Nonparametric Choice Modeling: Applications to Operations Management

by

Srikanth Jagabathula

S.M., Massachusetts Institute of Technology (2008)
B.Tech., Indian Institute of Technology Bombay (2006)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2011

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SEP 27 2011
LIBRARIES
ARCHIVES

© Massachusetts Institute of Technology 2011. All rights reserved.

Author .....
Department of Electrical Engineering and Computer Science
Aug 18, 2011

Certified by .....
Vivek F. Farias
Associate Professor
Thesis Supervisor

Certified by ....
Devavrat Shah
Associate Professor
Thesis Supervisor

Accepted by ....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Theses

Nonparametric Choice Modeling: Applications to Operations Management

by

Srikanth Jagabathula

Submitted to the Department of Electrical Engineering and Computer Science
on Aug 18, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

With the recent explosion of choices available to us in every walk of our life, capturing the choice behavior exhibited by individuals has become increasingly important to many businesses. At the core, capturing choice behavior boils down to being able to predict the probability of choosing a particular alternative from an offer set, given historical choice data about an individual or a group of "similar" individuals. For such predictions, one uses what is called a choice model, which models each choice occasion as follows: given an offer set, a preference list over alternatives is sampled according to a certain distribution, and the individual chooses the most preferred alternative according to the sampled preference list. Most existing literature, which dates back to at least the 1920s, considers parametric approaches to choice modeling. The goal of this thesis is to deviate from the existing approaches to propose a nonparametric approach to modeling choice. Apart from the usual advantages, the primary strength of a nonparametric model is its ability to scale with the data – certainly crucial to applications of our interest where choice behavior is highly dynamic. Given this, the main contribution of the thesis is to operationalize the nonparametric approach and demonstrate its success in several important applications.

Specifically, we consider two broad setups: (1) solving decision problems using choice models, and (2) learning the choice models. In both setups, data available corresponds to marginal information about the underlying distribution over rankings. So the problems essentially boil down to designing the 'right' criterion to pick a model from one of the (several) distributions that are consistent with the available marginal information.

First, we consider a central decision problem in operations management (OM): find an assortment of products that maximizes the revenues subject to a capacity constraint on the size of the assortment. Solving this problem requires two components: (a) predicting revenues for assortments and (b) searching over all subsets of a certain size for the optimal assortment. In order to predict revenues for an assortment, of all models consistent with the data, we use the choice model that results in the 'worst-case' revenue. We derive theoretical guarantees for the predictions, and

show that the accuracy of predictions is good for the cases when the choice data comes from several different parametric models. Finally, by applying our approach to real-world sales transaction data from a major US automaker, we demonstrate an improvement in accuracy of around 20% over state-of-the-art parametric approaches. Once we have revenue predictions, we consider the problem of finding the optimal assortment. It has been shown that this problem is provably hard for most of the important families of parametric of choice models, except the multinomial logit (MNL) model. In addition, most of the approximation schemes proposed in the literature are tailored to a specific parametric structure. We deviate from this and propose a general algorithm to find the optimal assortment assuming access to only a subroutine that gives revenue predictions; this means that the algorithm can be applied with any choice model. We prove that when the underlying choice model is the MNL model, our algorithm can find the optimal assortment efficiently.

Next, we consider the problem of learning the underlying distribution from the given marginal information. For that, of all the models consistent with the data, we propose to select the sparsest or simplest model, where we measure sparsity as the support size of the distribution. Finding the sparsest distribution is hard in general, so we restrict our search to what we call the 'signature family' to obtain an algorithm that is computationally efficient compared to the brute-force approach. We show that the price one pays for restricting the search to the signature family is minimal by establishing that for a large class of models, there exists a "sparse enough" model in the signature family that fits the given marginal information well. We demonstrate the efficacy of learning sparse models on the well-known American Psychological Association (APA) dataset by showing that our sparse approximation manages to capture useful structural properties of the underlying model. Finally, our results suggest that signature condition can be considered an alternative to the recently popularized Restricted Null Space condition for efficient recovery of sparse models.

Thesis Supervisor: Vivek F. Farias
Title: Associate Professor

Thesis Supervisor: Devavrat Shah
Title: Associate Professor

To my parents – Shanmukha Rao and Nagamani Jagabathula

Acknowledgments

This thesis is a product of my close collaboration with my research advisors Vivek Farias and Devavrat Shah. Words cannot express my feelings of gratitude towards both for their constant support, encouragement, and guidance. They are the main reason I can now look back and say that I have had an incredible last five years at MIT. I started working with Devavrat from the day I joined MIT. He has always been a constant source of guidance and inspiration. I am grateful to him for his care and patience, especially during my initial years at grad school. Life in graduate school can be trying at times and it is a testament to his patience that he always displayed genuine concern at every step. Part of the reason I decided to continue with my PhD after my Masters was the passion in reasearch he instilled in me. I started working with Vivek in the third year of my graduate school, and we never looked back. Vivek has such infectitious enthusiasm in everything he does that one can't help but get excited along with him. I am always amazed by his ability to quickly come up with a bird's-eye view of any problem or crisp explanations for any complex topic. In addition, he is always full of great ideas about anything under the sun – if you want advice on something, even completely unrelated to research, he is the one you should go to. I hope to have the pleasure of working with both on interesting problems going into the future.

I am also grateful to Prof. Gabriel Bitran and Prof. John Tsitsiklis for being on my thesis committee. I am greatly indebted to them for their constant encouragement and infectitious enthusiasm about my work. I also greatly appreciate the genuine concern, constant encouragement and help of Prof. Georgia Perakis and Prof. Retsef Levi. During the course of my stay at LIDS, I also had the good fortune of interacting with Prof. Munther Dahleh, Prof. Sanjoy Mitter, and Prof. Alan Willsky – they have all been highly inspirational in addition to being instrumental in making me feel part of the greater LIDS family.

I am also thankful to Lynne Dell and Jennifer Donovan for being such efficient and caring administrators. Their sweet and approachable personas have made life at

LIDS extremely enjoyable.

Life at MIT would not have been as much fun without the friends I cultivated. My labmates and good friends Ammar Ammar, Shreeshankar Bodas, Sewoong Oh, Tauhid Zaman, and Yuan Zhong have been an amazing set of colleagues and friends. Their interesting personality quirks have made all of our social gatherings incredibly memorable for me. Special thanks to Ammar Ammar and Tauhid Zaman for bearing me as an officemate. I cannot forget the amount of fun we had – making me want to spend more time at the office than at my room. I also feel priviledged to have had Vishal Doshi, Kyomin Jung, Urs Niesen, and Jinwoo Shin as my colleagues and labmates.

Also, special thanks to my roommates over the years – Abhinav Akhoury, Nitin Rao, Diwakar Shukla, and Tauhid Zaman. I sympathize with them about having to bear my several quirks as a roommate. Each of them taught me several things about life in general. I greatly value the relationships we have built over the years.

I am also thankful to all the friendships I have nurtured at greater MIT over the years. I thank them for their unconditional love, support, and the numerous dinners and movie nights.

Finally, I want to thank my family for their constant support all through my life – no achievement in my life would have been possible without them. I am forever indebted to my parents Shanmukha Rao and Nagamani. Their love, encouragement, and sacrifice at every step of the way have made me what I am. I can only hope to one day be as honest, loving, and hardworking they are. This thesis is dedicated to them. I also express a feeling of bliss for having a loving sister like Nagadeepthi. She has always been the older sister who is genuinely concerned and ever encouraging.

Last, but not the least, I thank my love Deepti for always being there for me. I really thank her for putting up with me during my ups and downs, acccommodating my whims, and showing constant love and care. She is the one person who truly knows and understands me. I thank her for all her faith in me.

Preliminaries

How to read this thesis

This thesis focuses on operationalizing a new nonparametric approach to choice modeling. The thesis is logically divided into three parts: (1) introduction (Chapter 1), (2) using choice models to make decisions (Chapters 3 and 4), and (3) learning choice models from historical transaction data (Chapter 5). The introduction sets the stage by providing the motivation for a nonparametric approach to choice modeling. It also provides a comprehensive overview of the problems we consider and interprets the main results we obtain. In order to make this thesis more widely accessible, an attempt has been made to explain the problems and results in the introduction with as little notation as possible. A reader is strongly advised to read the introduction first. The individual chapters provide precise details of the results we obtain. All the chapters are self-sufficient and can be read in any order.

Chapter 2 deals with background and provides a comprehensive overview of the rich history of choice models. Due to the fundamental nature of choice models, the literature on choice models goes back to at least the 1920s and spans several areas. Chapter 2 stitches together the diverse modeling approaches across the different areas through two common themes.

Chapters 3 and 4 are devoted to solving decision problems using choice models. They discuss the algorithmic solutions we provide and the guarantees we can prove for the algorithms.

Chapter 5 deals with question of learning sparse choice models from historical preference data. Finally, in order to provide a better flow, some of the implementation details and experimental setup details have been moved to the appendix.

Bibliographic notes

The preliminary results of Chapter 3 and Chapter 5 have appeared as conference papers by Farias et al. [2009] and Jagabathula and Shah [2008]. Part of the results

of Chapter 5 are published as the journal paper by Jagabathula and Shah [2011]. Most of results of Chapter 3 are submitted as a journal paper by Farias et al. [2010]. Finally, preprints of the results in Chapter 4 has been made available by Farias et al. [2010].

The paper containing preliminary results of Chapter 5 received the Best Student Paper Award at the Neural Information Processing Systems (NIPS) in December 2008. In addition, the paper with the results presented in Chapter 3 received the first place in the MSOM Student Paper Competition in 2010.

Contents

[1]This family of models is also referred to in the literature as Random parameter logit (RPL), Kernel/Hybrid logit.

List of Figures

List of Tables

Chapter 1

Introduction

The recent explosion of choices available for products, coupled with easy access to information through the internet, has fundamentally changed how we make purchases. Increasingly, purchase decisions are influenced not only by the individual merits of the products, but also by what else is on offer. An important implication of this 'choice behavior' is that the purchase decisions of customers are greatly affected by the selection of products that a seller offers. At a first cut, it seems that a good strategy for the seller is to offer as wide a selection of products as possible – after all, wouldn't a wide selection increase both sales and customer satisfaction? While this is a reasonable strategy on surface, it has in fact been shown *not* to be optimal – there *is* such a thing as too much choice, and too much variety can adversely affect both customer satisfaction and sales (see Iyengar and Lepper [2000]). Moreover, capacity constraints like limited shelf space for a retailer or limited screen real-estate for an e-tailer typically constrain the variety a seller can offer in practice. Thus, in order to optimize the offered selection of products, it has become increasingly important for the seller to gain a better understanding of the choice behavior of customers.

Added to the explosion of choices, there has been a corresponding explosion of raw data. Technological advances have made it extremely simple to collect and analyze large quantities of raw data (such as the point-of-sales scanner data), making available to businesses a wealth of information to understand customer choice behavior. While the data possesses a wealth of useful information, extracting it in a meaningful way is

far from trivial. Thus, along with opening several potential avenues, this 'data deluge' is posing challenges to businesses in handling huge quantities of data and extracting useful information from it effectively.

Therefore, we are at this critical juncture, where businesses increasingly need to gain a better understanding of customer choice behavior, and to do that, they have access to large quantities of data containing rich information about customer preferences. What is lacking though is a means to effectively capture the information in the data to gain a meaningful understanding of user choice behavior. The goal of this thesis is to fill this void through a nonparametric approach to choice modeling, where we take a data-centric view point and develop tools that enable practitioners to effectively channel the information into insights by off-loading subjective judgements to the data. Our approach is a deviation from the traditional parametric approaches to choice modeling. In order to understand the need for a new approach, we next take a closer look at the challenges that arise with modeling choice. Before we do a deep dive, it is worth mentioning that in addition to the applications in operations management, choice modeling naturally arises in the problem of *rank aggregation*, which is important to applications like web-search, recommendation systems, and polling systems. We elaborate on some of these applications in the later sections.

1.1 Parametric models and their limitations

Understanding and predicting choice behavior essentially boils down to understanding individual preferences – after all, when offered a set of alternatives, an individual chooses the alternative that is the most preferred of the offered choices. More precisely, a concrete goal in this context is to predict the probability that a particular alternative will be chosen from an offer set by using the historical data about the choices made either by an individual or by a group of "similar" individuals. For instance, in the retail context, historical choice data is available as sales transactions, and the goal would be to predict the probability of purchase of a particular type of product when a specific assortment of products is offered. Traditionally, a *choice*

model has been used for this purpose. Given the ubiquity of choice behavior, several variants of choice models – each to serve a different purpose – have been studied in diverse areas at least for the last four decades. But at the core, almost all 'choice models' model choice behavior through preference lists: each individual has a preference list of the alternatives, and when offered a set of alternatives, chooses the most preferred alternative. This view was proposed several decades ago by the economists (see Mas-Colell et al. [1995]), and forms the basis for the popular *utility maximization* model, where the preferences can be equivalently captured by a utility function (satisfying some properties) that maps each alternative to a real number. When we go from individuals to a group of "similar" individuals, we can talk about an aggregate choice model, which corresponds to a distribution over preference lists: the weight of each preference list corresponds to the fraction of people who make choices according to the particular preference list. A distribution over preference lists also makes sense to model the choice behavior at an individual level when a single preference list fails to capture the complexity of individual preferences. Therefore, the typical recipe in this context is to use the historical choice data to learn the distribution over preference lists and use the so obtained choice model to make predictions.

Now, the main challenge to learning and using a choice model (a distribution over preference lists) is the woefully limited data that is available in practice. Particularly, when there are N alternatives, we are trying to learn a function with $N!$ degrees of freedom with typically polynomially (in N) many data points. The traditional approach to overcoming this data limitation has been to limit the degrees of freedom by imposing a parametric structure on the distribution. This has been a very popular approach and has led to a rich body of work in diverse areas like statistics, marketing, and psychology, resulting in an extensive family of models for the distributions over preference lists dating back to at least the 1920s. Of these diverse modeling approaches, the most popular and extensively studied family of distributions is the Random Utility Maximization (RUM) family (a detailed exposition of the existing modeling approaches is given in Chapter 2). The RUM family of models has beginnings in the psychological literature, where the goal was to develop models that

were grounded in more basic processes. It is assumed in this model that choices are driven by various measurable attributes of the alternatives and the individual, which lead each individual to assign each alternative a utility value; once the utilities are assigned, the individual chooses the alternative with the maximum utility. This is essentially the view proposed in the classical utility maximization framework; the only difference is that in the classical setting, the assigned utilities are assumed to be real values, whereas in the RUM family, utilities are modeled as random variables[1]. With utilities modeled as random variables, different choices of parametric distributions for the random utilities give rise to different members of the RUM family. Once a parametric distribution is selected, choices are made as follows: in each decision instance, the individual samples utilities according to the selected distribution and chooses the alternative with the maximum of the sampled utilities. The choice of the parametric distribution is typically motivated by computational aspects and the application at hand.

It should be noted here that the RUM family has a rich history dating all the way back to the 1920s and has been successfully applied to real-world contexts. Nevertheless, it has its pros and cons. The strength of the RUM framework is that it often results in intuitively interpretable structures and also allows incorporation of fine-grained attribute-level information (through the utility function) about the individuals and the alternatives. The downside however is that selecting the "right" structure is far from trivial. Specifically, there are two challenges: Firstly, choosing the "right" parametric structure for the problem at hand requires making the optimal trade-off between the richness of the model and the learnability of the model – the chosen structure has to be simple enough, on the one hand, that it can be learned from the limited "amount" of information in the data available, and complex enough, on the other hand, that it can capture the richness of the choice problem being modeled; striking the right balance is challenging because formalizing the trade-off between model "complexity" and "amount" of information in the data is non-trivial.

[1]There are interesting interpretations of the randomness introduced into utilities. See the introduction of Chapter 2 for more details.

Secondly, even if one did choose an appropriate structure, parametric models do not scale with the "amount" of information at hand. The implication is that, once the structure is fixed, the model could either under-fit or over-fit as the "amount" of information in the data changes – which often happens in practice (an illustration of this fact in a real-world scenario is presented later).

1.2 The task of nonparametric choice modeling

In order to overcome the limitation of the parametric models to scale with data, this thesis studies[2] a nonparametric approach to choice modeling. One of the main application areas of our focus is the prediction and decision problems faced by operations managers. These applications possess three important characteristics:

1. Making accurate and fine-grained predictions is more important than understanding the underlying choice process; in fact, the decisions based on these predictions potentially impact millions of dollars.

2. The choice behavior exhibited by individuals is highly dynamic and rapidly varying across geographical locations and times.

3. Technological developments have made available enormous amounts of transaction data.

Given these characteristics, it is evident a data-driven modeling approach is not only a possibility, but also a necessity in this context. Motivated by these considerations, instead of fixing a structure a priori (not even an exponential structure), we start with the family of all distributions over preference lists, and off-load to the data the task of selecting an appropriate distribution from this general family. This data-driven approach naturally leads to models whose complexity adaptively scales with

[2]We note that another popular family is the the exponential family of choice models. This family does allow the model complexity (as measured by the number of parameters) scale with the data. However, they pose several computational challenges even to simply compute choice probabilities and hence are not very very popular in the applications of our interest; a detailed discussion is presented in Chapter 2.

the data. Our experiments with real-world data show that our approach is effective in reducing the chance of over-fitting and under-fitting. Our approach is particularly valuable to practitioners because it not only removes the burden of making subjective judgements, but also allows for the same implementation to be used in different settings by just changing the input data. The price one pays with this nonparametric approach is that the rich structure of parametric models – which is typically exploited to solve decisions problems efficiently – is no longer available. This is however not a huge price for the following reasons. Firstly, in many applications (like the ones we are considering in OM) accuracy is more important than computationally efficiency; solving the decision problem to optimality having assumed an incorrect model is clearly of not much use. Secondly, as elaborated later, except in the case of the simplest of the models – namely the MNL model – very little is already known about solving decision problems efficiently; therefore, most of the parametric models are already complicated enough that their use does not simplify the associated decision problems in any meaningful manner. We note here that, as elaborated later, our results in the nonparametric setting to a large extent subsume the results known for parametric models making this more general than the existing approaches.

Although a nonparametric approach to choice modeling is very appealing, solving the core problem of learning distributions over permutations from marginal information is a notoriously hard problem; the main challenge, both from a theoretical and practical standpoint, is the factorial (in the number of alternatives) blowup in their support size. In addition, their structure does not allow for representation through a compact graphical model. In fact, one of the main reasons for the popularity of parametric approaches is the difficultly associated with learning general distributions over permutations without imposing tractable structures. Very little is already understood in this setting. Therefore, in addition to the practical implications, solving the problem of learning a distribution over permutations from marginal information involves answering rich theoretical methodological questions. A common methodological thread in our work is to (a) understand what type of information can be extracted from marginal data, and (b) propose methods to effectively extract that

information. In each of the settings we consider, we obtain a formal characterization of the type of information that can be extracted from the given marginal data. We then propose computationally feasible methods to effectively extract that information from the marginal data. Our work also has rich connections to the area of compressive sensing that has recently become popular in the areas of coding theory, signal processing, and streaming algorithms, which we explore in detail in later sections.

Before we detail the concrete problems we consider and the main contributions we make in this thesis, we introduce some preliminaries of the model.

1.3 Model preliminaries

We describe our model in the context of a customer purchasing a product from an offered set of products; of course, the model extends trivially to the context of the choice of general alternatives. We consider a universe of N products or alternatives, denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$. Where required, we assume that a customer always has an 'outside' or 'no-purchase' option; the choice of the 'outside' option is equivalent to the customer not choosing anything from the offered set of products, and we often denote the 'outside' option by product 0. Each customer is associated with a preference list (or ranking) σ of the $N+1$ products in $\mathcal{N} \cup \{0\}$, and the customer with preference list σ prefers product i to product j if and only if $\sigma(i) < \sigma(j)$. Now, given any assortment of products $\mathcal{M} \subset \mathcal{N}$, the customer chooses the product that is the most preferred of the no-purchase option and the offered products in $\mathcal{M}$; in other words, the customer chooses the product $\underset{i \in \mathcal{M} \cup \{0\}}{\operatorname{argmin}} \sigma(i)$. This view of choice behavior is clearly intuitively very reasonable and very general.

