certain benchmarks improves once we switch to scenarios with higher noise levels. We see only a moderate decrease in prediction accuracy when increasing noise for consider-then-choose models.

- *Empirical analysis: applications in the context of online platforms.* We also analyze a dataset obtained from an online car-sharing platform to provide additional evidence that consider-then-choose models are generally robust to the noise in the offer set definition in comparison to classical choice models. After modeling the consideration set formation of customers with the linear-in-attributes utility, we demonstrated significant improvement of the proposed models in terms of predictive performance. However, the flexibility of our framework allows us to estimate the choice model with non-linear in product attributes formation of consideration sets (e.g., decision trees or random forests). As a result, applying a non-linear approach in consider-then-choose models gives us a further boost in prediction performance. For example, after calibrating and testing models with the industry partner dataset, we obtain that the random forest-based consider-then-choose model outperforms the benchmark by 43.3% in terms of the MAPE score, and by 53.7% in terms of the RMSE metrics.

3.2 Study based on the synthetic data: robustness to the noise in offer sets

In this section, we describe the results of an extensive simulation study, the main purpose of which is to demonstrate that choice models based on the consider-then-choose framework are more robust to the noise in offer sets than their classical counterparts. We consider the case when the offer sets are not perfectly observed, and we focus our analysis on understanding when the consider-then-choose type of models are better equipped to handle offer set noise than other popular models in the literature, such as the MNL model.

To streamline the analysis of this simulation study, we consider the following setting. Suppose that the ground truth model is the MNL choice model. Customers have perfect information on the offer set S, and they consider all the items on offer. Given the offer S, the customer chooses product a_j with probability $v_j / \left(1 + \sum_{a_i \in S} v_i\right)$, where the parameter $v_i > 0$ is the "weight" or the attraction value corresponding to product a_i. The modeler observes customers choices, but does not observe the offer set perfectly. In the presence of such noise, we compare the predictions of the MNL model against ICC model to understand the conditions under which the one outperforms the other.

In our setup, the benchmark MNL model does not suffer from model misspecfication – only the noise in the offer sets. Our model, on the other hand, suffers from model misspecification. For instance, it follows from Condition 1 in Propostion 2.2.5 that cannibalization is one directional in our model, whereas it is bi-directional in the ground-truth model. Comparing the predictions of our model to that of the benchmark MNL model should then allow us to quantify the effect of offer set noise, because in the absence of nosie, the benchmark MNL model should provide perfect predictions. The results, we obtain, can only be more favorable to our model when the offer sets are noisy, unlike what we are assuming in the ground-truth model.

3.2.1 Synthetic data generation process

Let $\gamma \in [0, 1]$ be the noise depth, such that each item in the product universe is exposed to the noise with probability γ, i.e., if n is the number of items in the product universe then γn is the average number of items that are exposed to the noise. Next, let η be the noise intensity, such that η is the conditional probability of an item to be in the offer set as a noisy observation, given that the item is exposed to the noise. The higher γ and η the more noise is added to the dataset. Deterministic utility values for items $a_j \in N$ are randomly chosen from the interval $[1, 2]$, i.e., $v_j \sim U[1, 2]$, assigning 0 to the utility from the outside option. We assume that we have $n = 15$ items in the product universe. Then, given the parameter values for γ, η, and parameters $\boldsymbol{v}$ of the MNL model, the data-simulation procedure consists of the following steps:

1. We randomly sample 100 offer sets, i.e., $\{S_m\}|_{m=1}^{100}$.

2. For each offer set S_m, we generate 10000 sales transactions according to the MNL model with parameter values $\boldsymbol{v}$.

3. We generate 100 sets $\{\bar{S}_m\}|_{m=1}^{100}$, such that each set $\bar{S}_m$ is obtained by tossing a coin for each product $a_j \in N$ and including it in $\bar{S}_m$ with probability γ.

4. We transform the sales transactions data by adding extra items with probability η, for every offer set S_m, if these items belong to the $\bar{S}_m$, i.e., we modify every offer set S_m of sales transactions data such that every item $a_j \in N \setminus S_m$ is added into the offer set with probability η if $a_j \in \bar{S}_m$. For instance, if $\eta = 0.5$ and $\bar{S} = N$ then for every transaction, characterized by the tuple (a_j, S_m), we modify offer set S_m by adding on average half of the items from the subset $N \setminus S_m$.

Using the procedure above, we generate both training and test synthetic datasets in order to test the hypothesis that ICC choice model is rather robust to the noise in the offer sets in comparison with the MNL model. We generate transaction data (i.e., both training and test datasets) for $\gamma \in \{0.05, 0.1, ..., 1\}$ and $\eta \in \{0.1, 0.2, ..., 1\}$. This parametrization leads to 200 different scenarios. We simulated 100 different instances of the sales transaction data for each of those scenarios.

3.2.2 Results and discussion

In Figure 3-1, we present the heatmap of the prediction scores under MNL model, where each column corresponds to a particular noise intensity η and each row corresponds to a particular noise depth γ. We focus on the MAPE and RMSE prediction scores in left and right panels, respectively. Recall that MNL is the ground truth model for this simulation study. As expected, MNL model captures the ground-truth choice probabilities almost exactly when $\gamma = 0.05$ and $\eta = 0.1$, i.e., there is only a small amount of noise added to the sales transactions data. We observe that MNL prediction scores increase with higher noise intensity for a given noise depth. Interestingly, it can also be seen that MNL prediction scores are not monotonic with respect to the noise depth, i.e., the scores first increase with higher noise depth and then decrease. In order to better explain the variation of prediction scores by the noise depth and noise intensity variables, in Figure 3-1, we run the following linear regression:

$$Score_{ij} = \beta_0 + \beta_1 \cdot Intensity_i + \beta_2 \cdot Intensity_i^2 + \beta_3 \cdot Asymm_j + \beta_4 \cdot Shared_j + \varepsilon_{ij}, \qquad (3.1)$$

where $Intensity_i$ is the noise intensity such that $Intensity_i \in \{0.1, 0.2, ..., 1\}$, $Asymm_j$ is probability that an item in the product universe is exposed to noise only in the test data set or only in the training dataset for $j \in \{1, 2, ..., 20\}$, and $Shared_j$ is probability that an item in the product universe is exposed to noise both in the test and training datasets for $j \in \{1, 2, ..., 20\}$. Note that $Asymm_j = \gamma_j(1-\gamma_j) + (1-\gamma_j)\gamma_j = 2\gamma_j(1-\gamma_j)$ and $Shared_j = \gamma_j^2$, where γ_j is the noise depth such that $\gamma_j \in \{0.05, 0.1, ..., 1\}$. The results for the regression (3.1) are reported in the last column in Table 3.1. It follows from the Table 3.1 that noise intensity deteriorates the predictive performance of the MNL model in non-linear way, with the coefficient for the linear term being positive, and the coefficient for the quadratic term being negative. The variables *Asymm* and *Shared* are positively correlated with the MAPE score, i.e., the prediction performance of the MNL model worsens as the number of items in the product universe, which are exposed to the noise, increases. Interestingly, the variable *Asymm* has more than seven times higher economic significance than the variable *Shared*, which indicates that the benchmark (i.e, MNL model) struggles the most in making accurate predictions when the impact of the noise is asymmetric between the training and test sales transactions.[1] Note that the independent variables in the regression model (3.1) explain most of the variation in the MAPE score under the MNL model, i.e., $R^2_{adj} = 0.93$.

In Figure 3-2, we present the heatmap of the prediction scores improvements under ICC model versus MNL model, where each column corresponds to a particular noise intensity η and each row corresponds to a particular noise depth γ. We focus on the MAPE and RMSE prediction scores in left and right panels, respectively. Non-surprisingly, ground-truth MNL model significantly outperforms ICC model when there is only a small amount of noise added to the sales transactions data, e.g., $\gamma = 0.05$ and $\eta = 0.1$. As expected, our model cannot capture the ground-truth choice

[1] Note that the highest noise asymmetry is achieved when γ is equal to 0.5. We have the worst performance under MNL model for this level of the noise depth γ, see Figure 3-1.

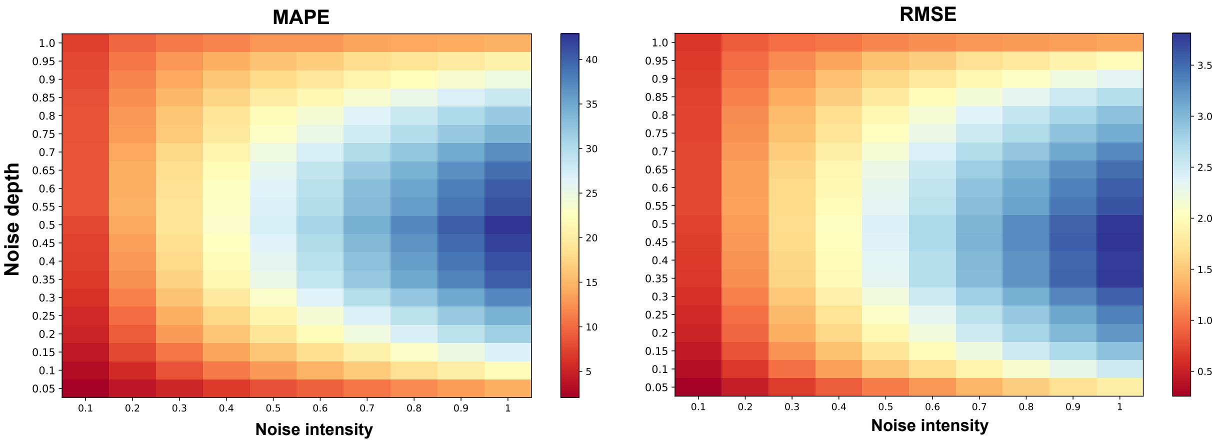

Figure 3-1: Heatmap of the prediction scores under MNL model.

probabilities exactly because of model misspecification.

Figure 3-2 reveals that the improvement of ICC prediction scores over MNL are not monotonic with respect to the noise depth (i.e., the scores first increase with higher noise depth and then decrease) and noise intensity (i.e., the scores first increase with higher noise intensity and then decrease). To better explain the variation of prediction improvements by the noise depth and noise intensity variables, in Figure 3-2, we run the following linear regression which is similar to the regression (3.1):

$$Score\_Impr_{ij} = \beta_0 + \beta_1 \cdot Intensity_i + \beta_2 \cdot Intensity_i^2 + \beta_3 \cdot Asymm_j + \beta_4 \cdot Shared_j + \varepsilon_{ij}. \quad (3.2)$$

The results for the regression (3.2) are presented in the last column in Table 3.2. It follows from the table that the improvement of ICC model over MNL increases with noise intensity non-linearly such that the coefficients for the linear and squared terms are positive and negative, respectively. Moreover, the improvement is positively correlated with number of items in the product universe that are exposed to the noise. Since the coefficient corresponding to the variable *Asymm* is more than twice higher than the coefficient corresponding to the variable *Shared*, we conclude that ICC model has a higher chance to outperform the benchmark in scenarios when the sets of items that are exposed to the noise in training and test datasets do not intersect. Note that in some of the real world scenarios, we are likely to have more noise in the hold-out sample than in the training dataset. Results in this section are robust to these scenarios as well (see Figure 3-3).

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
	Score	Score	Score	Score	Score
Intensity	27.336***				42.079***
	(17.997)				(14.246)
Intensity^2		22.929***			-13.403***
		(15.755)			(-5.122)
Asymm			34.662***		39.320***
			(8.677)		(28.338)
Shared				-2.590	5.385***
				(-1.143)	(8.015)
const	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 statistics in parentheses

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

Table 3.1: Regression models where the dependent variable is the MAPE score under the MNL model.

Interestingly, it can be inferred form Tables 3.1 and 3.2 that the dependent variables (i.e., *Intensity*, *Asymm*, and *Shared*) impact the improvement of ICC over MNL in the same way, qualitatively, as they impact the MNL prediction scores. As a result, it can be stated that the ability of ICC model to outperform MNL model is higher in scenarios when MNL model struggles to provide accurate predictions.

3.3 Empirical study on the IRI academic dataset

In the real world settings, noise is likely to result in an estimate of the offer set that is a superset of the true offer set. This type of noise implies that we may not know the offer set exactly, but we can always come up with a superset of the true offer set. In fact, this is true in retail settings, where stock out events mask the true offer set information.

In this section, we compare the predictive power of GCC, ICC, and the latent-class MNL (LC-MNL) benchmark models based on the household purchase panel and store data from the IRI Academic Dataset [12] under different real-world scenarios. This panel dataset keeps track

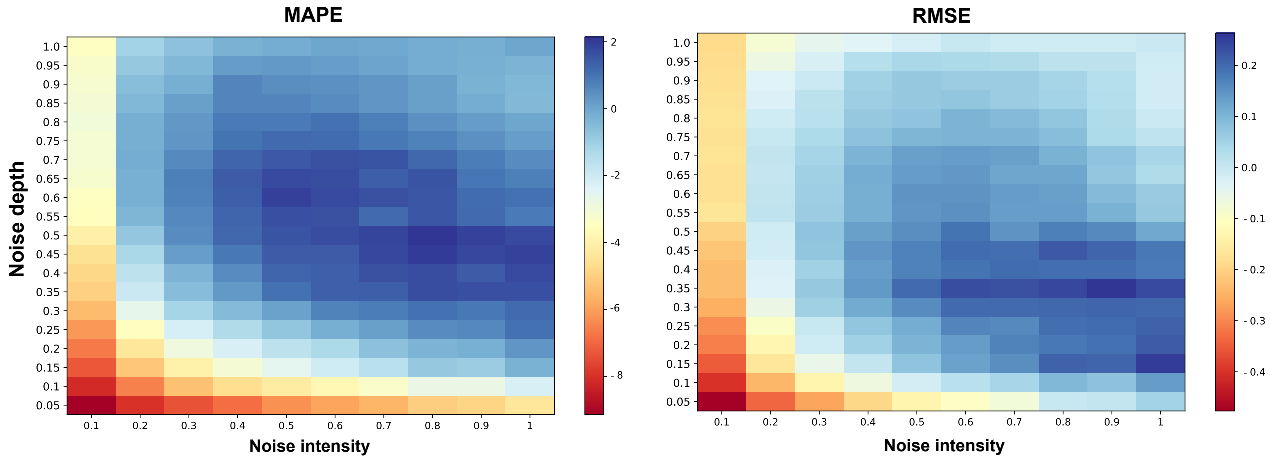

Figure 3-2: Heatmap of the prediction scores improvements under ICC model versus MNL.

of the household purchase histories for grocery and drug store chains, collected from 47 markets across the US over the years 2001-2011. Note that we calibrate GCC model by applying the EM algorithm with panel data, where we rank items in the product universe according to their popularity in the sales transaction data for every category (see Section 2.4.3 for details).

Overall, the main purpose of this empirical study is threefold: (a) provide various real-world scenarios based on the IRI dataset when we are likely to face a lot of noise in the offer set definitions while making the long-term demand predictions; (b) investigate the prediction performance of choice models under different noise regimes, e.g., quantify the improvement of GCC over LC-MNL under several real-world scenarios with various noise intensities; and (c) compare GCC model to ICC model, which is a restricted version of the GCC model.

According to the empirical study, the improvement of GCC versus MNL increases as we add more noise to the dataset. In other words the predictive performance of GCC model is robust to the noise in the sales transaction data in comparison with the classical LC-MNL choice model. We also find that GCC model significantly outperforms ICC model in prediction performance.

3.3.1 Data analysis

The dataset consists of weekly sales transactions. We analyze a total of 20 categories, presented in Table 3.3. We focus on sales transactions data from calendar year 2007. For every store visit, we are given the following information: the Universal Product Code (UPC) and price of the

	Model (1) Impr.	Model (2) Impr.	Model (3) Impr.	Model (4) Impr.	Model (5) Impr.
Intensity	4.162***				16.621***
	(7.883)				(17.214)
Intensity^2		3.025***			-11.327***
		(6.157)			(-13.241)
Asymm			7.311***		11.530***
			(7.112)		(25.421)
Shared				2.539***	4.878***
				(4.811)	(22.207)
const	-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 statistics in parentheses

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

Table 3.2: Regression models where the dependent variable is the MAPE score improvement of the ICC model over the MNL model.

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 the explicit information about the subset of items offered to each individual upon her store visit, we construct this subset by aggregating all of the transactions made in a particular store withing a given category during a particular week. We aggregate items with the same vendor code into a single product since we cannot work directly with each UPC because of data sparsity. We divide the sales transaction data into two parts: the training data, which consists of the first 26 weeks of the sales observations, and the test data, which consists of the last 26 weeks of the sales observations.

3.3.2 Results and discussion

We stress test the demand prediction models based on the IRI dataset under different scenarios in the retail setting. Overall, we focus on three scenarios:

1. *No extra noise.* We assume that the retailer has an accurate information on the product

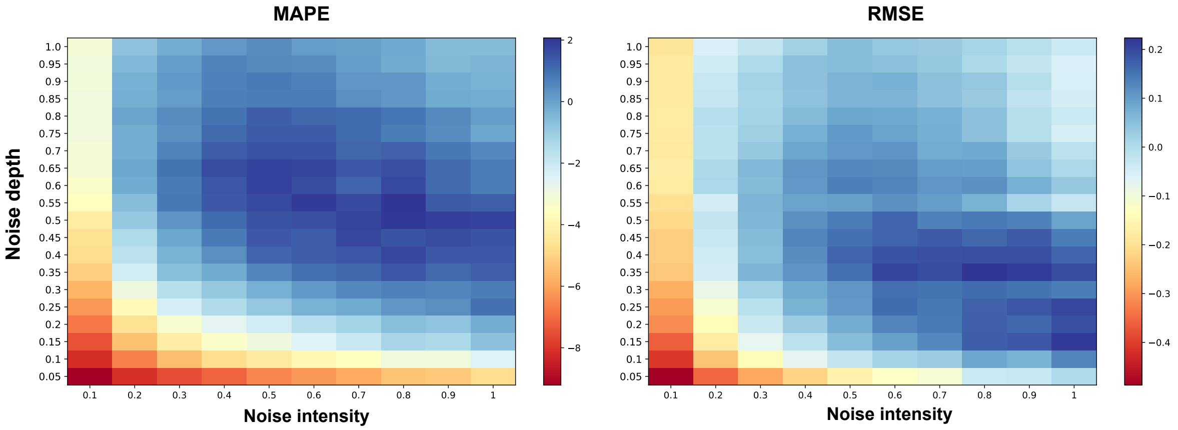

Figure 3-3: Heatmap of the prediction scores improvements under ICC model versus MNL. Noise depth in the training dataset is 10% lower than in the test dataset.

assortments. To this end, we make predictions by feeding the models with accurate offer sets, which are obtained from the test dataset.

2. *Store-based noise.* In order to be successful in major strategic and investment decisions, a retail company needs an accurate long-term demand forecasting. In this case, the robustness of the demand prediction models to the noise in offer sets is very crucial, as retailers usually do not have accurate information on product stockouts in the distant future. To this end, we assume that a store manager makes demand predictions given the overall assortment of products in their store. We obtain the product assortments for each store by taking the union of all the offer sets over the test data. We model this scenario by feeding the prediction models with the estimated product assortments for every store.

3. *Week-based noise.* 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 on the centralized level, without knowing the up-to-date information on product assortments in every store. Instead, the warehouse might know the estimate of the product assortments across all the stores over a specified time period. We model this scenario by feeding the prediction models with the union of product assortments over all the stores in the retail chain on weekly basis.

	Category Shorthand	Expanded Name	# vend	OS size	# cust.	# trans.
1	blades	Blades	10	4.18	703	1084
2	cigets	Cigarettes	18	7.14	452	2343
3	coffee	Coffee	73	19.80	3101	11526
4	coldcer	Cold cereal	45	17.66	4438	26701
5	deod	Deodorant	36	14.55	1345	2383
6	diapers	Diapers	8	3.30	337	919
7	fzpizza	Frozen pizza	47	15.50	3460	13431
8	hotdog	Hot dogs	44	16.81	3318	8886
9	laundet	Laundry detergent	24	10.08	3196	8698
10	margbutr	Margarine/Butter	19	10.35	3474	14596
11	mayo	Mayonnaise	19	6.86	3761	8676
12	mustketc	Mustard	60	17.07	3728	9238
13	peanbutr	Peanut butter	25	7.99	3153	8059
14	shamp	Shampoo	81	18.74	1466	2884
15	spagsauc	Spaghetti/Italian sauce	74	17.85	3473	11879
16	sugarsub	Sugar substitutes	17	5.05	750	1406
17	toitisu	Toilet tissue	13	7.66	3760	14411
18	toothbr	Toothbrushes	52	15.86	1115	1810
19	toothpa	Toothpaste	38	12.05	2110	4482
20	yogurt	Yogurt	32	9.84	3766	24096

Table 3.3: Summary statistics of the data used in IRI case study.

In Figure 3-4, we present scatter plots of the improvements of GCC model versus LC-MNL model across 20 product categories, under three scenarios discussed above: (1) no extra noise added, represented by green pluses; (2) store-based noise added, represented by blue crosses; and (3) week-based noise added, represented by red dots. In the left and right panels, we measure the predictive performance of the models based on the MAPE and RMSE metrics, respectively. We observe that GCC outperforms LC-MNL for around half of product categories even under the first scenario (i.e., no extra noise is added to the sales transaction data) and for almost all product categories under the second and third scenarios. Note that we have red dots located to the right of green pluses with blue crosses being in between, under both MAPE and RMSE scores and across most of the product categories. It reveals that the improvement of GCC over LC-MNL across product categories increases when we switch to the noisy regimes.

Then, Figure 3-5 exhibits MAPE (see left panel) and RMSE (see right panel) scores of GCC and LC-MNL models, averaging across 20 product categories, for three different noise regimes. We observe that the performance of the LC-MNL model deteriorates once we shift from the "no

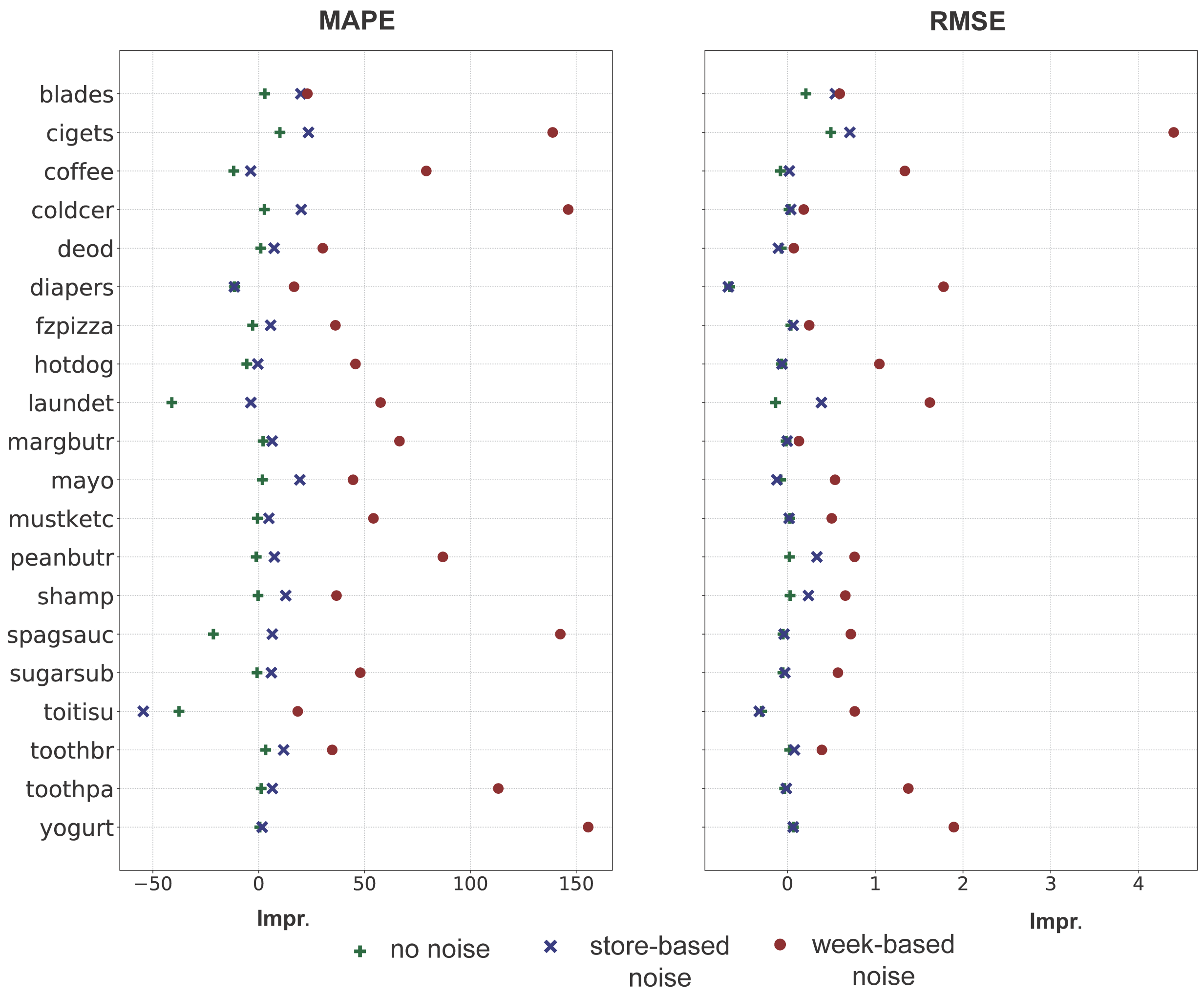

Figure 3-4: Scatter plots of the prediction score improvements of GCC model over LC-MNL.

extra noise" to the "store-based noise" and "week-based noise" scenarios. On the other hand, the predictive performance of the GCC model only moderately decreases once we switch to the noisy regimes, i.e., the performance stays rather flat for all three noise regimes. From the panels in Figure 3-5, we observe that the improvements of GCC over LC-MNL are -5.6% (-0.027%), 3.7% (0.061%), and 68.4% (0.973%) under the first, second, and third noise regimes, respectively, based on the MAPE (RMSE) score.

We notice that the improvement of GCC over MNL, in Figure 3-4, varies across product categories for a given noise regime. To better explain this variation, we regress the improvement of GCC over LC-MNL for each category against the noise intensity. We measure the noise

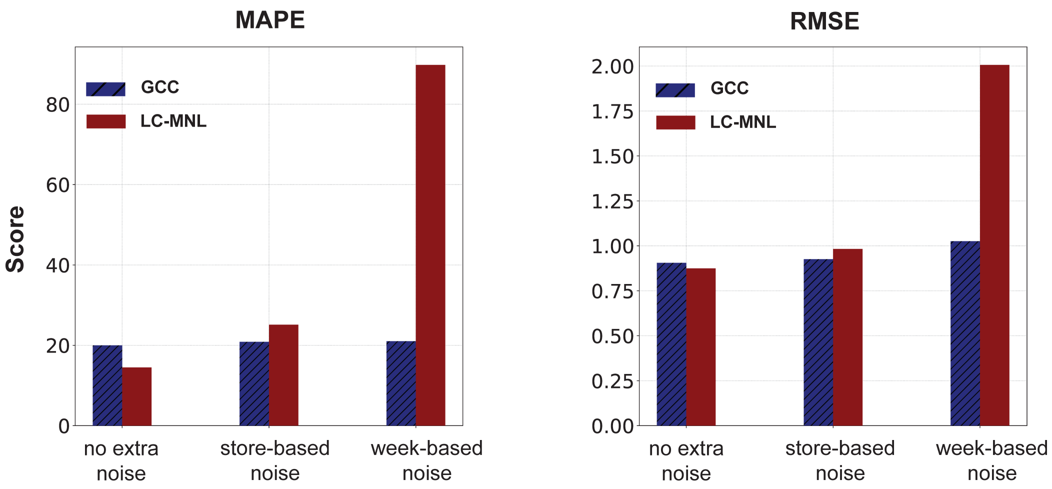

Figure 3-5: The average prediction scores under GCC and LC-MNL choice models.

intensity in the following way:

$$\text{noise intensity} = \frac{1}{T}\sum_{t=1}^{T}\frac{|\tilde{S}| - |S|}{|\tilde{S}|},$$

where S is the offer set from the actual sales transactions data and $\tilde{S}$ is a noisy offer set. Intuitively, the noise intensity under "store-based noise" scenario is equal to the average percentage of the items that are stocked out in a store for a given category. The noise intensity under "week-based noise" scenario is equal to the average percentage of the items that are available in a store over the total assortment of items that are available in the retail chain for a given week. The left and right panels in Figure 3-6 illustrate the regression under "store-based noise" and "week-based noise" regimes, respectively. We see a clear positive correlation between the improvement of GCC over MNL and noise intensity in both panels, which suggests that the improvement becomes more significant with higher noise intensity in the product category.