We model the aggregate behavior of a population of customers through a distribution λ over the space of all permutations of the products in $\mathcal{N} \cup \{0\}$; specifically, for any permutation σ, $\lambda(\sigma)$ corresponds to the fraction of customers in the population with preference list σ. It is easy to see that the distribution λ captures all the desired choice probabilities. Particularly, suppose the customer population described by distribution λ is offered an assortment of products $\mathcal{M} \subset \mathcal{N}$; then, the probability

that product i is purchased (the fraction of customers that purchase product i) from $\mathcal{M}$ is equal to the sum of the weights of λ over all permutations that result in the choice of i when offered $\mathcal{M}$. More precisely, the probability of purchase of i given $\mathcal{M}$ can be written as[3]

$$\mathbb{P}_\lambda(i|\mathcal{M}) = \sum_{\sigma \in \mathcal{S}_i(\mathcal{M})} \lambda(\sigma),$$

where $\mathcal{S}_i(\mathcal{M})$ is the set of permutations that result in the choice of i from $\mathcal{M}$; equivalently, $\mathcal{S}_i(\mathcal{M})$ is the set of permutations that prefer i to all the products in $\mathcal{M} \cup \{0\}$). More formally, we can write

$$\mathcal{S}_i(\mathcal{M}) = \{\sigma \colon \sigma(i) < \sigma(j) \text{ for all } j \in \mathcal{M} \cup \{0\}\}.$$

The distribution λ captures all the required choice probabilities and we call it the choice model. As mentioned above, using distributions over permutations to model choice behavior is very general and forms the basis for the general class of random utility choice models.

In OM/RM applications, the sales or revenues expected from offering an assortment $\mathcal{M}$ of products is an important quantity of interest. Given a choice model λ, one can compute the expected revenue as follows. Let p_i denote the price of product i or the revenue obtained from the sale of product i; as expected, we set p_0 to 0. Suppose assortment $\mathcal{M}$ is offered to customers from a population described by choice model λ. Then, the revenue expected from each arriving customer is given by

$$R(\mathcal{M}) = \sum_{i \in \mathcal{M}} p_i \mathbb{P}_\lambda(i|\mathcal{M}).$$

Setting $p_i = 1$ for all $i \in \mathcal{N}$ in the above expression yields the probability that an arriving customer will make a purchase, or what is called the 'conversion-rate', which refers to the probability of converting an arriving customer to a sale.

Broadly speaking, our goal in this thesis is to either learn the choice model λ or

[3]Sometimes we drop λ and simply use $\mathbb{P}(i|\mathcal{M})$ to denote the choice probability in cases where the choice model is irrelevant or implied by the context.

estimate a functional of the choice model such as the revenues $R(\mathcal{M})$. In order to perform either of these tasks, we use the data that is available at hand. We defer the precise definition of the type of data to later chapters, but we describe here how we intuitively think of data. Loosely speaking, we think of data as imposing constraints on the possible distributions λ that can describe the population. Specifically, without any other information, the choice behavior of a population could be described by any distribution λ. Now suppose we gather some data on how the customers make choices (maybe by offering the population members some assortments and noting down their choices). This new information now restricts the plausible distributions λ: particularly, disallows all distributions that are not consistent with the observed choices. With this view, the data available in practice allows us to identify a family of consistent distributions. Since λ has $(N+1)!$ degrees of freedom and data available is often limited, it is typically the case that there is more than one consistent distribution λ. For instance, data available in practice typically consists of choices (or purchases) made by individuals when offered different assortments of products; it is easy to see that this data does not contain information about the preferences of individuals over products that are not preferred to the no-purchase option. Because of that, roughly speaking, learning λ or estimating a functional of λ reduces to designing the "right" criterion to pick one of the consistent distributions.

Before we conclude the section on model description, we make the following remarks about the model.

- Note that we modeled each individual as making choices according to a dedicated preference list. On the surface, it seems that this assumption is restrictive because it requires individuals to exhibit transitive preferences i.e., if an individual prefers a to b and b to c, then the individual must prefer a to c. Although such rationality is assumed in most economic models, several empirical studies have invalidated the assumption in practice (see Anand [1993]). Contrary to what it seems on the surface, our model does not quite require such a strong rationality assumption. Specifically, an equivalent way to model a population of "similar" individuals is to assume that in any choice occasion each individual

samples a permutation σ according to distribution λ and makes choices according to the sampled permutation[4]. By modeling individuals as distributions λ over permutations, we allow the same individual to sample different permutations on different choice occasions leading one to prefer a to b and b to c, but prefer c to a, as long as the choices are on different occasions. This makes our model richer and more realistic.

- Although we model each individual as making choices according to a preference list over all the $N+1$ products, the individual need not be aware of one's own entire preference list; in any choice instance, the individual need only be aware of his/her preferences over the offered products.

1.4 Contributions of the thesis

The main contribution of this thesis is to operationalize such a nonparametric approach to choice modeling. We distinguish two broad setups based on the ultimate goals: (a) to learn the choice model, and (b) to use the choice model to make predictions or decisions. In the latter setup, as shall become evident later, the problem typically boils down to learning a linear functional of the choice model, rather than the entire distribution. The latter setup is clearly a special case of the former setup. However, since the second setup is simpler, we can obtain stronger results; further, these results have huge practical implications because a wide-range of practically important problems that are studied extensively in the area of OM reduce to prediction and decision problems using choice models.

Next, we detail our contributions by providing a brief overview of the problem we consider and the main results we obtain. Due to the nature of the results, we first detail our contributions relative to solving prediction and decision problems using choice models, and then we describe our results for the problem of learning the distribution over preference lists. The main application area of our focus is the

[4]Similarity arises from the fact that each individual in the population samples preference lists from the same distribution.

set of decision problems involving choice models faced by operations managers. In connection to this, we also describe a case-study applying our solutions to the real-world problems faced by a major US automaker. While the focus shall remain on problems in OM, we also briefly describe empirical studies with election data from the American Psychological Association (APA).

In the context of decision problems, we study two related problems that are central to the area of OM; their significance stems from the fact that they constitute crucial components in solving several important problems in the area of OM. Specifically, given historical sales transaction data, we consider the problems: (a) predict the expected revenues or sales from offering a particular assortment of products, and (b) determine the assortment that maximizes the expected revenues subject to a capacity constraint on the size of the assortment. The second problem is often referred to as *static assortment optimization problem* and needs revenue predictions as a subroutine.

As one can imagine, both problems are important in their own right. Their consideration is motivated by the fact that any improvements to existing solutions will have significant practical implications. Specifically, solutions to these two problems lead to a solution to the *single-leg, multiple fare-class yield management problem*; this problem is central to the area Revenue Management (RM) and deals with the allocation of aircraft seat capacity to multiple fare classes when customers exhibit choice behavior. In particular, consider an airline selling tickets to a single-leg aircraft. Assume that the airline has already decided the fare classes and is trying to dynamically decide which fare-classes to open as a function of the remaining booking time and the remaining number of seats. This dynamic decision problem can be cast in a reasonably straightforward manner as a dynamic program with one state variable. As explained later, the solution to the dynamic program reduces to solving a slight variant of the static assortment optimization problem. Thus, solution to the two problems effectively solves a central problem to the area of RM. We next present an overview of the solutions we propose for the two problems.

1.4.1 Revenue predictions

The problem of using historical sales transaction data to predict expected revenues or sales for different assortments is a basic – yet, non-trivial – problem faced by operations managers. An important special case is the prediction of the assortment-level 'conversion-rate' for each assortment – the probability of converting an arriving customer into a purchasing customer. As one can imagine, such predictions form critical inputs to several important business decisions, both operational and otherwise. Now, in order to predict revenues, we need to predict the purchase probability of each product, which in turn depends on all the products on offer due to substitution behavior: an arriving potentially substitutes an unavailable product with an available one. As a result, a choice model is used to capture this dependence resulting from substitution. However, most of the work in OM that deals with solving decision problems requiring revenue predictions take the choice model as given and assume access to accurate predictions; when it comes to actually predicting revenues, a popular parametric model like the multinomial logit (MNL) model is used. Unfortunately, as mentioned above, parametric models suffer from various limitations, the most important of which is that their model complexity does not scale with data because of which they fail to glean new structural information in additional data; put differently, parametric models tend to over-fit or under-fit resulting in poor accuracy in their predictions. In order to overcome this issue, we propose a nonparametric approach to predicting revenues and sales that affords the revenue manager the opportunity to avoid the challenging task of fitting an appropriate parametric model to historical data. As mentioned above, we start with the entire family of distributions over preference lists and let the data select the "right" choice model to make revenue predictions. Specifically, we make the following important contributions.

Given historical sales transaction data and a heretofore unseen assortment, we start with the goal of predicting the revenues expected from each arriving customer when offered the assortment. We first identify the family of distributions over preference lists that are consistent with the available sales data. This family typically

has more than one distribution because as noted above, there is only limited information available in practice, which is insufficient to specify a unique distribution. Thus, data limitation leads to an inherent uncertainty in specifying the underlying choice model. A popular approach to resolving uncertainty in model parameters is to use a 'robust' approach, where one plays out the worst-case scenario. Adopting this approach, we offer as the prediction the worst-case expected revenue possible assuming that the true model lies in the set of models consistent with the observed sales data. In other words, one can expect revenues that are at least as large as we predicted. We term our method the *robust approach.* A major advantage of this method from a practical standpoint is that it eliminates the need for any subjective judgements in choosing an appropriate parametric structure. This makes its practical implementation very simple: once the method is implemented, a practitioner can treat it as a black-box to which one feeds in historical sales data and the assortment to obtain a revenue prediction. The ability to use the same implementation is valuable especially when a retailer or a car manufacturer wants to make predictions at different stores or dealerships distributed across the country, even though choice behavior may change drastically with geographic location.

Two important questions arise at this juncture: (1) how good are the predictions in practice? and (2) how can the predictions be made in a computationally efficient manner (or quickly)? Before we provide answers to these questions, we make the following remark. Note that the choice model that is used to compute revenues (the one that yields the worst-case revenue) depends on the input sales data as well as the assortment for which revenues are being predicted. Dependence of the choice model on the data is desired. The dependence on the assortment however might seem unnatural. While that might be the case, we emphasize here that our ultimate goal is to predict revenues as accurately as possible, and not to learn the underlying choice model. Thus, the actual distribution used is irrelevant as long as the prediction is close the "true" value. Put differently, once the assortment is fixed, the revenue function maps each distribution to a scalar revenue prediction; thus, it is possible to have different distributions resulting in predictions that are close to the "true"

value, and as long as the distribution used results in a good prediction, we can safely ignore it. Therefore, we focus on the accuracy of predictions rather than the actual distribution used.

Quality of revenue predictions

In order to understand how good the revenue predictions are, we conduct empirical and theoretical analyses. In the empirical analysis, we gauged the practical value of our approach, both in terms of the absolute quality of the predictions produced, and also relative to using alternative parametric approaches. In the theoretical analysis, we provide guarantees on the revenue predictions relative to popular parametric approaches; we also obtain a characterization of the choice model used for revenue predictions to show that the model complexity indeed scales with the data that is input. We conducted two empirical studies:

(i) *Simulation Study:* We use the simulation study to demonstrate that the robust approach can effectively capture the structures of a number of different parametric models and produce good revenue predictions. The general setup in this study was as follows: We used a parametric model to generate synthetic transaction data. We then fed this synthetic transaction data to our revenue prediction procedure to predict expected revenues over a swathe of offer sets. Our experimental design permits us to compare these predictions to the corresponding 'ground truth'. The parametric families we considered included the multinomial logit (MNL), nested logit (NL), and mixture of multinomial logit (MMNL) models. In order to 'stress-test' our approach, we conducted experiments over a wide range of parameter regimes for these generative parametric choice models, including some that were fit to DVD sales data from Amazon.com. Even though the robust approach is agnostic to the underlying ground-truth, our results show that the predictions produced by our method in all the settings are remarkably close to the ground-truth. From this we conclude that the robust approach can successfully capture the underlying choice structure to produce accurate predictions just from limited sales transaction data. An important

implication of this conclusion is that even if the world did behave according to one of the popular parametric choice models, not much is lost in using the our robust approach.

(ii) *Empirical Study with Sales Data from a Major US Automaker:* In addition to the simulation study, we carried out an empirical case study with sales data from a major US automaker to test our approach in real-world scenarios. The purpose of the case study is two-fold: (1) to demonstrate how our setup can be applied with real-world data, and (2) to pit the robust method in a "horse-race" against the MNL and MMNL parametric families of models. For the case study, we used sales data collected daily at the dealership level over 2009 to 2010 for a range of small SUVs offered by a major US automaker for a dealership zone in the Midwest. We used a portion of this sales data as 'training' data. We made this data available to our robust approach, as well as to the fitting of an MNL model and an MMNL model. We tested the quality of 'conversion-rate' predictions (i.e. a prediction of the sales rate given the assortment of models on the lot) using the robust approach and the incumbent parametric approaches on the remainder of the data. We conducted a series of experiments by varying the amount of training data made available to the approaches. We conclude from our results that (a) the robust method improves on the accuracy of either of the parametric methods by about 20% (this is large) in all cases and (b) unlike the parametric models, the robust method is apparently not susceptible to under-fitting and over-fitting issues. In fact, we see that the performance of the MMNL model relative to the MNL model deteriorates as the amount of training data available decreases due to over-fitting.

In summary, both our empirical studies demonstrate the practical value of our method in producing quality revenue and sales predictions.

In addition to the empirical analysis, we provide insights into our approach through theoretical analysis. Specifically, the success of our empirical studies points to a characteristic of choice behavior observed in practice that enables accurate revenue

approximations from limited data. Our theoretical analysis sheds light on this characteristic. To understand this, first note that the accuracy of our revenue predictions depends on two factors: (a) the "complexity" of the underlying choice behavior, and (b) the "amount" of information in the available data. For a given complexity of underlying choice behavior, the more the data, the better the accuracy we expect. Similarly, for a given amount of information in the available data, the greater the complexity of the underlying choice behavior, the worse the accuracy we expect. The bounds we derive on the accuracy of our predictions confirm this intuition. We defer the details to later sections and describe the basic insight: robust approach with limited data produces accurate estimates if only a "few" products account for most of the cannibalization of the sales of each product. To elaborate on this, "similar" products in the offered assortment result in a cannibalization of (reduction in) the sales of each other; cannibalization of the sales of the notebooks offered by Apple Inc., by their iPads is a current example. For each product, if only a few products account for most of the cannibalization, then the robust approach can produce accurate predictions with limited data. For a given level of accuracy, there is an inverse relationship between the "amount" of data and the number of products where the cannibalization is concentrated. Using this insight, we derive general bounds on the relative error incurred by our approach. The bounds we derive can be computed using the partial data that is available and, hence, we can provide provable and computable guarantees for the estimates we produce in practice. We then specialize the error bounds to the MNL and the MMNL families of models. We observe that for the MNL case, the above characteristic of choice models that leads to small errors translates into the following MNL characteristic: sum of the weights of a few products must account for most of the sum of the weights of all the products in the assortment. Similarly, for the MMNL case, which is a mixture of MNL models, the characteristic roughly translates to having less heterogeneity in customer tastes, with the constituent MNL models possessing above MNL characteristic. Thus, the robust approach can be confidently used whenever these intuitive conditions are met.

Finally, we claimed above that an advantage of the nonparametric approach is that it scales the model complexity with the "amount" of data. We formally demonstrate this fact in this context. Specifically, we use *sparsity* or the support size of the distribution over preference lists as a measure of the complexity of the choice model and the dimension of the data vector as a measure of the "amount" of data; we shall show in later sections that transaction data can be represented as a vector in an appropriate dimensioned Euclidean space. With these metrics, we show that the sparsity of the choice model used by the robust method for revenue prediction scales with the dimension of the data vector.

Computation of revenue predictions

We now address the question of computing the predictions in a computationally efficient manner (quickly in practice). As shall be seen later, predicting revenues through the robust method boils down to solving a linear program (LP) with $(N+1)!$ variables. Since we only care about the optimal value (and not the optimal solution), we instead consider solving the dual problem. The number of the variables in the dual problem is of the order of the dimension of the data vector, and the number of constraints scales as $(N+1)!$. Now, solving the dual program efficiently is equivalent to finding an efficient separation oracle for the constraint polytope, whose structure depends on the type of the available data. We show that for certain types of data, it is indeed possible to obtain an efficient separation oracle – yielding an efficient method to predict revenues. Unfortunately, an efficient separating oracle does not exist for general types of data. In order to handle such cases, we propose two approaches:

1. First, we propose an extremely simple to implement approach that relies on sampling constraints in the dual that will, in general produce approximate solutions that are upper bounds to the optimal solution of our robust estimation problem. The approach of sampling constraints is an intuitive and a well-known heuristic; in fact, the recent theory developed by Calafiore and Campi [2005] provides a partial theoretical justification for the method. In our empirical analysis with both simulated and real-data we found the constraint sampling method to be

extremely simple to implement, while producing excellent approximations to the optimal value.

2. Next, we propose an approach that produces sequentially tighter relaxations of the constraint polytope. Although this approach is slightly more complex to implement, it yields sequentially tighter approximations to the robust problem. In certain special cases, this approach is provably efficient and optimal.

To summarize, we can either efficiently solve the LP in our robust method to optimality or obtain close approximations in practice. In addition, the methods we propose to solve the LP are easy to implement in practice.

In summary, our robust method presents a nonparametric way to use historical sales data to predict revenues or sales for assortments not been offered previously. Through extensive empirical studies, we establish the practical value of our revenue prediction method. Furthermore, the almost 20% improvements in prediction accuracy for conversion-rates obtained for data from a major US automaker demonstrate the huge impact our method can create in real-word applications. We also provide insights into the method through theoretical analysis. Specifically, we can provide guarantees for our revenue predictions. Specialization of our error bounds to specific parametric models provides insights into when we could expect the robust method to produce accurate predictions. Finally, we also formally establish that, as desired, the choice model used for revenue prediction scales with the "amount" of data that is available.

1.4.2 Assortment optimization

Next we consider the problem of static assortment optimization, where the goal is to find an assortment of products of size at most C from a larger universe of N products that maximizes the expected revenue. This is a classical decision problem that has straightforward applications to retail, online advertising, and revenue management. For instance, a retailer typically has a limited shelf capacity and is interested in offering a profit/sales maximizing assortment of products. In online advertising, the

capacity constraint would correspond to the availability of only a limited number of slots to show ads and a purchase would correspond to a click on an ad. The goal of the advertiser then would be to carefully select the assortment of ads to show to maximize profit. Finally, as mentioned above, the single-leg, multiple fare-class yield management problem, which is central to RM, reduces to solving a static assortment optimization problem at each stage.

For this problem, we assume access to a subroutine that can be called to get an estimate of the expected revenues of any assortment of products. At this stage, we shall leave the subroutine unspecified – it could either be the revenue prediction subroutine we propose (as described above) or some other existing subroutine. Given this, a straightforward solution to the decision problem is exhaustive search over all assortment of size at most C, which would require $O(N^C)$ calls to the revenue subroutine. In fact, in general, one cannot do better than exhaustive search. Whenever, N or C is small, exhaustive search is a reasonable solution; however, when N or C is large, an exhaustive search becomes computationally prohibitive in practice. Therefore, our goal is to propose an algorithm that can produce a "good" approximation to the optimal assortment with only a "few" calls to the revenue subroutine.

Since the worst-case requires exhaustive search, the design of a more efficient algorithm requires exploitation of problem-specific structure. In order to exploit structure, previous works considered this problem in the context of a specific parametric structure for the underlying choice model. The two parametric structures that have been considered are the multinomial logit (MNL) model and the mixture of MNL (MMNL) models. The problem is well-understood in the context of the MNL model.

Specifically, two main results have been obtained:

1. When the problem is un-capacitated, or equivalently when $C = N$, the optimal assortment belongs to one of the N assortments $S_1, S_2, \ldots, S_N$, where set S_i consists of the top-i products according to their prices/profits (see Talluri and van Ryzin [2004a]). Thus, the optimal assortment can be found by searching over at most N assortments.

2. In the capacitated case, Rusmevichientong et al. [2010a] specialize the idea proposed by Megiddo [1979] for solving combinatorial optimization problems with rational objectives to obtain an exact algorithm to find the optimal assortment. They show that the algorithm has a complexity of $O(NC)$. They also propose a more complicated implementation of the algorithm, which can reduce the complexity to $O(C \log N)$.

The first result constitutes a very intuitive characterization of the optimal assortment. The authors show that this characterization results in a practically pleasing nested policy for the single-leg, multiple fare-class yield management problem, where lower fare-classes are slowly closed as the remaining time and capacity decrease. The second result corresponds to the capacitated problem, which turns out to be a significantly harder problem than the un-capacitated one. The authors show that the intuitive characterization of the optimal assortment in the un-capacitated case does not extend to the capacitated case, requiring a more complicated algorithm. The characterization they obtain is more complicated and requires careful exploitation of the MNL structure. In the process, the authors derive several interesting and useful properties of the MNL structure.

The above two results are significant in that they provide a complete understanding of the static assortment optimization problem in the context of the MNL model. Unfortunately, it is not too far-fetched to say that beyond the MNL model, very little is known about the decision problem.

In the context of the MMNL model, it has been recently shown that even the un-capacitated decision problem is NP-complete even when there are only two components in the mixture (see Rusmevichientong et al. [2010b]). The authors also propose a PTAS for solving the un-capacitated decision problem. However, the running time of the PTAS scales exponentially with the number of mixture components.

Finally, in the context of the nested logit (NL) model, the authors Rusmevichientong et al. [2009] consider a slightly more complicated variant of the static assortment optimization problem, where they assume that each product has a fixed-cost of c_i and the goal is to find a revenue maximizing assortment such that the sum of the fixed

costs of the products in the assortment is at most C. Taking $c_i = 1$ for all products i yields the static assortment optimization problem. The authors show that this problem is NP-complete in the context of the NL model. Then, they propose a PTAS by reducing the decision problem to a sum-of-ratios problem.

In summary, we make two observations. First, except for the simple case of the MNL model, either not much is understood or the decision problem is known to be hard. Second, the algorithms that have been proposed to solve the decision problem (exactly or approximately) heavily exploit the sum-of-ratios structure of the objective; consequently, the algorithms – even without any guarantees – cannot be used with other choice models like the probit model or the mixture of MNL models with a continuous mixture.

Given these issues, our goal is to design a general optimization scheme that is (a) not tailored to specific parametric structures and (b) requires only a subroutine that gives revenue estimates for assortments. We make several contributions in relation to this problem.

First, we propose a general set-function optimization algorithm, which given a general function defined over sets, finds an estimate of the set (or assortment) where the function is maximized. This set-function optimization algorithm clearly applies to the static assortment optimization problem, thereby yielding the optimization scheme with the desired properties. Note that since we are considering a very general setup, there is not much structure to exploit. Hence, we adopt the greedy method – the general technique for designing heuristics for optimization problems. However, a naive greedy implementation algorithm fails even in the simple case of the MNL model. Specifically, consider the simpler un-capacitated decision problem. Here, a naive greedy implementation would start with the empty set and incrementally build the solution set by adding at each stage a product that results in the maximum increase in revenue; this process would terminate when addition of a product no longer results in an increase in revenue. It is easy to see that the naive implementation would succeed in solving the decision problem only if the optimal assortments exhibit a nesting property: the optimal assortment of size C_1 is a subset of the optimal assortment of

size C_2 whenever $C_1 < C_2$. Unfortunately, the nesting property does not hold even in the case of the MNL model. In order to overcome this issue, we allow for greedy "exchanges" in addition to greedy "additions." Particularly, at every stage, we allow a new product to be either added (which we call an "addition") to the solution set or replace an existing product (which we call an "exchange") in the solution set; the operation at each stage is chosen greedily. The termination condition now becomes an interesting question. As in the naive implementation, we could terminate the process when addition or exchange no longer results in an increase in revenue. However, since we never run out of products for exchanges, the algorithm may take an exponential (in the number of products) number of steps to terminate. We overcome this issue by introducing a control parameter that caps the number of times a product may be involved in exchanges. Calling that parameter b, we show that the algorithms calls the revenue subroutine $O(N^2bC^2)$ times for the capacitated problem. We thus obtain a general algorithm with the desired properties to solve the static assortment optimization problem.

Next, we establish the usefulness of the algorithm. For that, we first consider the case of the MNL model, where the decision problem is well-understood. Specifically, we assume that the underlying choice model is an instance of the MNL family and the revenue subroutine yields revenue estimates for assortments under the specific instance. We can show that the the algorithm we propose, when run with $b \geq C$, succeeds in finding the optimal assortment with $O(N^2C^3)$ calls to the revenue subroutine. Therefore, in the special case when the underlying choice model is the MNL model, our algorithm captures what is already known. It also provides a simpler alternative to the more complicated algorithm proposed by Rusmevichientong et al. [2010a]. We also consider the case when noise corrupts the available revenue estimates – a common practical issue. In this case, we show that our algorithm is robust to errors in the revenue estimates produced by the subroutine. Particularly, if the underlying choice model is the MNL model and the revenue estimate produced by the subroutine may not be exact but within a factor $1 - \varepsilon$ of the true value, then we can show that our algorithm finds an estimate of the optimal assortment with revenue that is within

$1 - f(\varepsilon)$ of the optimal value; here $f(\varepsilon)$ goes to zero with ε and also depends on C and the parameters of the underlying model. In summary, our theoretical analysis shows that our algorithm finds the exact optimal solution in the noiseless case or a solution with provable guarantees in the noisy case, whenever the underlying choice model is the MNL model. In this sense, our results subsume what is already known in the context of the MNL model.

In the context of the more complicated models like the nested logit (NL) and the mixtures of MNL models, the decision problem is provably hard. As discussed above, even obtaining a PTAS can be very complicated and requires careful exploitation of the structure. We however believe that it is possible to obtain "good" approximations to the optimal assortments in practice.

1.4.3 Learning choice models

In the problems discussed so far, our ultimate goal was to use a choice model to make a decision. As a result, we avoided the unnecessary indirection of actually learning the underlying choice model and developed methods to directly solve the decision problem. However, as elaborated below, explicit learning of choice models is required in many important applications.

For instance, consider the problem of 'customer segmentation.' It is central to several important applications and involves segmenting the customer population into groups of individuals that have similar preferences. A statement to the extent "your customers are mainly of 3 types, and their preferences are described by these preference lists" is of high value in these contexts. One way to usefully segment customers is to learn a choice model over K preference lists and then segment the customer population into K classes with the preferences of each class described by one of the K learned preference lists. Such segmentation is especially crucial to applications that need effective targeting of resources. The classical application that heavily uses customer segmentation is marketing, where it has long since been known that the marketing strategy needs to be effectively targeted to the specific customer type. Interestingly, a non-traditional application area that also relies on customer segmen-

tation is the area of the recently popular recommendation/discovery systems, where the recommendations (be it movies, books, or news stories) need to be tailored in a useful way to the specific customer types.

In addition to applications related to OM/RM, another broad class of problems where learning distributions over preference lists becomes important is 'rank aggregation'. As mentioned above, this is an important problem that arises in various contexts like web-search, polling, betting, and elections, in which the goal essentially is to come up with a final ranking given some partial preference information. The rank-aggregation problem has been studied extensively in the area of social choice theory, where extensive work has been done to determine the "right" final ranking given access to the entire distribution over rankings. Of course, in most practical applications, such a distribution is not readily available. What is readily available though is partial information about the distribution: for instance, in the context of web-search, clicks give information about which document is preferred from a set of documents that were shown; similarly, in the context of polling, one may have access to pairwise comparison information Given this, a reasonable approach to aggregating rankings is to learn a distribution over rankings that captures the underlying choice structure from marginal preference information and use any of the several methods developed in the social choice literature for aggregation.

Finally, there is a host of other applications in which distributions over rankings are compressed in order to store efficiently[5] by retaining only partial information (typically in the form a subset of Fourier coefficients; see Huang et al. [2008] and Kondor et al. [2007]). In these cases, the recovery of the compressed distribution essentially boils down to learning a distribution over permutations from partial information.

In summary, there is a wide-range of important applications that need learning of the underlying choice model from available marginal information. Now, given marginal information, we can identify a family of choice models that are consistent with the given information as explained above; the family is almost certainly not a

[5]Such compression is indeed a necessity given that distributions over rankings have a factorial (in N) blow-up.

singleton because the data is insufficient to specify a unique distribution. Therefore, the problem of learning the choice model reduces to the problem of finding an appropriate criterion to select one of the models consistent with the available data. Now, a popular statistical criterion for model selection that has been extensively used in many contexts is the criterion of *parsimony*, which encourages the selection of the 'most parsimonious' model that is consistent with the data.

The criterion of parsimony is justified in many ways. Philosophically speaking, this criterion is consistent with the *Occam's razor* philosophy, which roughly stated, suggests that under the absence of additional information, one should tend toward 'simpler' theories. Statistically speaking, parsimony is born out of the need not to over-fit. Finally, operationally speaking, parsimony is desired because parsimonious models are easier to handle in practice – both computationally and otherwise. Of course, parsimony is nuanced idea and it is not straightforward to operationalize the criterion. In parametric models, parsimony has often been translated into parsimony of parameters; for instance, an MNL model with N parameters can be considered 'more parsimonious' than an exponential family with N^2 parameters. In the nonparametric case, however, the sparsity (or the support size) of the distribution becomes a natural candidate for a measure of parsimony of the model. In addition, as described above in the context of choice models used for revenue predictions, sparsity is also naturally born out of the fact that only marginal information is available; more precisely, a distribution of sparsity no more than[6] $(N-1)^2+1$ is needed to describe the first-order information, which captures the probability that i is ranked at position r for all i and r. Finally, sparse models have found immense success (in both theory and practice) in the area of *compressive sensing*, which has gained recent popularity in the areas of signal processing, coding theory, and streaming algorithms (see Donoho [2006], Candes et al. [2006b,a]).

Given the considerations above, we propose to recover the underlying choice model by identifying the sparsest distribution that is consistent with the available data.

[6]This statement follows from Caratheodory's theorem that states that every point in a convex polytope of dimension d can be decomposed into a convex combination of at most $d+1$ extreme points, and the fact that doubly stochastic matrices have a dimension of $(N-1)^2$.

From an operational perspective, two main questions arise at this point: (1) how does one find the sparsest consistent distribution in an efficient manner? and (2) how "good" are sparse models in practice? In addition, from a theoretical standpoint, the question about the discriminative power of the sparsest-fit criterion arises. More precisely, it is useful to obtain a description of the the family of models that can be *identified* as the unique sparsest models consistent with the marginal information. Intuitively, we expect a characterization of the form: if the underlying model is "sparse enough", then it can be identified by as the unique sparsest solution consistent with the marginal information; the sparsity bound of course should depend on the dimension of the data, which can be treated as a measure of the "amount" of information that is available. Next, we describe the contributions we make to answering these questions.

First, we consider the issue of efficient learning of sparse models. Specifically, we consider two settings: the noiseless and the noisy settings. In the noiseless setting, we assume we have access to marginal information about the underlying distribution that is not corrupted by noise and seek to find the sparsest distribution that is exactly consistent with the marginal information. In the noisy setting – which is more practical – we assume that the marginal information is corrupted by noise and seek to find the sparsest distribution that is within ε of the marginal information under an appropriate norm. We also present our theoretical results on the discriminative power of the sparsest-fit criterion and why convex relaxations have no bite in the section on noiseless setting. After we describe our results for efficient learning of sparse models, we describe our results of an empirical study.

Noiseless setting

In the noiseless setting, we assume we have access to exact marginal information about the underlying distribution λ in the form of an $m \times 1$ vector y that is related to λ through a linear transform. Through a slight abuse of notation, we write $y = A\lambda$, where we think of the choice model λ as an $N! \times 1$ vector and A is an $m \times N!$ matrix with entries in the set $\{0, 1\}$. Most types of marginal information that are both

theoretically and practically interesting can be cast in this form.

Under this setting, our goal is to determine the sparsest distribution that is consistent with the given marginal information y. In other words, our interest is in solving the problem:

$$\begin{aligned} \underset{\lambda}{\text{minimize}} \quad & \|\lambda\|_0 \\ \text{subject to} \quad & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{aligned} \tag{1.1}$$

Two interesting questions arise at this point: (1) can we characterize the family of models that can be identified as the unique optimal solutions to the program (1.1)? and (2) how to solve the program in (1.1) efficiently? We answer both these questions by identifying a family of models $\mathcal{F}$, which can be *efficiently* recovered from marginal information as the *unique optimal solutions* to the program in (1.1).

Before we describe the family $\mathcal{F}$ of models, we quickly address the issue of solving (1.1) through convex relaxation i.e., by solving:

$$\begin{aligned} \underset{\lambda}{\text{minimize}} \quad & \|\lambda\|_1 \\ \text{subject to} \quad & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0, \end{aligned} \tag{1.2}$$

where $\|\lambda\|_1 \stackrel{\text{def}}{=} \sum_\sigma |\lambda(\sigma)|$. It is clear from the constraints in (1.2) that all feasible models have the same ℓ_1 norm (equal to 1). Moreover, it is hardly the case that the set of consistent models is a singleton. In fact, we can prove that for a marginal data vector y that comes from a randomly generated distribution of sparsity $K \geq 2$, the set of consistent distributions is not a singleton with a high probability (see Theorem 8). Thus, the ℓ_1 criterion is not resulting in any reduction of the family of consistent distributions, and we cannot guarantee the optimality of an arbitrarily selected model from the set of feasible models. In fact, as described in the the noisy setting (see discussion after Theorem 14), selecting an arbitrary consistent distribution can be

highly suboptimal; specifically, we can show that for the first-oder marginal information, selecting an arbitrary basic feasible solution that minimizes the ℓ_1 norm can have a sparsity of $\Theta(N^2)$, whereas the sparsest distribution can be shown to have a sparsity of $O(N)$.

Coming back to the family of models $\mathcal{F}$, we define it as consisting all distributions that belong to the signature family and satisfy linear independence conditions. We define each of these conditions in turn. The precise definition of the signature family of distributions depends on the type of data that is available. The definition is easiest to state for the first-order information, as given below:

> **Signature family.** A distribution (choice model) λ is said to belong the signature family if for each permutation σ that is in the support (i.e., $\lambda(\sigma) > 0$) there exist an pair i, r such that $\sigma(i) = r$ and $\sigma'(i) \neq r$ for any permutation σ' in the support. Equivalently, for every permutation σ in the support of λ, there exists a pair i, r such that σ ranks i at position r, but no other permutation in the support ranks i at position r.

In other words, whenever a distribution belongs to the signature family, every permutation σ in the support has a unique 'signature', which corresponds to the pair i, r such that $\sigma(i) = r$ and $\sigma'(i) \neq r$, for every other permutation σ' in the support. Given this, we call a distribution belonging to the signature family as satisfying the signature condition. The linear independence condition is precisely stated in Section 5.4 of Chapter 5 and roughly requires that the all the subset sums of the set of distribution values in the support $\{\lambda(\sigma_1), \lambda(\sigma_2), \ldots, \lambda(\sigma_K)\}$ are distinct. Given this, we can show that whenever there exists a model λ in $\mathcal{F}$ such that $y = A\lambda$, then λ is also the *unique sparsest* model consistent with y (see Theorem 7); Thus, the family $\mathcal{F}$ describes a set of sufficient conditions for the underlying model λ that guarantee that it can be recovered as the unique sparsest model consistent with the marginal data. These conditions are also essentially necessary because we can exhibit counter-examples of models that don't satisfy the conditions and are not the sparsest models consistent with the marginal data. Therefore, as far as identification through

the sparsest criterion goes, the family of models in $\mathcal{F}$ is the best we can do.

In addition, we can show that the models in $\mathcal{F}$ can be recovered efficiently. Specifically, we propose what we call the sparsest-fit algorithm in order to efficiently determine the sparsest distribution whenever the data vector y is generated from a model belonging to $\mathcal{F}$. When run with the marginal information y, the sparsest-fit algorithm terminates with either (a) the sparsest distribution if the data vector y is generated by a model in $\mathcal{F}$ or (b) a certificate that y is not generated by a model in $\mathcal{F}$. If the sparsest distribution belongs to $\mathcal{F}$ and has a sparsity of K, the the algorithm has a worst-case complexity of $O(m \log m + m2^K)$. In fact, under a random generative model, to be described shortly, the algorithm can be shown to have a complexity of $O(\exp(\log^2 m))$ with a high probability. Compare this to the worst-case complexity of $O(\exp(\Theta(KN \log N)))$, and it is evident that restriction to the family of models in $\mathcal{F}$ has resulted in a huge reduction in the computational complexity.

A natural concern at this point is how restrictive the family of distributions $\mathcal{F}$ is. In order to address this issue, we consider the following random generative model for distributions of with a support size K:

> **Random model.** Given $K \in \mathbb{Z}_+$ and an interval $\mathscr{C} = [a, b]$, $0 < a < b$, a random mode λ with sparsity K and values in $\mathscr{C}$ is generated as follows: choose K permutations independently and uniformly at random [7], say $\sigma_1, \ldots, \sigma_K$; select K values from $\mathscr{C}$ uniformly at random, say $p_1, \ldots, p_K$; then model λ is defined as
>
> $$\lambda(\sigma) = \begin{cases} p_i & \text{if} \quad \sigma = \sigma_i, \ 1 \leq i \leq K \\ 0 & \text{otherwise.} \end{cases}$$

Then we establish that as long as the sparsity K scales in a certain way with N, say $K = O(K(N))$, a choice model generated according to the random model belongs to the family $\mathcal{F}$ with a high probability as $N \to \infty$. The precise scaling of $K(N)$ depends on the type of marginal information that is available. We derive the precise

[7] Throughout, we will assume that the random selection is done *with* replacement.

scaling for a variety of important types of partial information. We state below our results for only three types of partial information; see Section 5.4 of Chapter 5 for more details. Specifically, we can show that

1. $K(N) = \log N$ for comparison information: probability that alternative i is preferred to alternative j for every pair of alternatives i, j;

2. $K(N) = \sqrt{N}$ for top-set information: comparison information and the probability that i is ranked first for all alternatives i.

3. $K(N) = N \log N$ for first-order information: probability that i is ranked at position r for all alternatives i and positions r.

See Theorems 9 for precise statements of the results. We establish similar scalings for other types of partial information. The following remarks are in order. It is easy to see that the family of choice models that can be recovered efficiently can be characterized in terms of their sparsity. As desired, the sparsity and hence the complexity of the models as measured in terms of sparsity, is scaling with the "amount" of information in the data. In this sense, our results formalize the intuitive notion that first-order information has "more information" than top-set information has "more information" than comparison information. Finally, it is clear that the family $\mathcal{F}$ describes a "large" family of models satisfying a certain sparsity constraint.

Noisy setting

In the noisy setting, we consider the case when the marginal information available to us may be corrupted by noise. Specifically, we assume that $\|y - A\lambda\| \leq \varepsilon$, where λ is the underlying choice model, y is the data vector that is corrupted by noise, and $\varepsilon > 0$ is a measure of the magnitude of noise. Given this, our goal is to find the sparsest model that is within ε of the data vector y. More precisely, our goal is to

solve the following program:

$$\begin{aligned} \underset{\lambda}{\text{minimize}} \quad & \|\lambda\|_0 \\ \text{subject to} \quad & \|A\lambda - y\| \leq \varepsilon, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{aligned} \tag{1.3}$$

Given the program in (5.4), a natural question that arises is whether we can solve it efficiently. Interestingly, a more fundamental question, which informs the question of efficient solvability of (5.4), is "how sparse can the sparsest solution be?" To elaborate further, first observe that for any choice model μ, the first-order marginal matrix $M(\mu)$ is doubly stochastic. Thus, it follows from $Y = M(\lambda) + \eta$ and $\|\eta\|_2 \leq \delta$ that solving the program in (5.4) is essentially equivalent to determining the convex decompositions of all doubly stochastic matrices that are within a ball of radius $\delta+\varepsilon$ of $M(\lambda)$ and choosing the sparsest convex decomposition. Now, it follows from Birkhoff-von Neumann's celebrated result (see Birkhoff [1946] and von Neumann [1953]) that a doubly stochastic matrix belongs to an $(N-1)^2$ dimensional polytope with the permutation matrices as the extreme points. Therefore Caratheodory's theorem tells us that it is possible to find a convex decomposition of any doubly stochastic matrix with at most $(N-1)^2+1$ extreme points, which in turn implies that the sparsest model consistent with the observations has a support of at most $(N-1)^2+1 = \Theta(N^2)$. We raise the following natural question at this point: given *any* doubly stochastic matrix M, does there exist a choice model λ with sparsity significantly smaller than $\Theta(N^2)$ such that $\|M(\lambda) - M\|_2 \leq \varepsilon$.