In Figure 3-7, we compare the prediction performance of GCC against ICC model based on the MAPE score across 20 product categories. We observe that the relative improvement of GCC model over ICC is 18.5%, 18.1%, and 8.9% under the first, second, and third noise scenarios, respectively, on average, across 20 product categories.

We emphasize two major findings here. First, we observe that the relative predictive performance of GCC over LC-MNL improves with the noise added to the dataset. This finding is

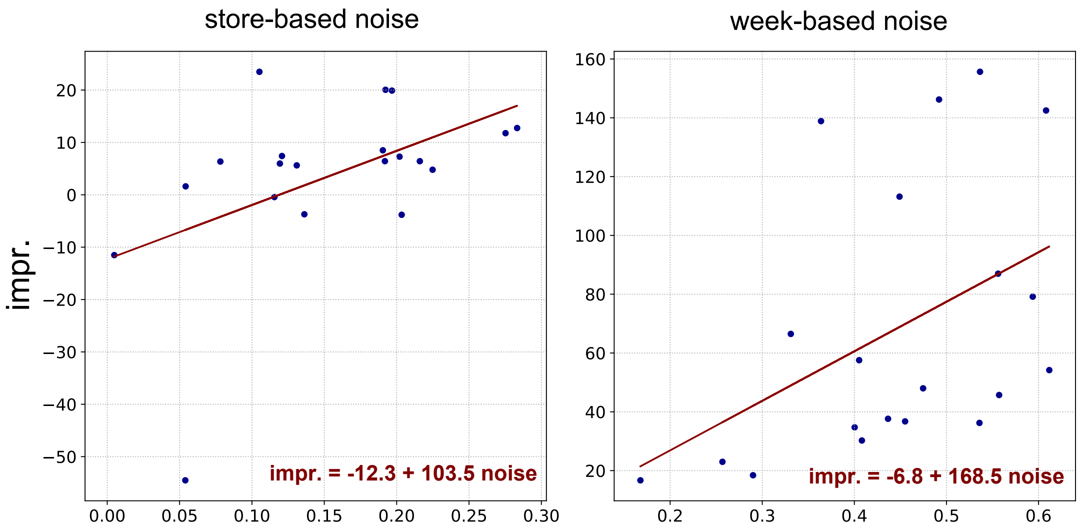

Figure 3-6: Scatter plots and liner regressions of the MAPE score improvement of ICC (vs. MNL model) over the noise intensity.

consistent with the analysis in Section 3.2 based on the synthetic dataset. Moreover, the improvements vary significantly across product categories and noise regimes. Second, we find that GCC significantly outperforms the ICC model in terms of the prediction performance as it can better capture heterogeneity of customer preferences. Overall, our main observation is that the predictive performance of GCC model is rather robust to the noise in the sales transaction data in comparison with standard LC-MNL model.

3.4 Case study on the car-sharing dataset: prediction analysis

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

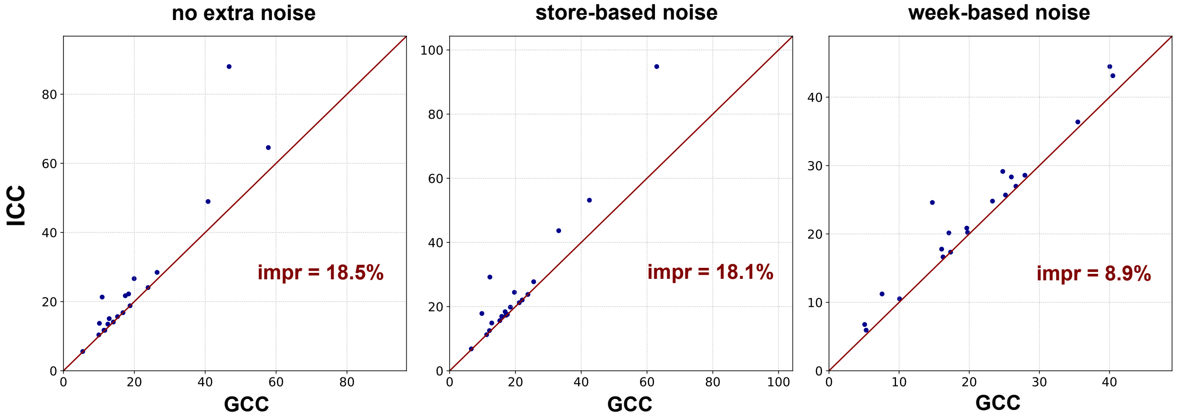

Figure 3-7: Scatter plots of the MAPE scores under the GCC vs. ICC models.

3.4.1 Industry partner and data analysis

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

For the empirical analysis in this section, we use a historical dataset, which includes a sample of the rentals completed in a major US city over a period of two years. Each observation in the dataset is a rental (i.e., a renter who books the listed car from a particular location given the set of available alternatives on a specific day/time). Our dataset includes 26.8K rentals from around five hundred car providers. For each rental, we have access to several observable features, such as car owner ID, hourly rental price, car access (i.e., open or closed), car location hours (i.e., 24 hours or restricted), car location type (i.e., garage, street, surface lot, or valet), car brand (e.g, BMW, Tesla, MINI), car types (i.e., economy, standard, fullsize, 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 Section 3.5.1, we examine the extent to which various features, specified above (e.g., hourly rental price), impact the consideration/competition set structure of renters. A detailed summary of the data is provided in Table 3.4.

3.4.2 Modeling assumptions

Our modeling assumptions are motivated by the desire to strike a balance between the flexibility of the feature-based consider-then-choose models and their tractability. In principle, we use a semiparametric approach in order to calibrate these two-stage models. In the first stage, the renters form their consideration set and we represent their utilities from considering a car a_j with linear-in-parameters function u_j, i.e., $u_j = \boldsymbol{\beta}^T x_j + \varepsilon_j$, or using non-linear methods from machine learning, e.g., decision trees or random forests. Then, we assume that during the second stage the renter chooses the most preferred car, among the considered ones, according to the preference order $\boldsymbol{\sigma}$ over the universe of car alternatives. Modeling the second stage choice process this way, we do not parameterize the ranking $\boldsymbol{\sigma}$ which implies that the cars are assumed to have the same attributes over time. However, according to our dataset, this assumption is justified (see Section 3.5.2 for the details).

Note that the dataset consists of the rental request observations such that for every transaction we know which car was reserved and we can infer the set of available cars, listed in the online platform at the time of the request, with their characteristics. The offer sets are approximately built by aggregating all of the listed and available cars within 0.3 miles distance form the location of the car, which was in fact rented. In general, in order to calibrate feature-based consider-then-choose models (e.g., GCC model) with our dataset, we need to estimate two types of parameters: the ranking $\boldsymbol{\sigma}$ over all the cars (i.e., in total 514 cars) listed in the online platform and parameters associated with consideration set formation of renters. In order to simplify the estimation procedure for this case study, we assume that the ranking $\boldsymbol{\sigma}$ is known a priori. Specifically, the cars are ranked according to their popularity among renters, defined as the number of times the vehicle was rented.

	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%

Table 3.4: Descriptive statistics, the car-sharing dataset.

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

Table 3.5: Additional descriptive statistics, the car-sharing dataset.

3.4.3 Feature-based predictive accuracy results

Next, we conduct an out-of-sample prediction testing of the models to quantify the predictive performance of consider-then-choose models versus the benchmark while taking into account the car attributes. We split the dataset into two parts: the first 80%, in-sample, rental observations, and the remaining 20%, out of sample transactions. Overall, we compare the predictive performance of the LC-MNL benchmark with three variants of our model class: GCC, Decision Tree-based CC (DT-CC), and Random Forest-based CC (RF-CC), on the accuracy of two prediction measures: MAPE and RMSE (see Section 2.3.4), where lower scores stand for better prediction.

The peer-to-peer car-sharing platform can calibrate the choice-based models to make demand predictions using the training dataset, which consists of car rentals with each rental observation being a tuple (a_{jt}, S_t), where a_{jt} is the chosen car and S_t is the set of cars available at the reservation time t. In order to optimize strategic and marketing decisions, the online platform need to make long-term (or medium-term) demand forecasts for the cars listed on the online application. In the real world settings, the company can not rely on the accurate data on car availabilities over time in the distant future, i.e., we can not test prediction power of choice models by using the offer sets from the test dataset described above. Instead, the company might divide the city into several geographical areas and make predictions based on the aggregate assortment of cars listed in each area. For our case study, we divide the city in 42 equal-spaced areas and estimate the assortments of cars by taking the superset of all the cars on offer at each area in

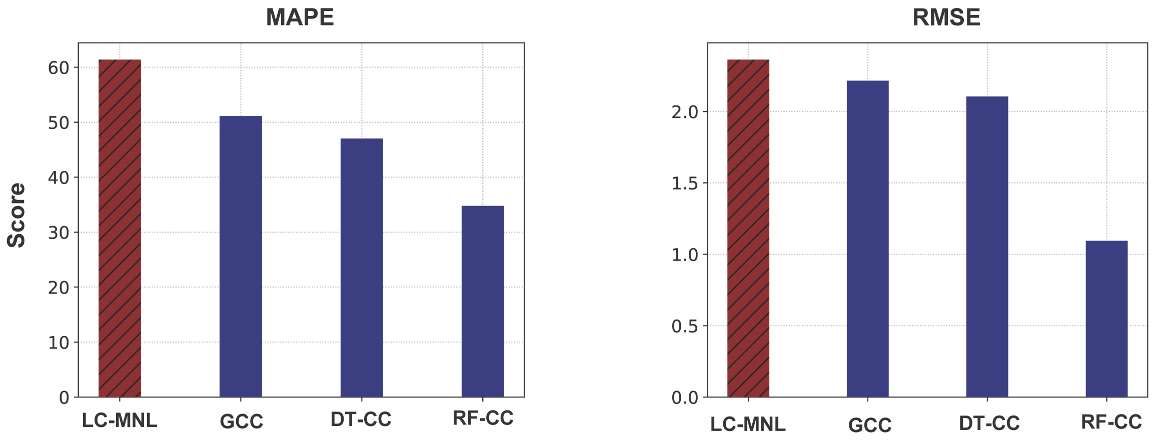

Figure 3-8: Prediction results under Consider-then-Choose (CC) and LC-MNL models with car features.

the hold-out data sample.

In Figure 3-8, we present the prediction performance results of the models based on MAPE and RMSE scores in the left and right panels, respectively, averaged across car brands. The MAPE score of consider-then-choose models exhibit an improvement of 16.7%, 23.4%, and 43.3%(RF-CC) over LC-MNL for GCC, DT-CC, and RF-CC models, respectively. We also observe that our model combinations obtain improvements of 6.2%, 10.9%, and 53.7% over LC-MNL for GCC, DT-CC, and RF-CC models, respectively, based on RMSE metrics.

Figure 3-9 exhibits the MAPE scores computed for every brand separately under the RF-CC and LC-MNL models, where the brands are ordered according to their popularity (i.e., percentage of the total number of reservations coming from every brand), e.g., Honda is the most popular brand while Mercury is the least popular brand in the dataset. The panels in Figure 3-9 illustrate that MAPE scores vary significantly across brands both for RF-CC and LC-MNL models. To analyze this variation, in the Figure 3-10, we regressed the improvement of RF-CC versus LC-MNL against the popularity of brands and MAPE score of LC-MNL model. We observe a clear positive correlation between MAPE score improvements and popularity of brands, which indicates that we can better predict the demand for more popular brands. We can also see a clear positive correlation between the improvements and MAPE score under LC-MNL model, which allows us to conclude that consider-then-choose type of models are especially relevant in prediction tasks (i.e., CC models dominate LC-MNL) when LC-MNL model provides a

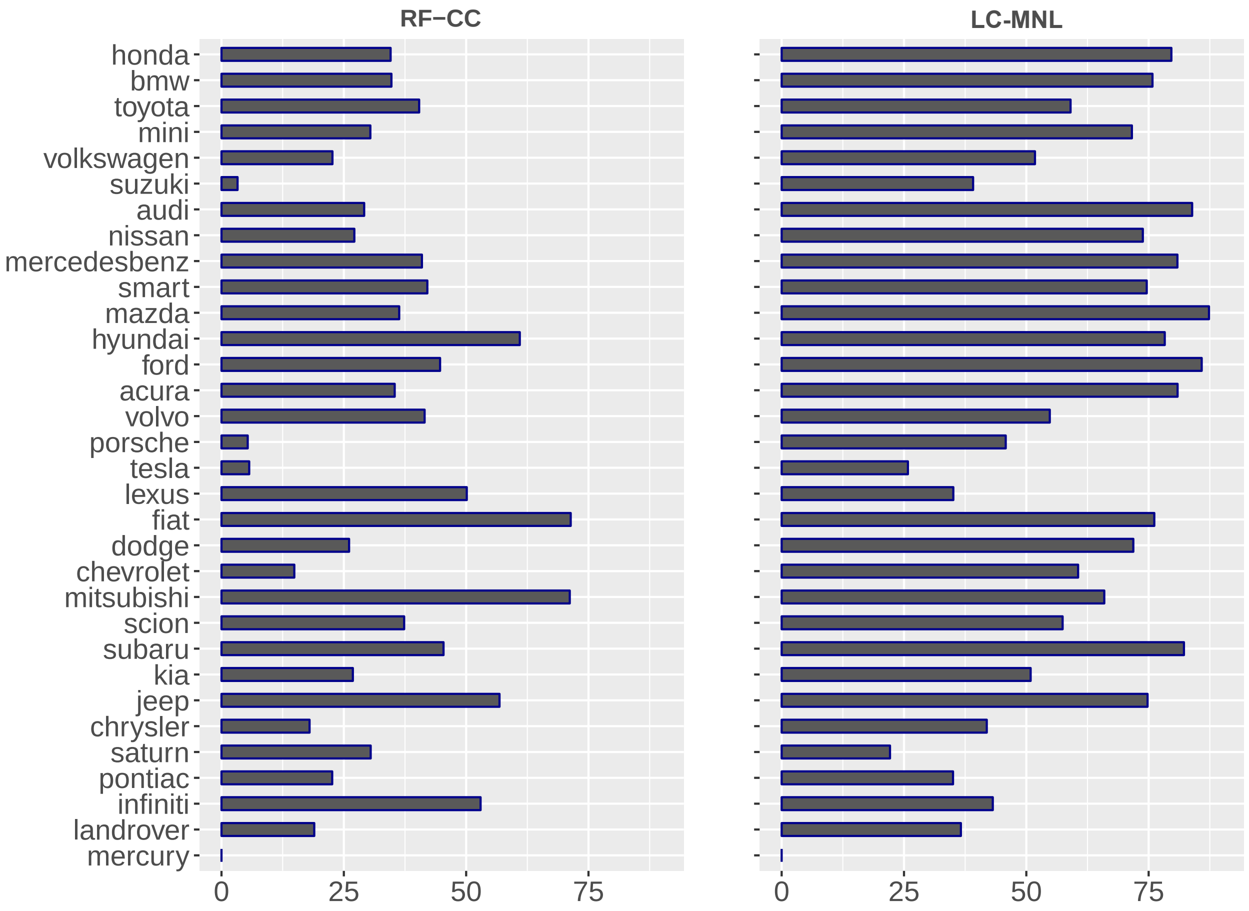

Figure 3-9: MAPE scores of car brands.

relatively bad prediction performance. To this end, the LC-MNL model provides a relatively bad prediction for the brands that are influenced by the noisy observations the most. Being robust to the noise, consider-then-choose models provide significantly better predictive performance under these circumstances. Note that these insights are consistent with our numerical study based on the synthetic dataset in Section 3.2.

The results above indicate that consider-then-choose models forecast customers' choices considerably better than the traditional LC-MNL model under both prediction scores. First of all, accounting for the consideration set formation with the linear-in-parameters GCC model with logistically distributed error term, we can better predict the choices of customers. This improvement can be attributed to the effectiveness of consider-then-choose models to alleviate the noise impact on the offer set definition from sales transaction data. Moreover, we can further boost

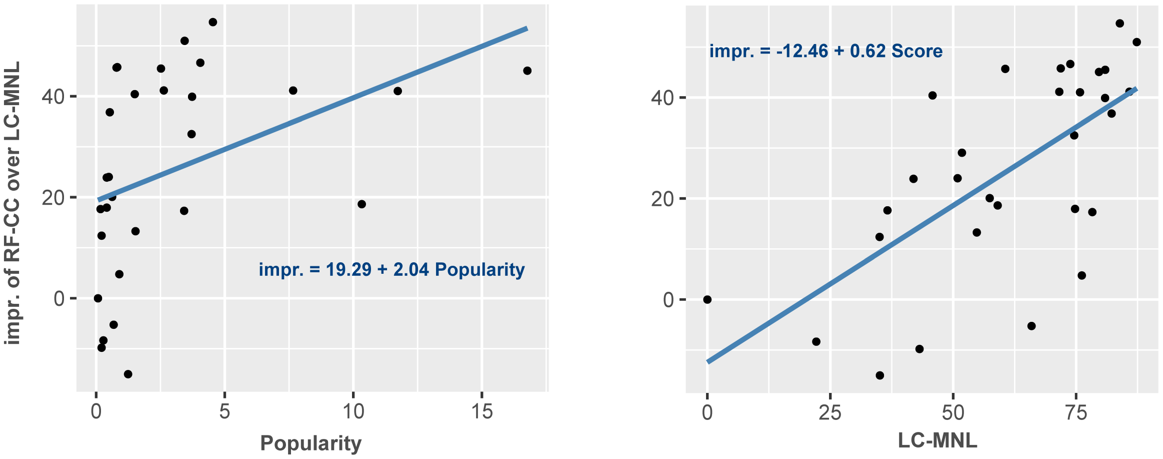

Figure 3-10: Scatter plot and linear regression of the percentage improvement of RF-CC versus LC-MNL over brand popularity and LC-MNL prediction accuracy.

the predictive performance of the two-stage models by modeling the consideration set formation in a non-linear-in-parameters way, with decision trees or random forests. We take it as a strong supportive evidence for the validity of inferring consideration sets from transaction data with consider-then-choose models. After calibrating DT-CC and RF-CC models we can get some insights of how customers form their consideration sets. In particular, Figure 3-11 illustrates an instance of the decision tree obtained after fitting the DT-CC model.

3.5 Case study on the car-sharing dataset: explanatory analysis

In this section, we calibrate Logistic-based Consider-then-Choose (L-CC) and MNL models, accounting for car features, and discuss the modeling assumptions. We also provide explanatory analysis of choice models in order to gain insights about 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.

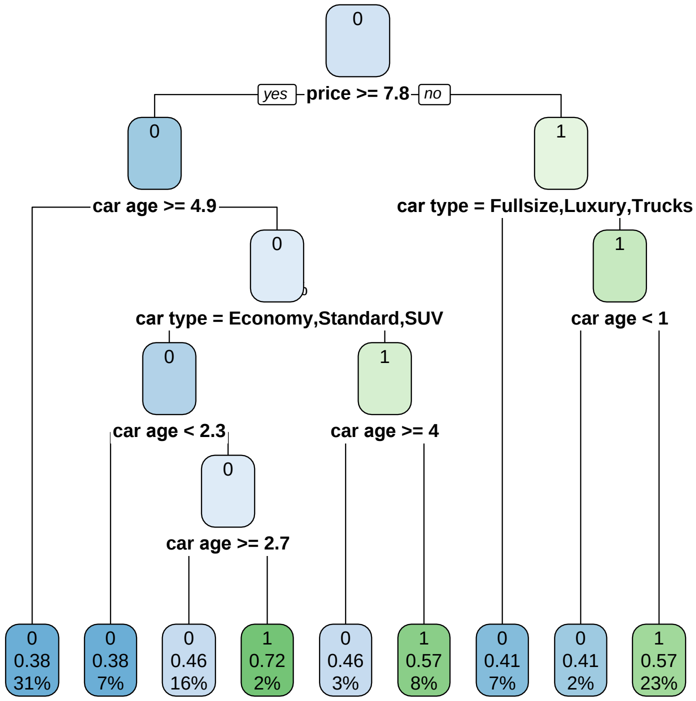

Figure 3-11: Decision tree for consideration set formation of the renters based on the car-sharing dataset.

3.5.1 Explanatory analysis

We start this section by calibrating the L-CC model with features to examine the extent to which various variables impact the consideration/competition set structure. Assuming that the cars are ranked according to their popularity among renters (see Section 3.4.2), the problem of fitting L-CC model is the one of estimating the coefficients $\boldsymbol{\beta}$. The car features available to the renters through the online platform are divided into three groups: (1) car brand; (2) car location type and accessibility: 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); (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 ε_j are logistically distributed, we estimate the $\boldsymbol{\beta}$ vector using logistic regression analysis.

The results for the L-CC model appear in the first column of Table 3.6. In the middle column, the table lists the average marginal effects (AME) of the L-CC model when all the covariates are at their mean. Then, we also calibrate the usual linear-in-parameters MNL model where the utility from reserving the car alternative j is represented with-linear-in parameters function U_j, i.e., $U_j = \boldsymbol{\beta}^T x_j + \varepsilon_j$. On the right, the Table 3.6 presents the estimates of the MNL model parameters. However, the interpretation of the β vector for L-CC and MNL models is different. The parameters of the L-CC model, listed in the left column of Table 3.6, show the estimated impact of exogenously imposed changes in car features on the consideration set formation. Rather, the parameters of the MNL model in the right column of Table 3.6 show the influence of car features on the customer's choices, i.e., revealed preferences. Notably, quite a few coefficients estimated based on L-CC and MNL models are not aligned, i.e., the covariates that increase (or decrease) the likelihood of considering the car under L-CC model might not necessarily increase (or decrease) the likelihood of booking the car under the MNL model, e.g., the utility of the renter from considering the car brand Jeep is higher by 0.98 ($t = 4.9$, $p < 0.01$) than the utility from considering the baseline brands while the utility of the renter from reserving the same brand under the MNL model is lower by 0.93 ($t =$ -6.2, $p < 0.01$) than the utility from reserving baseline brands. Also, some of the covariates that are statistically significant in explaining the choice of renters under MNL model might be statistically insignificant under L-CC model, e.g., the utility from choosing a car parked in the street is lower ($t =$ -8.33, $p < 0.01$) than the utility from choosing the car located in the valet parking area while the discrepancy between these two parking location types are insignificant ($t = 1.36$, $p > 0.10$) under the L-CC model. The price and the car age coefficients are statistically significant and negative for both L-CC and MNL models. However, the impact of additional \$1 increase in the car hourly rental price on the utility from considering the vehicle is equivalent to the car being 0.52 years older, while the impact of additional \$1 increase in the car hourly rental price on the utility from booking the vehicle under the MNL model is equivalent to the car being 3.75 years older. According to these findings, the car age plays relatively more important role during the formation of the consideration set in the L-CC model compared to its role in the choice process under the MNL model.

Next, we consider three types of car attributes (i.e., car brand, car location type and accessibility, and car type and features), with the objective of empirically verifying their impact on the consideration set formation under the L-CC model and on the choice probabilities under the MNL model. According to Table 3.7, the car type and features attributes are more statistically significant than car brand attributes under the L-CC model, whereas the opposite effect takes place under the MNL model. These findings are robust to the various measures of statistical significance and goodness-of-fit presented in Table 3.7 such as LL, AIC, BIC, Likelihood Ration (LR) statistics, and Wald statistics. Overall, it is implied that car location type and accessibility play the least important role both for the consideration set formation and for the final choice decision. However, the renters are likely to build their consideration sets based on car brands rather than on car properties, even though while evaluating alternatives regarding their overall choice, customers are likely to pay more attention on the car properties rather than the car brands.

3.5.2 Discussion of model estimation assumptions

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

Semiparametric approach

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

We start by analyzing the variation of the hourly price parameter over car alternatives. In Table 3.5, we report that the average coefficient of variation (CV) of the hourly price across all the

	L-CC		AME (L-CC)		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					

$^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 3.6: Logistics-based Consider-then-Choose (L-CC) and MNL model estimation results, the car-sharing dataset.

		Excluded groups	Log-like	AIC	BIC	LR	Wald
L-CC	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

Table 3.7: Statistical significance of three groups of car attributes.

car alternatives is around 5%, while owners of cars listed, on average, only around two different values of the price. Moreover, the most frequently used value of the hourly price corresponds to 78% of the car rentals and the second most frequently used value of the hourly price corresponds to 16% of the car rentals. The low variation of the rental price is explained by the policies of the online platform, for the time span of the dataset, that allows the owners to choose the price by themselves, i.e., the platform as a central agent did not dynamically adjust the listed rental price to efficiently match demand and supply as opposed to many ride-sharing platforms (e.g., Uber, Lyft), which optimize the price of the ride to match riders with drivers on-demand. Finally, in the same Table 3.5, 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).

Price endogeneity problem

Next, we want to address the concerns of potential price endogeneity in our empirical analysis. First of all, estimating the demand with personalized data significantly alleviates the price endogeneity problem, since each renter has only a trivial influence on the number of cars supplied and the market rental price, while the empirical work with aggregate level transaction data is more likely to face a very sever endogeneity issues. Nevertheless, having access to individual consumer data is not always a big advantage because individuals' demand could be correlated. For example, we might have unobservable demand or supply shocks if a local convention was organized in a particular day that might shift the demand curve. In this case, we need to use instrumental variables to address the endogeneity problem. The natural approach in this case would be to use the typical Hausman-style instrument [44], i.e., the average rental price of similar cars in other geographical locations. However, in our dataset we are highly unlikely to have any price endogeneity issues because the rental price variation of the listed cars is very insignificant as it was discussed above, i.e., the price does not react to any unobservable shocks (see Table 3.5).

3.6 Conclusion

In order to be successful in the long run, firms need to make accurate long-term demand predictions. Herein, robustness of predictions models is of great importance because we are likely to face significant noise in offer set definitions while making long-term demand predictions. In this chapter, we demonstrate that models that account for the consideration set formation of customers are robust to noise in the offer sets. Recall that, in the spirit of the consider-then-choose framework, we assume that customers make a purchasing decision in two stages. First, a boundedly rational consumer, who suffers from limited attention, forms her consideration/competition set, which is usually a small subset of substitutable items in the product category due to cognitive limitations. Secondly, the consumer evaluates all products in her consideration set and purchases the one that is most preferred.