Geometrically speaking, the question above translates to: given a ball of radius ε around M, is there a subspace spanned by K extreme points that intersects the ball, for *any* double stochastic matrix M and some K that is significantly smaller than $\Theta(N^2)$? Note that the answer to this question can give us an indication of whether a straightforward approach (such as convex relaxation) can produce good approximations to the sparsest solution. In particular, it is possible that for a general doubly stochastic matrix M, there do not exist models λ with sparsity significantly

smaller than $\Theta(N^2)$ such that $\|M(\lambda) - M\|_2 \leq \varepsilon$. In other words, it is possible that for a general M, there is no subspace of K extreme points that intersects the ε ball around M for $K \ll N^2$. If such is the case, then convex relaxations – which result in a models of sparsity $O(N^2)$ – can produce solutions close to the optimal, at least for general M. In that case, we could attempt to characterize the class of matrices M for which we can solve (5.4) through convex relaxation. If, on the other hand, we can prove that for a general M, the sparsest model can have sparsity significantly smaller than $\Theta(N^2)$, then using straightforward approaches like convex relaxations can result in a highly suboptimal solution. Interestingly, we can prove a result that shows that the "right" scaling in fact is $O(N/\varepsilon^2)$ (see Theorem 14), showing that the sparsest solution indeed has a sparsity that is significantly smaller than $\Theta(N^2)$. We thus expect to go beyond convex relaxations in order to efficiently recover the sparsest solution, leading us to our next question about efficient recovery of the sparsest solution.

As mentioned above, a brutre-force exhaustive search would result in a computational complexity of $\exp(\Theta(KN \log N))$. In order to obtain a computationally efficient solution, we restrict our search to the signature family of distributions. With this restriction, we can recast the program in (1.3) into an integer program (IP) with m^2 variables and polynomially (in N) constraints (we refrain from describing the IP here because that requires additional notation; refer to Section 5.8 of Chapter 5 for the details). While the restriction to the signature family has allowed us to reduce the IP in (1.3) with $N!$ variables into an IP with m^2 variables, we are unable to guarantee any improvements on the worst-case running time. However, the IP formulation with m^2 constraints affords us the ability to devise a multiplicative weight update heuristic based on the Plotkin-Shmoys-Tardos (PST) (see Plotkin et al. [1991]) framework. The heuristic is described in detail in Section 5.8 of Chapter 5. We are able to to provide the following guarantees for the heuristic: Assuming that we have access to first-order partial information and there exists a model λ^* of sparsity K in the signature family such that $\|y - A\lambda^*\| \leq \varepsilon$, we can guarantee that our multiplicative weight update heuristic will have a computational complexity of $O(\exp(\Theta(K \log N)))$ and will find a model $\hat{\lambda}$ such that $\|y - A\hat{\lambda}\| < \varepsilon$ and the sparsity of $\hat{\lambda}$ is $O(K \log N)$ (see

Theorem 15).

Several remarks are in order. Note that for the case of the first-order information, we have managed to reduce the computational complexity from $\exp(\Theta(KN \log N))$ to $\exp(\Theta(K \log N))$ at the cost of a factor $\log N$ in the sparsity; although this requires existence of a distribution of sparsity K in the signature family that is within ε of the data vector y, as will be seen shortly, we can guarantee the existence of such a distribution. Moreover, note that the distribution $\hat{\lambda}$ produced by our heuristic may not belong to the signature family. This in itself is not a matter of concern because restriction to the signature family was necessitated due to computational issues; in the case of the heuristic, the signature family plays an indirect role through λ^*. Finally, our guarantee applies only to the first-order information. The heuristic we propose can certainly be applied with other types of partial information. Conditional on the existence of a model λ^* with sparsity K, in the signature family such that $\|y - A\lambda^*\| < \varepsilon$, we can guarantee that a model with sparsity $O(K \log N)$ will be found. However, the computational complexity of the heuristic depends on the existence of an efficient representation of certain polytopes. For the first-order information, the polytope of interest is the Birkhoff polytope, which a widely known efficient representation. This is unfortunately not the case with for other types of partial information. Nevertheless, the heuristic could be used with "appropriate" relaxations.

For the first-order information, the result we stated above assumed the existence of a model λ^* such that $\|y - A\lambda^*\| < \varepsilon$. We justify this assumption by showing that the signature family is "dense". Specifically, we show that whenever y is a noisy observation of $A\lambda$ and λ is either the MNL model or the maximum-entropy distribution (with appropriate conditions on the parameters), then for N large enough, there exists a model λ^* in the signature family such that $\|y - A\lambda^*\| < 2\varepsilon$ and the sparsity of λ^* is $O(N/\varepsilon^2)$ (see Theorem 16). Thus, even with the restriction to the signature family, the sparsity scaling is still $O(N/\varepsilon^2)$ implying that we are not losing much in terms of sparsity by the restriction to the signature family.

Although we can provide impressive guarantees for our multiplicative weights update heuristic – which is trying to solve essentially a very hard combinatorial op-

timization problem – its practical applicability is somewhat limited. This is because our heuristic is not provably efficient for any type of partial information except the first-order information; even for the first-order information, the implementation of the heuristic is slightly involved requiring solving a host of LPs. Due to these considerations, we generalize our sparsest-fit algorithm described above to what we call the greedy sparsest-fit algorithm so that it *always* outputs a valid distribution. The greedy sparsest-fit algorithm reduces to the sparsest-fit algorithm whenever the underlying choice model belongs to the signature family and there is no noise. Similar to the sparsest-fit algorithm, for the first-order information and the types of information obtained from sales transactions, the greedy sparsest-fit algorithm it has a running time complexity of $O(m \log m + m2^K)$, where K is the sparsity of the solution obtained. Moreover, the greedy sparsest-fit algorithm can be used with a host of different types of partial information. We demonstrate the effectiveness of the greedy sparsest-fit algorithm through a simulation study, whose results are summarized above.

Empirical study

In order to demonstrate that sparse models are effective in capturing the underlying structure of the problem, we conducted an empirical study with the well-known APA (American Psychological Association) dataset that was first used by Diaconis [1989]. The APA dataset comprises the ballots collected for electing the president of APA. Each member expresses her/his preferences by rank ordering the candidates contesting the election. In the year under consideration, there were five candidates contesting the election and a total of 5,738 votes that were complete rankings. This information yields a distribution mapping each permutation to the fraction voters who vote for it. Given all the votes, the winning candidate is determined using the *Hare* system (see Fishburn and Brams [1983] for details about the Hare system).

A common issue in such election systems is that it is a difficult cognitive task for voters to rank order all the candidates even if the number of candidates is only five. This, for example, is evidenced by the fact out of more than 15,000 ballots cast in the APA election, only 5,738 of them are complete.One way to overcome

this issue is to design an election system that collects only partial information from members. If we collect only partial information, then we are faced with the issue of using this partial information to determine the winner. Having complete distribution affords us the flexibility of using any of the several rank aggregation systems out there. In order to retain this flexibility, our approach is to fit a sparse distribution to the partial information and use the distribution with the favorite rank aggregation system to determine the "winning" ranking. Such an approach would be of value if the sparse distribution can capture the underlying structural information of the problem at hand. Therefore, with the aim to understand the type of structure sparse models can capture, we used the first-order marginal matrix M (5×5 matrix with entry i, j equal to the probability that i is ranked at position r) that is obtained from 5,738 complete rankings with the greedy sparsest-fit algorithm to determine a sparse model of sparsity 6. Note that the support size of 6 is a significant reduction from the full support size of $5! = 120$ of the underlying distribution. The average relative error in the approximation of M by the first-order marginals $\hat{M}$ of the sparse model is less than 0.075, where the relative error of entry i, j is defined as $|M_{ij} - \hat{M}_{ij}|/M_{ij}$. The main conclusion we can draw from the small relative error we obtained is that the greedy sparsest-fit algorithm can successfully find sparse models that are a good fit to the data in interesting practical cases.

In addition, we used the *Hare* system to determine the winning ranking. When applied to both the distributions, the winning ranking obtained from the original distribution was 23145 and from the sparse distribution was 13245, where the candidates are ordered from the first rank to the last. Given these outcomes, we make the following observations. As is not surprising, the rankings obtained are clearly different, but the sparse model manages to capture the ranking of the candidates 4 and 5. The sparse model declares candidate 1 as the winner, whereas the original ranking declared candidate 2 the winner. We argue that declaring candidate 1 as the winner is not totally unreasonable, and in fact arguably the better choice. Specifically,it was observed in Diaconis [1989] that the data shows that candidate 1 has strongest 'second' position vote and the least "hate" vote or last position vote

(see Table 5.1. Moreover, as observed in Diaconis [1989], the APA has two groups of voters with different political inclinations: *academics* and *clinicians*. The authors conclude that candidates $\{2,3\}$ and $\{4,5\}$ belong to different groups with candidate 1 somewhat neutral. From these two observations, one could argue that candidate 1 is a better choice than candidate 2. Furthermore, it is from the winning ranking of $\hat{\lambda}$ the grouping of candidates 2,3 and 4,5 are evident, leading us to conclude that the sparse model $\hat{\lambda}$ is able to capture the partitioning of the APA into two groups.

In conclusion, our empirical study shows that the greedy sparsest-fit algorithm can find sparse distributions that not only fit the given data well, but also manage to capture the underlying structure from only marginal information.

1.5 Organization of the thesis

The rest of the thesis is organized as follows. Chapter 2 deals with background and provides a comprehensive overview of the rich history of choice models. Due to the fundamental nature of choice models, the literature on choice models goes back to at least the 1920s and spans several areas. Chapter 2 stitches together the diverse modeling approaches across the different areas through two common themes.

Chapters 3 and 4 are devoted to solving decision problems using choice models. Chapter 3 discusses the problem of revenue predictions and Chapter 4 presents our optimization algorithm.

Chapter 5 deals with question of learning sparse choice models from historical preference data. Furthermore, in order to provide a better flow, some of the implementation details and experimental setup details have been moved to the appendix.

Finally, Chapter 6 presents the conclusions of the thesis and discusses future directions.

Chapter 2

Overview of choice models

This chapter is devoted to a brief overview of popular parametric choice models that have been extensively studied in the literature and successfully applied in several practical settings. As the name suggests, choice models attempt to capture the very basic act of human free will: choice. It is then not surprising that the study of choice models has spanned diverse fields ranging from psychology, sociology, statistics, transportation, marketing, and operations management dating back to at least the 1920s. Although choice has been the basic building block, the ultimate goal for the study of choice models has been different in different fields; consequently, the modeling treatments in different fields have emphasized different aspects of choice behavior depending on the ultimate goal. Due to this diversity of treatments and applications, developing a comprehensive summary of the modeling approaches is an almost hopeless task. The common theme of course is that all approaches eventually induce a distribution over the permutations, but such a characterization is too general. Fortunately, one can obtain a reasonable condensation of the varied approaches. Specifically, two broad frameworks standout from the vast collection of seemingly different modeling approaches. Historically, most of the initial modeling approaches to choice were ad-hoc and driven by the specific application at hand. In the process of formalizing the modeling approaches, two dominant frameworks emerged and most of the initial ad-hoc models were later shown to be consistent with these frameworks. The frameworks yield two broad families of choice models: the *Random Utility Max-*

imization (RUM) family and the *exponential* family. The RUM family of models has beginnings in the psychological literature, where the goal was to develop models that were grounded in more basic processes. The exponential family, on the other hand, has origins in statistics and emphasizes data availability rather than basic processes driving choice.

To elaborate further, the RUM family of models assume that choices are driven by the various measurable attributes of the alternatives and the individual. Specifically, each indiviudal is assumed to assign to each alternative a utility value that is a function of the various attributes that influence choice; once the utilities are assigned, the individual chooses the alternative with the maximum utility. This is essentially the view proposed in the classical utility maximization framework; the only difference is that in the classical setting, the assigned utilities are assumed to be real values, whereas in the RUM family, utilities are modeled as random variables. Introducing randomness into utilities clearly results in a natural generalization of the classical utility maximization framework. Interestingly, randomness was initially necessiated by the empirical need to fit the model to real-world data. As McFadden [2000] explains, it was soon realized that randomness in utilities is an inherent component that naturally arises because of heterogeneity in user preferences. More precisely, when modeling a group of "similar" individuals, the differences in their "tastes" make them assign "slightly" different utility values to the same alternatives, which can be captured through randomness. Moreover, even when modeling just one individual, the individual may assign different utilities to the same alternative on different choice occasions because of a variation in the attributes (like mood) known to the individual but not the modeler. With utilities modeled as random variables, different choices of parametric distributions for the random utilities give rise to different members of the RUM family. Once a parametric distribution is selected, choices are made as follows: in each decision instance, the individual samples utilities according to the selected distribution and chooses the alternative with the maximum of the sampled utilties. This model clearly induces an appropriate distribution over the space of permutations. The choice of the parametric distribution is typically motivated by

computational aspects and the application at hand. RUM family of models has been found to be very rich and several ad-hoc models grounded in basic processes have later been shown to be special instances of the RUM family.

The exponential family of models naturally arises in contexts where it is relatively easier to collect summary statistics (or marginal information) of the full set of data. Such a situation is common when dealing with ranked data. For instance, consider a ranked election setting where N candidates are running for an election and each vote is a ranking of the N candidates. In this case, instead of collecting data about the fraction of voters who vote for each ranking, it is definitely easier to collect (or estimate) what are called first-order summary statistics: the fraction of people who rank candidate i to position r. Given only summary statistics, it is natural to consider families of models that are completely described by such statistics. Interestingly, the classical Fisher-Koopman-Pitman-Darmois (see Koopman [1936]) theorem establishes that the exponential family of models is the richest family of models that are completely described by a given set of summary statistics. Equivalently, the given summary statistics are also the sufficient statistics for the distribution. The exponential family of models is very rich and has been extensively studied not only in the context of ranked data, but also in the context of both discrete and continuous random vectors. Especially in the ranked data setting where the underlying distribution has $N!$ degrees of freedom and only summary statistics are naturally available in practice, exponentially families provide a rigorous way to scale model complexity with the data. In this sense, they provide a nonparametric way to model choice data. However, especially with ranked data, exponential families are quite complicated to deal with: learning the parameters of an exponential family is usually computationally challenging (see Crain [1976], Beran [1979], and Wainwright and Jordan [2008]); in addition, they are present computational challenges[1] just to compute choice probabilities even with all the parameters given.

Next, we provide more details about the two broad families of models. We start

[1]The computational challenge arises because computing the choice probabilities requires computing the marginals of a distribution that has an $N!$ support size (see Wainwright and Jordan [2008]).

with a brief historical account of methodological developments in choice theory. After that we provide an overview of the RUM framework and go into the details of some important special instances like the multinomial logit (MNL), nested logit (NL), cross-nested logit (CNL), and mixtures of MNL (MMNL) models. We then formally describe the exponential family of models and discuss the max-entropy distributions and distance-based models. Our treatment below is self-contained, however, not exhaustive; where required, we provide interested readers with pointers to relevant literature.

2.1 Historical account

We provide below a very brief historical account of choice models, without attempting to be exhaustive. For a more complete overview we refer interested readers to Diaconis [1988] and McFadden [2000]. Some of the origins of modern choice models can be traced back to psychological literature. In 1927, Thurstone [1927] introduced his *law of comparative judgement*, which models the comparison of imperfectly perceived intensities of physical stimuli like weights of objects, temperatures of objects, loudness of tones, etc. To be concrete, consider a simple experiment as described in Diaconis [1988], where individuals are asked to rank a set of tones according to their perceived loudness. It is an established empirical fact that due to various imperfections, the same individual on different occasions gives different answers. In order to account for this, Thurstone introduced randomness and modeled the perceived intensity of each stimulus i as $\mu_i + \xi_i$, where μ_i is unobservable "true" stimulus and ξ_i are i.i.d random variables with normal distribution; consequently, on each choice occasion an individual perceives tone i to be louder than tone j if $\mu_i + \xi_i$ is larger than $\mu_j + \xi_j$. Thurstone derived the probability that tone i is perceived to be louder than tone j, and showed that it has a form that we now call binomial probit. This model has then be shown to be a good fit for several tasks. Luce and Suppes [1965] provides a good review of the experimental validation of Thurstone's model. It is easy to see that Thurstone's model is a member of the RUM family where utilities are the intensities of various

stimuli. Marschak [1959] and Block and Marschak [1960] generalized Thurstone's model to random utility maximization over multiple alternatives.

In an alternate line of work, Luce [1959] published an axiomatic approach to choice modeling. Loosely speaking, Luce postulated that the ratio of choice probabilities of two alternatives is independent of the choice set that includes them. This is called the Independence of Irrelevant Alternatives (IIA) property. More precisely, Luce's axiom states that given any choice set $\mathcal{M}$, a subset $\mathcal{A} \subset \mathcal{M}$, and an alternative $i \in \mathcal{A}$, $\mathbb{P}(i|\mathcal{M}) = \mathbb{P}(i|\mathcal{A})\mathbb{P}(\mathcal{A}|\mathcal{M})$, where $\mathbb{P}(\mathcal{A}|\mathcal{M})$ denotes the probability that subset $\mathcal{A}$ is chosen from the assortment $\mathcal{M}$. If this axiom holds, Luce showed that each alternative i can be assigned a positive weight w_i such that the probability of the choice of alternative i from any choice set is proportional to w_i. This model induces a distribution over ranked lists through the following generative model: the first-ranked alternative is chosen from all the alternatives with probability proportional to its weights, say w_i; once i is chosen, the second-ranked alternative is chosen from the remaining alternatives $\mathcal{N} \setminus \{i\}$ again with probability that is proportional to its weight, and so on. Specifically, the probability of choosing any permutation σ is

$$\mathbb{P}_w(\sigma) = \prod_{r=1}^{N} \frac{w_{\sigma^{-1}(r)}}{w_{\sigma^{-1}(r+1)} + w_{\sigma^{-1}(r+2)} + \cdots + w_{\sigma^{-1}(N)}},$$

where $\sigma^{-1}(r)$ denotes the product that is ranked at position r by σ.

Empirical work demonstrated that with the appropriate choice of parameters, the choice probabilities predicted by Thurstone's model and Luce's model were very close Luce [1977]. The first major insight into understanding the relationship between the two models came due to Block and Marschak [1960], who showed that the Luce model is equivalent to the Thurstone model when the random variables are assumed to be independent double exponential or Type I Extreme Valued; E. Holman and A. A. J. Marley (reported in Luce and Suppes [1965]) gave a more direct proof of this fact. It was later established by McFadden [1973] – and independently by Yellott [1977] under less restrictive conditions – that the double exponential distribution condition is indeed necessary. Specifically, it was established that whenever $N \geq 3$, a

necessary condition for the Thurstone model with independent random variables to be equivalent to the Luce model is that the random variables are i.i.d double exponential. Interestingly, as Yellott [1977] shows, the double exponential distribution condition is not necessary when $N = 2$. Thus, an equivalence was established between two modeling approaches that were perceived to be different. It should be noted here that the equivalence is by no means obvious and definitely not easy to prove.

As Luce [1977] points out, an interesting outcome of Yellott [1977] work is that it brought to prominence the double exponential distribution or the extreme value distribution, which had arisen in statistics many years ago (see Fisher and Tippett [1928], Gumbel [1966]). The extreme value distribution arises because of its appealing *max-stable* property that the maximum of two independent extreme value random variables with the same scale parameter is also extreme value with the same scale factor. Interestingly, the following asymptotic property has also been established: the maximum of N i.i.d random variables converges in distribution to the extreme value distribution as $N \to \infty$. This result can be thought of as the counterpart of the Central Limit Theorem with sum replaced by maximum. With this result the double exponential distribution can be justified by assuming that the utility assigned is influenced by the maximum of a number of "small" independent factors.

Independent of all the work above and motivated by modeling the results of horse races, Plackett [1975] developed family of distributions over permutations of increasing "complexity". The first-order model posited by Plackett is the same as the Luce model because of which the Luce model is often also referred to as the Plackett-Luce model. Interestingly, Plackett's approach is the basis for many of the systems designed for beating the races (see Ziemba and Hausch [1984]).

It should be noted that the Luce model is the same as the multinomial logit (MNL) model – a workhorse for applications in marketing and econometrics. Particularly in the area of marketing, the work by Guadagni and Little [1983] paved the way for rich work in fitting choice models to scanner panel data.

In terms of empirical applications, after initial applications in psychology, travel demand analysis has been a major area of application for choice models. Many

empirical statistical tools for empirical use of choice models have been developed in relation to the travel demand analysis. Daniel McFadden, an econometrician who shared the 2000 Nobel Memorial Prize in Economic Sciences with James Heckman, pioneered the application of discrete choice models to travel demand analysis. Driven by the practical issues that arise with applying choice models, the work related to travel demand analysis has led to several rich developments in the field of RUM models. See McFadden [2000] for an overview of the role of RUM models in travel demand analysis.

The Luce model has been empirically tested especially by mathematical psychologists and was found to be restrictive due to the IIA property mentioned above. The limitation of the IIA property is commonly explained through the famous "red-bus, blue-bus" example. Suppose a commuter initially faces a decision between two modes of operation – a car and a red bus – and is indifferent between them. Then, the commuter would choose each of these options with probability 1/2. Now suppose we add a blue bus to the mix. Assuming that commuters do not care about the color of the bus, we would expect the probability of choosing the car to remain 1/2, while the probability of choosing the red and blue buses to become a 1/4 each. However, the IIA property, which dictates that ratio of the choice probabilities of the car and red bus should be preserved, results in new choice probabilities of a 1/3 each. Intuitively, the reason for this failure is that the IIA axiom does not take into account the existence of "perfect" substitutes.

In order to overcome the IIA limitation of the Luce model, several extensions have been proposed. Some of the fixes have focused on modifying the model so that it better reflects the behavioral aspects of the choice process. For instance, Tversky [1972] describes a choice by a hierarchical elimination process, where it is assumed that individuals make choices according to features or attributes of the alternatives through a process of hierarchical elimination. This process is best explained through an example. Consider the choice of a restaurant for dinner. The choice might be made by eliminating alternatives based on the cuisine, location, online reviews, price, etc. until only one alternative is left. Others fixes have focused on relaxing the

independence assumption of the random variables in the RUM models. Popular among them is the nested-logit (NL). The NL model was first derived by Ben-Akiva [1973] in connection with applications to travel demand analysis. Roughly speaking, in an NL model, the alternatives are partitioned into groups or nests and decision maker first chooses the nest and then one of the products within the nest according to an MNL model. This model partially addresses the IIA problem. In particular, in the above example, the red bus and the blue bus would belong to one nest and the car to another nest; both nests would have the same probability of being chosen and with the 'bus-nest', both the buses would have the same probability of being chosen. However, the IIA property still holds within the nest.

In summary, choice models have a rich history dating back to almost a century and spanning several fields like psychology, statistics, and transportation. As is evident from the discussion above, there is immense diversity in the approaches taken to choice modeling. Our historical account has been an attempt at covering the highlights of choice modeling approaches in connection to the RUM framework. Our account has indeed not been exhaustive and we refer the interested readers to McFadden [2000] and Diaconis [1988] for a more comprehensive treatment.

In the next two sections, we provide an overview of the RUM models and the exponential family of models.

2.2 Random Utility Maximization (RUM) models

In this section, we provide a brief overview of the RUM family of models. We start with the basic form form of the RUM models and then go into the details of some important special instances: the multinomial logit (MNL), the nested logit (NL), cross-nested logit (CNL), and the mixture of MNL models. The descriptions are self-contained but brief; we refer an interested reader to Ben-Akiva and Lerman [1985] for more details.

In its most general form, the RUM model models the utilities assigned by the decision maker as a random utility vector $U = (U_1, U_2, \ldots, U_N)$, where U_i corre-

sponds to the utility of alternative i. In each decision instance, the decision maker samples a utility vector $u = (u_1, u_2, \ldots, u_N)$ from an appropriately chosen joint distribution and from an offer set $\mathcal{M}$ chooses the alternative with the maximum utility i.e., $\text{argmin}_{i \in \mathcal{M}} u_i$. Different choices of sampling distributions for the utility vectors yields different instances of RUM models. For instance, a natural choice for the sampling distribution could be a multivariate Gaussian distribution; this choice yields what is called a multinomial probit (MNP) model. Although, Gaussian distribution is natural to consider, the MNP model is hard handle. Specifically, there is no closed-form expression for the choice probabilities. The reason is that the Gaussian distribution is not closed under maximization, and the computation of choice probabilities requires evaluating tails probabilities of the maximum of the utilities. Choosing the extreme-value distribution or the double exponential distribution alleviates this issue because, as noted above, independent extreme-value random variables with the same scale parameter are closed under maximization. Choosing different forms of extreme-value distributions yields different families of choice models as we describe below.

2.2.1 Multinomial logit (MNL) family

The MNL model is a popular and most commonly used parametric model in economics, marketing and operations management (see Ben-Akiva and Lerman [1985], Anderson et al. [1992]) and is the canonical example of an RUM model. In the MNL model, the utility of the customer from product j takes the form $U_j = V_j + \xi_j$, where V_j is the deterministic component and the error terms $\xi_1, \xi_2, \ldots, \xi_N$ are i.i.d. random variables having a Gumbel distribution with location parameter 0 and scale parameter 1. Let w_j denote e^{μ_j}; then, according to the MNL model, the probability that product j is purchased from an assortment $\mathcal{M}$ is given by

$$\mathbb{P}(j|\mathcal{M}) = w_j / \sum_{i \in \mathcal{M}} w_i.$$

A major advantage of the MNL model is that it is analytically tractable. In particular, it has a closed form expression for the choice probabilities. However, it

has several shortcomings. As explained above, a major limitation of the MNL model is that it exhibits Independent of Irrelevant Alternatives (IIA) property i.e., the relative likelihood of the purchase of any two given product variants is independent of the other products on offer. This property may be undesirable in several practical contexts where some product are 'more like' other products so that the randomness in a given customers utility is potentially correlated across products. There are other – more complicated – variants that have been proposed to alleviate the IIA issue – the most popular being the NL model, which we describe next.

2.2.2 Nested logit (NL) family

The nested logit (NL) family of models, first derived by Ben-Akiva [1973], was designed to explicitly capture the presence of shared unobserved attributes among alternatives. In particular, the universe of products is partitioned into L mutually exclusive subsets called *nests* denoted by $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_L$ such that

$$\mathcal{N} = \bigcup_{\ell=1}^{L} \mathcal{N}_\ell \qquad \text{and} \quad \mathcal{N}_\ell \cap \mathcal{N}_m = \emptyset, \text{ for } m \neq \ell.$$

The partitioning is such that products sharing unobserved attributes lie in the same nest. Each customer has utility U_j for product j given by $U_j = V_j + \xi_\ell + \xi_{j,\ell}$; here, ξ_ℓ is the error term shared by all the products in nest $\mathcal{N}_\ell$, and $\xi_{j,\ell}$ is the error term that is product specific and assumed to be i.i.d across different products. In this logit case, $\xi_{j,\ell}$ are assumed to be i.i.d standard Gumbel distributed with location parameter 0 and scale parameter 1. The nest specific error terms $\xi_1, \xi_2, \ldots, \xi_L$ are assumed i.i.d., distributed such that for each ℓ, j, $\xi_\ell + \xi_{j,\ell}$ is Gumbel distributed with location parameter 0 and scale parameter $\rho < 1$. Let w_j denote e^{μ_j} and let

$$w(\ell, \mathcal{M}) \stackrel{\text{def}}{=} \sum_{i \in \mathcal{N}_\ell \cap \mathcal{M}} w_i.$$

Then, with the above assumptions on the error terms, it can be shown that (see Ben-Akiva and Lerman [1985]) the probability that product j is purchased when offered

assortment $\mathcal{M}$ is

$$\mathbb{P}(j|\mathcal{M}) = \mathbb{P}(\mathcal{N}_\ell|\mathcal{M})\,\mathbb{P}(j|\mathcal{N}_\ell,\mathcal{M}) = \frac{(w(\ell,\mathcal{M}))^\rho}{\sum_{m=1}^{L}(w(m,\mathcal{M}))^\rho}\ \frac{w_j}{w(\ell,\mathcal{M})}. \tag{2.1}$$

Nested logit models alleviate the issue of IIA exhibited by the MNL models. Further, they have a closed form expression for the choice probabilities, which makes them computationally tractable. However, these models still exhibit IIA property within a nest. Furthermore, in practice, it is often a challenging task to partition the products into different nests. Another limitation of the NL model is that it requires each product to be placed in exactly one nest. There are several settings where this can be restrictive. For instance, a common issue in practice is the placement of the 'no-purchase' option. Although it can be placed in a nest of its own, it is intuitively more appealing to place it in every nest because it is correlated with all the products. In order to capture such situations, an NL model has been extended to what is called a cross-nested logit (CNL) model, which describe next.

2.2.3 Cross-nested logit (CNL) family

A cross-nested logit (CNL) model is an extension of the NL model in that it allows each product to belong to multiple nests. The name cross-nested seems to be due to Vovsha [1997] and Vovsha's model is similar to the Ordered GEV model proposed by Small [1987]. We discuss below only a special form of the CNL model in which only one product is assumed to belong to multiple nests; the more general form in which multiple products belong to multiple nests can be obtained from the special form in a straightforward manner. In the setting where only one product belongs to multiple nests, the probability of purchase of product j is given by (2.1) (see Ben-Akiva and Lerman [1985]), where $w(\mathcal{M},\ell)$ is now defined as

$$w(\ell,\mathcal{M}) \stackrel{\text{def}}{=} \alpha_\ell w_0 + \sum_{i\in(\mathcal{N}_\ell\cap\mathcal{M})\setminus\{0\}} w_i.$$

Here α_ℓ is the parameter capturing the level of membership of the no-purchase option in nest ℓ. The following conditions are imposed on the parameters α_ℓ, $\ell = 1, 2, \ldots, L$

$$\sum_{\ell=1}^{L} \alpha_\ell^\rho = 1 \quad \text{and} \quad \alpha_\ell \geq 0, \text{ for } \ell = 1, 2, \ldots, L.$$

The first condition is a normalization condition that is imposed because it is not possible to identify all the parameters.

While the CNL model overcomes the limitations of the NL model, it is less tractable and Marzano and Papola [2008] showed that it cannot capture all possible types of correlations among products. Furthermore, the MNL, the NL, and the CNL models don't account for heterogeneity in customer tastes. The MMNL family of models, described next, explicitly account for heterogeneity in customer tastes.

2.2.4 Mixed multinomial logit (MMNL)[2] family

The mixed multinomial logit (MMNL) family of models is a very general class of choice models. In fact, it is considered to be the most widely used and the most promising of the discrete choice models currently available Hensher and Greene [2003]. It was introduced by Boyd and Mellman [1980] and Cardell and Dunbar [1980].

In this model, the utility of customer c from product j is given by $U_{c,j} = \langle \beta_c, x_j \rangle + \varepsilon_{c,j}$, where x_j is the vector of *observed* attributes of product j; β_c is the vector of regression coefficients that are *stochastic* and not fixed[3] to account for the *unobserved* effects that depend on the *observed explanatory variables*; $\varepsilon_{c,j}$ is the stochastic term to account for the rest of the unobserved effects; and $\langle \beta_c, x_j \rangle$ denotes the dot product of β_c and x_j. In this logit context, it is assumed that the variables $\varepsilon_{c,j}$ are i.i.d across customers and products and distributed according to the Gumbel distribution of location parameter 0 and scale parameter 1. The distribution chosen for β_c depends on the application at hand and the variance of the components of β_c accounts for the heterogeneity in customer tastes. Assuming that β has a distribution $G(\beta; \theta)$ param-

[2]This family of models is also referred to in the literature as Random parameter logit (RPL), Kernel/Hybrid logit.

[3]This is unlike in the MNL model where the coefficients are assumed to be fixed, but unknown.

eterized by θ, probability that a particular product j is purchased from assortment $\mathcal{M}$ is

$$\mathbb{P}(j|\mathcal{M}) = \int \frac{\exp\left\{\beta^T x_j\right\}}{\sum_{i\in\mathcal{M}} \exp\left\{\beta^T x_i\right\}} G(d\beta;\theta).$$

The MMNL family is a very rich class of models. McFadden and Train [2000] show that under mild regularity conditions, an MMNL model can approximate arbitrarily closely the choice probabilities of *any* discrete choice model that belongs to the class of RUM models. This is a strong result showing that MMNL family of models is very rich and models can be constructed to capture various aspects of consumer choice. In particular, it also overcomes the IIA limitation of the MNL and nested MNL (within a nest) families of models. However, MMNL models are far less computationally tractable than both the MNL and nested MNL models. In particular, there is in general no closed form expression for choice probabilities, and thereby the estimation of these models requires the more complex simulation methods. In addition – and more importantly – while the MMNL family can in principle capture highly complex consumer choice behavior, appropriate choice of features and distributions of parameters is highly application dependent and is often very challenging Hensher and Greene [2003].

2.3 Exponential family of models

The exponential family is a family of general statistical models that have been developed to fit ranking data independent of the application. Even beyond ranked data, exponential families have been extensively studied, and several classical distributions like the normal, exponential, gamma, and binomial distribution are special instance of the exponential family. The exponential family of distributions naturally arise in contexts where summary statistics are easier to gather than the entire set of data. Specifically, the Koopman-Pitman-Darmois theorem establishes that the exponential family of models is the richest family of models that are completely described by a given set of summary statistics.

More precisely, in independent works, Darmois [1935], Koopman [1936], and Pitman [1936], subsequent to Fisher [1922] own less explicit indication of the result, established the following result. Loosely stated: suppose $\mathbb{P}_\theta(x)$ is a family of distributions on the real line with the parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$. Let $y_1, y_2, \ldots, y_n$ be n samples generated independently from the distribution $\mathbb{P}_\theta(\cdot)$ and let $y = (y_1, y_2, \ldots, y_n)$. Then, the n dimensional joint density $\prod_{i=1}^n \mathbb{P}_\theta(y_i)$ has a sufficient statistic $\phi(y) = (\phi_1(y), \phi_2(y), \ldots, \phi_s(y))$ of dimension $s < n$ *if and only if* $\mathbb{P}_\theta$ belongs to the exponential family. Particularly, $\mathbb{P}_\theta$ is of the form

$$\mathbb{P}_\theta(y) \propto \exp\left(\sum_{j=1}^{r} b_j(\theta)\psi_j(y)\right),$$

where $r \leq s$ and for some functions $\psi_j(\cdot)$ that are related to the sufficient statistic $\phi(\cdot)$. In other words, exponential family of models is the richest family of models that are completely described by a given set of summary statistics. As mentioned above, this is of immense importance to the set of ranked data because in practice, it is reasonable to collect only summary statistics.

Next, we discuss two special instances of models that belong to the exponential family: max-entropy distributions and distance-based ranking models.

2.3.1 Max-entropy distributions

The exponential family of distributions naturally arise maximum entropy distributions. Specifically, given marginal information of the form $\mathbb{E}[t_i(X)]$ with $i = 1, 2, \ldots, m$ for any random variable X, it can be shown that the distribution with the maximum entropy is a member of the exponential family. More precisely, consider the setup of ranked data. Suppose we have access to first-order marginal information i.e., for each alternative i and position r, we have access to the probability of i that is ranked at position r; we denote this quantity by μ_{ir}. It is easy to see that taking $t_{ir}(\sigma) = \mathbb{1}_{\{\sigma(i)=r\}}$ for any permutation σ, where $\mathbb{1}_{\mathcal{A}}$ takes the value 1 when the event $\mathcal{A}$ is true and 0 when $\mathcal{A}$ is false, we can write $\mu_{ir} = \mathbb{E}[t_{ir}(\sigma)]$. It can then be shown

that the maximum entropy distribution is of the form

$$\mathbb{P}_{\theta^*}(\sigma) \propto \exp\left(\sum_{1\leq i,r\leq N} \theta^*_{ir} t_{ir}(\sigma)\right),$$

where the parameter θ^* is such that $\mathbb{E}_{\mathbb{P}_{\theta^*}}[t_{ir}(\sigma)] = \mu_{ir}$ for all $1 \leq i, r \leq N$.

2.3.2 Distance-based ranking models

Distance-based ranking models constitute a popular sub-class of the exponential family of distributions. In this class, each permutations σ is assigned a probability given by

$$\mathbb{P}_{\theta,\sigma_0}(\sigma) \propto \exp\left(-\theta d(\sigma, \sigma_0)\right),$$

where $\theta \geq 0$, $d(\sigma, \sigma_0)$ measures the distance of σ from σ_0, and σ_0 is the permutation around which the distribution is concentrated. The distance $d(\cdot, \cdot)$ is typically assumed to be right-invariant i.e., it is invariant to re-labeling of the items being ranked – more precisely, $d(\sigma_1 \circ \pi, \sigma_2 \circ \pi) = d(\sigma_1, \sigma_2)$, for any permutations σ_1, σ_2, π, where $\sigma \circ \pi(i) = \sigma(\pi(i))$. In addition, $d(\cdot, \cdot)$ is also assumed to be a metric satisfying the usual metric axioms of positivity, symmetry, and triangle-inequality.

Different choices of the distance $d(\cdot, \cdot)$ yield different instances of the distance-based ranking models. Some of the widely used metrics in applied settings according to Diaconis Diaconis [1988] are as follows:

1. Kendall's tau metric: $d(\sigma, \pi) = \sum_{i<j} \mathbb{1}_{(\sigma(i)-\sigma(j))(\pi(i)-\pi(j))<0}$ i.e., the number of pairs of items that are ranked in reverse-order by σ and π;

2. Spearman's rho metric: $d(\sigma, \pi) = (\sum_i (\sigma(i) - \pi(i))^2)^{1/2}$;

3. Spearman's footrule metric: $d(\sigma, \pi) = \sum_i |\sigma(i) - \pi(i)|$;

4. Hamming metric: $d(\sigma, \pi) = \sum_i \mathbb{1}_{\sigma(i)\neq\pi(i)}$;

5. Cayley's metric: $d(\sigma, \pi)$ = number of minimal transpositions required to go from σ to π;

6. Ulam's metric: $d(\sigma, \pi) = N -$ (maximal number of items ranked in the same relative order by σ and π.

Taking $d(\cdot, \cdot)$ to be the Kendall-tau distance yields the popular Mallows' model (see Mallows [1957]). We note here that the Kendall-tau distance $T(\sigma, \pi)$ between a permutation σ and the identity permutation id two permutations can be written as

$$T(\sigma, \mathrm{id}) = \sum_i V_i(\sigma),$$

where $V_i(\sigma) = \sum_{j>i} \mathbb{1}_{\sigma(j)<\sigma(i)}$ i.e, the number of elements in the set $\{i+1, i+2, \ldots, N\}$ that are ranked before i. Since $T(\cdot, \cdot)$ it is easy to see that all the distances are determined by the distance to the identity id. Hence, Mallow's model can also be written as

$$\mathbb{P}_{\theta,\sigma_0} \propto \exp\left(-\theta \sum_i V_i(\sigma \circ \sigma_0^{-1})\right).$$

Mallow's model has also been extended to extended to yield the Generalized Mallow's model:

$$\mathbb{P}_{\theta,\sigma_0} \propto \exp\left(-\sum_i \theta_i V_i(\sigma \circ \sigma_0^{-1}))\right).$$

The Mallow's and the Generalized Mallow's models are versatile ranking models that provide effective means to model distributions concentrated around a central ranking σ_0. The size of the parameter θ controls the "dispersion" of the model around the central ranking σ_0. One can obtain finer control over the dispersion in the Generalized model; specifically, by taking the values θ_i to decrease with i one can emphasize the greater importance of ranking the top-ranked items correctly i.e., according the central ranking σ_0.