In this chapter, we first demonstrate the robustness of our approach based on the synthetic dataset. We explore different noise regimes (noise scenarios) when consider-then-choose type models are better equipped to handle offer set noise than other popular models in the literature, such as the MNL model. Then, we analyze two real world settings – a retail operation and a car-sharing platform. Our empirical results suggest that the predictive performance of consider-then-choose models is significantly more accurate than state-of-the-art benchmarks widely used in marketing and economics literature. Moreover, we show that the relative improvement of consider-then-choose models in predictive performance becomes even more significant with increased noise in the consideration set definition embedded in the data. These results lead to a promising methodology for researchers interested in choice-based demand estimation, where identification of consideration sets is of significant importance. Moreover, we demonstrate that the consider-then-choose type of choice model can provide important managerial insights about the consideration set formation.

3.6.1 Model extension

The GCC model, proposed in this, chapter might suffer from its one-directional cannibalization property, i.e., preferences of individuals are characterized by the unique ranking in the GCC model. Even though the same one directional cannibalization helps in reducing the impact of noise in the offer set, the GCC model might be worse off when there is no noise in the sales transaction data and cannibalization in the data generation process is bi-directional. Therefore, we propose general consideration - then - general choice (GCGC) model as an extension of the GCC model where we allow customers to be heterogeneous in their preferences. First, we assume that there is a distribution $\mu : \mathscr{S}_n \rightarrow [0,1]$ over $\mathscr{S}_n$, which is the set of all full rankings or permutations of products in N^+ with cardinality $(n+1)!$. According to this model, before making a choice, customers, first, sample both consideration set $C \subset N$ and the preference order $\sigma \in \mathscr{S}_n$ from the general distributions over the consideration sets λ and rankings μ, respectively. Then, customers choose the most preferred alternative in C in accordance with the preference order σ. This model can be estimated in a similar way as GCC model with EM algorithm (see Section 3.6.2 for the details) by dividing customers into K segments such that for every segment $h \in \{1, ..., K\}$ a customer considers an arbitrary subset of items $C \subseteq N$ with likelihood $\lambda(C) = \prod_{a_j \in C} \theta_{hj} \prod_{a_j \notin C} (1 - \theta_{hj})$ (recall that θ_{hj} is the probability to include item a_j in the consideration set in the segment h) and make choices according to the preference order σ_h. For a sufficiently large K, this heuristic calibration would be exact. If we have sparsity in customer segments, then using this estimation technique with a relatively small K would provide the exact calibration of the GCGC model.

We compare the prediction performance of GCC model versus GCGC based on the IRI dataset under different noise scenarios in the retail industry (see Section 3.3): (1) no extra noise, (2) store-based noise, and (3) week-based noise. We estimated both GCC and GCGC models for $K = 1, 2, ..., 5$ and report the best performance measure from these 5 variants for MAPE and RMSE metrics. Figure 3-12 illustrates MAPE (see left panel) and RMSE (see right panel) scores of GCC and GCGC models, averaging across 20 product categories, for three different noise regimes. We observe that the predictive performance of both models is rather robust to the noise in the offer sets, i.e., the performances stay rather flat for all three noise regimes. From

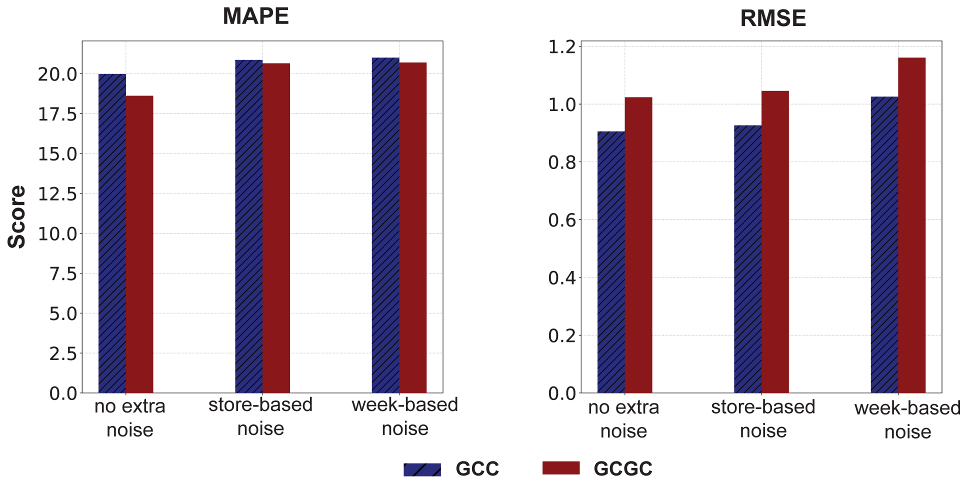

Figure 3-12: The average prediction scores under GCC and GCGC choice models.

the panels in Figure 3-12, we can see that neither GCC nor GCGC can dominate the other, i.e., GCC dominates GCGC based on the RMSE metrics while GCGC outperforms the GCC model based on the MAPE metrics. In Figure 3-13, we present scatter plots of the improvements of GCC model versus GCGC model across 20 product categories under three noise regimes discussed above: (1) no extra noise added, represented by green pluses, (2) store-based noise added, represented by blue crosses, and (3) week-based noise added, represented by red dots. In the left and right panels, we measure the predictive performance of the models based on the MAPE and RMSE, respectively. Note that the improvement of GCC over GCGC varies across product categories and noise regimes. The figure further supports the claim that we do not have a clear winner between these two models by revealing that neither of these two models dominate the other one for most of the product categories, noise regimes, or score metrics.

3.6.2 GCGC model: estimation methodology

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

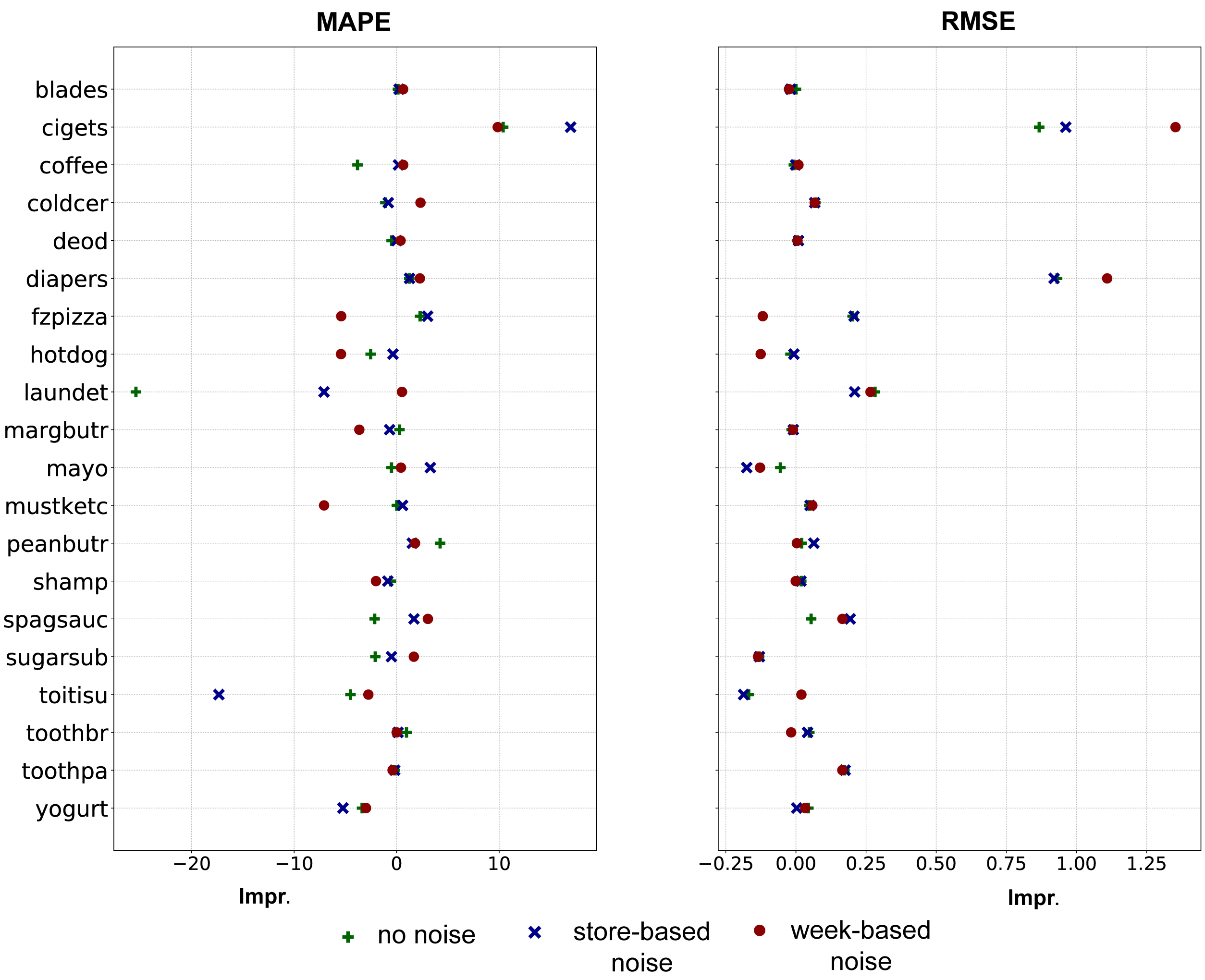

Figure 3-13: Scatter plots of the prediction score improvements of GCC model over GCGC.

Estimation with aggregate level data

Similarly to the Section 2.4.3, we calibrate GCGC model by dividing transactions into K segments such that customers in segment h sample their consideration sets based on the attention parameters $\boldsymbol{\theta}_h$ and have their preferences characterized by the ranking σ_h. Then, the log-likelihood function can be represented in the following way:

$$\log \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) = \sum_{t=1}^{T} \log \left(\sum_{h=1}^{K} \gamma_h \theta_{h,j_t} \prod_{\substack{a_j \in S_t: \\ a_j \succ_h a_{jt}}} (1 - \theta_{hj}) \right), \tag{3.3}$$

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

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

Initialization: we randomly allocate sales transaction to one of the K classes, resulting in an initial allocation $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$, which form a partition of the collection of all the transactions. Consequently, we set $\gamma_h^{(0)} = |\mathcal{D}_h|/(\sum_{d=1}^{K} |\mathcal{D}_d|)$. Then, $\succ_h$ (i.e., σ_h) and $\theta_{hj}^{(0)}$, for all $h \in \{1, ..., K\}$ and $a_j \in N^+$, are obtained by solving the following optimization problem:

$$\max_{\succ_h, \boldsymbol{\theta}_h} \sum_{t \in \mathcal{D}_h} \Big(\log \theta_{h,j_t} + \sum_{\substack{a_j \in S_t: \\ a_j \ \succ_h \ a_{j_t}}} \log(1 - \theta_{hj}) \Big),$$

which is solved by using the outer-approximation algorithm for the ICC model in Section 2.4.2. *E-step:* we compute $P_{ht}^{(q)}$, which is the membership probability of every transaction at time t to belong to the segment h based on the parameter estimates $\{\boldsymbol{\succ}^{(q-1)}, \boldsymbol{\theta}^{(q-1)}, \boldsymbol{\gamma}^{(q-1)}\}$ and the purchasing transactions data $(a_{j_t}, S_t)|_{t=1}^{T}$:

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

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

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

and then optimize the conditional expected value of the log-likelihood function, obtained in the previous step, in terms of $\boldsymbol{\theta}_h$ and $\succ_h$ for all $h \in \{1, ..., K\}$:

$$\max_{\succ_h, \boldsymbol{\theta}_h} \sum_{t=1}^{T} P_{ht}^{(q)} \log \Big(\theta_{h,j_t} \prod_{\substack{a_j \in S_t:\\ a_j \succ_h a_{j_t}}} (1-\theta_{hj})\Big),$$

which is solved by using the outer-approximation algorithm for the ICC model in Section 2.4.2.

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

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

Estimation with panel data

We update the EM algorithm above in the following way:

Initialization: we randomly allocate individuals to one of the K classes, resulting in an initial allocation $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$. Consequently, we set $\gamma_h^{(0)} = |\mathcal{D}_h|/(\sum_{d=1}^{K} |\mathcal{D}_d|)$. Then, $\succ_h$ (i.e., σ_h) and $\theta_{hj}^{(0)}$, for all $h \in \{1, ..., K\}$ and $a_j \in N^+$, are obtained by solving the following optimization problem:

$$\max_{\succ_h, \boldsymbol{\theta}_h} \sum_{i \in \mathcal{D}_h} \Big(\log \theta_{h,j_{it}} + \sum_{\substack{a_j \in S_{it}:\\ a_j \succ_h a_{j_{it}}}} \log(1-\theta_{hj})\Big),$$

which is solved by using the outer-approximation algorithm for the ICC model in Section 2.4.2.

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

$$P_{hi}^{(q)} = \frac{\gamma_h^{(q-1)} \prod_{t=1}^{T_i} \left[\theta_{h,j_{it}}^{(q-1)} \prod_{\substack{a_j \in S_{it}: \\ a_j \succ_h^{(q-1)} a_{j_{it}}}} (1-\theta_{hj}^{(q-1)}) \right]}{\sum_{r=1}^{K} \prod_{t=1}^{T_i} \left[\gamma_h^{(q-1)} \left(\theta_{r,j_{it}}^{(q-1)} \prod_{\substack{a_j \in S_{it}: \\ a_j \succ_h^{(q-1)} a_{j_{it}}}} (1-\theta_{rj}^{(q-1)}) \right) \right]}.$$

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

$$\gamma_h^{(q)} = \frac{\sum_{i=1}^{m} P_{ht}^{(q)}}{m},$$

and then optimize the conditional expected value of the log-likelihood function, obtained in the previous step, in terms of $\boldsymbol{\theta}_h$ and $\succ_h$ for all $h \in \{1, ..., K\}$:

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

which is solved by using the outer-approximation algorithm for the ICC model in Section 2.4.2.

Conclusion and Future Directions

This thesis begins by considering the implementation of personalized operational decisions in the retail industry motivated by the availability of individual-level transaction data and recent technological advances by brick-and-mortar stores. Using panel data, we demonstrate how the retailer can first infer customer-level preferences for items within the category of analysis and then decide the optimal subset of products to promote (if any) for each individual customer's visit to the store with the objective of maximizing immediate revenue from this visit. Next, we propose a methodology to estimate consideration sets of customers from sales transaction data. Theoretically, we derive necessary and sufficient conditions for a collection of observed choice probabilities to be consistent with our model. In particular, we provide a closed-form expression for computing the distribution over consideration sets from observed choice probabilities. Finally, we analyze different operational applications of a consider-then-choose framework motivated by the problems faced by online platforms. We conclude that models accounting for consideration sets of customers are generally more robust to noise in the offer set definition than classical choice models. Our work opens up opportunities and directions for future research. We begin with future directions which address certain limitations of our proposed personalized promotions algorithm:

- *Effects of stockpiling.* Our method focuses on the short-term objective of immediate profit maximization in deciding the set of promoted products. As such, it ignores the stockpiling behavior of customers, whereby customers take advantage of the discounted prices to "stockpile" or purchase more than their immediate consumption need. Such stockpiling behavior impacts future purchase incidences from the customer and the long-term revenue for the retailer. We argue that when compared to mass promotions, personalized promotions mitigate the negative effect of stockpiling. The reason is that stockpiling typically occurs when a brand is promoted to a customer who would purchase the product at full price anyway; the price discounts only end up enticing the customer to stockpile, shifting her future purchases to the current period. Such erroneous price discounts are more likely to occur as part of mass promotions when compared to personalized promotions. Nevertheless, personalization may not completely eliminate stockpiling, in which case our formulation can be embedded within a dynamic programming (DP) framework to incorporate the effects of stockpiling and any other long-term effects of current period promotion decision (c.f., [103], [56]). A state variable that keeps track of last purchase incidence and corresponding quantity could help towards the optimal timing of future promotions.

- *Reference price effect.* Another long-term effect of price promotions is the reference price effect. Repeated price reductions have the potential for lowering the reference price of the brands for the customers [54]. Such a recalibration of the reference price reduces the future impact of price discounts because customers now evaluate the discounts with respect to the lower reference price as opposed to the full price of the product. The reference price effect can be incorporated by embedding our promotion optimization framework within a DP. The state variable of the DP keeps track of the reference price and the current promotion decision affects the future profit (value-to-go) through its effect on the reference price.

- *Purchase quantity estimation.* In our implementation, we estimated the purchase quantity of a brand conditional on its purchase (for both promoted and non-promoted copies) by taking a simple average of the observed purchase quantities in the training data. This estimation technique might suffer from endogeneity bias. For instance, retailers may strategically promote products expecting higher quantity purchases, say, during holidays. In practice, this endogeneity bias may be corrected through detailed structural modeling to obtain a more precise estimate of the purchase quantity of a brand as a function of its promotion status. Such a correction will most certainly improve the predictive performance of our model.

In the second part of this dissertation, we proposed GCGC model (see Section 3.6.1) where we extended the GCC model by allowing customers to be heterogeneous in their preferences. An important area of the future research would be to further study GCGC model and its application in practice: (i) find the optimal number of segments in GCGC model to strike a balance between capturing heterogeneity of customers in their preferences and robustness to noise in sales transaction data; (ii) address the problem of identifiability of GCGC model (i.e., what is the maximum number of segments for the model to be identified?); and (iii) propose a more efficient way to estimate the model for large-scale problems.

Appendices

Chapter 4

Proofs and Supplementary Materials for Chapter 1

4.1 Preliminaries on DAGs

For completeness, we summarize the relevant notation from Chapter 1 and also introduce additional notation. Let $\mathcal{N} = \{a_1, \ldots, a_n\}$ denote the universe of n products. For the purposes of this chapter, we ignore the no-purchase option and information about product promotions in order to simplify exposition. This assumption is without loss of generality since we can explicitly account for promotions by expanding the product universe and include the no-purchase option as one more item. This is further developed in Section 1.6 in the main body of this thesis.

A DAG D is a subset of pairwise preferences, $\{(a_j, a_{j'}) \colon 1 \leq j, j' \leq n\}$. We visualize a DAG D as a directed graph with nodes as products and a directed edge from a to b if the ordered pair $(a, b) \in D$. We abuse notation and let D denote both the directed graph and the collection of pairwise preferences. We let E_D denote the set of pairwise preferences in the transitive reduction of D.

Let $\mathscr{S}_n$ denote the collection of all possible $n!$ rankings or permutations of the products in $\mathcal{N}$. For any ranking $\sigma \in \mathscr{S}_n$, we let $\sigma(a)$ denote the preference ranking of product a under ranking σ. We adopt the convention that lower ranked products are preferred over higher ranked ones, which means that product a_j is preferred over product $a_{j'}$ under σ if and only if $\sigma(a_j) < \sigma(a_{j'})$. Given a DAG D, let S_D denote the subset of rankings that are consistent with D; that is,

$S_D := \{\sigma \in \mathscr{S}_n \colon \sigma(a) < \sigma(b) \text{ whenever } (a,b) \in D\}$.

For any product a_j and DAG D, the reachability set $\Psi_D(a)$ comprises the set of all nodes that can be reached from a through a directed path in D. Formally, $\Psi_D(a) := \{b\colon$ there is a directed path from a to b in $D\}$. We assume that a is reachable from itself, so $a \in \Psi(a)$. The set $\Theta_D(a)$ comprises the nodes *from which* a can be reached, i.e., $\Theta_D(a) := \{b\colon$ there is a directed path from b to a in $D\}$. To be consistent, we also include a in $\Theta_D(a)$. When the DAG D is clear from the context, we drop D from the notation and simply write $\Psi(a)$ and $\Theta(a)$.

For any subset $S \subseteq \mathcal{N}$, suppose π is a ranking of the products in S possibly including less than n products. Then, $\sigma(\pi)$ denotes the set of all complete rankings of the products in $\mathcal{N}$ that are consistent with π, i.e., $\sigma(\pi) = \{\sigma \in \mathscr{S}_n \colon \sigma(a) < \sigma(b) \text{ whenever } \pi(a) < \pi(b)\}$.

4.2 Technical results

4.2.1 Propositions and proofs in Section 1.2

Preference graph decycling

Here we argue that the decycling procedure prioritizes retaining as many candidate edges as possible (measured by the sum of the weights involved). To that end, we show that the weight of the candidate edges in DAG D after preference graph G decycling is equal to the weight of the candidate edges in DAG D^* such that $D^* \subseteq G$ has maximum weight of the candidate edges.

Let $cw(D)$ denote the weight of the candidate edges in DAG D and let $iw(D)$ denote the weight of the implicit candidate edges in DAG D. Let $tw(D)$ denote the total weight of DAG D, i.e., $tw(D) = cw(D) + iw(D)$. Let us define D^* as a DAG in G with the maximum weight of the candidate edges, i.e., $D^* = \arg\max\{cw(D) : D \subseteq G\}$. Let D denote the DAG obtained from G after solving MIP (1.2). The next result follows.

Proposition 1. *Candidate weight of DAG D, obtained after MIP (1.2) decycling applied over G is equal to the candidate weight of DAG D^* such that $D^* = \max\{cw(D) : D \subseteq G\}$, i.e.,*

$$cw(D) = cw(D^*).$$

Proof. of Proposition 1: Assume by contradiction that $cw(D) < cw(D^*)$. Because $cw(\cdot)$ is always an integer, it follows that $cw(D^*) - cw(D) \geq 1$. Further, note that D^* is a feasible solution to the optimization problem MIP (1.2), which implies that $tw(D) \geq tw(D^*)$. It now follows from the definition of $tw(\cdot)$ that

$$cw(D) + iw(D) \geq cw(D^*) + iw(D^*) \quad \Longrightarrow \quad iw(D) \geq cw(D^*) - cw(D) + iw(D^*).$$

Because $iw(D^*) \geq 0$ and $cw(D^*) - cw(D) \geq 1$, we obtain that $iw(D) \geq 1$. This, however, is not possible because of the scaling factor $1/(n^2T)$. Specifically, note that the maximum number of implicit candidate edges is $n(n-1)$ and the maximum possible weight of each implicit edge is $T/(n^2T) = 1/n^2$. Therefore, the aggregate implicit weight is always bounded above by $n(n-1)/n^2$, which is strictly less than 1. We have thus arrived at a contradiction. □

Likelihood of the DAG-based choice model

Proposition 2. *For a given set of parameters $\boldsymbol{\beta}$ that characterize the distribution λ, the likelihood function of the DAG-based choice model is given by*

$$\log \mathcal{L}(\textit{Panel Data}) = \sum_{i=1}^{m} \log \lambda(D_i) = \sum_{i=1}^{m} \log \left(\sum_{\sigma \in S_{D_i}} \lambda(\sigma) \right).$$

Proof. of Proposition 2: For each individual i and transaction t, let $C_{it} \subseteq S_{it}$ denote the consideration set of the individual. Because products outside the consideration set do not affect the individual's choices, the data log-likelihood only depends on the choices and the consideration sets, rather than choices and offer sets. Letting $f(a_{j_{it}}, C_{it}, D_i)$ denote the probability of purchasing product $a_{j_{it}}$ from consideration set C_{it} for individual i with DAG D_i, we can write

$$\begin{aligned}
\log \mathcal{L}(\text{Panel Data}|\boldsymbol{\beta}) &\triangleq \sum_{i=1}^{m} \log \Pr\left[(a_{j_{it}}, C_{it})|_{t=1}^{t=T_i}, D_i \middle| \boldsymbol{\beta}\right] \\
&= \sum_{i=1}^{m} \log\left(\Pr(D_i|\boldsymbol{\beta}) \cdot \Pr\left[(a_{j_{it}}, C_{it})|_{t=1}^{t=T_i} \middle| \boldsymbol{\beta}, D_i\right]\right) \\
&= \sum_{i=1}^{m} \log\left(\Pr(D_i|\boldsymbol{\beta}) \cdot \prod_{t=1}^{T_i} \Pr\left[(a_{j_{it}}, C_{it}) \middle| \boldsymbol{\beta}, D_i\right]\right) \\
&= \sum_{i=1}^{m} \log \Pr(D_i|\boldsymbol{\beta}) + \sum_{i=1}^{m} \sum_{t=1}^{T_i} \log f(a_{j_{it}}, C_{it}, D_i)\} \qquad (4.1)
\end{aligned}$$

The second equality follows from a straightforward application of the conditional probability formula. The third equality follows because conditioning on the DAG D_i, individual i's purchase probabilities can be computed independently.

We now focus on the term $f(a_{j_{it}}, C_{it}, D_i)$. Note that we only observe the offer sets S_{it}. The consideration sets C_{it} are latent. Nevertheless, given D_i, we can constrain C_{it} sufficiently to allow for the computation of $f(a_{j_{it}}, C_{it}, D_i)$. There are two cases. First, we consider the case when none of the edges in the set $\{(a_{j_{it}}, a_k) \colon a_k \in S_{it} \setminus \{a_{j_{it}}\}\}$ was deleted in the decycling step. In this case, the customer always prefers product $a_{j_{it}}$ over all the other products in the offer set S_{it}, and therefore, chooses product $a_{j_{it}}$ irrespective of the consideration set. This implies that $f(a_{j_{it}}, C_{it}, D_i) = 1$ for all $C_{it} \subseteq S_{it}$ such that $a_{j_{it}} \in C_{it}$.

Next, we consider the case when some of the edges in the set $\{(a_{j_{it}}, a_k) \colon a_k \in S_{it} \setminus \{a_{j_{it}}\}\}$ are deleted in the decycling step. Because the decycling procedure deletes the smallest possible weight of edges, the edge $(a_{j_{it}}, a_k)$ for some $a_k \in S_{it} \setminus \{a_{j_{it}}\}$ is deleted only if there is a directed path from a_k to $a_{j_{it}}$ in the final DAG D_i. Now, because a_k is preferred over $a_{j_{it}}$, the only way the customer would choose $a_{j_{it}}$ when a_k was also on offer is if a_k was not considered. We can thus conclude that $a_k \notin C_{it}$ for all $a_k \in S_{it} \setminus \{a_{j_{it}}\}$ that are preferred over $a_{j_{it}}$. Equivalently, $a_{j_{it}}$ is preferred over a_k for all $a_k \in C_{it}$, which implies that $f(a_{j_{it}}, C_{it}, D_i) = 1$.

We have thus shown that $f(a_{j_{it}}, C_{it}, D_i) = 1$ for all individuals i and all transactions t. It now follows from (4.1) that

$$\log \mathcal{L}(\text{Panel Data}|\boldsymbol{\beta}) = \sum_{i=1}^{m} \log \Pr(D_i|\boldsymbol{\beta}) = \sum_{i=1}^{m} \log \lambda(D_i) = \sum_{i=1}^{m} \log \left(\sum_{\sigma \in S_{D_i}} \lambda(\sigma) \right).$$

The result of the proposition now follows. □

4.2.2 Propositions and proofs in Section 1.3

The following auxiliary results will be used in the proofs of the results in the main body of this thesis. We start quoting Lemma A1 from [52]:

Lemma A1 in [52] *Consider two subsets $S_1, S_2 \subset \mathcal{N}$ with $S_1 \cap S_2 = \emptyset$. Let π_1 be a ranking over S_1, and π_2 be a ranking over S_2. Assume w.l.o.g. that $\pi_1 = (a_{1,1}, a_{1,2}, \ldots, a_{1,k_1})$ and $\pi_2 = (a_{2,1}, a_{2,2}, \ldots, a_{2,k_2})$. For a fixed i, $0 \leq i \leq k_1$, let $\mathcal{S}_i(\pi_1, \pi_2)$ be the set of rankings in $\mathscr{S}_n$ consistent with both π_1 and π_2, where the head of π_2 is located after the ith element of π_1, i.e.,*

$$\mathcal{S}_i(\pi_1, \pi_2) = \{\sigma \in \mathscr{S}_n : \sigma \in \sigma(\pi_1) \cap \sigma(\pi_2),\ \textit{with } \sigma(a_{2,1}) \geq i+1\}.$$

Then,

$$\lambda(\mathcal{S}_i(\pi_1, \pi_2)) = \prod_{j=1}^{i} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'} + \sum_{j'=1}^{k_2} v_{2,j'}} \prod_{j=i+1}^{k_1} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'}} \lambda(\pi_2).$$

The first new result refers to an expression for the probability of DAG D that is split in two independent terms by breaking the bottom part of D in two pieces.

Lemma 4.2.1. *Given a DAG D, let $a_y \in \mathcal{N}$ be a node such that every node in $\Psi_D(a_y) \setminus \{a_y\}$ has at most one incoming edge and the subgraph $D[a_y]$ induced in D by the set of nodes $\Psi_D(a_y)$ is a directed tree; see Figure 4-1. Further, let $\bar{D}[a_y]$ denote the subgraph induced in D by the set of nodes $\big(\mathcal{N} \setminus \Psi_D(a_y)\big) \cup \{a_y\}$. Then, under the MNL distribution λ, we have that*

$$\lambda(D) = \lambda(D[a_y]) \cdot \lambda_y(\bar{D}[a_y]),$$

where λ_y is the distribution over rankings obtained by replacing the MNL weight v_y of product a_y

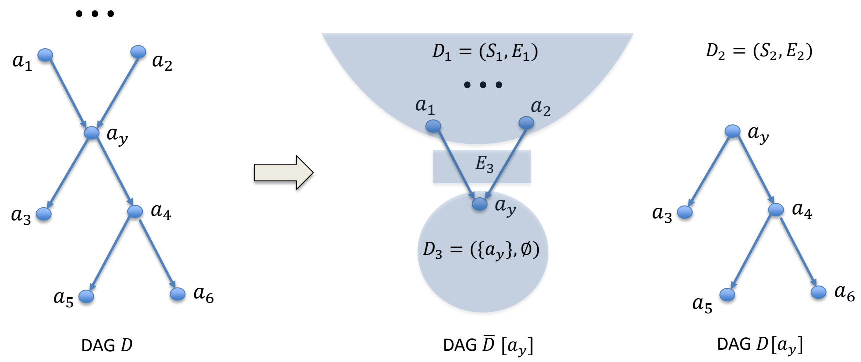

Figure 4-1: Illustration for the proof in Lemma 4.2.1. The bottom part of a DAG is split in two independent terms.

with $v_{\Psi_D(a_y)} = \sum_{a_i \in \Psi_D(a_y)} v_i$.

Proof. of Lemma 4.2.1: Note that $D[a_y]$ is the tree "hanging" from the node a_y in DAG D. We establish the result of this lemma by showing that the term $\lambda(D[a_y])$ factors out from the expression for $\lambda(D)$.

For that, let S_1 denote $\mathcal{N} \setminus \Psi_D(a_y)$ and S_2 denote $\Psi_D(a_y)$. It is clear that $S_1 \cap S_2 = \emptyset$ and $S_1 \cup S_2 = \mathcal{N}$. For any ranking σ and position $1 \leq r \leq n$, let $\sigma^{-1}(r)$ denote the product ranked at position r under σ. Let D_1 and D_2 denote the subgraphs in D induced by the sets S_1 and S_2, respectively. It follows from our notation that $D_2 = D[a_y]$. Let E_1 and E_2 denote the edges in the transitive reductions of D_1 and D_2, respectively.

With this notation, we first establish the following result.

Claim: The set E_D of edges in the transitive reduction of D can be partitioned as

$$E_D = E_1 \cup E_2 \cup E_3, \text{ where } E_3 = \{(a, a_y) \colon (a, a_y) \in E_D\} \text{ and } E_i \cap E_j = \varnothing \; \forall\, 1 \leq i \neq j \leq 3 \quad (4.2)$$

Proof. We first note that $E_1 \cup E_2 \cup E_3 \subseteq E_D$ since it follows by definition that $E_i \subseteq E_D$ for all $1 \leq i \leq 3$. To show that $E_D \subseteq E_1 \cup E_2 \cup E_3$, consider an edge $(a, b) \in E_D$. We note that if $a \in S_2$, then b must belong to S_2. The reason is that since $S_2 = \Psi_D(a_y)$, if $a \in S_2$, then a is reachable from a_y and because b is reachable from a, it must be that b is also reachable from a_y, which implies that $b \in \Psi_D(a_y) = S_2$. Therefore, there are two cases to consider: (i) both a and b belong to S_1 or S_2 and (ii) $a \in S_1$ and $b \in S_2$. In case (i), it follows by definition that (a, b)

belongs to E_1 or E_2. In case (ii), since every node in $S_2 \setminus \{a_y\}$ can have at most one incoming edge and every node in $S_2 \setminus \{a_y\}$ already has an incoming edge from a node in S_2, the only way there can be an edge from $a \in S_1$ to b is if $b = a_y$. It now follows that $(a, b) = (a, a_y) \in E_3$. We have thus shown that $E_D = E_1 \cup E_2 \cup E_3$. The fact that the three sets E_1, E_2, and E_3 are mutually disjoint follows immediately from noting that the sets S_1 and S_2 are disjoint.

With the above decomposition of the edges of E_D, we now show that the set of rankings S_D can be decomposed in a convenient manner. Consider the following definitions:

- Let π_1 (of length k_1) and π_2 (of length k_2) be rankings of products in the sets S_1 and S_2, respectively. Note that $k_1 + k_2 = n$.

- Let X be the set of tuples (π_1, π_2) such that π_1 and π_2 are consistent with DAGs D_1 and D_2, respectively. In other words, $X = \{(\pi_1, \pi_2) \colon \sigma(\pi_1) \subset S_{D_1}, \sigma(\pi_2) \subset S_{D_2}\}$.

- For any $1 \leq i \leq k_1$, let $\mathcal{S}_i(\pi_1, \pi_2)$ denote the set of rankings in $\mathscr{S}_n$ consistent with both π_1 and π_2 where the head of π_2 is located after the ith element of π_1, i.e. $\mathcal{S}_i(\pi_1, \pi_2) = \left\{\sigma \in \mathscr{S}_n : \sigma \in \sigma(\pi_1) \cap \sigma(\pi_2), \text{ with } \sigma(\pi_2^{-1}(1)) \geq \sigma(\pi_1^{-1}(i)) + 1\right\}$.

- Further, let $i(\pi_1)$ denote the position of the least preferred item in $\Theta_D(a_y)$ in the ranking π_1, i.e., $i(\pi_1) = \max\{\pi_1(a) \colon a \in \Theta_D(a_y)\}$.

Claim: The set of rankings S_D is obtained by taking a tuple $(\pi_1, \pi_2) \in X$ and combining them such that the head of π_2 occurs after the $i(\pi_1)$th element of π_1. More precisely, we claim that

$$S_D = \bigcup_{(\pi_1,\pi_2) \in X} \mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2), \quad \text{where } \mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2) \cap \mathcal{S}_{i(\pi_1')}(\pi_1', \pi_2') = \emptyset \text{ for } (\pi_1, \pi_2) \neq (\pi_1', \pi_2') \tag{4.3}$$

Proof. We first show that $\mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2) \subseteq S_D$ for all $(\pi_1, \pi_2) \in X$. For that, consider $\sigma \in \mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)$ and consider an edge $(a, b) \in E_D$. It is sufficient to show that $\sigma(a) < \sigma(b)$. It follows from (4.2) that (a, b) is in either E_1 or E_2 or E_3. If (a, b) is in E_1, then we must have that $\pi_1(a) < \pi_1(b)$ because π_1 is consistent with D_1. Since σ is consistent with π_1, we have that $\sigma(a) < \sigma(b)$. Using a symmetric argument, we can similarly show that $\sigma(a) < \sigma(b)$ when $(a, b) \in E_2$.

Now suppose that $(a, b) \in E_3$. We then have that $b = a_y$. Since π_2 is consistent with D_2 and a_y is preferred over every product in $S_2 \setminus \{a_y\}$ under the partial order D_2, it follows that a_y must be the head of π_2, i.e., $\pi_2(a_y) = 1$. Now, let a^* denote the least preferred element under π_1 in the set $\Theta_D(a_y)$. Since $(a, a_y) \in E_D$, we have that $a \in \Theta_D(a_y)$, implying that $\sigma(a) = \pi_1(a) \leq \pi_1(a^*) = \sigma(a^*)$, with equality when $a = a^*$. Note that both rankings σ and π_1 must coincide until position $i(\pi_1)$. It also follows by our definitions that $\pi_1(a^*) = i(\pi_1)$ and $\sigma(a_y) = \sigma(\pi_2^{-1}(1)) > \sigma(\pi_1^{-1}(i(\pi_1)) = \sigma(a^*)$, where the inequality follows from the definition of $\mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)$ and the fact that $\sigma \in \mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)$. We have thus shown that $\sigma(a) \leq \sigma(a^*) < \sigma(a_y) = \sigma(b)$, establishing the result that $\mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2) \subseteq S_D$ for all $(\pi_1, \pi_2) \in X$, which implies that $\bigcup_{(\pi_1,\pi_2)\in X} \mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2) \subseteq S_D$.

We now show that $S_D \subseteq \bigcup_{(\pi_1,\pi_2)\in X} \mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)$. For that consider $\sigma \in S_D$ and let π_1 and π_2 denote the rankings σ induced on the set of products S_1 and S_2, respectively. We show that $\sigma \in \mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)$. It follows by the definitions of π_1 and π_2 that $\sigma \in \sigma(\pi_1) \cap \sigma(\pi_2)$. Therefore, it is sufficient to show that $\sigma(\pi_2^{-1}(1)) \geq \sigma(\pi_1^{-1}(i(\pi_1)) + 1$. Using the arguments above, it readily follows that $\pi_2^{-1}(1) = a_y$. Since σ is consistent with D, we have that $\sigma(a) < \sigma(a_y)$ for all $a \in \Theta_D(a_y)$, and in particular, $\sigma(a^*) < \sigma(a_y)$, where a^* is the least preferred product under π_1 from the set $\Theta_D(a_y)$. Since $i(\pi_1) = \pi_1(a^*)$ by definition, we have shown that $\sigma(\pi_1^{-1}(i(\pi_1))) < \sigma(a_y)$. We have thus established that $S_D \subseteq \bigcup_{(\pi_1,\pi_2)\in X} \mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)$.

The arguments above establish that $S_D = \bigcup_{(\pi_1,\pi_2)\in X} \mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)$. The disjointness of $\mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)$ and $\mathcal{S}_{i(\pi_1')}(\pi_1', \pi_2')$ for $(\pi_1, \pi_2) \neq (\pi_1', \pi_2')$ readily follows from the disjointness of $\sigma(\pi_1) \cap \sigma(\pi_2)$ and $\sigma(\pi_1') \cap \sigma(\pi_2')$ for $(\pi_1, \pi_2) \neq (\pi_1', \pi_2')$. We have thus established the claim in (4.3).

We can then write from (4.3) that

$$\lambda(D) = \sum_{\sigma \in S_D} \lambda(\sigma) = \sum_{(\pi_1,\pi_2)\in X} \lambda\big(\mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)\big) = \sum_{\pi_1 \colon \sigma(\pi_1) \subset S_{D_1}} \sum_{\pi_2 \colon \sigma(\pi_2) \subset S_{D_2}} \lambda\big(\mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)\big).$$

Now consider $(\pi_1, \pi_2) \in X$. Without loss of generality, suppose that $\pi_1 = (a_{1,1}, \ldots, a_{1,k_1})$ and

$\pi_2 = (a_{2,1}, \ldots, a_{2,k_2})$. Then, invoking [52, Lemma A1], we can write

$$\begin{aligned}
\lambda\big(\mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)\big) &= \left[\prod_{j=1}^{i(\pi_1)} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'} + \sum_{j'=1}^{k_2} v_{2,j'}} \prod_{j=i(\pi_1)+1}^{k_1} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'}}\right] \cdot \lambda(\pi_2) \\
&= \left[\prod_{j=1}^{i(\pi_1)} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'} + v_{\Psi_D(a_y)}} \prod_{j=i(\pi_1)+1}^{k_1} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'}}\right] \cdot \lambda(\pi_2) \\
&= g(\pi_1) \cdot \lambda(\pi_2),
\end{aligned}$$

where the second equality follows from the fact that $S_2 = \Psi_D(a_y)$, and where we define

$$g(\pi_1) = \prod_{j=1}^{i(\pi_1)} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'} + v_{\Psi_D(a_y)}} \prod_{j=i(\pi_1)+1}^{k_1} \frac{v_{1,j}}{\sum_{j'=j}^{k_1} v_{1,j'}}.$$

We now have

$$\lambda(D) = \sum_{\pi_1 :\, \sigma(\pi_1) \subset S_{D_1}} \sum_{\pi_2 :\, \sigma(\pi_2) \subset S_{D_2}} \lambda\big(\mathcal{S}_{i(\pi_1)}(\pi_1, \pi_2)\big) = \left[\sum_{\pi_1 :\, \sigma(\pi_1) \subset S_{D_1}} g(\pi_1)\right] \cdot \left[\sum_{\pi_2 :\, \sigma(\pi_2) \subset S_{D_2}} \lambda(\pi_2)\right].$$

Noting that

$$\sum_{\pi_2 :\, \sigma(\pi_2) \subset S_{D_2}} \lambda(\pi_2) = \sum_{\sigma \in S_{D_2}} \lambda(\sigma) = \lambda(D[a_y]),$$

we have shown that

$$\lambda(D) = \lambda(D[a_y]) \cdot \left[\sum_{\pi_1 :\, \sigma(\pi_1) \subset S_{D_1}} g(\pi_1)\right]. \tag{4.4}$$

It now suffices to show that

$$\lambda_y(\bar{D}[a_y]) = \sum_{\pi_1 :\, \sigma(\pi_1) \subset S_{D_1}} g(\pi_1).$$

For that, consider the distribution λ_y in which the weight v_y is replaced by $v_{\Psi_D(a_y)}$, and repeat the above set of arguments for the DAG $\bar{D}[a_y]$ with the nodes of the DAG decomposed into sets S_1 as defined above and $S_3 = \{a_y\}$. For any ranking π_1 of the products in set S_1, note that $g(\pi_1)$

remains the same under both distributions λ and λ_y. As a result, following (4.4), we can write

$$\lambda_y(\bar{D}[a_y]) = \lambda_y(D_3) \cdot \left[\sum_{\pi_1 \colon \sigma(\pi_1) \subset S_{D_1}} g(\pi_1) \right],$$

where D_3 is the DAG induced in D by S_3. Since S_3 is a singleton, the DAG D_3 will be empty, implying that $\lambda_y(D_3) = 1$. We have thus shown that

$$\lambda_y(\bar{D}[a_y]) = \sum_{\pi_1 \colon \sigma(\pi_1) \subset S_{D_1}} g(\pi_1).$$

The result of the lemma now follows. □

To establish the next result, we first quote the so-called *internal consistency* property referred to in [52], i.e., the probability of a ranking in a subset of products only depends on the relative order of the elements in the subset.

Proposition A1.2 in [52]. *For any subset S of k products, let π be a ranking over the elements of S. Let $\sigma(\pi)$ denote the set of full rankings over $\mathcal{N}$ defined as*

$$\sigma(\pi) = \{\sigma \colon \sigma(a_i) < \sigma(a_j) \text{ whenever } \pi(a_i) < \pi(a_j) \text{ for all } a_i, a_j \in S\}.$$

Then, it must hold that

$$\lambda(\pi) = \sum_{\sigma \in \sigma(\pi)} \lambda(\sigma) = \prod_{r=1}^{k} \frac{v_{\pi_r}}{\sum_{j=r}^{k} v_{\pi_j}}.$$

We also need the following notation. Given a DAG D and product a_y, let λ_y^{aug} denote the distribution of rankings on the expanded product universe $\mathcal{N} \cup \{a'_y\}$, where a'_y is a copy of the product a_y, with the weight v_y of product a_y replaced with $v_{\Psi_D(a_y)}$ and the copy a'_y also assigned the weight $v_{\Psi_D(a_y)}$.

We say that a node a in DAG D is a v-node if it has more than one incoming edge. We define the v-degree of a DAG D as $\sum_{a_j \text{ is a v-node}} (d_j^{\text{in}} - 1)$, where d_j^{in} is the in-degree of node a_j. We now establish the following result.

Lemma 4.2.2. *Suppose that a leaf node a_y in the DAG D has at least two incoming edges. Then there exists DAG D^{split} whose* v*-degree is one less than that of D, such that*

$$\lambda_y^{\text{aug}}(D^{\text{split}}) \leq \lambda(D).$$

The inequality is strict when all the parameters under the MNL model are positive.

Furthermore the approximate likelihoods of the DAGs D and D^{split} are equal, i.e., $\tilde{\lambda}_y^{\text{aug}}(D^{\text{split}}) = \tilde{\lambda}(D) = \prod_{a\in\mathcal{N}} \frac{v_a}{\sum_{a'\in\Psi_D(a)} v_{a'}}$.

Proof. of Lemma 4.2.2: Since the leaf node a_y in DAG D has at least two incoming edges, suppose w.l.o.g. that $(a_1, a_y), (a_2, a_y) \in E_D$. Let D_1 denote the DAG obtained by adding the isolated copy a'_y to D. Let D^{split} denote the DAG obtained by erasing the edge (a_1, a_y) and adding the edge (a_2, a'_y) to D_1; in other words, $E_{D^{\text{split}}} = E_D \setminus \{(a_1, a_y)\} \cup \left\{(a_2, a'_y)\right\}$. Figure 4-2 illustrates these DAGs. Note that by construction, the v-degree of D^{split} is one less than that of D because the in-degree of node a_y has been reduced by 1.

We need the following intermediate result.

Claim: $\lambda(D) = \lambda^{\text{aug}}(D_1)$.

Proof. Note that since a_y is a leaf node in D, we have that $v_{\Psi_D(a_y)} = v_y$. Therefore, the distribution λ_y^{aug} is defined on the expanded universe $\mathcal{N} \cup \left\{a'_y\right\}$ with the weights of the products in $\mathcal{N}$ remaining the same as in λ and the weight v_y assigned to product a'_y.

Now, for any ranking $\pi \in S_D$ (including only products in set $\mathcal{N}$), let $\sigma(\pi) \subset S_{D_1}$ denote the set of rankings of the products in the set $\mathcal{N} \cup \left\{a'_y\right\}$ that are consistent with π. By invoking [52, Proposition A1.2], it holds that $\lambda(\pi) = \sum_{\sigma\in\sigma(\pi)} \lambda_y^{\text{aug}}(\sigma)$. We can now write

$$\lambda_y^{\text{aug}}(D_1) = \sum_{\sigma\in S_{D_1}} \lambda_y^{\text{aug}}(\sigma) = \sum_{\pi\in S_D} \sum_{\sigma\in\sigma(\pi)} \lambda_y^{\text{aug}}(\sigma) = \sum_{\pi\in S_D} \lambda(\pi) = \lambda(D).$$

Therefore, in order to establish that $\lambda_y^{\text{aug}}(D^{\text{split}}) \leq \lambda(D)$, it is sufficient to show that $\lambda_y^{\text{aug}}(D^{\text{split}}) \leq \lambda_y^{\text{aug}}(D_1)$. For that, consider the three DAGs I_1, I_2, and I_3 (see Figure 4-3), defined over the set

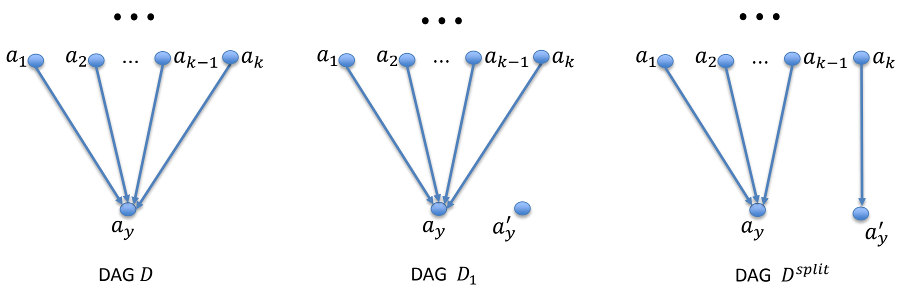

Figure 4-2: Bottom parts of DAGs from Lemma 4.2.2.

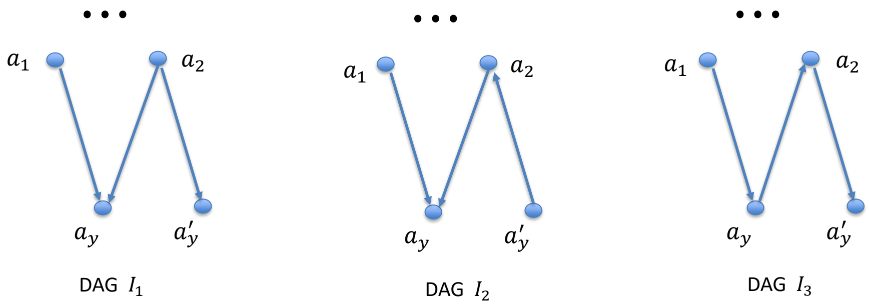

Figure 4-3: Bottom parts of DAGs from Lemma 4.2.2.

of products in $\mathcal{N} \cup \{a'_y\}$ such that

$E_{I_1} = E_{D_1} \cup \left\{(a_2, a'_y)\right\}, E_{I_2} = E_{D_1} \cup \left\{(a'_y, a_2)\right\},$ and $E_3 = \left(E_{D_1} \setminus \{(a_2, a_y)\}\right) \cup \left\{(a_y, a_2), (a_2, a'_y)\right\}.$

It is then follows by the definitions that

$$\lambda_y^{\text{aug}}(D_1) = \lambda_y^{\text{aug}}(I_1) + \lambda_y^{\text{aug}}(I_2)$$
$$\lambda_y^{\text{aug}}(D^{\text{split}}) = \lambda_y^{\text{aug}}(I_1) + \lambda_y^{\text{aug}}(I_3).$$

Thus, to show that $\lambda_y^{\text{aug}}(D^{\text{split}}) \leq \lambda_y^{\text{aug}}(D_1)$, it is sufficient to show that $\lambda_y^{\text{aug}}(I_3) \leq \lambda_y^{\text{aug}}(I_2)$.

To show that $\lambda_y^{\text{aug}}(I_3) \leq \lambda_y^{\text{aug}}(I_2)$, define the mapping $h\colon I_3 \to I_2$ such that for any $\sigma \in I_3$, which is of the form $\sigma = (\ldots, a_1, \ldots, a_y, \ldots, a_2, \ldots, a'_y, \ldots)$, we map it to σ' in I_2, of the form $\sigma' = (\ldots, a_1, \ldots, a'_y, \ldots, a_2, \ldots, a_y, \ldots)$ obtained by swapping the positions of the products a_y and a'_y. Now, it can be verified that the mapping $h(\cdot)$ is an injection, i.e., $h(\sigma) \neq h(\sigma')$ whenever $\sigma \neq \sigma'$. Then, since attraction parameters for nodes a_y and a'_y are the same, i.e., $v_{a_y} = v_{a'_y}$, it follows that for any $\sigma \in I_3$, $\lambda_y^{\mathrm{aug}}(\sigma) = \lambda_y^{\mathrm{aug}}\big(h(\sigma)\big)$. As a result, we obtain

$$\lambda_y^{\mathrm{aug}}(I_3) = \sum_{\sigma \in I_3} \lambda_y^{\mathrm{aug}}(\sigma) = \sum_{\sigma \in I_3} \lambda_y^{\mathrm{aug}}(h(\sigma)) \leq \sum_{\sigma' \in I_2} \lambda_y^{\mathrm{aug}}(\sigma') = \lambda_y^{\mathrm{aug}}(I_2), \tag{4.5}$$

where the inequality holds because there could be additional rankings $\sigma' \in I_2$ that do not have a pre-image in I_3 through $h(\cdot)$. We have thus shown that $\lambda_y^{\mathrm{aug}}(D^{\mathrm{split}}) \leq \lambda_y^{\mathrm{aug}}(D_1)$, which implies that $\lambda_y^{\mathrm{aug}}(D^{\mathrm{split}}) \leq \lambda(D)$.

The inequality in $\lambda_y^{\mathrm{aug}}(D^{\mathrm{split}}) \leq \lambda(D)$ is strict when all the parameters under the MNL model are positive. We establish this result by showing that there exists $\sigma' \in I_2$ such that $h(\sigma) \neq \sigma'$ for all $\sigma \in I_3$. It then follows that the inequality in (4.5) is strict, implying that $\lambda_y^{\mathrm{aug}}(D^{\mathrm{split}}) < \lambda(D)$. Consider the ranking $\sigma' = (\ldots, a'_y, \ldots, a_1, \ldots, a_2, \ldots, a_y, \ldots)$ such that $\sigma' \in I_2$. As noted above, any $\sigma \in I_3$ is of the form $\sigma = (\ldots, a_1, \ldots, a_y, \ldots, a_2, \ldots, a'_y, \ldots)$, so that it gets mapped to $h(\sigma) = (\ldots, a_1, \ldots, a'_y, \ldots, a_2, \ldots, a_y, \ldots)$. Therefore, we have $h(\sigma)(a_1) < h(\sigma)(a'_y)$ for all $\sigma \in I_3$, whereas $\sigma'(a_1) > \sigma'(a'_y)$. Thus, we have that $\sigma' \neq h(\sigma)$ for all $\sigma \in I_3$, establishing the claim.

We are now left with showing that $\tilde{\lambda}_y^{\mathrm{aug}}(D^{\mathrm{split}}) = \tilde{\lambda}(D)$. Since the reachability weights $v_{\Psi_D(a)}$ for all $a \in \mathcal{N}$ under λ, and $v_{\Psi_{D^{\mathrm{split}}}(a)}$ for all $a \in \mathcal{N} \cup \{a'_y\}$ under λ_y^{aug}, are equal by definition, and the approximations $\tilde{\lambda}$ and $\tilde{\lambda}_y^{\mathrm{aug}}$ only depend on the reachability weights, then the equality $\tilde{\lambda}_y^{\mathrm{aug}}(D^{\mathrm{split}}) = \tilde{\lambda}(D)$ immediately follows. □

We can now proceed to prove the results in Section 1.3.

Proof of Proposition 1.3.1

We show the result, $\tilde{\lambda}(D) \leq \lambda(D)$, by induction on the v-degree, k, of DAG D.

Base case: $k = 0$. When $k = 0$, DAG D does not have any v-nodes. Then, D is a forest of directed trees each with a unique root. It follows from [52, Proposition 3.2] that $\tilde{\lambda}(D) = \lambda(D) = \prod_{a \in \mathcal{N}} \frac{v_a}{\sum_{a' \in \Psi_D(a)} v_{a'}}$, establishing the base case.

Induction hypothesis: Suppose $\tilde{\mu}(D) \leq \mu(D)$ for any DAG D with v-degree less than or equal to p, for some $p \geq 0$, for all distributions μ under the PL model.