The distance-based ranking models are popular and have been extensively studied in the literature and applied in several settings Critchlow et al. [1991].

In addition to the RUM and exponential families discussed above, there is a wealth of other models that have been studied in the literature in various contexts. Critchlow et al. [1991] provides a comprehensive overview of the various probabilistic models over rankings.

2.4 Chapter summary and discussion

This chapter provided a broad overview of the popular choice models that have been studied in the literature. Due to the diversity of modeling approaches in the literature, our attempt was not be exhaustive. Despite the diversity, two broad frameworks standout from the vast collection of seemingly different modeling approaches. The frameworks yield two broad families of choice models: the *Random Utility Maximization* (RUM) family and the *exponential* family. The RUM family of models has beginnings in the psychological literature, where the goal was to develop models that were grounded in more basic processes. RUM framework is a generalization of the classical utility maximization framework studied extensively in Economics. The exponential family, on the other hand, has origins in statistics and emphasizes data availability rather than basic processes driving choice. We provided a brief historical account of of the methodological developments in choice theory. We then provided an overview of the RUM framework with details of some important special instances like the multinomial logit (MNL), nested logit (NL), cross-nested logit (CNL), and mixtures of MNL (MMNL) models. Finally, we formally described the exponential family of models and discussed the max-entropy distributions and distance-based models.

From the discussion of the different modeling approaches, we highlight the following two important aspects.

First, it is evident from our discussion above that there is a vast diversity in the choice modeling approaches that have been taken. In fact, despite the long history of work dating all the way back to the 1920s, there is neither a convergence to a consensus nor an emergence of a dominant approach; of course, given the diversity of the application areas, this fragmentation of modeling approaches is not surprising. This fragmentation of approaches now makes the task of model selection challenging, especially for a practitioner. Specifically, different modeling approaches have been proposed with different objectives and different application domains in mind, and each comes with its own set of strengths and weaknesses – both in terms of the structural assumptions it makes and the computational issues it raises. Thereby,

selecting the "right" model for an application at hand requires subjective judgements and hard to get expert input.

Second, note that different modeling approaches view data differently. In the RUM framework, the choice process takes the center stage and data is typically an after-thought. More precisely, choice models are designed to capture the individual decision making behavior by accounting for the various factors that influence the choice behavior; for instance, the mean utilities in an MNL model are typically modeled as linear models of the attributes of the decison-maker and the alternatives, allowing one to make use of fine-grained inputs. In the case of exponential families, on the other hand, data (in the form of summary statistics) takes the center stage; here, instead of explicitly modeling the underlying choice process, the "richest" model that can be fit to the given data is chosen. Note that this statement is especially true for the family of max-entropy distributions[4]. Consequently, in the case of max-entropy distributions, the burden of picking the right model maybe off-loaded to the data.

Given the above observations, it is evident that the RUM models allow for fine-grained inputs at the cost of making subjective judgements and at the risk of producing inaccurate predictions due to the selection of incorrect structures. The max-entropy distributions, on the other hand, off-load the burden of picking the model to the data at the cost of not being able to include fine-grained information. Depending on the application at hand, one or the other approach may be appropriate.

Finally, we note here that our approach is more along the lines of max-entropy distributions in that we want to off-load the burden of model selection to the data. The applications of our interest possess three important characteristics:

1. Making accurate and fine-grained predictions is more important than understanding the underlying choice process; in fact, the decisions based on these predictions potentially impact millions of dollars.

2. The purchase/choice behavior exhibited by customers is highly dynamic and

[4]The distance-based models make an attempt to capture some application context by considering distributions that are concentrated around a central permutation. However, the structure imposed is very coarse and does not capture the fine-level attribute information.

rapidly varying across geographical locations and times.

3. Technological developments have made available enormous amounts of transaction data.

Given these characteristics, it is evident a data-driven modeling approach is not only a possibility, but also a necessity for these applications. Now, based on the observations we made above, it is natural to consider the exponential families for these applications. However, as mentioned above, learning the parameters of a general exponential family is a computationally challenging task. Moreover, even if one did learn the parameters, the $N!$ support size of the distribution makes just the computation of the choice probabilities challenging. Therefore, there is a need for a new nonparametric approach, which we propose in the next chapter.

Chapter 3

Decision problems: revenue prediction

This chapter deals with the application of choice models to make decisions. There are several important practical applications where the end-goal is to make a decision, and a choice model is a critical component to making that decision. The main application area of our focus is the set of decision problems faced by operations managers. In this context, a central decision problem is the *static assortment optimization* problem in which the goal is to find the optimal assortment: the assortment of products with the maximum revenue subject to a constraint on the size of the assortment. Solving the decision problem requires two components: (a) a subroutine that uses historical sales transaction data to predict the expected revenues from offering each assortment of products, and (b) an optimization algorithm that uses the subroutine to find the optimal assortment. The present chapter focuses on designing a revenue prediction subroutine, and the next chapter deals with the optimization algorithm.

As one can imagine, the problems of predicting revenues and finding the optimal assortment are important in their own right, and their consideration is motivated by the fact that any improvements to existing solutions will have significant practical implications. Specifically, solutions to these two problems lead to a solution to the *single-leg, multiple fare-class yield management problem*; this problem is central to the area Revenue Management (RM) and deals with the allocation of aircraft seat ca-

pacity to multiple fare classes when customers exhibit choice behavior. In particular, consider an airline selling tickets to a single-leg aircraft. Assume that the airline has already decided the fare classes and is trying to dynamically decide which fare-classes to open as a function of the remaining booking time and the remaining number of seats. This dynamic decision problem can be cast in a reasonably straightforward manner as a dynamic program with one state variable. As shown in Talluri and van Ryzin [2004a], the solution to the dynamic program reduces to solving a slight variant of the static assortment optimization problem. Thus, solution to the two problems effectively solves the single-leg, multiple fare-class yield management problem âŁ" a central problem to RM with huge practical implications.

In the context of static assortment optimization problem, a choice model is required to predict revenues. In particular, a choice model fit to historical data predicts choice probabilities: the probability that a particular product is chosen from an offer set of products. Assuming that the prices (or revenues obtained from each sale) of products is known, the predicted choice probabilities can be used to predict revenues. Most of the work in OM that deals with solving decision problems requiring revenue predictions take the choice model as given and assume access to accurate predictions. When it comes to predicting revenues, a popular parametric model like the multinomial logit (MNL) model is used. As mentioned in the first chapter, parametric models suffer from various limitations the primary of which is that their complexity does not scale with data because of which they fail to glean new structural information in additional data. Put differently, parametric models tend to over-fit or under-fit resulting in poor accuracy in their predictions. In order to overcome this issue, we propose a nonparametric approach to predicting revenues and sales that affords the revenue manager the opportunity to avoid the challenging task of fitting an appropriate parametric model to historical data. Specifically, we start with the entire family of distributions over preference lists and let the data select the "right" choice model to make revenue predictions.

Next, we describe in detail the solutions we propose to the above problems. We begin with a brief overview of the related work in Section 3.1. We then give a formal

description of the model and the precise descriptions of the problems in Section 3.2. Section 3.3 describes our revenue prediction subroutine and discusses various methods to implement the subroutine in a computationally efficient manner in practice. Sections 3.4, 3.5, 3.6 are devoted to establish the quality of the revenue predictions: Section 3.4 discusses a simulation study that demonstrates that our approach can produce accurate predictions by effectively capturing various parametric structures; Section 3.5 describes a case-study with real-world data from a major US automaker in which our method achieved an improvement over existing approaches of around 20% in the accuracy of revenue predictions; finally, Section 3.6 derives error guarantees for our approach when the underlying choice model is either the MNL or the MMNL model. We finish our discussion about revenue predictions in Section 3.7, where we obtain a characterization of the choice models used for revenue predictions in terms of the sparsity of the distribution. Finally, Section 3.8 concludes with a summary and some thoughts on future directions.

3.1 Relevant literature

The study of choice models and their applications spans a vast literature across multiple fields including at least Marketing, Operations and Economics. In disciplines such as marketing learning a choice model is an interesting goal unto itself given that it is frequently the case that a researcher wishes to uncover "why" a particular decision was made. Within operations, the goal is frequently more application oriented with the choice model being explicitly used as a predictive tool within some larger decision model. Since our goals are aligned with the latter direction, our literature review focuses predominantly on OM; we briefly touch on key work in Marketing. We note that our consideration of 'sparsity' as an appropriate non-parametric model selection criterion is closely related to the burgeoning statistical area of compressive sensing; we discuss those connections in a later section.

The vast majority of decision models encountered in operations have traditionally ignored substitution behavior (and thereby choice modeling) altogether. Within air-

line RM, this is referred to as the "independent demand" model (see Talluri and van Ryzin [2004b]). Over the years, several studies have demonstrated the improvements that could be obtained by incorporating choice behavior into operations models. For example, within airline RM, the simulation studies conducted by Belobaba and Hopperstad [1999] on the well known passenger origin and destination simulator (PODS) suggested the value of corrections to the independent demand model; more recently, Ratliff et al. [2008] and Vulcano et al. [2010] have demonstrated valuable average revenue improvements from using MNL choice-based RM approaches using real airline market data. Following such studies, there has been a significant amount of research in the areas of inventory management and RM attempting to incorporate choice behavior into operations models.

The bulk of the research on choice modeling in both the areas has been optimization related. That is to say, most of the work has focused on devising optimal decisions *given* a choice model. Talluri and van Ryzin [2004a], Gallego et al. [2006], van Ryzin and Vulcano [2008], Mahajan and van Ryzin [1999], Goyal et al. [2009] are all papers in this vein. Kök et al. [2008] provides an excellent overview of the state-of-the-art in assortment optimization. Rusmevichientong et al. [2010a] consider the multinomial logit (MNL) model and provide an efficient algorithm for the static assortment optimization problem and propose an efficient policy for the dynamic optimization problem. A follow on paper, Rusmevichientong and Topaloglu [2009], considers the same optimization problem but where the mean utilities in the MNL model are allowed to lie in some arbitrary uncertainty set. Saure and Zeevi [2009] propose an alternative approach for the dynamic assortment optimization problem under a general random utility model.

The majority of the work above focuses on optimization issues given a choice model. Paper such as Talluri and van Ryzin [2004a] discuss optimization problems with *general* choice models, and as such our revenue estimation procedure fits in perfectly there. In most cases, however, the choice model is assumed to be given and of the MNL type. Papers such as Saure and Zeevi [2009] and Rusmevichientong and Topaloglu [2009] loosen this requirement by allowing some amount of *parametric*

uncertainty. In particular, Saure and Zeevi [2009] assume unknown mean utilities and learn these utilities, while the optimization schemes in Rusmevichientong and Topaloglu [2009] require knowledge of mean utilities only within an interval. In both cases, the structure of the model (effectively, MNL) is *fixed* up front.

The MNL model is by far the most popular choice model studied and applied in OM. The origins of the MNL model date all the way back to the Plackett-Luce model, proposed independently by Luce [1959] and Plackett [1975]. Before becoming popular in the area of OM, the MNL model found widespread use in the areas of transportation (see seminal works of McFadden [2000], Ben-Akiva and Lerman [1985]) and marketing (starting with the seminal work of Guadagni and Little [1983], which paved the way for choice modeling using scanner panel data). See Wierenga [2008], Chandukala et al. [2008] for a detailed overview of choice modeling in the area of Marketing. The MNL model is popular because its structure makes it tractable both in terms of estimating its parameters and solving decision problems. However, the tractability of the MNL model comes at a cost: it is incapable of capturing any heterogeneity in substitution patterns across products (see Debreu [1960]) and suffers from Independent of Irrelevant Alternatives (IIA) property (see Ben-Akiva and Lerman [1985]), both of which limit its practical applicability.

Of course, these issues with the MNL model are well recognized, and far more sophisticated models of choice have been suggested in the literature (see, for instance, Ben-Akiva and Lerman [1985], Anderson et al. [1992]); the price one pays is that the more sophisticated models may not be easily identified from sales data and are prone to over-fitting. It must be noted that an exception to the above state of affairs is the paper by Rusmevichientong et al. [2006] that considers a general nonparametric model of choice similar to the one considered here in the context of an assortment pricing problem. The caveat is that the approach considered requires access to samples of entire customer preference lists which are unlikely to be available in many practical applications.

Our goal relative to all of the above work is to *eliminate* the need for structural assumptions and thereby, the associated risks as well. We provide a means of going

directly from raw sales transaction data to revenue or sales estimates for a given offer set. While this does not represent the entirety of what can be done with a choice model, it represent a valuable application, at least within the operational problems discussed.

3.2 Choice model and problem formulations

We consider a universe of $N+1$ products, $\mathcal{N} = \{1, 2, \ldots, N\}$. We assume that a customer always has an 'outside' or the 'no-purchase' option; the choice of the 'outside' option is equivalent to the customer not choosing anything from the offered set of products, and we denote the 'outside' option by product 0. A customer is associated with a permutation (or ranking) σ of the products in $\mathcal{N} \cup \{0\}$; the customer prefers product i to product j if and only if $\sigma(i) < \sigma(j)$. When offered an assortment of products $\mathcal{M} \subset \mathcal{N}$, the customer purchases the product that is the most preferred of the no-purchase option and the products offered in $\mathcal{M}$. In particular, she purchases

$$\underset{i \in \mathcal{M} \cup \{0\}}{\operatorname{argmin}} \sigma(i). \tag{3.1}$$

This view of choice behavior is clearly intuitively very reasonable, and its origins date back to Economists in the 1950s (see Mas-Colell et al. [1995]). In fact, is view of preference lists or permutations is the basis for the utility maximization theory, where for tractability reasons, a utility function that maps products to real numbers is used to capture customer preferences.

3.2.1 Aggregate choice model

We model the aggregate behavior of a population of customers through a distribution λ over the space of all permutations of the products in $\mathcal{N} \cup \{0\}$; specifically, for any permutation σ, $\lambda(\sigma)$ corresponds to the fraction of customers in the population with preference list σ. It is easy to see that the distribution λ captures all the desired choice probabilities. Particularly, suppose the customer population described by distribution

λ is offered an assortment of products $\mathcal{M} \subset \mathcal{N}$; then, the probability that product i is purchased (the fraction of customers that purchase product i) from $\mathcal{M}$ is equal to the sum of the weights of λ over all permutations that result in the choice of i when offered $\mathcal{M}$. More precisely, the probability of purchase of i given $\mathcal{M}$ can be written as[1]

$$\mathbb{P}_\lambda(i|\mathcal{M}) = \sum_{\sigma \in \mathcal{S}_i(\mathcal{M})} \lambda(\sigma),$$

where $\mathcal{S}_i(\mathcal{M})$ is the set of permutations that result in the choice of i from $\mathcal{M}$; equivalently, $\mathcal{S}_i(\mathcal{M})$ is the set of permutations that prefer i to all the products in $\mathcal{M} \cup \{0\}$). More formally, we can write

$$\mathcal{S}_i(\mathcal{M}) = \{\sigma \colon \sigma(i) < \sigma(j) \text{ for all } j \in \mathcal{M} \cup \{0\}\}.$$

The distribution λ captures all the required choice probabilities and we call it the choice model. As mentioned above, using distributions over permutations to model choice behavior is very general and forms the basis for the general class of random utility choice models.

In applications of our interest, the sales or revenues expected from offering an assortment $\mathcal{M}$ of products is an important quantity of interest. Given a choice model λ, one can compute the expected revenue as follows. Let p_i denote the price of product i or the revenue obtained from the sale of product i; as expected, we set p_0 to 0. Suppose assortment $\mathcal{M}$ is offered to customers from a population described by choice model λ. Then, the revenue expected from each arriving customer is given by

$$R(\mathcal{M}) = \sum_{i \in \mathcal{M}} p_i \mathbb{P}_\lambda(i|\mathcal{M}).$$

Setting $p_i = 1$ for all $i \in \mathcal{N}$ in the above expression yields the probability that an arriving customer will make a purchase, or what is called the 'conversion-rate', which refers to the probability of converting an arriving customer to a sale.

[1]Sometimes we drop λ and simply use $\mathbb{P}(i|\mathcal{M})$ to denote the choice probability in cases where the choice model is irrelevant or implied by the context.

3.2.2 Data

The class of choice models we work with is quite general and imposes a minimal number of assumptions on customers a-priori. That said the data available in practice to calibrate such a model is typically be limited – specifically, corresponds to historical sales transaction data from usually a small collection of assortments. In order to capture a wide-range of practically relevant cases, we introduce a general compact representation of the available data. It will be quickly seen that the abstract notion we posit is relevant to data one might obtain from sales information.

We assume that the data observed by the seller is given by an m-dimensional 'partial information' vector $y = A\lambda$, where $A \in \{0,1\}^{m \times N!}$ makes precise the relationship between the observed data and the underlying choice model. Typically we anticipate $m \ll N!$ signifying, for example, the fact that we have sales information for only a limited number of assortments. Before understanding how transactional data observed in practice relates to this formalism, we consider, for the purposes of illustration a few simple concrete examples of data vectors y; we subsequently introduce a type of data relevant to our experiments and transaction data observed in the real world.

- *Comparison Data:* This data represents the fraction of customers that prefer a given product i to a product j. The partial information vector y is indexed by i,j with $0 \leq i,j \leq N; i \neq j$. For each i,j, $y_{i,j}$ denotes the fraction of customers that prefer product i to j. The matrix A is thus in $\{0,1\}^{N(N-1) \times N!}$. A column of A, $A(\sigma)$, will thus have $A(\sigma)_{ij} = 1$ if and only if $\sigma(i) < \sigma(j)$.

- *First-order Data:* This data represents the fraction of customers that rank a given product i as their rth choice. Here the partial information vector y is indexed by i,r with $0 \leq i,r \leq N$. For each i,r, y_{ri} is thus the fraction of customers that rank product i at position r. The matrix A is then in $\{0,1\}^{N^2 \times N!}$. For a column of A corresponding to the permutation σ, $A(\sigma)$, we will thus have $A(\sigma)_{ri} = 1$ iff $\sigma(i) = r$.

- *Top Set Data:* This data refers to a concatenation of the "Comparison Data" above and information on the fraction of customers who have a given product i as their topmost choice for each i. Thus $A^\top = [A_1^\top A_2^\top]$ where A_1 is simply the A matrix for comparison data, and $A_2 \in \{0,1\}^{N \times N!}$ has $A_2(\sigma)_i = 1$ if and only if $\sigma(i) = 1$.

Transaction Data: More generally, in the retail context, historical sales records corresponding to displayed assortments might be used to estimate the fraction of purchasing customers who purchased a given product i when the displayed assortment was $\mathcal{M}$. We might have such data for some sequence of test assortments say $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M$. This type of data is consistent with our definition (i.e. it may be interpreted as a linear transformation of λ) and is, in fact, closely related to the comparison data above. In particular, denoting by y_{im}, the fraction of customers purchasing product i when assortment $\mathcal{M}_m$ is on offer, our partial information vector, $y \in [0,1]^{N.M}$, may thus be indexed by i, m with $0 \leq i \leq N, 1 \leq m \leq M$. The matrix A is then in $\{0,1\}^{N.M \times N!}$. For a column of A corresponding to the permutation σ, $A(\sigma)$, we will then have $A(\sigma)_{im} = 1$ iff $i \in \mathcal{M}_m$ and $\sigma(i) < \sigma(j)$ for all products j in assortment $\mathcal{M}_m$.