Induction step: Assuming that the induction hypothesis is true, we prove the result for $k = p+1$. It is clear that there exists a v-node $a_y \in \mathcal{N}$ satisfying the conditions in Lemma 4.2.1, i.e., every node in $\Psi_D(a_y) \setminus \{a_y\}$ has at most one incoming edge and the subgraph $D[a_y]$, induced in D by the set of nodes $\Psi_D(a_y)$ is a directed tree with unique root. As in Lemma 4.2.1, let $\bar{D}[a_y]$ denote the subgraph induced in D by the set of nodes $(\mathcal{N} \setminus \Psi_D(a_y)) \cup \{a_y\}$. Now consider

$$\begin{aligned}
\tilde{\lambda}(D) &= \prod_{j\in\mathcal{N}} \frac{v_j}{\sum_{j'\in\Psi_D(a_j)} v_{j'}} \\
&= \left(\prod_{j\in\Psi_D(a_y)} \frac{v_j}{\sum_{j'\in\Psi_D(a_j)} v_{j'}} \right) \cdot \left(\prod_{j\in\mathcal{N}\setminus\Psi_D(a_y)} \frac{v_j}{\sum_{j'\in\Psi_D(a_j)} v_{j'}} \right) \\
&= \tilde{\lambda}(D[a_y]) \cdot \left(\prod_{\substack{j\in\mathcal{N}\setminus\Psi_D(a_y),\\ a_y\in\Psi_D(a_j)}} \frac{v_j}{\sum_{j'\in\Psi_D(a_j)} v_{j'}} \right) \cdot \left(\prod_{\substack{j\in\mathcal{N}\setminus\Psi_D(a_y),\\ a_y\notin\Psi_D(a_j)}} \frac{v_j}{\sum_{j'\in\Psi_D(a_j)} v_{j'}} \right) \\
&= \lambda(D[a_y]) \cdot \left(\prod_{\substack{j\in\mathcal{N}\setminus\Psi_D(a_y),\\ a_y\in\Psi_D(a_j)}} \frac{v_j}{v_{\Psi_D(a_y)} + \sum_{j'\in\Psi_{\bar{D}[a_y]}(a_j)\setminus\{a_y\}} v_{j'}} \right) \cdot \left(\prod_{\substack{j\in\mathcal{N}\setminus\Psi_D(a_y),\\ a_y\notin\Psi_D(a_j)}} \frac{v_j}{\sum_{j'\in\Psi_{\bar{D}[a_y]}(a_j)} v_{j'}} \right) \\
&= \lambda(D[a_y]) \cdot \tilde{\lambda}_y(\bar{D}[a_y]),
\end{aligned}$$

where the fourth equation follows because $D[a_y]$ is a directed tree with a unique root, which implies that $\lambda(D[a_y]) = \tilde{\lambda}(D[a_y])$ [52, Proposition 3.2], and the fact that $\Psi_D(a_j) = \Psi_{\bar{D}[a_y]}(a_j)$ for all $j \in \mathcal{N}$ such that $a_y \notin \Psi_D(a_j)$. We now have

$$\begin{aligned}
\tilde{\lambda}(D) &= \lambda(D[a_y]) \cdot \tilde{\lambda}_y(\bar{D}[a_y]) \\
&= \lambda(D[a_y]) \cdot \tilde{\lambda}_y^{\text{aug}}(D_y^{\text{split}}) \\
&\leq \lambda(D[a_y]) \cdot \lambda_y^{\text{aug}}(D_y^{\text{split}}) && \text{[by the ind. hypoth.]} \\
&\leq \lambda(D[a_y]) \cdot \lambda_y(\bar{D}[a_y]), \text{ with strict inequality if } v_j > 0 \; \forall \; a_j \in \mathcal{N} && \text{[by Lemma 4.2.2]} \\
&= \lambda(D) && \text{[by Lemma 4.2.1]},
\end{aligned}$$

where the second equality holds by Lemma 4.2.2 taking $\bar{D}[a_y]$ here as D there, and D_y^{split} here as D^{split} there; and the first inequality follows from induction hypothesis with distribution $\mu = \lambda_y^{\text{aug}}$ since the v-degree of D_y^{split} is equal to p. The result of the proposition now follows. □

Proof of Proposition 1.3.2

We must have that $S_D \subset S_{\bar{D}}$ since if σ is consistent with D, i.e., $\sigma \in S_D$, then it must also be consistent with $\bar{D}$, i.e., $\sigma \in S_{\bar{D}}$. It now follows that

$$\lambda(D) = \sum_{\sigma \in S_D} \lambda(\sigma) \leq \sum_{\sigma \in S_{\bar{D}}} \lambda(\sigma) = \lambda(\bar{D}).$$

We now show that $\lambda(D) < \lambda(\bar{D})$ when all the PL parameters are strictly positive by exhibiting a ranking $\sigma \in S_{\bar{D}}$ such that $\sigma \notin S_D$. Suppose, by contradiction, we have that $S_{\bar{D}} = S_D$. Then, any subgraph of $\bar{D}$ which has less edges than D, is a transitive reduction of D, which results in contradiction (recall that DAG D is assumed to be its unique transitive reduction). As a result, there is $\sigma \in S_{\bar{D}}$ such that $\sigma \notin S_D$. Then, we have

$$\lambda(\bar{D}) - \lambda(D) \geq \lambda(\sigma) > 0,$$

which holds because all the parameters of λ are positive. □

For the proof of Proposition 1.3.3, we recall here some notation. Let $R(D, \bar{D})$ denote the ratio of the upper bound to the lower bound $\lambda(\bar{D})/\tilde{\lambda}(D)$, for any DAG $\bar{D} \subset D$. Let ℓ denote the size of the largest reachability set in DAG D, i.e., $\ell = \max_{a \in \mathcal{N}} |\Psi_D(a)|$, and let p denote the number of nodes with v-nodes in their reachability sets, i.e., $p = |\{a \in \mathcal{N} \colon \exists \text{v-node } b \in \Psi_D(a)\}|$. Further, let $\Delta := \max_a \max_{b \in \Psi_D(a) \setminus \{a\}} v_b / v_a$ be the maximum ratio between the weights of nodes within the same directed path in the DAG.

Proof of Proposition 1.3.3 Define $\Phi(D) \subset D$ as a DAG with each node having a unique parent, such that for any distribution λ, $\lambda(\Phi(D)) \geq \lambda(D)$. Recall that $\bar{D}$ is a forest of directed trees obtained by deleting arcs from D in order to break the v-nodes. Thus, set $\bar{D} = \Phi(D)$ so that each node has a unique parent, verifying

$$\lambda(\bar{D}) = \prod_{a \in \mathcal{N}} \frac{v_a}{\sum_{a_j \in \Psi_{\Phi(D)}(a)} v_j}.$$

In turn, $\tilde{\lambda}(D)$ is the lower bound obtained by treating D as a forest of directed trees with unique root. That is,

$$\tilde{\lambda}(D) = \prod_{a \in \mathcal{N}} \frac{v_a}{\sum_{a_j \in \Psi_D(a)} v_j}.$$

We have from Propositions 1.3.1 and 1.3.2:

$$\tilde{\lambda}(D) \leq \lambda(D) \leq \lambda(\bar{D}). \tag{4.6}$$

Then,

$$\log R(D, \bar{D}) = \log \frac{\lambda(\bar{D})}{\tilde{\lambda}(D)} = \log \left(\prod_{a \in \mathcal{N}} \frac{\sum_{a_j \in \Psi_D(a)} v_j}{\sum_{a_j \in \Psi_{\mathbf{\Phi}(D)}(a)} v_j} \right)$$

Let $\mathcal{F}_{\mathcal{D}}$ be the set of nodes in D with more than one incoming edge. Continuing the sequence of equalities above:

$$\begin{aligned}
\log R(D, \bar{D}) &= \log \left(\prod_{a \in \mathcal{N}} \frac{\sum_{a_j \in \Psi_D(a)} v_j}{\sum_{a_j \in \Psi_{\mathbf{\Phi}(D)}(a)} v_j} \right) = \sum_{a \in \mathcal{N}} \log \left(\frac{\sum_{a_j \in \Psi_D(a)} v_j}{\sum_{a_j \in \Psi_{\mathbf{\Phi}(D)}(a)} v_j} \right) \\
&= \sum_{a \in \mathcal{N}} \log \left(1 + \frac{\sum_{a_j \in \Psi_D(a) \setminus \Psi_{\mathbf{\Phi}(D)}(a)} v_j}{\sum_{a_j \in \Psi_{\mathbf{\Phi}(D)}(a)} v_j} \right) \\
&= \sum_{a \in \mathcal{N}} \mathbb{I}[\mathcal{F}_D \cap \Psi_D(a) \neq \varnothing] \cdot \log \left(1 + \frac{\sum_{a_j \in \Psi_D(a) \setminus \Psi_{\mathbf{\Phi}(D)}(a)} v_j}{\sum_{a_j \in \Psi_{\mathbf{\Phi}(D)}(a)} v_j} \right) \\
&\leq \sum_{a \in \mathcal{N}} \mathbb{I}[\mathcal{F}_D \cap \Psi_D(a) \neq \varnothing] \cdot \log \left(1 + \frac{\sum_{a_j \in \Psi_D(a) \setminus \{a\}} v_j}{v_a} \right) \\
&= \sum_{a \in \mathcal{N}} \mathbb{I}[\mathcal{F}_D \cap \Psi_D(a) \neq \varnothing] \cdot \log \left(1 + \sum_{a_j \in \Psi_D(a) \setminus \{a\}} \frac{v_j}{v_a} \right) \\
&\leq \sum_{a \in \mathcal{N}} \mathbb{I}[\mathcal{F}_D \cap \Psi_D(a) \neq \varnothing] \cdot \log \left(1 + \sum_{a_j \in \Psi_D(a) \setminus \{a\}} \Delta \right) \\
&\leq \sum_{a \in \mathcal{N}} \mathbb{I}[\mathcal{F}_D \cap \Psi_D(a) \neq \varnothing] \cdot \log(1 + \ell \cdot \Delta) \leq p \cdot \log(1 + \ell \cdot \Delta),
\end{aligned}$$

where the fourth equality follows since the surviving terms are the once where $\Psi_D(a) \setminus \Psi_{\mathbf{\Phi}(D)}(a) \neq \emptyset$, i.e., there are nodes in $\Psi_D(a)$ with more than one incoming edge; and the first inequality holds because $a \in \Psi(a)$, and we add terms in the numerator and take out terms from the denominator. The last three inequalities follow from the definitions of Δ, ℓ, and p, respectively. From (4.6), we have that

$$0 \leq \lim_{n\to\infty} \log \frac{\lambda(D)}{\tilde{\lambda}(D)} \leq \lim_{n\to\infty} \log R(D,\bar{D}) \leq \lim_{n\to\infty} p \cdot \log(1+\ell\cdot\Delta) \leq \lim_{n\to\infty} n\cdot\log(1+n\cdot\Delta)$$
$$= \lim_{n\to\infty} \log(1+n\cdot\Delta)^{\frac{n^2\cdot\Delta}{n\cdot\Delta}} = \lim_{n\to\infty} (\Delta n^2)\cdot\log(1+n\cdot\Delta)^{\frac{1}{n\cdot\Delta}} = 0,$$

since as $n \to \infty$, it can be shown that $n\Delta \in o(n^{-1})$ and $n^2\Delta \in o(1)$. $\square$

Proof of Proposition 1.3.4

Recall that the merged DAG $D \uplus C(a_j, S)$ is obtained by taking the union of the graphs D and $C(a_j, S)$. From the definitions of the bounds in (1.4), it must hold that

$$\begin{aligned}
\log \frac{\overline{f}(a_j, S, D)}{\underline{f}(a_j, S, D)} &= \log\left(\frac{\lambda(\overline{D \uplus C(a_j, S)})}{\tilde{\lambda}(D \uplus C(a_j, S))} \cdot \frac{\lambda(\overline{D})}{\tilde{\lambda}(D)} \right) \\
&= \log\left(\frac{\lambda(\overline{D \uplus C(a_j, S)})}{\tilde{\lambda}(D \uplus C(a_j, S))} \right) + \log\left(\frac{\lambda(\overline{D})}{\tilde{\lambda}(D)} \right) \\
&\leq 2 \cdot p \cdot \log(1 + \ell \cdot \Delta),
\end{aligned}$$

where the last inequality follows from Proposition 1.3.3. As argued in the proof of Proposition 1.3.3, $p \log(1 + \ell\Delta) \to 0$ as $n \to \infty$ when $\Delta n^2 \in o(1)$, as $n \to \infty$. □

4.3 Heuristic for preference graph decycling

In Section 1.2.3, we formulated MILP (1.2) for preference graph decycling (Phase 3) in the DAG construction process. Since solving the MILP to optimality could be challenging (e.g., if we have thousands of products or brands), we propose a tractable, greedy heuristic to decycle the preference graph.

Let $FindPath(a_k, a_j, G)$ denote the output of Dijkstra's algorithm on a directed graph G, which finds the shortest path between nodes a_k and a_j and returns the set of weighted edges comprising this path (potentially, the empty set). The Dijkstra's algorithm runs in $O(|V_G|^2)$ time, where V_G is the set of nodes in G.

Taking advantage of the polynomial running time of Dijkstra's, our heuristic proceeds as follows: For a directed graph G, we run Dijkstra's between all pair of nodes a_k and a_j, in both directions. In case both paths exist, then there is a cycle containing a_k and a_j, and the edge with minimum weight is removed. As a result, since there are $O(|V_G|^2)$ pairs of nodes, the preference graph decycling can be implemented with $O(|V_G|^4)$ computational complexity. The steps are described in Algorithm 3 below.

In order to validate the effectiveness of Algorithm 3, we compare empirical prediction results on the actual sales dataset obtained from DAGs decycled via MIP (1.2) and the analogous results

Algorithm 3 Preference graph G decycling

```
procedure DECYCLE($G$)    ▷ Where $G$ is a graph with set of nodes $V_G$ and set of weighted edges $E_G$.
    for $a_k$ in $V_G$ do
        for $a_j$ in $V_G \setminus \{a_k\}$ do
            while $FindPath(a_k, a_j, G) \neq \emptyset$ & $FindPath(a_j, a_k, G) \neq \emptyset$ do
                $Cycle \leftarrow FindPath(a_k, a_j, G) \cup FindPath(a_j, a_k, G)$
                Remove the edge $(a_x, a_y)$ with minimum weight in $Cycle$ from the set $E_G$
    return DAG $D = G$
```

from DAGs decycled via Algorithm 3. A description of the sales data is provided in Section 1.4.1 in the main body of the thesis.

In Figure 4-4 (left panels), we observe that using MIP (1.2) we delete 6.27% fewer edges than when using Algorithm 3 and obtain 0.4% denser DAGs on average across 27 product categories. In the middle and right panels we represent the scatter plot over 27 product categories of the average miss rate and χ^2 scores, respectively. It follows that by using MIP (1.2) we obtain 1.17% lower miss rate and 2.75% lower χ^2 score than by using Algorithm 3. These results provide good support for the use of the greedy heuristic as an alternative to the exact solution of MIP (1.2) in cases where the number of products is large. We highlight here though that in all our experiments reported in the main body of the thesis we used MIP (1.2) limited to a max time of 30 seconds, and retaining the best feasible solution when optimality was not reached. According to Table 1.1, the largest category contains 95 products in our case.

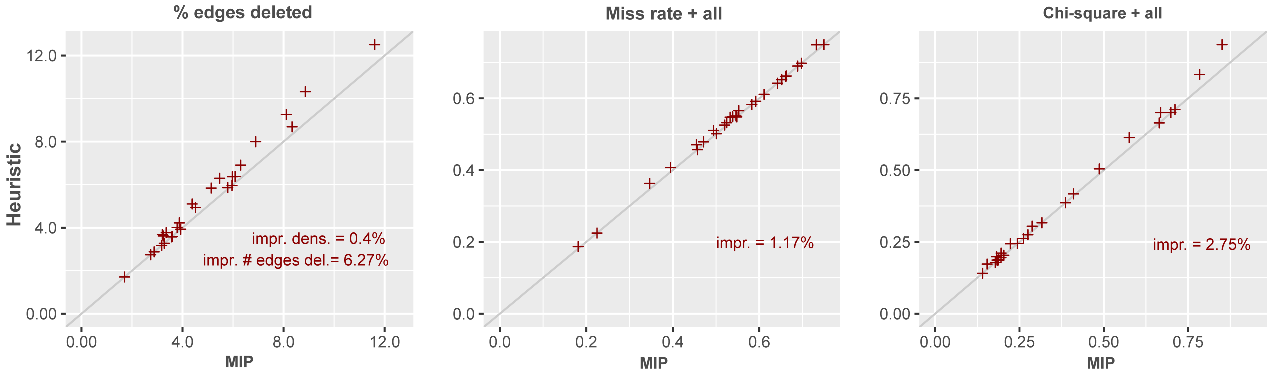

Figure 4-4: Comparison of the performance of heuristic algorithm 3 vs. MIP (1.2).

4.4 Benchmark models

4.4.1 LC-MNL model

LC-MNL model captures heterogeneity among customers by allowing them to belong to K different classes with some probability. Customers from class $h \in \{1,..,K\}$ make choices according to the single class MNL model with a parameter value $\beta^0_{hj_{it}} + I_{j_{it}}\beta_{hj_{it}}$ of product $j_{it} \in \{1, 2, ..., n\}$, where $I_{j_{it}} = 1$ if product j_{it} is under promotion at time t for individual i, and 0 otherwise. A prior probability of a customer to belong to the class h is $\gamma_h \geq 0$ such that $\sum_{h=1}^{K} \gamma_h = 1$. The regularized maximum likelihood problem under K class LC-MNL model can be formulated as follows:

$$\max_{\boldsymbol{\beta},\boldsymbol{\gamma}:\ \beta^0_{h1}=0, h\in[K]} \sum_{i=1}^{m} \log\left(\sum_{h=1}^{K} \gamma_h \prod_{t=1}^{T_i} \frac{\exp\big(\beta^0_{hj_{it}} + I_{j_{it}}\beta_{hj_{it}}\big)}{\sum_{a_\ell\in S_{it}} \exp(\beta^0_{h\ell} + I_{\ell_{it}}\beta_{h\ell})}\right) - \alpha \sum_{h=1}^{K} \left(\left\|\boldsymbol{\beta}^0_h\right\|_1 + \left\|\boldsymbol{\beta}_h\right\|_1\right)$$

When the value of α is fixed and $K = 1$, it can be shown that this optimization problem is globally concave and therefore can be solved efficiently [92]. Note that we tuned the value of α by 5-fold cross-validation. Since the problem is nonconcave for $K > 1$, the EM technique is used to fit the model (see Appendix A2.1.1 in [52]). Specifically, we initialize the EM with a random allocation of customers to one of the K classes, resulting in an initial allocation $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$, which form a partition of the collection of all the customers. Then, we set $\gamma_h^{(0)} = |\mathcal{D}_h| / \big(\sum_{d=1}^{K} |\mathcal{D}_d|\big)$. In order to get a parameter vector $(\boldsymbol{\beta}_h^{(0)}, \boldsymbol{\beta}_h)$, we fit a single class MNL model to each subset of customers. Based on each customer i's purchase history $(a_{j_{it}}, S_{it})$ for $1 \leq t \leq T_i$, we can estimate their posterior membership probabilities $\forall\ h \in \{1,..,K\}$:

$$\hat{\gamma}_{ih} = \frac{\gamma_h \prod_{t\in T_i} \left[\exp\big(\beta^0_{hj_{it}} + I_{j_{it}}\beta_{hj_{it}}\big) / \Big(\sum_{a_\ell\in S_{it}} \exp(\beta^0_{h\ell} + I_{\ell_{it}}\beta_{h\ell})\Big)\right]}{\sum_{d=1}^{K} \gamma_d \prod_{t\in T_i} \left[\exp\big(\beta^0_{dj_{it}} + I_{j_{it}}\beta_{dj_{it}}\big) / \Big(\sum_{a_\ell\in S_{it}} \exp(\beta^0_{h\ell} + I_{\ell_{it}}\beta_{h\ell})\Big)\right]},$$

and the prediction can be made as follows:

$$f(j_{it}, S_{it}) = \sum_{h=1}^{K} \hat{\gamma}_{ih} \frac{\exp\big(\beta^0_{hj_{it}} + I_{j_{it}}\beta_{hj_{it}}\big)}{\sum_{a_\ell\in S_{it}} \exp(\beta^0_{h\ell} + I_{\ell_{it}}\beta_{h\ell})},$$

where $f(j_{it}, S_{it})$ is a probability to choose an item j_{it} from the offer set S_{it}.

4.4.2 RPL model

In this model, we assume that $\boldsymbol{\beta}$ is sampled from multivariate normal distribution, i.e, $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ is the mean, and Σ is the covariance matrix, which is assumed to be diagonal. Then the log-likelihood of the sequence of purchases of all individuals $i \in \{1,..,m\}$ for $t = \{1, ..., T_i\}$ is equal to $\sum_{i=1}^{m} \log\left(\int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T_i} \frac{\exp\left(\beta^0_{j_{it}} + I_{j_{it}}\beta_{j_{it}}\right)}{\sum_{a_\ell \in S_{it}} \exp\left(\beta^0_\ell + I_{\ell_{it}}\beta_\ell\right)}\right] \phi(\boldsymbol{\beta}) d\boldsymbol{\beta}\right)$ such that $\beta^0_{j_{it}} + I_{j_{it}}\beta_{j_{it}}$ is a parameter value of product $j_{it} \in \{1, 2, ..., n\}$, where $I_{j_{it}} = 1$ if product j_{it} is under promotion at time t for individual i, and 0 otherwise. Model parameters are estimated through maximum simulated likelihood estimation (MSLE) where we use the simulated probabilities to approximate the following log-likelihood function:

$$\max_{\boldsymbol{\mu}, \Sigma:\ \mu_1 = 0} \sum_{i=1}^{m} \log\left(\int_{-\infty}^{+\infty} \left[\prod_{t=1}^{T_i} \frac{\exp\left(\beta^{0r}_{j_{it}} + I_{j_{it}}\beta^r_{j_{it}}\right)}{\sum_{a_\ell \in S_{it}} \exp(\beta^{0r}_\ell + I_{\ell_{it}}\beta^r_\ell)}\right] \phi(\boldsymbol{\beta}) d\boldsymbol{\beta}\right),$$

where for any random draw $r = 1, 2..., R$ of a random vector $\boldsymbol{\xi}^r$, that is sampled as $2n-$dimensional multivariate standard normal, we have that $\beta^r_\ell = \mu_\ell + \xi^r_\ell \sigma_\ell$, for any $\ell = 1, 2, ..., n$, and $\beta^{0r}_{\ell - n} = \mu_\ell + \xi^r_\ell \sigma_\ell$, $\ell = n+1, n+2, ..., 2n$. The above optimization problem is nonconcave. To solve the problem, we choose R = 400 and use a general non-linear solver to converge to a stationary point (see Appendix A2.1.1 in [52]). Then we make predictions as follows:

$$f(j_{it}, S_{it}) = \int \frac{\exp\left(\beta^0_{j_{it}} + I_{j_{it}}\beta_{j_{it}}\right)}{\sum_{a_\ell \in S_{it}} \exp(\beta^0_\ell + I_{\ell_{it}}\beta_\ell)} \hat{\phi}(\boldsymbol{\beta} | \mathcal{H}_i; \boldsymbol{\mu}, \Sigma) d\boldsymbol{\beta},$$

where $f(j_{it}, S_{it})$ is the probability to choose an item j_{it} from the offer set S_{it} for individual i, $\hat{\phi}(\boldsymbol{\beta} | \mathcal{H}_i; \boldsymbol{\mu}, \Sigma)$ is the posterior distribution of parameter vector $\boldsymbol{\beta}$ for customer i, conditioning on population prior and observed choices of customer i, i.e., $\mathcal{H}_i = \{(a_{j_{it}}, S_{it}) : 1 \leq t \leq T_i\}$.

4.5 Evaluation of analytical bounds

In this section, we focus on the PO-MNL Promotion model and illustrate the behavior and quality of the bounds proposed in Section 1.3.

4.5.1 Bounds on the probability of a DAG

The presence of v-nodes (i.e., nodes with more than one incoming edge) in DAGs of individuals complicates the maximum likelihood estimation of parameter values under PO-based choice models. The left panels in Figure 4-5 illustrate that individuals without cycles in their preference graph have on average 17.09 v-nodes in their DAG whereas individuals with cycles in their preference graph have on average 19.93 v-nodes in their DAG.

A tractable approximation of the likelihood of a DAG D is given by

$$\tilde{\lambda}(D) = \prod_{a_j \in \mathcal{N}} \frac{\upsilon_j}{\sum_{a_k \in \Psi_D(a_j)} \upsilon_k},$$

where $\upsilon_j = \exp(\beta_j)$, $\forall\ a_j \in \mathcal{N}$ and $\Psi_D(a_j)$ denotes the reachability function such that $\Psi_D(a_j) = \{a_k : a_k \text{ is reachable from } a_j \text{ in } D\}$. Note that $\Psi_D(a_j)$ is always nonempty, since we assume that each node a_j is reachable from itself. The approximation $\tilde{\lambda}(D)$ of the likelihood of DAG D is exact when D is a forest of directed trees, each with a unique root. We show in Proposition 1.3.1 that $\tilde{\lambda}(D)$ is a lower bound for the likelihood of DAG D.

Next, in order to find the upper bound approximation of DAG D likelihood, let us denote $\bar{D}$ the DAG obtained from D where for every node with more than one incoming edge we delete all the incoming edges but one. Instead of deleting an arbitrary set of edges, we can determine edges to delete to make the approximation as tight as possible. Finding the tightest upper bound is challenging in general. In order to ease the computational process, we develop a greedy-type heuristic $\mathbf{\Phi}(D)$ (see Algorithm 4) to obtain a tight upper bound of DAG D likelihood, i.e., $\bar{D} = \mathbf{\Phi}(D)$ and $\lambda(D) \leq \lambda(\bar{D})$.

For a collection of panel data represented by X and a given set of parameters $\boldsymbol{\beta}$, let $\log \overline{\mathcal{L}}(X, \boldsymbol{\beta})$ denote the upper bound approximation of the log-likelihood function under PO-MNL Promotion model defined as $\log \overline{\mathcal{L}}(X, \boldsymbol{\beta}) = \sum_{i=1}^{m} \log \lambda(\bar{D}_i)$. Then, letting $\bar{\boldsymbol{\beta}}^*$ be the solution to the maximization problem of the upper bound of the likelihood function, i.e., $\bar{\boldsymbol{\beta}}^* = \arg\max_{\boldsymbol{\beta}} \log \overline{\mathcal{L}}(X, \boldsymbol{\beta})$, we have that the maximum value of the exact log-likelihood function $\log \mathcal{L}(X, \boldsymbol{\beta}^*)$ satisfies:

$$\log \mathcal{L}(X, \boldsymbol{\beta}^*) \leq \log \overline{\mathcal{L}}(X, \boldsymbol{\beta}^*) \leq \log \overline{\mathcal{L}}(X, \bar{\boldsymbol{\beta}}^*).$$

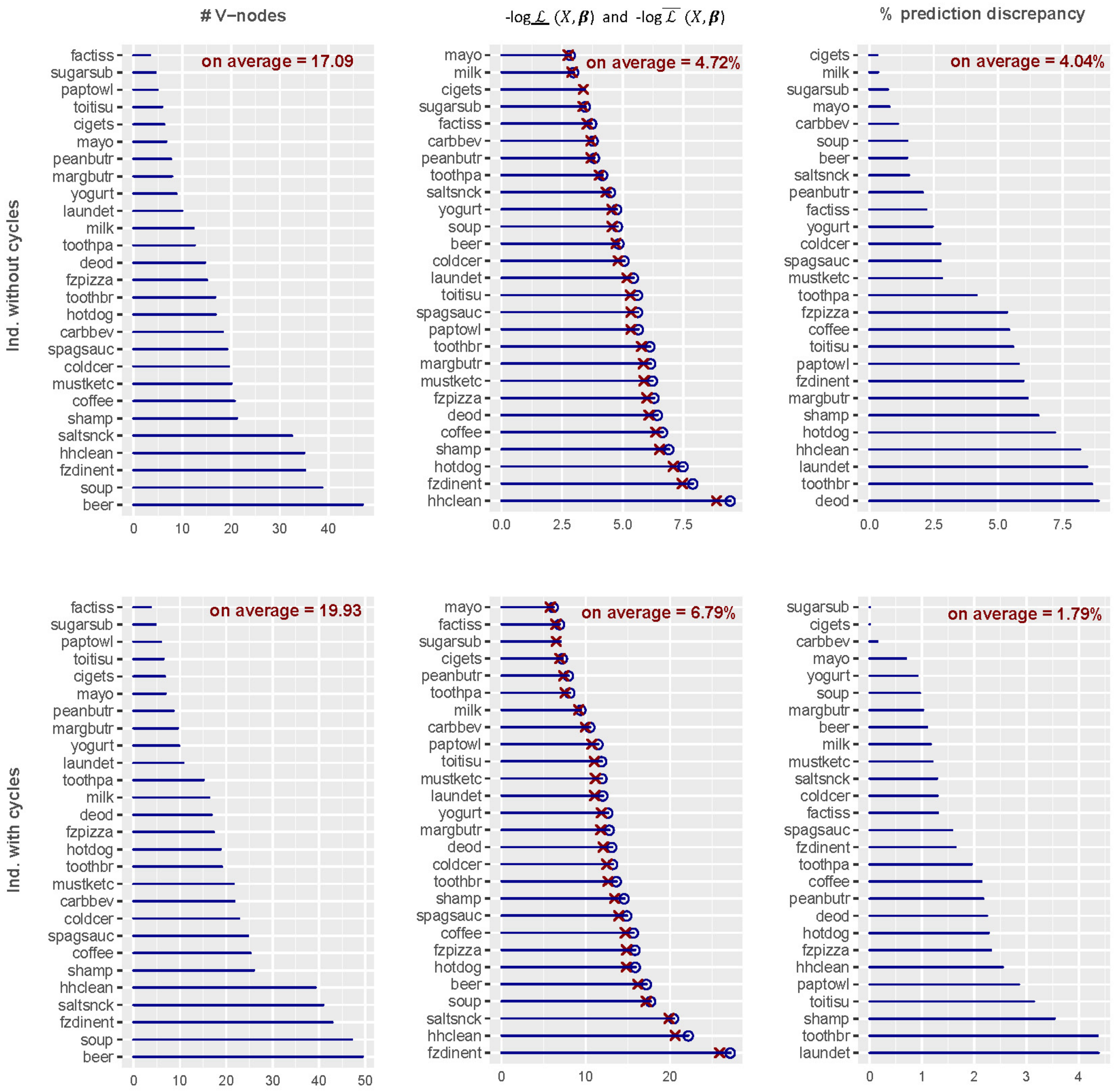

Figure 4-5: Analysis of bounds for the probability of a DAG.

Similarly, let $\log \underline{\mathcal{L}}(X, \boldsymbol{\beta}) = \sum_{i=1}^{m} \log \tilde{\lambda}(D_i)$ denote the lower bound approximation of the log-likelihood function under PO-MNL Promotion model, with optimal values $\underline{\boldsymbol{\beta}}^*$. Then,

$$\log \underline{\mathcal{L}}(X, \underline{\boldsymbol{\beta}}^*) \leq \log \mathcal{L}(X, \underline{\boldsymbol{\beta}}^*) \leq \log \mathcal{L}(X, \boldsymbol{\beta}^*).$$

Algorithm 4 DAG D transformation to find its upper bound likelihood

```
procedure $\mathbf{\Phi}(D)$,    ▷ where $\mathbf{\Phi}(D)$ is the DAG with each node having a unique parent s.t. $\lambda(\mathbf{\Phi}(D)) \geq \lambda(D)$
    $A \leftarrow \mathcal{F}_D$    ▷ $\mathcal{F}_D$ is the set of nodes in $D$ with more than one incoming edge
    for $a_i$ in $\mathcal{F}_D$ do
        $D'$ is obtained from $D$: $V_{D'} = V_D$, and $E_{D'} = E_D \setminus B_i$,    ▷ where $B_i$ is the set of incoming edges into node $a_i$
        $D \leftarrow D'$
    while $A \neq \emptyset$ : do
        $(a_x, a_y) = \underset{(a_i,a_j)\in B_i}{\arg\min}\ \lambda(D')$ s.t. $D'$: $V_{D'} = V_D$, $E_{D'} = E_D \cup (a_i, a_j)$ and $a_j \in A$
        $D \leftarrow D'$
        $A \leftarrow A \setminus \{a_y\}$
    return DAG $\bar{D} = \mathbf{\Phi}(D)$
```

A natural question that arises is about the size of the gap between both easy-to-compute bounds (lower and upper). The middle column of the panels in Figure 4-5 illustrates that the upper bound of the log-likelihood function (i.e., $\log \overline{\mathcal{L}}(X, \boldsymbol{\beta})$) is higher than the lower bound of the log-likelihood function (i.e., $\log \underline{\mathcal{L}}(X, \boldsymbol{\beta})$) by 4.72% for individuals without cycles in their preference graph, and by 6.79% for individuals with cycles in their preference graph, on average across 27 product categories. This observation provides good support to use any of the bounds as an approximation for the estimation problem under the exact likelihood of the DAGs. In particular, we used the lower bound $\log \underline{\mathcal{L}}(X, \underline{\boldsymbol{\beta}}^*)$.

4.5.2 Bounds on the probability of purchase

Next, we illustrate the behavior and quality of the bounds we have developed for posterior probabilities of purchase when customers make choices consistently with their partial orders. In particular, we propose the approximate probability of choosing product a_j from offer set S assuming that the sampled preference list is consistent with DAG D:

$$\hat{f}(a_j, S, D) = \begin{cases} \frac{\tilde{\lambda}(D \uplus C(a_j,S))}{\tilde{\lambda}(D)}, & \text{if } a_j \in h_D(S), \\ 0, & \text{otherwise.} \end{cases}$$

Letting $\underline{f}(a_j, S, D)$ denote the lower bound of the purchase probability and $\overline{f}(a_j, S, D)$ denote the upper bound of the purchase probability such that $\underline{f}(a_j, S, D) = \frac{\tilde{\lambda}(D \uplus C(a_j,S))}{\lambda(\overline{D})}$ if $a_j \in h_D(S)$ and 0, otherwise; $\overline{f}(a_j, S, D) = \frac{\lambda(\overline{D \uplus C(a_j,S)})}{\tilde{\lambda}(D)}$ if $a_j \in h_D(S)$, and 0, otherwise; we have the following inequalities (see Corollary 1.3.1 in Section 1.3):

$$\underline{f}(a_j, S, D) \leq \hat{f}(a_j, S, D) \leq \overline{f}(a_j, S, D),$$

and for the exact and hard-to-compute probability of purchase $f(a_j, S, D)$,

$$\underline{f}(a_j, S, D) \leq f(a_j, S, D) \leq \overline{f}(a_j, S, D).$$

The right column of Figure 4-5 illustrates that the percentage of transactions when the prediction of the item to be chosen is made using the upper bound posterior probability of purchase, i.e., $\overline{f}(a_j, S, D)$, is different from the prediction of the item to be chosen using the lower bound posterior probability of purchase, i.e., $\underline{f}(a_j, S, D)$, in only 4.04% of the instances for individuals without cycles in their preference graph, and in only 1.79% of the instances for individuals with cycles in their preference graph. In both cases, our prediction is the item with the highest probability of being purchased. This empirical observation provides good support for the use of $\hat{f}(a_j, S, D)$ as a proxy for the true and hard-to-compute probability of purchase $f(a_j, S, D)$.

In our reported results in Section 1.4 we use the following tractable formula to compute the posterior probabilities of purchase:

$$\tilde{f}(a_j, S, D) = \begin{cases} \frac{v_{\Psi_D(a_j)}}{\sum_{a_k \in h_D(S)} v_{\Psi_D(a_k)}}, & \text{if } a_j \in h_D(S), \\ 0, & \text{otherwise.} \end{cases}$$

This expression is intended to be a good approximation for the alternative approximation $\hat{f}(a_j, S, D)$, which we already know is a good approximation for the exact $f(a_j, S, D)$. We verify this in Figure 4-6. Therein, we compare the choice prediction results made with $\tilde{f}(a_j, S, D)$ vs. the choice prediction results made with $\hat{f}(a_j, S, D)$ for individuals with and without cycles under "chi-square" score and miss rate (see description of the metrics in Section 1.4.3). For all the panels in Figure 4-6, the average MAE (Mean Absolute Error) is below 0.5%, which indicates that the posterior probability approximation $\tilde{f}(a_j, S, D)$ is very close to the posterior probability based on the lower bound of the DAG likelihood $f(a_j, S, D)$ in terms of predictive performance.

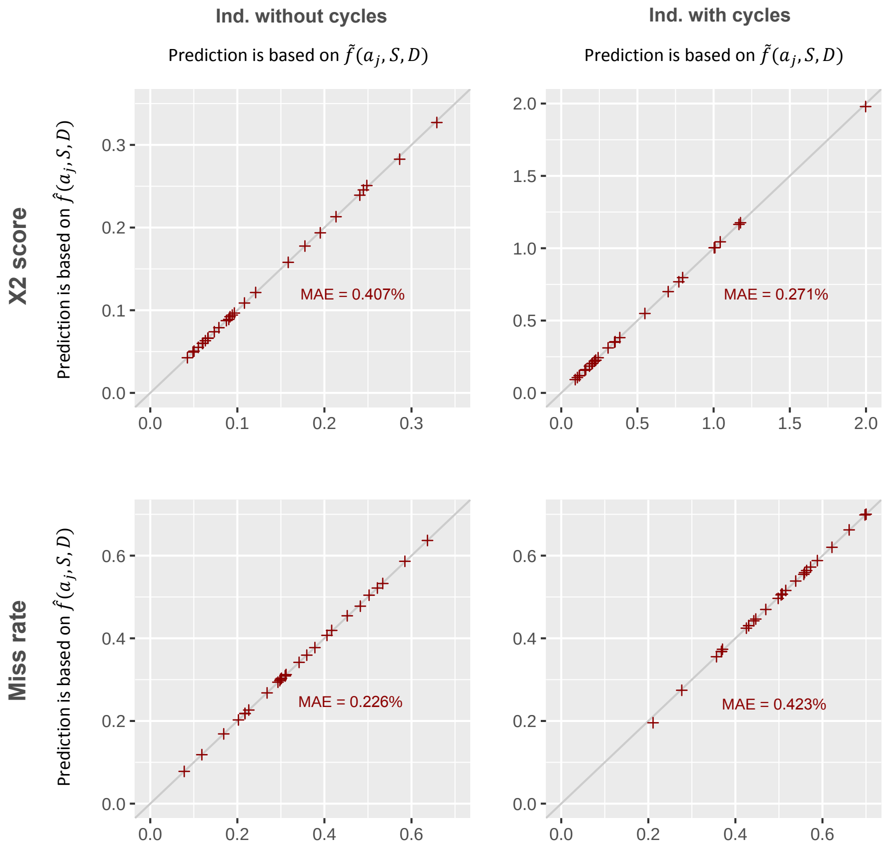

Figure 4-6: Comparison of approximations for purchase probabilities.

4.6 Optimization of personalized promotions

We now show that the set of constraints (1.16)–(1.19) ensures that $\boldsymbol{p}$'s are normalized attraction values. Recall that $z_j = 1$ for all products a_j in the set $h_D(S(\boldsymbol{y}))$ of heads in the subgraph of the transitive closure of D restricted to the set $S(\boldsymbol{y})$.

Lemma 4.6.1. *Suppose* $(0 \le p_j \le 1 \colon a_j \in S_{\mathcal{A}})$ *satisfy* (1.16)–(1.19), *then*

$$p_j = \begin{cases} 0, & \textit{if } z_j = 0, \\ \frac{v_{\Psi_D(a_j)}}{1+\sum_{k:\ a_k \in S_{\mathcal{A}}} v_{\Psi_D(a_k)}}, & \textit{if } z_j = 1. \end{cases}$$

Proof. Proof: To simplify notation, let w_j denote $v_{\Psi_D(a_j)}$ for each $a_j \in S_{\mathcal{A}}$. For convenience, we reproduce the set of constraints (1.16)–(1.19) below:

$$p_j \leq z_j \quad \forall \ a_j \in S_{\mathcal{A}}, \tag{4.7}$$

$$p_0 + \sum_{j \colon a_j \in S_{\mathcal{A}}} p_j = 1, \tag{4.8}$$

$$0 \leq p_j \leq w_j p_0, \quad \forall \ a_j \in S_{\mathcal{A}} \tag{4.9}$$

$$p_0 + z_j - 1 \leq p_j / w_j \quad \forall \ a_j \in S_{\mathcal{A}}. \tag{4.10}$$

Let S denote the set $\{a_j \colon z_j = 1\}$, consisting of all the product indices such that $z_j = 1$. It immediately follows from (4.7) that $p_j = 0$ for all $a_j \in S_{\mathcal{A}} \setminus S$ (since $z_j = 0$ therein). Then, for all $a_j \in S_{\mathcal{A}} \setminus S$, (4.9) trivially holds and (4.10) reduces to $p_0 \leq 1$ (which also trivially holds).

Now, for any j such that $a_j \in S$ (and hence, $z_j = 1$), we have from (4.9) and (4.10) that

$$0 \leq p_j \leq w_j p_0 \text{ and } p_0 \leq p_j / w_j.$$

It thus follows that $p_j = p_0 w_j$ for all $a_j \in S$. We now obtain from (4.8) that

$$p_0 + \sum_{a_j \in S_{\mathcal{A}}} p_j = 1 \implies p_0 + \sum_{a_j \in S} p_j = 1 \implies p_0 + \sum_{a_j \in S} p_0 w_j = 1 \implies p_0 = 1 \Big/ \left(1 + \sum_{a_j \in S} w_j \right),$$

where the first implication follows because $p_j = 0$ for all $a_j \in S_{\mathcal{A}} \setminus S$ and the second implication follows because $p_j = p_0 w_j$ for all $a_j \in S$. Since $w_j > 0$ for all a_j, it follows that $p_0 \leq 1$, as needed. We have thus obtained that

$$p_j = 0 \text{ for all } a_j \in S_{\mathcal{A}} \setminus S \text{ and } p_j = w_j p_0 = w_j \Big/ \left(1 + \sum_{a_k \in S} w_k \right) \text{ for all } a_j \in S.$$

In other words, we have that $p_j = w_j z_j / \left(1 + \sum_{a_k \in S_{\mathcal{A}}} w_k z_k\right)$, which follows from the definition of S. The result of the lemma now holds. □

Chapter 5

Proofs and Supplementary Materials for Chapter 2

5.1 Preliminaries on consider-then-choose models

For completeness, we summarize the relevant notation from Chapter 2 and also introduce additional notation. We consider a universe N of n products $\{a_1, a_2, \ldots, a_n\}$. We let a_0 denote the 'no-purchase' or the 'outside' option. A customer is presented with a subset $S \subseteq N$ of products and the customer chooses either one of the products in S or the outside option a_0. We let $\mathbb{P}_j(S)$ denote the probability that a customer chooses product $a_j \in S$ and $\mathbb{P}_0(S)$ the probability that the customer chooses the outside option. We use S^+ to denote the set $S \cup \{a_0\}$. Let $\lambda\colon 2^N \to [0,1]$ define a distribution over consideration sets such that $\sum_{C \subseteq N} \lambda(C) = 1$. The preference relation $\succ$ specifies a rank ordering σ over $n+1$ items which consist of the products in N plus 'no-purchase' option a_0 with $\sigma(a_i)$ denoting the preference rank of product a_i. The lower the rank of the product, the higher the preference, so that a customer's ranking σ indicates that product a is preferred to product b if and only if $\sigma(a) < \sigma(b)$, or equivalently $a \succ_\sigma b$. We assume that there is a distribution $\mu : \mathscr{S}_n \to [0,1]$ over $\mathscr{S}_n$, which is the set of all full rankings or permutations of products in N^+ with cardinality $(n+1)!$.

To simplify the exposition, we also let $\bar{X} := N \setminus X$, $X^+ := X \cup \{a_0\}$, and $\mathbb{P}_i(X) = \Pr(a_i | X^+)$. Let $\langle S \rangle$ denote the power set of S, i.e., $\langle S \rangle = 2^S$, and let $A \uplus B$ denote $\{a \cup b : a \in A, b \in B\}$ for any sets A, B.

5.2 Proofs of technical results

Proof. of Proposition 2.2.1: First, we argue that GCC class of models describing customer choice behavior is consistent with RUM. Indeed, it is straightforward to verify that GCC choice model with underlying preference order σ is equivalent to the rank-based (i.e., RUM) model where all the preference lists are obtained from a common permutation σ. In particular, consider the RUM model with the probability over rankings μ such that $\forall\ C \subseteq N$, $\mu(\sigma[C]) = \lambda(C)$, and $\mu(X) = 0$ for $X \notin \{\sigma[C] : C \subseteq N\}$, where $\sigma[C]$ is the full preference order with items in C at the top positions consistently with σ (e.g., the most preferred item in C is placed on the top position, the worst preferred item in C is placed on k-*th* position, where k is the cardinality of C) followed by the items that are not in C. It's easy to verify that the RUM with the defined distribution over rankings μ results into the same probabilities of purchases as GCC model. Then, it remains to show that RUM model class is not a specific case of GCC model class. To this end, we provide a particular example of RUM model class resulting in customers' choice frequencies that are inconsistent with GCC choice rule.

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

$$\mathbb{P}_1(\{a_1, a_2\}) = \mu_1,\ \mathbb{P}_1(\{a_1\}) = 1 \Rightarrow\ a_2 \text{ is preferred to } a_1, \text{ by GCC definition,}$$
$$\mathbb{P}_2(\{a_1, a_2\}) = 1 - \mu_1,\ \mathbb{P}_2(\{a_2\}) = 1 \Rightarrow\ a_1 \text{ is preferred to } a_2, \text{ by GCC definition.}$$

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

Lemma 5.2.1. *For any sets $Z \subseteq N$ and $Y \subseteq Z$, and the function $f : 2^N \to \mathbb{R}$, we have*

$$\sum_{P \subseteq Y} \sum_{X \subseteq P} (-1)^{|P|-|X|} \cdot f(Z \setminus X) = f(Z \setminus Y). \tag{5.1}$$

Proof. Proof: First consider the inclusion-exclusion principle stated by [37] in the following form. Let N be a finite set and $g : 2^N \to \mathbb{R}$ be a real-valued function defined on the subsets of N. Define the function $h : 2^N \to \mathbb{R}$ by $h(X) := \sum_{Y \subseteq X} g(Y)$, then $g(X) := \sum_{Y \subseteq X} (-1)^{|X|-|Y|} h(Y)$.

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

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

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

$$\begin{aligned} f(Z \setminus Y) = g(Y) = \sum_{P \subseteq Y} h(P) = \sum_{P \subseteq Y} (-1)^{|P|} \sum_{X \subseteq P} (-1)^{|X|} \cdot g(X) &= \sum_{P \subseteq Y} \sum_{X \subseteq P} (-1)^{|P|-|X|} \cdot g(X) \\ = \sum_{P \subseteq Y} \sum_{X \subseteq P} (-1)^{|P|-|X|} \cdot f(Z \setminus X). \end{aligned}$$

□

Lemma 5.2.2. *The combinatorial identity below is valid*

$$-\sum_{\beta=0}^{\min(r,w)} \mathcal{C}_{\beta}^{w} \cdot \left[\sum_{\alpha=r+1-\beta}^{u-\beta} (-1)^{\alpha} \mathcal{C}_{\alpha}^{u-\beta} \right] = \begin{cases} 1, & \text{if } w = u, \\ 0, & \text{if } w < u, \end{cases} \tag{5.2}$$

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

Proof. Proof: Let us consider two cases:

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

$$\begin{aligned} -\sum_{\beta=0}^{r} \mathcal{C}_{\beta}^{w} \cdot \left[\sum_{\alpha=r+1-\beta}^{u-\beta} (-1)^{\alpha} \mathcal{C}_{\alpha}^{u-\beta} \right] &= -\sum_{\beta=0}^{r} \mathcal{C}_{\beta}^{u} \cdot \left[\sum_{\alpha=r+1-\beta}^{u-\beta} (-1)^{\alpha} \mathcal{C}_{\alpha}^{u-\beta} \right] \\ = -\sum_{\beta=0}^{r} \sum_{\alpha=r+1-\beta}^{u-\beta} (-1)^{\alpha} \cdot \frac{u!}{\alpha!\beta!(u-\alpha-\beta)!} = 1, \end{aligned}$$

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

Base case: $s = 1$.

$$-\sum_{\beta=0}^{r}\sum_{\alpha=r+1-\beta}^{u-\beta}(-1)^{\alpha}\cdot\frac{u!}{\alpha!\beta!(u-\alpha-\beta)!} = -\sum_{\beta=0}^{u-1}(-1)^{u-\beta}\mathcal{C}_{\beta}^{u} = (-1)^{u+1}\cdot\sum_{\beta=0}^{u-1}(-1)^{\beta}\mathcal{C}_{\beta}^{u}$$

$$= (-1)^{u+1}\cdot((1-1)^{u}-(-1)^{u}) = 1.$$

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

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

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

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

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

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

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

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

$$= 1.$$

Case 2: $w < u$. the last equality is proved by induction on $s = u - r$:

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

$$-\sum_{\beta=0}^{\min(r,w)}\mathcal{C}_{\beta}^{w}\cdot\left[\sum_{\alpha=r+1-\beta}^{u-\beta}(-1)^{\alpha}\mathcal{C}_{\alpha}^{u-\beta}\right] = -\sum_{\beta=0}^{w}\mathcal{C}_{\beta}^{w}\cdot\left[\sum_{\alpha=u-\beta}^{u-\beta}(-1)^{\alpha}\mathcal{C}_{\alpha}^{u-\beta}\right]$$

$$= (-1)^{1+u}\sum_{\beta=0}^{w}(-1)^{\beta}\mathcal{C}_{\beta}^{w} = 0.$$

Induction hypothesis: $s = p$.

Induction step: $s = p + 1$.

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

$$\begin{aligned}
&-\sum_{\beta=0}^{w}\mathcal{C}_{\beta}^{w}\cdot\left[\sum_{\alpha=r+1-\beta}^{u-\beta}(-1)^{\alpha}\mathcal{C}_{\alpha}^{u-\beta}\right] = -\sum_{\beta=0}^{w}\mathcal{C}_{\beta}^{w}\cdot\left[\sum_{\alpha=u-p-\beta}^{u-\beta}(-1)^{\alpha}\mathcal{C}_{\alpha}^{u-\beta}\right] \quad [\text{since } r = u-p-1]\\
&= -\sum_{\beta=0}^{w}\mathcal{C}_{\beta}^{w}\cdot\left[\sum_{\alpha=u-p+1-\beta}^{u-\beta}(-1)^{\alpha}\mathcal{C}_{\alpha}^{u-\beta}\right] - \sum_{\beta=0}^{w}\mathcal{C}_{\beta}^{w}\cdot(-1)^{u-p-\beta}\cdot\mathcal{C}_{u-\beta-p}^{u-\beta}\\
&= -\sum_{\beta=0}^{w}\mathcal{C}_{\beta}^{w}\cdot(-1)^{u-p-\beta}\cdot\mathcal{C}_{u-\beta-p}^{u-\beta}, \quad [\text{ by induction hypothesis, } r = u-p]\\
&= (-1)^{1+u-p}\cdot\sum_{\beta=0}^{w}(-1)^{\beta}\cdot\mathcal{C}_{\beta}^{w}\cdot\mathcal{C}_{u-\beta-p}^{u-\beta} = (-1)^{1+u-p}\cdot\frac{w!}{p!}\cdot\sum_{\beta=0}^{w}(-1)^{\beta}\cdot\frac{(u-\beta)!}{\beta!(w-\beta)!(u-p-\beta)!}.
\end{aligned}$$

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

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

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

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

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

$u - w = m + 1$:

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

□

Proof. of Proposition 2.2.2: For every $C \subseteq N$ we define boolean functions $\chi_C : 2^N \to \mathbb{R}$ and $\psi_C : 2^N \to \mathbb{R}$ by

$$\begin{aligned}
\chi_C(X) &= (-1)^{|C|}\cdot\mathbf{I}[C \subseteq X],\\
\psi_C(X) &= (-1)^{|X|}\mathbf{I}[X \subseteq C],
\end{aligned}$$

where $\mathbf{I}[A]$ is an indicator function which is equal to 1, it condition A is satisfied, and 0 otherwise. Then for all $C_1, C_2 \subseteq N$ we claim that

$$\sum_{X\subseteq N}\chi_{C_1}(X)\cdot\psi_{C_2}(X) = \begin{cases}1, & \text{if } C_1 = C_2,\\ 0, & \text{otherwise,}\end{cases} \tag{5.3}$$

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

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

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

$$\begin{aligned}
\sum_{X \subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) &= \sum_{X \subseteq N} \mathbf{I}[C_1 \subseteq X] \cdot (-1)^{|C_1|+|X|} \mathbf{I}[X \subseteq C_2] \\
&= (-1)^{|C_1|} \cdot \sum_{X \subseteq N} (-1)^{|X|} \mathbf{I}[C_1 \subseteq X \subseteq C_2] \\
&= (-1)^{|C_1|} \cdot (-1)^{|C_1|} \cdot \sum_{k=0}^{|C_2|-|C_1|} (-1)^k \mathcal{C}_k^{|C_2|-|C_1|}, \text{ where } \mathcal{C}_k^n = \frac{n!}{k!(n-k)!}
\end{aligned}$$

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

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

Consequently, the probability to choose the "no-purchase" option a_0 from the offer set $\{N \setminus X\}^+$ is given by

$$\begin{aligned}
\mathbb{P}_0(N \setminus X) &= \sum_{C \subseteq X} \lambda(C) = \sum_{C \subseteq N} \lambda(C) \cdot (-1)^{2|C|} \cdot \mathbf{I}[C \subseteq X] \qquad (5.4) \\
&= \sum_{C \subseteq N} \lambda(C) \cdot (-1)^{|C|} \cdot \chi_C(X).
\end{aligned}$$

Then it follows that

$$\begin{aligned}
\sum_{X\subseteq C}(-1)^{|C|-|X|}\cdot\mathbb{P}_0(N\setminus X) &= \sum_{X\subseteq N}\mathbb{P}_0(N\setminus X)\cdot(-1)^{|C|+|X|}\mathbf{I}[X\subseteq C] \\
&= (-1)^{|C|}\cdot\sum_{X\subseteq N}\mathbb{P}_0(N\setminus X)\cdot\psi_C(X) \\
&= (-1)^{|C|}\cdot\sum_{X\subseteq N}\sum_{C_1\subseteq N}\lambda(C_1)\cdot(-1)^{|C_1|}\cdot\chi_{C_1}(X)\cdot\psi_C(X) \\
&\qquad\qquad\qquad\qquad \Big[\text{by Equation (5.4)}\Big] \\
&= (-1)^{|C|}\cdot\sum_{C_1\subseteq N}\lambda(C_1)\cdot(-1)^{|C_1|}\cdot\sum_{X\subseteq N}\chi_{C_1}(X)\cdot\psi_C(X) \\
&= (-1)^{|C|}\cdot\lambda(C)\cdot(-1)^{|C|}, \quad \Big[\text{ by Equation (5.3) }\Big] \\
&= \lambda(C).
\end{aligned} \tag{5.5}$$