3.2.3 Problem formulations

Imagine we have a corpus of transaction data, summarized by an appropriate data vector y as described in Section 3.2.2. Our goal is to use *just* this data to make predictions about the revenue rate (i.e the expected revenues garnered from a random customer) for some given assortment, say $\mathcal{M}$, that has never been encountered in past data.

We propose accomplishing this by solving the following program:

$$\begin{aligned} \underset{\lambda}{\text{minimize}} \quad & R(\mathcal{M}) \\ \text{subject to} \quad & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{aligned} \tag{3.2}$$

In particular, the optimal value of this program is our prediction for the revenue rate. In words, the feasible region of this program describes the set of all choice models consistent with the observed data y. The optimal objective value consequently corresponds to the *minimum* revenues possible for the assortment $\mathcal{M}$ under any choice model consistent with the observed data. Since the family of choice models we considered was *generic* this prediction relies on simply the data and basic economic assumptions on the customer that are tacitly assumed in essentially any choice model.

In the next sections, we describe our solutions to solving this problem.

3.3 Revenue predictions: computation

In the previous section we formulated the task of computing revenue estimates via a non-parametric model of choice and any available data as the mathematical program (3.2), which we repeat below, in a slightly different form for clarity:

$$\begin{aligned} \underset{\lambda}{\text{minimize}} \quad & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\ \text{subject to} \quad & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{aligned}$$

The above mathematical program is a linear program in the variables λ. Interpreting the program in words, the constraints $A\lambda = y$ ensure that any λ assumed in making a revenue estimate is *consistent* with the observed data. Other than this consistency requirement, writing the probability that a customer purchases $j \in \mathcal{M}, \mathbb{P}(j|\mathcal{M})$, as

the quantity $\lambda_j(\mathcal{M}) \triangleq \sum_{\sigma \in \mathcal{S}_j(\mathcal{M})} \lambda(\sigma)$ assumes that the choice model satisfies the basic structure laid out in Section 3.2.1. We make no other assumptions outside of these, and ask for the lowest expected revenues possible for $\mathcal{M}$ under *any* choice model satisfying these requirements.

Thus, while the assumptions implicit in making a revenue estimate are something that the user need not think about, the two natural questions that arise are:

1. How does one solve this conceptually simple program in practice given that the program involves an intractable number of variables?

2. Even if one did succeed in solving such a program are the revenue predictions produced useful or are they too loose to be of practical value?

This section will focus on the first question. In practical applications such a procedure would need to be integrated into a larger decision problem and so it is useful to understand the computational details which we present at a high level in this section. The second, 'so what' question will be the subject of the next two sections where we will examine the performance of the scheme on simulated transaction data, and finally on a real world sales prediction problem using real data.

3.3.1 The dual to the robust problem

At a high level our approach to solving (3.2) will be to consider the dual of that program and then derive efficient exact or approximate descriptions to the feasible regions of these programs. We begin by considering the dual program to (3.2). In preparation for taking the dual, let us define

$$\mathcal{A}_j(\mathcal{M}) \triangleq \{A(\sigma) : \sigma \in \mathcal{S}_j(\mathcal{M})\},$$

where recall that $\mathcal{S}_j(\mathcal{M}) = \{\sigma \in S_N : \sigma(j) < \sigma(i), \forall i \in \mathcal{M}, i \neq j\}$ denotes the set of all permutations that result in the purchase of $j \in \mathcal{M}$ when the offered assortment is $\mathcal{M}$. Since $S_N = \cup_{j \in \mathcal{M}} \mathcal{S}_j(\mathcal{M})$ and $\mathcal{S}_j(\mathcal{M}) \cap \mathcal{S}_i(\mathcal{M}) = \emptyset$ for $i \neq j$, we have implicitly specified a partition of the columns of the matrix A. Armed with this notation, the

dual of (3.2) is:

$$\begin{aligned} \underset{\alpha,\nu}{\text{maximize}} \quad & \alpha^\top y + \nu \\ \text{subject to} \quad & \max_{x^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top x^j + \nu\right) \leq p_j, \quad \text{for each } j \in \mathcal{M}. \end{aligned} \tag{3.3}$$

where α and ν are dual variables corresponding respectively to the data consistency constraints $A\lambda = y$ and the requirement that λ is a probability distribution (i.e. $\mathbf{1}^\top \lambda = 1$) respectively. Of course, this program has a potentially intractable number of constraints. We explore two approaches to solving the dual:

1. An extremely simple to implement approach that relies on sampling constraints in the dual that will, in general produce approximate solutions that are upper bounds to the optimal solution of our robust estimation problem.

2. An approach that relies on producing effective representations of the sets $\mathcal{A}_j(\mathcal{M})$, so that each of the constraints $\max_{x^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top x^j + \nu\right) \leq p_j$, can be expressed efficiently.This approach is slightly more complex to implement but in return can be used to sequentially produce tighter approximations to the robust estimation problem. In certain special cases, this approach is provably efficient and optimal.

3.3.2 First approach: constraint sampling

The following is an extremely simple to implement approach to approximately solve the problem (3.3):

1. Select a distribution over permutations, ψ.

2. Sample n permutations according to the distribution. Call this set of permutation $\hat{\mathcal{S}}$.

3. Solve the program:

$$\begin{aligned} \underset{\alpha,\nu}{\text{maximize}} \quad & \alpha^\top y + \nu \\ \text{subject to} \quad & \alpha^\top A(\sigma) + \nu \leq p_j, \quad \text{for each } j \in \mathcal{M}, \sigma \in \hat{\mathcal{S}} \end{aligned} \tag{3.4}$$

Observe that (3.4) is essentially a 'sampled' version of the problem (3.3), wherein constraints of that problem have been sampled according to the distribution ψ and is consequently a relaxation of that problem. A solution to (3.4) is consequently an upper bound to the optimal solution to (3.3).

The question of whether the solutions thus obtained provide meaningful approximations to (3.3) is partially addressed by recent theory developed by Calafiore and Campi [2005]. In particular, it has been shown that for a problem with m variables and given $n = O((1/\varepsilon)(m\ln(1/\varepsilon) + \ln(1/\delta))$ samples, we must have that with probability at least $1 - \delta$ the following holds: An optimal solution to (3.4) violates at most an ϵ fraction of constraints of the problem (3.3) under the measure ψ. Hence, given a number of samples that scales only with the number of variables (and is independent of the number of constraints in (3.3), one can produce an solution to (3.3) that satisfies all but a small fraction of constraints. The theory does not provide any guarantees on how far the optimal cost of the relaxed problem is from the optimal cost of the original problem.

The heuristic nature of this approach notwithstanding, it is extremely simple to implement, and in the experiments conducted in the next section, provided close to optimal solutions.

3.3.3 Second approach: efficient representations of $\mathcal{A}_j(\mathcal{M})$

We describe here one notion of an efficient representation of the sets $\mathcal{A}_j(\mathcal{M})$, and assuming we have such a representation, we describe how one may solve (3.3) efficiently. We will deal with the issue of actually coming up with these efficient representations in Appendix A.2, where we will develop an efficient representation for ranking data and demonstrate a generic procedure to sequentially produce such representations.

Let us assume that every set $\mathcal{S}_j(\mathcal{M})$ can be expressed as a disjoint union of D_j sets. We denote the dth such set by $\mathcal{S}_{jd}(\mathcal{M})$ and let $\mathcal{A}_{jd}(\mathcal{M})$ be the corresponding set of columns of A. Consider the convex hull of the set $\mathcal{A}_{jd}(\mathcal{M})$, $\text{conv}\{\mathcal{A}_{jd}(\mathcal{M})\} \triangleq \bar{\mathcal{A}}_{jd}(\mathcal{M})$. Recalling that $A \in \{0,1\}^{m\times N}$, $\mathcal{A}_{jd}(\mathcal{M}) \subset \{0,1\}^m$. $\bar{\mathcal{A}}_{jd}(\mathcal{M})$ is thus a

polytope contained in the m-dimensional unit cube, $[0,1]^m$. In other words,

$$\bar{\mathcal{A}}_{jd}(\mathcal{M}) = \{x^{jd} : A_1^{jd}\, x^{jd} \geq b_1^{jd}, \quad A_2^{jd}\, x^{jd} = b_2^{jd}, \quad A_3^{jd}\, x^{jd} \leq b_3^{jd}, \quad x^{jd} \in \mathbb{R}_+^m\} \quad (3.5)$$

for some matrices $A_\cdot^{jd}$ and vectors $b_\cdot^{jd}$. By a canonical representation of $\mathcal{A}_j(\mathcal{M})$, we will thus understand a partition of $\mathcal{S}_j(\mathcal{M})$ and a polyhedral representation of the columns corresponding to every set in the partition as given by (3.5). If the number of partitions as well as the polyhedral description of each set of the partition given by (3.5) is polynomial in the input size, we will regard the canonical representation as efficient. Of course, there is no guarantee that an efficient representation of this type exists; clearly, this must rely on the nature of our partial information i.e. the structure of the matrix A. Even if an efficient representation did exist, it remains unclear whether we can identify it. Ignoring these issues for now, we will in the remainder of this section demonstrate how given a representation of the type (3.5), one may solve (3.3) in time polynomial in the size of the representation.

For simplicity of notation, in what follows we assume that each polytope $\bar{\mathcal{A}}_{jd}(\mathcal{M})$ is in standard form,

$$\bar{\mathcal{A}}_{jd}(\mathcal{M}) = \{x^{jd} : A^{jd}\, x^{jd} = b^{jd}, \quad x^{jd} \geq 0.\}.$$

Now since an affine function is always optimized at the vertices of a polytope, we know:

$$\max_{x^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top x^j + \nu\right) = \max_{d, x^{jd} \in \bar{\mathcal{A}}_{jd}(\mathcal{M})} \left(\alpha^\top x^{jd} + \nu\right).$$

We have thus reduced (3.3) to a 'robust' LP. Now, by strong duality we have:

$$\begin{array}{lll} \underset{x^{jd}}{\text{maximize}} & \alpha^\top x^{jd} + \nu & \\ \text{subject to} & A^{jd}\, x^{jd} = b^{jd} & \equiv \\ & x^{jd} \geq 0. & \end{array} \qquad \begin{array}{ll} \underset{\gamma^{jd}}{\text{minimize}} & b^{jd\top} \gamma^{jd} + \nu \\ \text{subject to} & \gamma^{jd\top} A^{jd} \geq \alpha \end{array} \quad (3.6)$$

We have thus established the following useful equality:

$$\left\{\alpha,\nu : \max_{x^j \in \bar{\mathcal{A}}_j(\mathcal{M})} \left(\alpha^\top x^j + \nu\right) \leq p_j\right\} = \left\{\alpha,\nu : b^{jd\top}\gamma^{jd} + \nu \leq p_j, \gamma^{jd\top}A^{jd} \geq \alpha, d = 1,2,\ldots,D_j\right\}.$$

It follows that solving (3.2) is equivalent to the following LP whose complexity is polynomial in the description of our canonical representation:

$$\begin{array}{lll} \underset{\alpha,\nu}{\text{maximize}} & \alpha^\top y + \nu & \\ \text{subject to} & b^{jd\top}\gamma^{jd} + \nu \leq p_j & \text{for all } j \in \mathcal{M}, d = 1,2,\ldots,D_j \\ & \gamma^{jd\top}A^{jd} \geq \alpha & \text{for all } j \in \mathcal{M}, d = 1,2,\ldots,D_j. \end{array} \tag{3.7}$$

As discussed, our ability to solve (3.7) relies on our ability to produce an efficient canonical representation of $\mathcal{S}_j(\mathcal{M})$ of the type (3.5). In Appendix A.2, we first consider the case of ranking data, where an efficient such representation may be produced. We then illustrate a method that produces a sequence of 'outer-approximations' to (3.5) for general types of data, and thereby allows us to produce a sequence of improving lower bounding approximations to our robust revenue estimation problem, (3.2). This provides a general procedure to address the task of solving (3.3), or equivalently, (3.2).

We end this section with a brief note on noisy observations. In particular, in practice, one may see a 'noisy' version of $y = A\lambda$. Specifically, as opposed to knowing y precisely, one may simply know that $y \in \mathcal{E}$, where $\mathcal{E}$ may, for instance, represent an uncertainty ellipsoid, or a 'box' derived from sample averages of the associated quantities and the corresponding confidence intervals. In this case, one seeks to solve the problem:

$$\begin{array}{ll} \underset{\lambda, y \in \mathcal{E}}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\ \text{subject to} & A\lambda = y, \\ & 1^\top \lambda = 1, \\ & \lambda \geq 0. \end{array}$$

Provided $\mathcal{E}$ is convex, this program is essentially no harder to solve than the variant of the problem we have discussed and similar methods to those developed in this section

apply.

3.4 Revenue predictions: data-driven computational study

In this section, we describe the results of an extensive simulation study, the main purpose of which is to demonstrate that the robust approach can capture various underlying parametric structures and produce good revenue predictions. For this study, we pick a range of random utility parametric structures used extensively in current modeling practice.

The broad experimental procedure we followed is the following:

1. Pick a structural model. This may be a model derived from real-world data or a purely synthetic model.

2. Use this structural model to simulate sales for a set of test assortments. This simulates a data set that a practitioner likely has access to.

3. Use this transaction data to estimate marginal information y, and use y to implement the robust approach.

4. Use the implemented robust approach to predict revenues for a distinct set of assortments, and compare the predictions to the *true* revenues computed using the 'ground-truth' structural model chosen for benchmarking in step 1.

Notice that the above experimental procedure lets us isolate the impact of structural errors from that of finite sample errors. Specifically, our goal is to understand how well the robust approach captures the underlying choice structure. For this purpose, we ignore any estimation errors in data by using the 'ground-truth' parametric model to compute the *exact* values of any choice probabilities and revenues required for comparison. Therefore, if the robust approach has good performance across an interesting spectrum of structural models that are believed to be good fits to data

observed in practice, we can conclude that the robust approach is likely to offer accurate revenue predictions with no additional information about structure across a wide-range of problems encountered in practice.

3.4.1 Benchmark models and nature of synthetic Data

The above procedure generates data sets using a variety of 'ground truth' structural models. We pick the following 'random utility' models as benchmarks. A self-contained and compact exposition on the foundations of each of the benchmark models below may be found in Chapter 2.

Multinomial logit family (MNL): For this family, we have:

$$\mathbb{P}(j|\mathcal{M}) = w_j / \sum_{i \in \mathcal{M}} w_i.$$

where the w_i are the parameters specifying the models. See Appendix 2.2.1 for more details.

Nested logit family (NL): This model is a first attempt at overcoming the 'independence of irrelevant alternatives' effect, a shortcoming of the MNL model. For this family, the universe of products is partitioned into L mutually exclusive subsets, or 'nests', denoted by $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_L$ such that

$$\mathcal{N} = \bigcup_{\ell=1}^{L} \mathcal{N}_\ell \qquad \text{and} \quad \mathcal{N}_\ell \cap \mathcal{N}_m = \emptyset, \text{ for } m \neq \ell.$$

This model takes the form:

$$\mathbb{P}(j|\mathcal{M}) = \mathbb{P}(\mathcal{N}_\ell|\mathcal{M})\,\mathbb{P}(j|\mathcal{N}_\ell, \mathcal{M}) = \frac{(w(\ell, \mathcal{M}))^\rho}{\sum_{m=1}^{L} (w(m, \mathcal{M}))^\rho} \; \frac{w_j}{w(\ell, \mathcal{M})}. \tag{3.8}$$

where $\rho < 1$ is a certain scale parameter, and

$$w(\ell, \mathcal{M}) \stackrel{\text{def}}{=} \alpha_\ell w_0 + \sum_{i \in (\mathcal{N}_\ell \cap \mathcal{M}) \setminus \{0\}} w_i.$$

Here α_ℓ is the parameter capturing the level of membership of the no-purchase option

in nest ℓ and satisfies, $\sum_{\ell=1}^{L} \alpha_\ell^\rho = 1, \quad \alpha_\ell \geq 0$, for $\ell = 1, 2, \ldots, L$. In cases when $\alpha_\ell < 1$ for all ℓ, the family is called the *Cross nested logit (CNL)* family. For a more detailed description including the corresponding random utility function and bibliographic details, see Appendix 2.2.2

Mixed multinomial logit family (MMNL): This model accounts specifically for customer heterogeneity. In its most common form, the model reduces to:

$$\mathbb{P}(j|\mathcal{M}) = \int \frac{\exp\left\{\beta^T x_j\right\}}{\sum_{i\in\mathcal{M}} \exp\left\{\beta^T x_i\right\}} G(d\beta; \theta).$$

where x_j is a vector of observed attributes for the jth product, and $G(\cdot, \theta)$ is a distribution parameterized by θ selected by the econometrician that describes heterogeneity in taste. For a more detailed description including the corresponding random utility function and bibliographic details, see Appendix 2.2.4.

Transaction Data Generated: Having selected (and specified) a structural model from the above list, we generated sales transactions as follows:

1. Fix an assortment of two products, i, j.

2. Compute the values of $P(i|\{i,j,0\}), P(j|\{i,j,0\})$ using the chosen parametric model.

3. Repeat the above procedure for all pairs, $\{i,j\}$, and single item sets, $\{i\}$.

The above data is succinctly summarized as an $N^2 - N$ dimensional data vector y, where $y_{i,j} = P(i|\{i,j,0\})$ for $0 \leq i,j \leq N-1$, $i \neq j$. Given the above data, the precise specialization of the robust estimation problem (3.2) that we solve may be found in Appendix A.2.3.

3.4.2 Experiments conducted

With the above setup we conducted two broad sets of experiments. In the first set of experiments, we picked specific models from the MNL, CNL, and MMNL model

classes; the MNL model was constructed using DVD shopping cart data from Amazon.com, and the CNL and MMNL models were obtained through slight 'perturbations' of the MNL model. In order to avoid any artifacts associated with specific models, in the second set of experiments, we conducted 'stress tests' by generating a number of instances of models from each of the MNL, CNL, and MMNL models classes. We next present the details of the two sets of experiments.

The Amazon Model: We considered an MNL model fit to Amazon.com DVD sales data collected between 1 July 2005 to 30 September 2005 [2],where an individual customer's utility for a given DVD, j is given by:

$$U_j = \theta_0 + \theta_1 x_{j,1} + \theta_2 x_{j,2} + \xi_j;^{3}$$

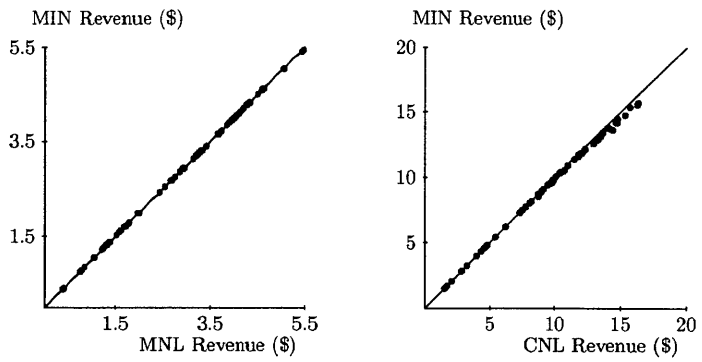
here $x_{j,1}$ is the the price of the package j divided by the number of physical discs it contains, and $x_{j,2}$ is the total number of helpful votes received by product j and ξ_j is a standard Gumbel. The model fit to the data has $\theta_0 = -4.31$, $\theta_1 = -0.038$ and $\theta_2 = 3.54 \times 10^{-5}$. See Table A.1 for the attribute values taken by the 15 products we used for our experiments. We will abbreviate this model AMZN for future reference.

We also considered the following synthetic perturbations of the AMZN model:

1. AMZN-CNL: We derived a CNL model from the original AMZN model by partitioning the products into 4 nests with the first nest containing products 1 to 5, the second nest containing products 6 to 9, the third containing products 10 to 13, and the last containing products 14 and 15. We choose $\rho = 0.5$. We assigned the no-purchase option to every nest with nest membership parameter $\alpha_\ell = (1/4)^{(1/\rho)} = 1/16$.