Now it remains to prove the uniqueness of probability distribution function λ obtained from purchasing transactions data under GCC choice model. Note that Equation (5.4) relates probability distribution λ over consideration sets to the choice frequencies $\mathbb{P}_0(N\setminus X)$ through the system of linear equations:

$$\mathbb{P}_0(N\setminus X)=\sum_{C\subseteq N}\lambda(C)\cdot(-1)^{|C|}\cdot\chi_C(X),\ \forall\ X\subseteq N \Longleftrightarrow \boldsymbol{y}=A\cdot\boldsymbol{\lambda}, \tag{5.6}$$

where $\boldsymbol{y}=(y_X)_{X\subseteq N}$ denotes the $\left|2^N\right|\times 1$ vector of choice fractions and $\boldsymbol{\lambda}=(\lambda_C)_{C\subseteq N}$ denotes the $\left|2^N\right|\times 1$ vector that represents the probability distribution function over consideration sets. A is the $\left|2^N\right|\times\left|2^N\right|$ matrix such that A's entry corresponding to the row X and column C is equal to $(-1)^{|C|}\cdot\chi_C(X)$. Therefore, the relation between the choice frequencies and the underlying model can be represented in a compact form as $\boldsymbol{y}=A\cdot\boldsymbol{\lambda}$. Then the proof of uniqueness of λ reduces to showing that $\det(A)\neq 0$. From Equation (5.5) we have

$$\lambda(C)=(-1)^{|C|}\cdot\sum_{X\subseteq N}\mathrm{Pr}_0(N\setminus X)\cdot\psi_C(X),\ \forall C\subseteq N \Longleftrightarrow \boldsymbol{\lambda}=B\cdot\boldsymbol{y},$$

which establishes alternative linear relationship between choice frequencies $\mathrm{Pr}_0(N\setminus X)$ and the model parameters λ in a compact form as $\boldsymbol{\lambda} = B \cdot \boldsymbol{y}$, where B is the $\left|2^N\right| \times \left|2^N\right|$ matrix such that B's entry corresponding to the row C and column X is equal to $(-1)^{|C|} \cdot \psi_C(X)$. Therefore, we get

$$\begin{aligned}
&\boldsymbol{\lambda} = B \cdot \boldsymbol{y} = B \cdot A \cdot \boldsymbol{\lambda}, \quad \Big[\text{by Equation (5.6)}\Big] \\
&\implies I = B \cdot A \implies \det(I) = \det(B) \cdot \det(A) \\
&\implies 1 = \det(B) \cdot \det(A) \implies \det(A) \neq 0.
\end{aligned}$$

□

Proof. of Proposition 2.2.3: Assume by contradiction that $\sigma(a_j) > \sigma(a_i)$ if $\mathbb{P}_i(\{a_i\}) > \mathbb{P}_j(\{a_i, a_j\})$. Then it follows from GCC model definition that $\mathbb{P}_i(\{a_i\}) = \mathbb{P}_i(\{a_i, a_j\})$, which leads to contradiction. □

Proof. of Proposition 2.2.4: It follows from the proposition that

$$\begin{aligned}
\lambda(C) &= \sum_{X \subseteq N} \sum_{Y \supseteq X \cup C} (-1)^{1+|Y|-|X \Delta C|} \cdot \mathbb{P}_0(X) \cdot \mathbf{I}[|X \cup C| \leq k < |Y|] \\
&= \sum_{X \subseteq N} \sum_{Y \supseteq X \cup C} \mathbb{P}_0(X) \cdot (-1)^{1+|X \cap C|} \cdot (-1)^{|Y|-|X \cup C|} \cdot \mathbf{I}[|X \cup C| \leq k < |Y|] \\
&= \sum_{X \subseteq N} \mathbb{P}_0(X) \cdot (-1)^{1+|X \cap C|} \cdot \mathbf{I}[|X \cup C| \leq k] \cdot \sum_{Y \supseteq X \cup C} (-1)^{|Y|-|X \cup C|} \cdot \mathbf{I}[|Y| > k] \\
&\Big[\text{since the expression depends only on the cardinality of sets } Y\text{, the summation over } Y \\
&\text{is reduced to the summation over the cardinality of sets } Y\Big] \\
&= \sum_{X \subseteq N} \mathbb{P}_0(X) \cdot (-1)^{1+|X \cap C|} \cdot \mathbf{I}[|X \cup C| \leq k] \cdot \sum_{\alpha = k+1-|X \cup C|}^{n-|X \cup C|} (-1)^{\alpha} \mathcal{C}_{\alpha}^{n-|X \cup C|}, \\
&\hspace{8cm} \left[\text{where } \mathcal{C}_k^n = \frac{n!}{k!(n-k)!}.\right]
\end{aligned}$$

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

$$\begin{aligned}
&\chi_C(X) = \mathbf{I}[C \subseteq \bar{X}, |C| \leq k],\\
&\psi_C(X) = (-1)^{1+|X\cap C|} \cdot \mathbf{I}[|X \cup C| \leq k] \cdot \sum_{\alpha=k+1-|X\cup C|}^{n-|X\cup C|} (-1)^{\alpha} \mathcal{C}_{\alpha}^{n-|X\cup C|}.
\end{aligned}$$

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

$$\mathbb{P}_0(X) = \sum_{C\subseteq N} \lambda(C) \cdot \mathbf{I}[C \subseteq \bar{X}, |C| \leq k] = \sum_{C\subseteq N} \lambda(C) \cdot \chi_C(X). \tag{5.7}$$

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

$$\sum_{X\subseteq N} \chi_{C_1}(X) \cdot \psi_{C_2}(X) = \begin{cases} 1, & \text{if } C_1 = C_2, \\ 0, & \text{otherwise.} \end{cases} \tag{5.8}$$

Consequently, it follows from the claim that

$$\begin{aligned}
&\sum_{X\subseteq N} \mathbb{P}_0(X) \cdot (-1)^{1+|X\cap C|} \cdot \mathbf{I}[|X \cup C| \leq k] \cdot \sum_{\alpha=k+1-|X\cup C|}^{n-|X\cup C|} (-1)^{\alpha} \mathcal{C}_{\alpha}^{n-|X\cup C|} \\
&= \sum_{X\subseteq N} \mathbb{P}_0(X) \cdot \psi_C(X) = \sum_{X\subseteq N} \sum_{C_1\subseteq N} \lambda(C_1) \cdot \chi_{C_1}(X) \cdot \psi_C(X) \\
&= \sum_{C_1\subseteq N} \lambda(C_1) \cdot \sum_{X\subseteq N} \chi_{C_1}(X) \cdot \psi_C(X) = \lambda(C), \quad \left[\text{by Equation (5.8)}\right].
\end{aligned} \tag{5.9}$$

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

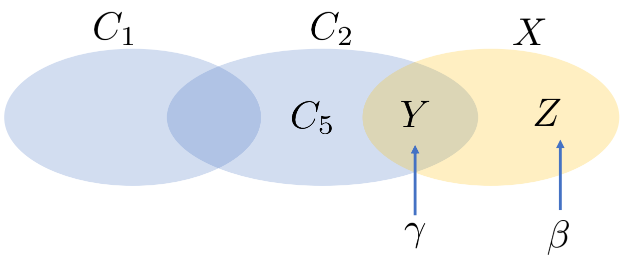

Figure 5-1: The case when $C_1 \not\subset C_2$ and $n - |C_1| - |C_5| > 0$.

Case 1: $C_2 \subseteq C_1$.

$$\begin{aligned}\sum_{X\subseteq N} \chi_{C_1}(X)\cdot\psi_{C_2}(X) &= \sum_{X\subseteq N}(-1)^1\cdot \mathbf{I}[|C_1|\le k]\cdot \mathbf{I}[X\cap C_1=\varnothing]\cdot\mathbf{I}[|X|\le k-|C_2|] \\ &\qquad\qquad \times \sum_{\alpha=k+1-|X|-|C_2|}^{n-|X|-|C_2|}(-1)^\alpha \mathcal{C}_\alpha^{n-|X|-|C_2|}\end{aligned}$$

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

$$= -\sum_{X\subseteq N}\mathbf{I}[|C_1|\le k]\cdot\mathbf{I}[X\cap C_1=\varnothing]\cdot\mathbf{I}[|X|\le k-|C_2|]\cdot\sum_{\alpha=k-|C_2|+1-|X|}^{n-|C_2|-|X|}(-1)^\alpha\mathcal{C}_\alpha^{n-|C_2|-|X|},$$

$\Big[$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$\Big]$

$$= -\sum_{\beta=0}^{\min(k-|C_2|,n-|C_1|)}\mathcal{C}_\beta^{n-|C_1|}\cdot\left[\sum_{\alpha=k-|C_2|+1-\beta}^{n-|C_2|-\beta}(-1)^\alpha\mathcal{C}_\alpha^{n-|C_2|-\beta}\right],\ \Big[\text{ where } \beta \text{ - cardinality of set } X\Big]$$

$$= \begin{cases} 1, & \text{if } C_1 = C_2, \\ 0, & \text{if } C_1 \subset C_2, \end{cases}$$

where the last equality follows by invoking Lemma 5.2.2, where $w = n - |C_1|$, $r = k - |C_2|$, and $u = n - |C_2|$.

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

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

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

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

$$=\begin{cases}\sum_{\gamma=0}^{|C_5|}(-1)^{1+\gamma}\cdot\mathcal{C}_{\gamma}^{|C_5|}\cdot\Big[\sum_{\beta=0}^{\min(k-|C_2|,n-|C_1|-|C_5|)}\mathcal{C}_{\beta}^{n-|C_1|-|C_5|}\cdot\Big[\sum_{\alpha=k+1-\beta-|C_2|}^{n-\beta-|C_2|}(-1)^{\alpha}\mathcal{C}_{\alpha}^{n-\beta-|C_2|}\Big]\Big], \\ \qquad\qquad\text{if } C_1\not\subset C_2 \text{ and } n-|C_1|-|C_5|>0, \text{see Figure 5-1} \\ \sum_{\gamma=0}^{|C_5|}(-1)^{1+\gamma}\cdot\mathcal{C}_{\gamma}^{|C_5|}\cdot\Big[\sum_{\beta=0}^{\min(k-|C_2|,n-|C_2|)}\mathcal{C}_{\beta}^{n-|C_2|}\cdot\Big[\sum_{\alpha=k+1-\beta-|C_2|}^{n-\beta-|C_2|}(-1)^{\alpha}\mathcal{C}_{\alpha}^{n-\beta-|C_2|}\Big]\Big] \\ \qquad\qquad\text{if } C_1\subset C_2, \text{see Figure 5-2} \\ \sum_{\gamma=0}^{|C_5|}(-1)^{1+\gamma}\cdot\mathcal{C}_{\gamma}^{|C_5|}\cdot\Big[\sum_{\alpha=k+1-|C_2|}^{n-|C_2|}(-1)^{\alpha}\mathcal{C}_{\alpha}^{n-|C_2|}\Big], \\ \qquad\qquad\text{if } C_1\not\subset C_2 \text{ and } n-|C_1|-|C_5|=0, \text{see Figure 5-3}\end{cases}$$

$\Big[$ where $C_5=C_2\setminus\{C_1\cap C_2\}\Big]$

$$=\begin{cases}\Big[-\sum_{\gamma=0}^{|C_5|}(-1)^{\gamma}\cdot\mathcal{C}_{\gamma}^{|C_5|}\Big]\cdot\Big[\sum_{\beta=0}^{\min(k-|C_2|,n-|C_1|-|C_5|)}\mathcal{C}_{\beta}^{n-|C_1|-|C_5|}\cdot\Big[\sum_{\alpha=k+1-\beta-|C_2|}^{n-\beta-|C_2|}(-1)^{\alpha}\mathcal{C}_{\alpha}^{n-\beta-|C_2|}\Big]\Big], \\ \qquad\qquad\text{if } C_1\not\subset C_2 \text{ and } n-|C_1|-|C_5|>0, \\ \Big[-\sum_{\gamma=0}^{|C_5|}(-1)^{\gamma}\cdot\mathcal{C}_{\gamma}^{|C_5|}\Big]\cdot\Big[\sum_{\beta=0}^{\min(k-|C_2|,n-|C_2|)}\mathcal{C}_{\beta}^{n-|C_2|}\cdot\Big[\sum_{\alpha=k+1-\beta-|C_2|}^{n-\beta-|C_2|}(-1)^{\alpha}\mathcal{C}_{\alpha}^{n-\beta-|C_2|}\Big]\Big] \\ \qquad\qquad\text{if } C_1\subset C_2, \\ \Big[-\sum_{\gamma=0}^{|C_5|}(-1)^{\gamma}\cdot\mathcal{C}_{\gamma}^{|C_5|}\Big]\cdot\Big[\sum_{\alpha=k+1-|C_2|}^{n-|C_2|}(-1)^{\alpha}\mathcal{C}_{\alpha}^{n-|C_2|}\Big], \quad \text{if } C_1\not\subset C_2 \text{ and } n-|C_1|-|C_5|=0,\end{cases}$$

$$=0,$$

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

In order to complete the proof, we show the uniqueness of probability distribution function λ in our setting. First, note that Equation (5.7) relates probability distribution λ over consideration

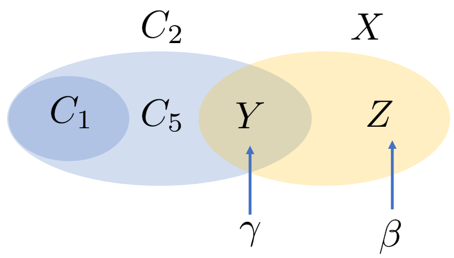

Figure 5-2: The case when $C_1 \subset C_2$.

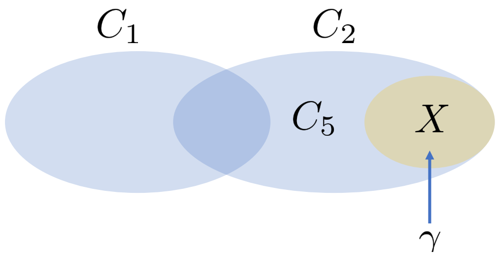

Figure 5-3: The case when $C_1 \not\subset C_2$ and $n - |C_1| - |C_5| = 0$.

sets to the choice frequencies $\Pr\left(a_0 \middle| X\right)$ through the system of linear equations:

$$\mathbb{P}_0(X) = \sum_{C \subseteq N} \lambda(C) \cdot \chi_C(X), \ \forall \ X \subseteq N \Longleftrightarrow \boldsymbol{y} = A \cdot \boldsymbol{\lambda}, \tag{5.10}$$

where $\boldsymbol{y} = (y_X)_{X \subseteq N}$ denotes the $\left|2^N\right| \times 1$ vector of choice fractions and $\boldsymbol{\lambda} = (\lambda_C)_{C \subseteq N}$ denotes the $\left|2^N\right| \times 1$ vector that represents the probability distribution function over consideration sets. A is the $\left|2^N\right| \times \left|2^N\right|$ matrix such that A's entry corresponding to the row X and column C is equal to $\chi_C(X)$. As a result, the relation between the choice frequencies and the underlying model can be represented in a compact form as $\boldsymbol{y} = A \cdot \boldsymbol{\lambda}$. Then the proof of uniqueness of λ reduces to showing that $\det(A) \neq 0$. It follows from Equation (5.9) that

$$\lambda(C) = \sum_{X \subseteq N} \mathbb{P}_0(X) \cdot \psi_C(X), \ \forall C \subseteq N \Longleftrightarrow \boldsymbol{\lambda} = B \cdot \boldsymbol{y},$$

which provides another relationship between choice frequencies $\mathbb{P}_0(X)$ and the model parameters λ in a linear form as $\boldsymbol{\lambda} = B \cdot \boldsymbol{y}$, where B is the $\left|2^N\right| \times \left|2^N\right|$ matrix such that B's entry corresponding to the row C and column X is equal to $\psi_C(X)$. Therefore, we get

$$\begin{aligned}
\boldsymbol{\lambda} = B \cdot \boldsymbol{y} = B \cdot A \cdot \lambda, \quad & \left[\text{by Equation (5.10)}\right] \\
\implies I = B \cdot A \implies \ & \det(I) = \det(B) \cdot \det(A) \\
\implies 1 = \det(B) \cdot \det(A) \ & \implies \ \det(A) \neq 0.
\end{aligned}$$

□

Lemma 5.2.3. *Assume that for all consideration sets $C \subseteq N$ we have that*

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

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

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

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

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

$$\begin{aligned}
\sum_{B\subseteq C}(-1)^{|C|-|B|}\mathbb{P}_0(S\setminus B) &= \sum_{B\subseteq C}(-1)^{|C|-|B|}\cdot\mathbb{P}_0(\{N\setminus\bar{S}\}\setminus B)\\
&= \sum_{B\subseteq C}(-1)^{|C|-|B|}\cdot\mathbb{P}_0(\{N\setminus B\}\setminus\bar{S})\\
&= \sum_{B\subseteq C}\sum_{A\subseteq\bar{S}}\sum_{D\subseteq A}(-1)^{|A|-|D|}\cdot(-1)^{|C|-|B|}\cdot\mathbb{P}_0(\{N\setminus B\}\setminus D)\\
&\Big[\text{by invoking Lemma 5.2.1 for every } B\subseteq C\text{, where } Z=N\setminus B,\ Y=\bar{S},\\
&P=A,\ \text{and } f(Z\setminus Y)=(-1)^{|C|-|B|}\cdot\mathbb{P}_0(\{N\setminus B\}\setminus\bar{S})\Big]\\
&= \sum_{A\subseteq\bar{S}}\sum_{D\subseteq A}\sum_{B\subseteq C}(-1)^{|C|+|A|-|D|-|B|}\cdot\mathbb{P}_0(\{N\setminus B\}\setminus D)\\
&= \sum_{A\subseteq\bar{S}}\sum_{X\in\langle A\cup C\rangle}(-1)^{|C|+|A|-|X|}\cdot\mathbb{P}_0(N\setminus X)\\
&\Big[\text{where } X=D\cup B\text{, since } A\cap C=\varnothing\Big]\\
&= \sum_{A\subseteq\bar{S}}\sum_{X\in\langle A\cup C\rangle}(-1)^{|C\cup A|-|X|}\cdot\mathbb{P}_0(N\setminus X)\quad\Big[\text{since } A\cap C=\varnothing\Big]\\
&= \sum_{A\subseteq\bar{S}}\sum_{X\subseteq C'}(-1)^{|C'|-|X|}\cdot\mathbb{P}_0(N\setminus X)\quad\Big[\text{where } C'=A\cup C\Big]\\
&\geq 0,\ \text{with strict inequality when } |C|\leq 3,\ \Big[\text{by assumptions of the Lemma,}\\
&\text{since } \sum_{X\subseteq C'}(-1)^{|C'|-|X|}\cdot\mathbb{P}_0(N\setminus X)\geq 0 \text{ with strict inequality when } |C'|\leq 3\Big].
\end{aligned}$$

□

Lemma 5.2.4. *If a sample of sales transaction data satisfies Conditions 1, 2, and 3, then for all $a_i\in S_1\cap S_2$ where $S_1,S_2\subseteq N$ and $S_1\subseteq S_2$ we have that $\mathbb{P}_i(S_1)\geq\mathbb{P}_i(S_2)$. .*

Proof. Proof: Prove the result by induction on the n=$|S_2|-|S_1|$. We consider $a_1\in S_1\cap S_2$ and $S_1\subseteq S_2$. For the base case $n=0$ we have that $S_1=S_2$ and $\mathbb{P}_1(S_1)=\mathbb{P}_1(S_2)$. Assume that the result holds for $n=k$, i.e., $S_2=S$ and $|S|-|S_1|=k$. Then we prove it for $n=k+1$. Let us suppose w.l.o.g. that $S_2=S\cup\{a_2\}$ and $a_2\notin S$. Next, assume, by contradiction, that

$\mathbb{P}_1(S_2 \setminus \{a_2\}) < \mathbb{P}_1(S_2)$. Consequently, by Condition 2 it follows that $\mathbb{P}_1(\{a_1\}) < \mathbb{P}_1(\{a_1, a_2\})$. Then by Condition 1 we have that

$$\mathbb{P}_2(\{a_2\}) = \mathbb{P}_2(\{a_1, a_2\}). \tag{5.11}$$

It now follows that

$$\begin{aligned}
&\mathbb{P}_1(\{a_1\}) - \mathbb{P}_1(\{a_1, a_2\}) \\
&= \Big(1 - \mathbb{P}_0(\{a_1\})\Big) - \Big(1 - \mathbb{P}_0(\{a_1, a_2\}) - \mathbb{P}_2(\{a_1, a_2\})\Big), \text{ [by standard probability property]} \\
&= \Big(1 - \mathbb{P}_0(\{a_1\})\Big) - \Big(1 - \mathbb{P}_0(\{a_1, a_2\}) - \mathbb{P}_2(\{a_2\})\Big), \Big[\text{by Equation (5.11)}\Big] \\
&= \Big(1 - \mathbb{P}_0(\{a_1\})\Big) - \Big(\mathbb{P}_0(\{a_2\}) - \mathbb{P}_0(\{a_1, a_2\})\Big), \text{ [by standard probability property]} \\
&= 1 - \mathbb{P}_0(\{a_1\}) - \mathbb{P}_0(\{a_2\}) + \mathbb{P}_0(\{a_1, a_2\}) > 0, \\
&\Big[\text{by Condition 3 and Lemma 5.2.3, when } C = S = \{a_1, a_2\}\Big],
\end{aligned}$$

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

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

Therefore, the result now follows by induction. □

Lemma 5.2.5. *Consider $a_1, a_2 \in S$, $S \subseteq N$, $a_1 \neq a_2$. Then GCC choice model, with strict preference list σ and distribution over consideration sets λ where $\lambda(C) > 0$ if $|C| \leq 3$, implies the following list of implications:*

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

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

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

Proof. Proof: a) Suppose that $\mathbb{P}_1(S \setminus \{a_2\}) > \mathbb{P}_1(S)$. Assume, by contradiction, that $a_1 \succ a_2$. Then it can be inferred from purchase probability definition under GCC, see Equation (2.1), that $\mathbb{P}_1(S \setminus \{a_2\}) = \mathbb{P}_1(S)$, which leads to contradiction. As a result, we have that $a_2 \succ a_1$ since preferences are strict and asymmetric. Then $\forall S' \subseteq N$ $s.t.$ $a_1, a_2 \in S'$ we establish that

$$\begin{aligned}\mathbb{P}_1(S' \setminus \{a_2\}) - \mathbb{P}_1(S') &\geq \lambda(\{a_1, a_2\}), \quad \Big[\text{by Equation (2.1)}\Big]\\ &> 0, \quad \Big[\text{by Assumption that } \lambda(C) > 0 \text{ if } |C| \leq 3\Big].\end{aligned}$$

b) Suppose that $\mathbb{P}_1(S \setminus \{a_2\}) = \mathbb{P}_1(S)$. Assume, by contradiction, that $a_2 \succ a_1$. Then it follows that

$$\begin{aligned}\mathbb{P}_1(S \setminus \{a_2\}) - \mathbb{P}_1(S) &\geq \lambda(\{a_1, a_2\}), \quad \Big[\text{by Equation (2.1)}\Big]\\ &> 0, \quad \Big[\text{by Assumption that } \lambda(C) > 0 \text{ if } |C| \leq 3\Big].\end{aligned}$$

which contradicts to the assumption above. As a result, we have that $a_1 \succ a_2$, since preferences are strict and asymmetric. Then by Equation (2.1) we have that $\forall S' \subseteq N$ $s.t.$ $a_1, a_2 \in S'$: $\mathbb{P}_1(S' \setminus \{a_2\}) = \mathbb{P}_1(S')$.

c) Suppose that $\mathbb{P}_1(S \setminus \{a_2\}) \neq \mathbb{P}_1(S)$. Then it is straightforward to verify that $\mathbb{P}_1(S \setminus \{a_2\}) > \mathbb{P}_1(S)$, since the following inequality holds from the Lemma 5.2.4: $\mathbb{P}_1(S \setminus \{a_2\}) \geq \mathbb{P}_1(S)$. Consequently, invoking the implication from part a), we have $a_2 \succ a_1$, and by Equation (2.1) we obtain that $\mathbb{P}_2(S \setminus \{a_1\}) = \mathbb{P}_2(S)$. □

Proof. of Proposition 2.2.5: *Necessity:* if purchasing transactions data is consistent with GCC choice model with strict preference list σ and distribution over consideration sets λ where $\lambda(C) > 0$ if $|C| \leq 3$, then we claim that three axioms Condition 1, Condition 2, and Condition 3 are satisfied. First, it follows from Proposition 2.2.2 that Condition 3 is satisfied. Then Condition 1 and Condition 2 are satisfied by Lemma 5.2.5.

Sufficiency: we claim that the choice rule that satisfies Condition 1, Condition 2, and Condition 3 is a GCC choice model with the strict preference list σ where no-purchase option is the least preferred item, and probability distribution function λ over consideration sets such that $\lambda(C) > 0$ if $|C| \leq 3$.

Define a binary relation δ_{ij} between products $a_i, a_j \subseteq N, a_i \neq a_j$, where $\delta_{ij} = 1$ if $\mathbb{P}_j(S \setminus$

$\{a_i\}) > \mathbb{P}_j(S)$ for some $S \subseteq N$ s.t. $a_i, a_j \in S$ (note, by Condition 2 it implies that $\mathbb{P}_j(S \setminus \{a_i\}) > \mathbb{P}_j(S)$ for all $S \subseteq N$ s.t. $a_i, a_j \in S$), and zero otherwise. We claim that δ_{ij} is complete, asymmetric, and transitive binary relation.

First, we prove that this binary relation is complete, i.e., either $\delta_{ij} = 1$ or $\delta_{ji} = 1$. Suppose that $\mathbb{P}_j(S \setminus \{a_i\}) \leq \mathbb{P}_j(S)$ for some $S \subseteq N$, i.e., $\delta_{ij} = 0$. Then it follows from the Lemma 5.2.4 that $\mathbb{P}_j(S \setminus \{a_i\}) = \mathbb{P}_j(S)$. Moreover, by Condition 2 we have that $\mathbb{P}_j(\{a_j\}) = \mathbb{P}_j(\{a_i, a_j\})$. We can now establish the following chain of equalities:

$$\begin{aligned}
&\mathbb{P}_i(\{a_i\}) - \mathbb{P}_i(\{a_i, a_j\}) \\
&= \Big(1 - \mathbb{P}_0(\{a_i\})\Big) - \Big(1 - \mathbb{P}_0(\{a_i, a_j\}) - \mathbb{P}_j(\{a_i, a_j\})\Big), \text{ [by standard probability property]} \\
&= \Big(1 - \mathbb{P}_0(\{a_i\})\Big) - \Big(1 - \mathbb{P}_0(\{a_i, a_j\}) - \mathbb{P}_j(\{a_j\})\Big), \Big[\text{by Condition 2, see above}\Big] \\
&= \Big(1 - \mathbb{P}_0(\{a_i\})\Big) - \Big(\mathbb{P}_0(\{a_j\}) - \mathbb{P}_0(\{a_i, a_j\})\Big), \text{ [by standard probability property]} \\
&= 1 - \mathbb{P}_0(\{a_i\}) - \mathbb{P}_0(\{a_j\}) + \mathbb{P}_0(\{a_i, a_j\}) > 0, \\
&\Big[\text{by Condition 3 and Lemma 5.2.3, where } C = S = \{a_i, a_j\}\Big],
\end{aligned}$$

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

Second, we establish that the defined binary relation $\boldsymbol{\delta}$ is asymmetric, i.e., if $\delta_{ij} = 1$ then $\delta_{ji} = 0$. Suppose that $\mathbb{P}_j(S \setminus \{a_i\}) > \mathbb{P}_j(S)$ for some $S \subseteq N$, i.e., $\delta_{ij} = 1$. Then by Condition 1 we have that $\mathbb{P}_i(S \setminus \{a_j\}) = \mathbb{P}_i(S)$ (note, by Condition 2 we have that for all $S' \subseteq N$ s.t. $a_1, a_2 \in S'$: $\mathbb{P}_i(S' \setminus \{a_j\}) = \mathbb{P}_i(S')$), which further implies that $\delta_{ji} = 0$. As a result, asymmetry of binary relation δ_{ij} now follows.

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

$$\begin{align}
\delta_{ij} = 1 &\Rightarrow \mathbb{P}_j(S \setminus \{a_i\}) > \mathbb{P}_j(S), \quad \big[\text{ for some } S \subseteq N\ \big] \notag\\
&\Rightarrow \mathbb{P}_j(\{a_j, a_k\}) > \mathbb{P}_j(\{a_i, a_j, a_k\}), \quad \Big[\text{by Condition } 2\Big] \notag\\
&\Rightarrow \mathbb{P}_i(\{a_i, a_k\}) = \mathbb{P}_i(\{a_j, a_i, a_k\}), \quad \Big[\text{by Condition } 1\Big] \tag{5.12}\\
&\Rightarrow \mathbb{P}_i(\{a_i\}) = \mathbb{P}_i(\{a_j, a_i\}), \quad \Big[\text{by Condition } 2\Big], \tag{5.13}\\
\delta_{jk} = 1 &\Rightarrow \mathbb{P}_k(S \setminus \{a_j\}) > \mathbb{P}_k(S), \quad \big[\text{ for some } S \subseteq N\ \big] \notag\\
&\Rightarrow \mathbb{P}_k(\{a_i, a_k\}) > \mathbb{P}_k(\{a_i, a_j, a_k\}), \quad \Big[\text{by Condition } 2\Big] \notag\\
&\Rightarrow \mathbb{P}_j(\{a_i, a_j\}) = \mathbb{P}_j(\{a_i, a_j, a_k\}), \quad \Big[\text{by Condition } 1\Big] \tag{5.14}\\
&\Rightarrow \mathbb{P}_j(\{a_j\}) = \mathbb{P}_j(\{a_j, a_k\}), \quad \Big[\text{by Condition } 2\Big], \tag{5.15}\\
\delta_{ik} = 0 &\Rightarrow \mathbb{P}_k(S \setminus \{a_i\}) \leq \mathbb{P}_k(S), \quad \big[\text{ for some } S \subseteq N\ \big] \notag\\
&\Rightarrow \mathbb{P}_k(S \setminus \{a_i\}) = \mathbb{P}_k(S), \quad \Big[\text{by Lemma 5.2.4}\Big] \tag{5.16}\\
&\Rightarrow \mathbb{P}_k(\{a_k\}) = \mathbb{P}_k(\{a_i, a_k\}), \quad \Big[\text{by Condition } 2\Big]. \tag{5.17}
\end{align}$$

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

$$\begin{align}
\text{For } S_1 = \{a_i, a_j\} : \mathbb{P}_i(S_1) + \mathbb{P}_j(S_1) + \mathbb{P}_0(S_1) = 1 \notag\\
&\Rightarrow \mathbb{P}_i(\{a_i\}) + \mathbb{P}_j(S_1) + \mathbb{P}_0(S_1) = 1, \quad \Big[\text{by Equation (5.13)}\Big] \notag\\
&\Rightarrow \mathbb{P}_i(\{a_i\}) + \mathbb{P}_j(S_4) + \mathbb{P}_0(S_1) = 1, \quad \Big[\text{by Equation (5.14)}\Big] \notag\\
&\Rightarrow \mathbb{P}_j(S_4) = \mathbb{P}_0(\{a_i\}) - \mathbb{P}_0(S_1), \quad [\text{by standard probability property}]. \tag{5.18}
\end{align}$$

$$\begin{aligned}
\text{For } S_2 = \{a_j, a_k\} : \mathbb{P}_j(S_2) + \mathbb{P}_k(S_2) + \mathbb{P}_0(S_2) = 1& \\
&\Rightarrow \mathbb{P}_j(\{a_j\}) + \mathbb{P}_k(S_2) + \mathbb{P}_0(S_2) = 1, \ \left[\text{by Equation (5.15)}\right] \\
&\Rightarrow \mathbb{P}_j(\{a_j\}) + \mathbb{P}_k(S_4) + \mathbb{P}_0(S_2) = 1, \ \left[\text{by Equation (5.16)}\right] \\
&\Rightarrow \mathbb{P}_k(S_4) = \mathbb{P}_0(\{a_j\}) - \mathbb{P}_0(S_2), \ \left[\text{by standard probability property}\right].
\end{aligned} \tag{5.19}$$

$$\begin{aligned}
\text{For } S_3 = \{a_i, a_k\} : \mathbb{P}_k(S_3) + \mathbb{P}_i(S_3) + \mathbb{P}_0(S_3) = 1& \\
&\Rightarrow \mathbb{P}_k(\{a_k\}) + \mathbb{P}_i(S_3) + \mathbb{P}_0(S_3) = 1, \ \left[\text{by Equation (5.17)}\right] \\
&\Rightarrow \mathbb{P}_k(\{a_k\}) + \mathbb{P}_i(S_4) + \mathbb{P}_0(S_3) = 1, \ \left[\text{by Equation (5.12)}\right] \\
&\Rightarrow \mathbb{P}_i(S_4) = \mathbb{P}_0(\{a_k\}) - \mathbb{P}_0(S_3), \ \left[\text{by standard probability property}\right].
\end{aligned} \tag{5.20}$$

$$\begin{aligned}
\text{For } S_4 = \{a_i, a_j, a_k\} : \mathbb{P}_i(S_4) + \mathbb{P}_j(S_4) + \mathbb{P}_k(S_4) + \mathbb{P}_0(S_4) = 1& \\
&\Rightarrow 0 = \mathbb{P}_0(\varnothing) - \mathbb{P}_0(\{a_i\}) - \mathbb{P}_0(\{a_j\}) - \mathbb{P}_0(\{a_k\}) + \mathbb{P}_0(S_1) + \mathbb{P}_0(S_2) \\
&+ \mathbb{P}_0(S_3) - \mathbb{P}_0(S_4), \ \left[\text{since } \mathbb{P}_0(\varnothing) = 1, \text{ and by Equations (5.18)-(5.20)}\right] \\
&> 0, \ \left[\text{by Condition 3 and Lemma 5.2.3, where } C = S = S_4\right],
\end{aligned}$$

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

$$\mathbb{P}_0(\varnothing) - \mathbb{P}_0(\{a_i\}) > 0, \left[\text{by Condition 3 and Lemma 5.2.3, where } C = S = \{a_i\}\right],$$

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

Next, we prove that

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

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

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

$$\begin{aligned}\mathbb{P}_r(S) &= \mathbb{P}_r(S \setminus \{a_j\}),\ \big[\text{by Condition 1}\ \big] \\ &= \mathbb{P}_0(S' \setminus \{a_r\}) - \mathbb{P}_0(S'),\ \ \big[\text{by induction hypothesis,} \\ &\text{and note that } a_j \notin S' \text{ since } a_r \succ a_j\big].\end{aligned}$$

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

$$
\begin{aligned}
\mathbb{P}_r(S) &= 1 - \mathbb{P}_0(S) - \sum_{i=1}^{p-1} \mathbb{P}_i(S) \\
&= -\mathbb{P}_0(S) + \mathbb{P}_0(\varnothing) - \sum_{i=1}^{p-1} \mathbb{P}_i(\{a_r, a_1, a_2..., a_{p-1}\}) \\
&= -\mathbb{P}_0(S) + \mathbb{P}_0(\varnothing) - \sum_{i=1}^{p-1} \mathbb{P}_i(\{a_i, a_{i+1}..., a_{p-1}\}), \; \left[\text{ by Condition 1}\right] \\
&= -\mathbb{P}_0(S) + \mathbb{P}_0(\varnothing) - \sum_{i=1}^{p-1} \Big(\mathbb{P}_0(\{a_{i+1}..., a_{p-1}\}) - \mathbb{P}_0(\{a_i, a_{i+1}..., a_{p-1}\}) \Big), \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad [\text{by induction hypothesis}] \\
&= -\mathbb{P}_0(S) + \mathbb{P}_0(\{a_1, a_2..., a_{p-1}\}) \\
&= \mathbb{P}_0(\{a_1, a_2..., a_{p-1}\}) - \mathbb{P}_0(\{a_r, a_1, a_2..., a_{p-1}\}) = \mathbb{P}_0(S' \setminus \{a_r\}) - \mathbb{P}_0(S').
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{P}_r(S) &= \mathbb{P}_0(S' \setminus \{a_r\}) - \mathbb{P}_0(S') \\
&= \mathbb{P}_0(S' \setminus \{a_r\}) + \left(\sum_{C \subseteq \hat{S}} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X) - \mathbb{P}_0(N \setminus \hat{S}) \right) - \mathbb{P}_0(S') \\
&\left[\text{by invoking Lemma 5.2.1, where } Z = N,\ Y = \hat{S},\ P = C, \text{and } f(Z \setminus Y) = \mathbb{P}_0(N \setminus \hat{S}) \right] \\
&= \sum_{C \subseteq \hat{S}} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X) - \mathbb{P}_0(S') \left[\text{ since } N \setminus \hat{S} = S' \setminus \{a_r\} \right] \\
&= \sum_{C \subseteq \hat{S}} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X) - \left(\sum_{C \subseteq \bar{S}'} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X) \right. \\
&\qquad\qquad\qquad\qquad \left. - \mathbb{P}_0(N \setminus \bar{S}') \right) - \mathbb{P}_0(S') \\
&\left[\text{by invoking Lemma 5.2.1, where } Z = N,\ Y = \bar{S}',\ P = C, \text{and } f(Z \setminus Y) = \mathbb{P}_0(N \setminus \bar{S}') \right] \\
&= \sum_{C \subseteq \hat{S}} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X) - \sum_{C \subseteq \bar{S}'} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X) \left[\text{ since } N \setminus \bar{S}' = S' \right] \\
&= \sum_{C \in \langle \bar{S}' \cup \{a_r\} \rangle} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X) - \sum_{C \subseteq \bar{S}'} \sum_{X \subseteq C} (-1)^{|C|-|X|} \cdot \mathbb{P}_0(N \setminus X) \\
&\left[\text{ since } \hat{S} = \bar{S}' \cup \{a_r\} \right] \\
&= \sum_{C \in \langle \bar{S}' \rangle \uplus \{a_r\}} \sum_{X \subseteq C} (-1)^{|C|-|X|} \mathbb{P}_0(N \setminus X) \\
&= \sum_{C \in \langle \bar{S}' \rangle \uplus \{a_r\}} \lambda(C), \text{ where } \lambda(C) = \sum_{X \subseteq C} (-1)^{|C|-|X|} \mathbb{P}_0(N \setminus X) \\
&= \sum_{C \subseteq N} \lambda(C) \cdot \mathbf{I}[a_r \in C] \cdot \mathbf{I}[C \in \langle \bar{S}' \rangle \uplus \{a_r\}] \\
&= \sum_{C \subseteq N} \lambda(C) \cdot \mathbf{I}[a_r \in C] \cdot \mathbf{I}[a_r \succ a_k \ \forall a_k \in S \cap C, a_k \neq a_r] \\
&= \sum_{C \subseteq N} \lambda(C) \cdot \mathbf{I}[a_r \in S \cap C] \cdot \mathbf{I}[a_r \succ a_k \ \forall a_k \in S \cap C, a_k \neq a_r] \left[\text{ since we assume that } a_r \in S, \right. \\
&\qquad\qquad\qquad\qquad \left. \text{otherwise the choice probability is } 0 \right],
\end{aligned}
$$

which is exactly the equation to compute the probability to purchase $a_r \in S$ for the offer set $S \subseteq N$ under GCC choice model. As a result, we also have $\mathbb{P}_0(S) = \sum_{C \subseteq N} \lambda(C) \cdot \mathbf{I}[S \cap C = \varnothing]$ becasue of the standard probability law, i.e., $\mathbb{P}_0(S) = 1 - \sum_{a_r \in S} \mathbb{P}_r(S)$. Note that above chain of equations specifies probability distribution function λ over consideration sets. Moreover, it follows from Proposition 2.2.2 that λ is defined uniquely. In order to complete the proof, we show that the preference relation $\succ$ is also defined uniquely. Suppose, by contradiction, there is another strict preference order $\succ'$ such that $\succ' \neq \succ$ and $\mathbb{P}_{\cdot}(\cdot)_{\succ',\lambda} = \mathbb{P}_{\cdot}(\cdot)_{\succ,\lambda}$. Therefore there exist items $a_i, a_j \in N$ s.t. $a_i \succ a_j$ and $a_j \succ' a_i$. By definition of GCC choice rule, we have

$$\mathbb{P}_i(\{a_i, a_j\})_{\succ,\lambda} = \sum_{C \subseteq N} \mathbf{I}[a_i \in C] \cdot \lambda(C),$$

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

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

$$\mathbb{P}_i(\{a_i, a_j\})_{\succ,\lambda} - \mathbb{P}_i(\{a_i, a_j\})_{\succ',\lambda} \geq \lambda(\{a_i, a_j\}) > 0, \ \left[\text{by Condition 3}\right],$$

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

Bibliography

[1] Albuquerque, P., and Bronnenberg, B. J. Measuring the impact of negative demand shocks on car dealer networks. *Marketing Science 31*, 1 (2012), 4–23.

[2] Allon, G., Federgruen, A., and Pierson, M. How much is a reduction of your customers' wait worth? an empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management 13*, 4 (2011), 489–507.

[3] Aouad, A., Farias, V. F., and Levi, R. Assortment optimization under consider-then-choose choice models. Working paper, MIT Sloan, 2015.

[4] Aouad, A., Farias, V. F., Levi, R., and Segev, D. The approximability of assortment optimization under ranking preferences. Working paper, MIT Sloan, 2015.

[5] Ben-Akiva, M. E., and Lerman, S. R. *Discrete choice analysis: theory and application to travel demand*, vol. 9. MIT press, 1985.

[6] Berbeglia, G., Garassino, A., and Vulcano, G. A comparative empirical study of discrete choice models in retail operations. Working paper, Melbourne Business School, Melbourne, Australia, 2018.

[7] Birghtwell, G., and Winkler, P. Counting linear extensions is# p-complete. In *Proceedings of the 23rd ACM Symposium on Theory of Computation* (1991), pp. 175–181.

[8] Blattberg, R. C., and Neslin, S. A. *Sales promotion: Concepts, methods, and strategies.* Prentice Hall, 1990.

[9] Blattberg, R. C., and Wisniewski, K. J. Price-induced patterns of competition. *Marketing science 8*, 4 (1989), 291–309.

[10] Block, H. D., Marschak, J., et al. Random orderings and stochastic theories of responses. *Contributions to probability and statistics 2* (1960), 97–132.

[11] Brinks, S. Indiscriminate promotions cost retailers. Forrester Consulting, April 2018.

[12] Bronnenberg, B. J., Kruger, M. W., and Mela, C. F. Database paper—the IRI marketing data set. *Marketing science 27*, 4 (2008), 745–748.

[13] Cachon, G. P., Terwiesch, C., and Xu, Y. On the effects of consumer search and firm entry in a multiproduct competitive market. *Marketing Science 27*, 3 (2008), 461–473.

[14] Campbell, B. M. *The existence of evoked set and determinants of its magnitude in brand choice behavior.* PhD thesis, Columbia University, 1969.

[15] Caplin, A., Dean, M., and Leahy, J. Rational inattention, optimal consideration sets and stochastic choice. Tech. rep., Working paper, 2016.

[16] Chandukala, S. R., Kim, J., Otter, T., Rossi, P. E., Allenby, G. M., et al. Choice models in marketing: Economic assumptions, challenges and trends. *Foundations and Trends® in Marketing 2*, 2 (2008), 97–184.

[17] Cohen, M., Jagabathula, S., and Mitrofanov, D. Customer loyalty in ride-hailing: Empirical evidence. *Working Paper* (2020).

[18] Cohen, M., and Mitrofanov, D. Lyft and uber ipo: Before and after. *Under Review* (2019).

[19] Cohen, M. C., Kalas, J., and Perakis, G. Optimizing promotions for multiple items in supermarkets. Working paper, Leonard N. Stern School of Business, New York University, 2017.

[20] Cohen, M. C., Leung, N.-H. Z., Panchamgam, K., Perakis, G., and Smith, A. The impact of linear optimization on promotion planning. *Operations Research 65*, 2 (2017), 446–468.

[21] Cooper, L. G. Competitive maps: The structure underlying asymmetric cross elasticities. *Management Science 34*, 6 (1988), 707–723.

[22] Davis, P. Spatial competition in retail markets: movie theaters. *The RAND Journal of Economics 37*, 4 (2006), 964–982.

[23] Dekimpe, M. G., Hanssens, D. M., and Silva-Risso, J. M. Long-run effects of price promotions in scanner markets. *Journal of Econometrics 89*, 1 (1998), 269–291.

[24] DelVecchio, D., Henard, D. H., and Freling, T. H. The effect of sales promotion on post-promotion brand preference: A meta-analysis. *Journal of Retailing 82*, 3 (2006), 203–213.

[25] Denman, T. Lowes foods launches personalized promotions. *RIS news* (2016), (April 18), https://risnews.com/lowes–foods–launches–personalized–promotions.

[26] Duran, M. A., and Grossmann, I. E. An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical programming 36*, 3 (1986), 307–339.

[27] Eliaz, K., and Spiegler, R. Consideration sets and competitive marketing. *The Review of Economic Studies 78*, 1 (2011), 235–262.

[28] Farhnam, A. Prices now pegged to your buying history at some markets. *ABC News* (2013), (December 2), http://abcnews.go.com/Business/supermarkets–introduce–personalized–pricing/story?id=21010246.

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

[30] Feldman, J., Paul, A., and Topaloglu, H. Assortment optimization with small consideration sets. Technical note, forthcoming in *Operations Research*, 2017.

[31] Feldman, J., and Topaloglu, H. Assortment optimization under the multinomial logit model with nested consideration sets. Tech. rep., Tech. rep., Working Paper, 2015.

[32] Foekens, E. W., Leeflang, P. S., and Wittink, D. R. Varying parameter models to accommodate dynamic promotion effects. *Journal of Econometrics 89*, 1 (1998), 249–268.

[33] Gallego, G., and Li, A. Attention, consideration then selection choice model. Working paper, HKUST, 2017.

[34] Gedenk, K., Neslin, S. A., and Ailawadi, K. L. Sales promotion. In *Retailing in the 21st Century*. Springer, 2006, pp. 345–359.

[35] Gilbride, T. J., and Allenby, G. M. A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science 23*, 3 (2004), 391–406.

[36] Goh, C. Y., Yan, C., and Jaillet, P. Estimating primary demand in bike-sharing systems. *Available at SSRN 3311371* (2019).

[37] Graham, R. L. *Handbook of combinatorics*. Elsevier, 1995.

[38] Guadagni, P. M., and Little, J. D. A logit model of brand choice calibrated on scanner data. *Marketing Science 2*, 3 (1983), 203–238.

[39] Hauser, J. Consideration-set heuristics. *Journal of Business Research 67*, 8 (2014), 1688–1699.

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

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

[42] Hauser, J. R., Ding, M., and Gaskin, S. P. Non-compensatory (and compensatory) models of consideration-set decisions. In *2009 Sawtooth Software Conference Proceedings, Sequin WA* (2009).

[43] Hauser, J. R., and Wernerfelt, B. An evaluation cost model of consideration sets. *Journal of consumer research 16*, 4 (1990), 393–408.

[44] Hausman, J. A. Valuation of new goods under perfect and imperfect competition. In *The economics of new goods*. University of Chicago Press, 1996, pp. 207–248.

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

[46] Hotelling, H. Stability in competition. *The economic journal 39*, 153 (1929), 41–57.

[47] Howard, J. A., and Sheth, J. N. The theory of buying behavior. *New York* (1969).

[48] Hoyer, W. D. An examination of consumer decision making for a common repeat purchase product. *Journal of consumer research 11*, 3 (1984), 822–829.

[49] Jagabathula, S., Mitrofanov, D., and Vulcano, G. Personalized retail promotions through a dag-based representation of customer preferences. *Available at SSRN 3258700* (2018).

[50] Jagabathula, S., Mitrofanov, D., and Vulcano, G. Inferring consideration sets from sales transaction data. *Available at SSRN 3410019* (2019).

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

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

[53] Jeuland, A. Brand choice inertia as one aspect of the notion of brand loyalty. *Management Science 25*, 7 (1979), 671–682.

[54] Kalyanaram, G., and Winer, R. S. Empirical generalizations from reference price research. *Marketing science 14*, 3_supplement (1995), G161–G169.

[55] Karp, R. Reducibility among combinatorial problems. In: Miller R.E., Thatcher J.W., Bohlinger J.D. (eds) Complexity of Computer Computations. The IBM Research Symposia Series. Springer, Boston, MA, 1972.

[56] Khan, R., Lewis, M., and Singh, V. Dynamic customer management and the value of one-to-one marketing. *Marketing Science 28*, 6 (2009), 1063–1079.

[57] Khandelwal, A., Ma, W., and Simchi-Levi, D. Predicting user choice in video games. Presentation at the INFORMS RMP Conference 2017, Amsterdam, NL, 2017.

[58] Kharif, O. Supermarkets offer personalized pricing. *Bloomberg* (2013), (November 15), https://www.bloomberg.com/news/articles/2013–11–14/2014–outlook–supermarkets–offer–personalized–pricing.

[59] L. McFadden, D. *Conditional Logit Analysis of Qualitative Choice Behavior*, vol. 8. Academic Press, 01 1974, pp. 105–142.

[60] Ladd, B. Technology startup apricart wants to revolutionize the grocery business for consumers and retailers. *Forbes* (2019), (February 11), https://www.forbes.com/sites/brittainladd/2019/02/11/technology–startup–apricart–wants–to–revolutionize–the–grocery–business–for–consumers–and–retailers/479ed1df4b45.

[61] Lederman, R., Olivares, M., and Van Ryzin, G. Identifying competitors in markets with fixed product offerings. Working paper, University of Chile, 2014.

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

[63] Li, J., and Netessine, S. Who are my competitors? let the customer decide. Working paper, Michigan Ross, 2012.

[64] Luce, R. D. *Individual Choice Behavior a Theoretical Analysis.* John Wiley and sons, 1959.

[65] Lynch, J. G. Memory and decision making, 1991.

[66] Mahajan, S., and van Ryzin, G. Stocking retail assortments under dynamic consumer substitution. *Operations Research 49* (2001), 334–351.

[67] Mahajan, S., and Van Ryzin, G. Stocking retail assortments under dynamic consumer substitution. *Operations Research 49*, 3 (2001), 334–351.

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

[69] Marden, J. I. Analyzing and modeling rank data. Monographs on Statistics and Applied Probability, Vol. 64, 1995.

[70] Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. Revealed attention. In *Behavioral Economics of Preferences, Choices, and Happiness.* Springer, 2016, pp. 495–522.

[71] Matejka, F., and McKay, A. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review 105*, 1 (2015), 272–98.

[72] Mela, C. F., Gupta, S., and Lehmann, D. R. The long-term impact of promotion and advertising on consumer brand choice. *Journal of Marketing Research 34*, 2 (1997), 248–261.

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

[74] Murphy, K. P. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[75] Newman, J. P., Ferguson, M. E., Garrow, L. A., and Jacobs, T. L. Estimation of choice-based models using sales data from a single firm. *Manufacturing & Service Operations Management 16*, 2 (2014), 184–197.

[76] Olivares, M., and Cachon, G. P. Competing retailers and inventory: An empirical investigation of general motors' dealerships in isolated us markets. *Management Science 55*, 9 (2009), 1586–1604.

[77] Pinkse, J., Slade, M. E., and Brett, C. Spatial price competition: a semiparametric approach. *Econometrica 70*, 3 (2002), 1111–1153.

[78] Pounder, J. For what it's worth –the future of personalized pricing. *The Guardian* (2015), (November 6), https://www.theguardian.com/media–network/2015/nov/06/personalised–pricing–future–online–offline–retail.

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

[80] Ratliff, R., Rao, B., Narayan, C., and Yellepeddi, K. A multi-flight recapture heuristic for estimating unconstrained demand from airline bookings. *Journal of Revenue and Pricing Management 7* (2008), 153–171.

[81] Roberts, J. H., and Lattin, J. M. Development and testing of a model of consideration set composition. *Journal of Marketing Research* (1991), 429–440.

[82] Roberts, J. H., and Lattin, J. M. Consideration: Review of research and prospects for future insights. *Journal of Marketing Research* (1997), 406–410.

[83] Rusmevichientong, P., Van Roy, B., and Glynn, P. W. A nonparametric approach to multiproduct pricing. *Operations Research 54*, 1 (2006), 82–98.

[84] Sher, I., Fox, J. T., Bajari, P., et al. Partial identification of heterogeneity in preference orderings over discrete choices. Tech. rep., National Bureau of Economic Research, 2011.

[85] Smith, N. Big data might lead to higher prices. *Bloomberg* (2018), (March 9), https://www.bloomberg.com/opinion/articles/2018–03–09/big–data–might–tell–retailers–which–consumers–to–charge–more.

[86] Srinivasan, S., Pauwels, K., Hanssens, D. M., and Dekimpe, M. G. Do promotions benefit manufacturers, retailers, or both? *Management Science 50*, 5 (2004), 617–629.

[87] Strauss, D. Some results on random utility models. *Journal of Mathematical Psychology 20*, 1 (1979), 35–52.

[88] Suh, J.-C. The role of consideration sets in brand choice: the moderating role of product characteristics. *Psychology & Marketing 26*, 6 (2009), 534–550.

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

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

[91] Thomadsen, R. Product positioning and competition: The role of location in the fast food industry. *Marketing Science 26*, 6 (2007), 792–804.

[92] Train, K. *Discrete choice methods with simulation.* Cambridge University Press, 2009.

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

[94] Tversky, A., and Kahneman, D. Judgment under uncertainty: Heuristics and biases. *science 185*, 4157 (1974), 1124–1131.

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

[96] Villas-Boas, J. M., and Winer, R. S. Endogeneity in brand choice models. *Management science 45*, 10 (1999), 1324–1338.

[97] Vujanic, A., and Goldstein, N. U.S. consumers want more personalized retail experience and control over personal information. Accenture Survey, March 9, 2015.

[98] Vulcano, G., Van Ryzin, G., and Chaar, W. Om practice—choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing & Service Operations Management 12*, 3 (2010), 371–392.

[99] Wang, R., and Sahin, O. The impact of consumer search cost on assortment planning and pricing. *Management Science* (2017).

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

[101] Wierenga, B., van Bruggen, G. H., and Althuizen, N. A. Advances in marketing management support systems. In *Handbook of Marketing Decision Models*. Springer, 2008, pp. 561–592.

[102] Wright, P., and Barbour, F. *Phased decision strategies: Sequels to an initial screening.* Graduate School of Business, Stanford University, 1977.

[103] Zhang, J., and Krishnamurthi, L. Customizing promotions in online stores. *Marketing science 23*, 4 (2004), 561–578.