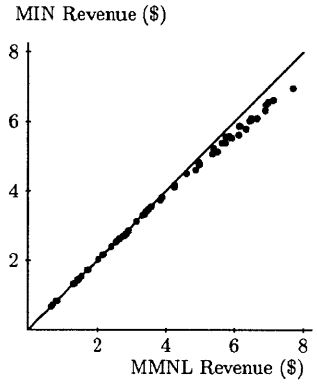
2. AMZN-MMNL: We derived an MMNL model from the original AMZN model by replacing each θ_i parameter with the random quantity $\beta_i = (1 + \eta_{i,j})\theta_i$, for $i = 0, 1, 2$ with $\eta_{i,j}$ is a customer specific random variable distributed as a zero mean normal random variable with standard deviation 0.25.

[2]The specifics of this model were shared with us by the authors of Rusmevichientong et al. [2010a].
[3]The corresponding weights w_j are given by $w_j = \exp(\theta_0 + \theta_1 x_{j,1} + \theta_3 x_{j,2})$.



(a) AMZN

(b) AMNN-CNL



(c) AMZN-MMNL

Figure 3-1: Robust revenue estimates (MIN) vs. true revenues for the AMZN, AMZN-CNL and AMZN-MMNL models. Each of the 60 points in a plot corresponds to (true revenue, MIN) for a randomly drawn assortment.

Figure 3-1 shows the results of the generic experiment for each of the three models above. Each experiment queries the robust estimate on sixty randomly drawn assortments of sizes between one and seven and compares these estimates to those under the respective true model for each case.

Synthetic Model Experiments: The above experiments considered structurally diverse models, each for a *specific* set of parameters. Are the conclusions suggested by Figure 3-1 artifacts of the set of parameters? To assuage this concern, we performed 'stress' tests by considering each structural model in turn, and for each model generating a number of instances of the model by drawing the relevant parameters from a generative family. For each structural model, we considered the following generative families of parameters:

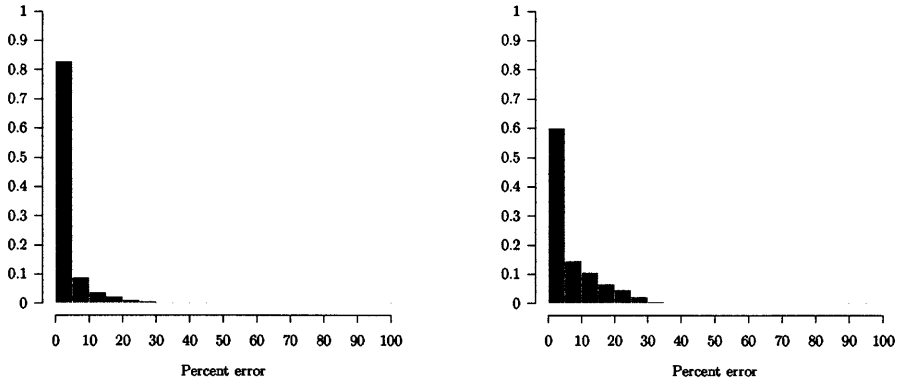
1. MNL Random Family: 20 randomly generated models on 15 products, each generated by drawing mean utilities, $\ln w_j$, uniformly between -5 and 5.

2. CNL Random Family: We maintained the nests, selection of ρ and α_j as in the AMZN-CNL model. We generated 20 distinct CNL models, each generated by drawing $\ln w_j$ uniformly between -5 and 5.

3. MMNL Random Family: We preserved the basic nature of the AMZN-MMNL model. We considered 20 randomly generated MMNL models. Each model differs in the distribution of the parameter vector β. The random coefficients β_j in each case are defined as follows: $\beta_j = (1 + \eta_{i,j})\theta_j$ where $\eta_{i,j}$ is a $N(\mu_j, 0.25)$ random variable. Each of the 20 models corresponds to a single draw of $\mu_j, j = 0, 1, 2$ form the uniform distribution on $[-1, 1]$.

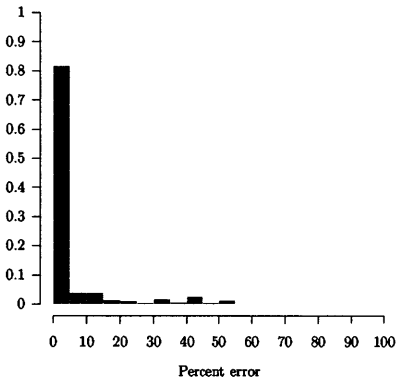
For each of the 60 structural model instances described above, we randomly generated 20 offer sets of sizes between 1 and 7. For a given offer set $\mathcal{M}$, we queried the robust procedure and compared the revenue estimate produced to the true revenue for that offer set; we can compute the latter quantity theoretically. In particular, we measured the relative error, $\varepsilon(\mathcal{M}) \stackrel{\text{def}}{=} \frac{R^{\text{true}}(\mathcal{M}) - R^{\text{MIN}}(\mathcal{M})}{R^{\text{MIN}}(\mathcal{M})}$. The three histograms in Figure 3-2 below represent distributions of relative error for the three generative families described above. Each histogram consists of 400 test points; a given test point corresponds to one of the 20 randomly generated structural models in the relevant family, and a random assortment.

In the above 'stress' tests, we kept the standard deviation of the MMNL models fixed at 0.25. The standard deviation of the MMNL model can be treated as a measure of the heterogeneity or the "complexity" of the model. Naturally, if we keep the "amount" of transaction data fixed and increase the standard deviation – and hence the complexity of the underlying model – we expect the accuracy of robust estimates to deteriorate. To give a sense of the sensitivity of the accuracy of robust revenue estimates to changes in the standard deviation, we repeated the above stress tests with the MMNL model class for three values of standard deviation: 0.1, 0.25, and 0.4. Figure 3-3 shows the comparison of the density plots of relative errors for



(a) MNL-Rand

(b) CNL-Rand



(c) MMNL-Rand

Figure 3-2: Relative error across multiple instances of the MNL, CNL and MMNL structural models.

the three cases.

We draw the following broad conclusion from the above experiments:

- Given limited marginal information for distributions over permutations, λ, arising from a number of commonly used structural models of choice, the robust approach effectively captures diverse parametric structures and provides close revenue predictions under range of practically relevant parametric models.

- With the type of marginal information y *fixed*, the accuracy of robust revenue predictions deteriorates (albeit mildly) as the complexity of the underlying

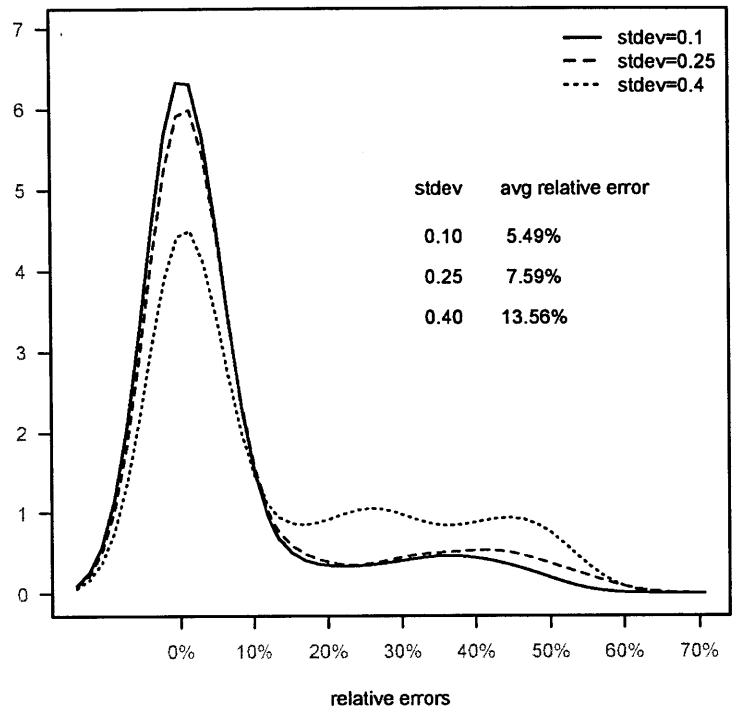


Figure 3-3: The accuracy of robust revenue estimates deteriorates with increase in model complexity, as measured here by variance of the MMNL model. The densities were estimated through kernel density estimation. The density estimates go below zero as a result of smoothing.

model increases; this is evidenced by the deterioration of robust performance as we go from the MNL to the MMNL model class, and similarly as we increase the standard deviation for the MMNL model while keeping the 'amount' of data fixed.

- The design of our experiments allows us to conclude that in the event that a given structural model among the types used in our experiments predicts revenue rates accurately, the robust approach is likely to be just as good *without* knowledge of the relevant structure. In the event that the structural model used is a poor fit, the robust approach will continue to provide meaningful guarantees on revenues under the mild condition that it is tested in an environment where

the distribution generating sales is no different from the distribution used to collect marginal information.

3.5 Revenue predictions: case-study with a major US automaker

In this section, we present the results of a case study conducted using sales transaction data from the dealer network of a major US automaker. Our goal in this study is to use historical transaction data to predict the sales rate or 'conversion rate' for any given offer set of automobiles on a dealer lot. This conversion-rate is defined as the probability of converting an arriving customer into a purchasing customer. The purpose of the case study is two-fold: (1) To demonstrate how the prediction methods developed in this paper can be applied in the real-world and the quality of the predictions they offer in an absolute sense, and (2) To pit the robust method for revenue predictions in a 'horse-race' against parametric approaches based on the MNL and MMNL families of choice models. In order to test the performance of these approaches in different regimes of calibration data, we carried out cross-validations with varying 'amounts' of training/calibration data. The results of the experiments conducted as part of the case study provide us with the evidence to draw two main conclusions:

1. The robust method predicts conversion rates more accurately than either of the parametric methods. In our case study, the improvement in accuracy was about 20% across all regimes of calibration data.

2. Unlike the parametric methods we study, the robust approach is apparently *not* susceptible to over-fitting and under-fitting.

The 20% improvement in accuracy is substantial. The second conclusion has important implications as well: In practice, it is often difficult to ascertain whether the data available is "sufficient" to fit the model at hand. As a result, parametric structures

are prone to over-fitting or under-fitting. The robust approach, on the other hand, *automatically* scales the complexity of the underlying model class with data available, so in principle one should be able to avoid these issues. This is borne out by the case study. In the remainder of this section we describe the experimental setup and then present the evidence to support the above conclusions.

3.5.1 Setup

Appendix A.3 provides a detailed description of our setup; here we provide a higher level discussion for ease of exposition. We collect data comprising purchase transactions of a specific range of small SUVs offered by a major US automaker over 16 months. The data is collected at the dealership level (i.e the finest level possible) for a network of dealers in the Midwest. Each transaction contains information about the date of sale, the identity of the SUV sold, and the identity of the other cars on the dealership lot at the time of sale. Here by 'identity' we mean a unique model identifier that collectively identifies a package of features, color and invoice price point. We make the assumption that purchase behavior within the zone can be described by a single choice model. To ensure the validity of this assumption, we restrict attention to a specific dealership zone, defined as the collection of dealerships within an appropriately defined geographical area with relatively homogeneous demographic features.

Our data consisted of sales information on 14 distinct SUV identities (as described above). We observed a total of $M = 203$ distinct assortments (or subsets) of the 14 products in the dataset, where each assortment $\mathcal{M}_i$, $i = 1, 2, \ldots, M$, was on offer at some point at some dealership in the dealership zone. We then converted the transaction data into sales rate information for each of the assortments as follows:

$$y_{j\mathcal{M}_i} = \frac{\text{num of sales of product } j \text{ when } \mathcal{M}_i \text{ was on offer}}{\text{num of customer arrivals when } \mathcal{M}_i \text{ was on offer}}, \quad \text{for } j \in \mathcal{M}_i, i = 1, 2, \ldots, M.$$

Note that the information to compute the denominator in the expression for $y_{j\mathcal{M}_i}$ is not available because the number of arriving customers who purchase nothing is not

known. Such data 'censoring' is common in practice and impacts both parametric methods as well as our approach. A common approximation here is based on demographic information relative to the location of the dealership. Given the data at our disposal, we are able to make a somewhat better approximation to overcome this issue. In particular, we assume a daily arrival rate of α_d for dealership d and measure the number of arrivals of assortment $\mathcal{M}$ as

$$\text{num of customer arrivals when } \mathcal{M} \text{ was on offer } = \sum_d \alpha_d \, \text{days}_d(\mathcal{M}),$$

where $\text{days}_d(\mathcal{M})$ denotes the number of days for which $\mathcal{M}$ was on offer at dealership d. The arrival rate to each dealership clearly depends on the size of the market to which the dealership caters. Therefore, we assume that $\alpha_d = f \times \text{size}_d$, where size_d denotes the "market size" for dealership d and f is a "fudge" factor. We use previous year total sales at dealership d for the particular model class as the proxy for size_d and tune the parameter f using cross-validation (more details in the appendix).

3.5.2 Experiments and results

We now describe the experiments we conducted and the present the results we obtained. In order to test the predictive performance of the robust, the MNL, and the MMNL methods, we carried out k-fold cross-validations with $k = 2, 5, 10$. In k-fold cross-validation (see Mosteller and Tukey [1987]), we arbitrarily partition the collection of assortments $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M$ into k partitions of about equal size, except may be the last partition. Then, using $k-1$ partitions as training data to calibrate the methods, we test their performance on the k^{th} partition. We repeat this process k times with each of the k partitions used as test data exactly once. This repetition ensures that each assortment is tested at least once. Note that as k decreases, the number of training assortments decreases resulting in more limited data scenarios. Such limited data scenarios are of course of great practical interest.

We measure the prediction accuracy of the methods using the relative error metric. In particular, letting $\hat{y}(\mathcal{M})$ denote the conversion-rate prediction for test assortment

$\mathcal{M}$, the incurred relative error is defined as $|\hat{y}(\mathcal{M}) - y(\mathcal{M})|/y(\mathcal{M})$, where

$$y(\mathcal{M}) := \frac{\text{num of customers who purchase a product when } \mathcal{M} \text{ is on offer}}{\text{num of customer arrivals when } \mathcal{M}_i \text{ was on offer}}.$$

In the case of the parametric approaches, $\hat{y}(\mathcal{M})$ is computed using the choice model fit to the training data. In the case of the robust approach, we solve an appropriate mathematical program. A detailed description of how $\hat{y}(\mathcal{M})$ is determined by each method is given in the appendix.

We now present the results of the experiments. Figure 3-4 shows the comparison of the relative errors of the three methods from k-fold cross-validations for $k = 10, 5, 2$. Table 3.1 shows the mean relative error percentages of the three methods and the percent improvement in mean relative error achieved by the robust method over the MNL and MMNL methods for the three calibration data regimes of $k = 10, 5, 2$. It is clear from the definition of k-fold cross-validation that as k decreases, the "amount" of calibration data decreases, or equivalently calibration data sparsity increases. Such sparse calibration data regimes are of course of great practical interest.

Table 3.1: Mean relative errors in percentages of different methods

k	MNL	MMNL	Robust	Percent Improvement over MNL	Percent Improvement over MMNL
10	43.43	43.39	34.79	19.89%	19.80%
5	43.25	45.73	35.79	17.23%	21.62%
2	45.65	46.61	36.83	19.33%	20.99%

The immediate conclusion we draw from the results is that the prediction accuracy of the robust method is better than those of both MNL and MMNL methods in all calibration data regimes. In particular, using the robust method results in close to 20% improvement in prediction accuracy over the MNL and MMNL methods. We also note that while the prediction accuracy of the more complex MMNL method is marginally better than that of the MNL method in the high calibration-data regime of $k = 10$, it quickly becomes worse as the amount of calibration data available decreases. This behavior is a consequence of over-fitting caused due to the complexity of the MMNL model. The performance of the robust method, on the other hand, remains

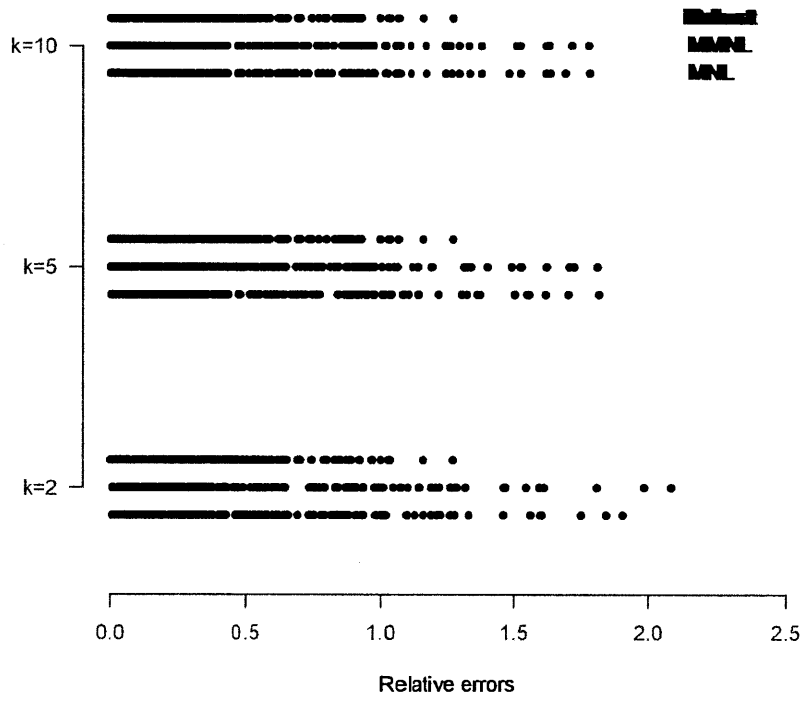


Figure 3-4: Robust method outperforms both MNL and MMNL methods in conversion-rate predictions across various calibration data regimes. The figure compares relative errors of the three methods in k-fold cross-validations for $k = 10, 5, 2$. Each point corresponds to the relative error for a particular test assortment.

stable across the different regimes of calibration-data.

3.6 Revenue predictions: theoretical guarantees

The results from our simulation study (Figs. 3-1, 3-2) demonstrate high accuracy of the robust approach for a wide-range of practically relevant scenarios. In this section, we explain the success of the robust approach by deriving analytical bounds for the relative error. The error bound we derive can be computed using *only* the available partial information, thereby providing readily computable guarantees for our estimates in practical applications. For the analysis in this section, we restrict

ourselves to the case when we have access to complete k^{th}-order partial information: for *every* assortment $\mathcal{M}$ of size at most k, we have access to $\mathbb{P}(j \mid \mathcal{M} \cup \{0\})$ for all products $j \in \mathcal{M}$; note that $k = 2$ is the 'censored' variant of comparison information used above for experiments. This restriction simplifies analysis, provides insights, and the analysis can easily be generalized to general types of partial information.

As expected, the accuracy of our revenue predictions depends on two factors: (a) the "complexity" of the underlying choice behavior, and (b) the "amount" of information in the available data. For a given complexity of underlying choice behavior, the more the data, the better the accuracy we expect. Similarly, for a given amount of information in the available data, the greater the complexity of the underlying choice behavior, the worse the accuracy we expect. The bounds we derive on the accuracy of our predictions confirm this intuition. Before we provide the precise details, we describe the basic insight: robust approach with limited data produces accurate estimates if only a "few" products account for most of the cannibalization of the sales of each product. To elaborate on this, we note that "similar" products in the offered assortment result in a cannibalization of (reduction in) the sales of each other; cannibalization of the sales of the notebooks offered by Apple Inc., by their iPads is a current example. For each product, if only a few other products account for most of the cannibalization, then the robust approach can produce accurate predictions with limited data. For a given level of accuracy, there is an inverse relationship between the "amount" of data and the number of products where the cannibalization is concentrated. Using this insight, we derive general bounds on the relative error incurred by our approach. The bounds we derive can be computed using the partial data that is available and, hence, we can provide provable and computable guarantees for the estimates we produce in practice. We then specialize the error bounds to the MNL and the MMNL families of models. We observe that for the MNL case, the above characteristic of choice models that leads to small errors translates into the following MNL characteristic: sum of the weights of a few products must account for most of the sum of the weights of all the products in the assortment. Similarly, for the MMNL case, which is a mixture of MNL models, the characteristic roughly translates

to having less heterogeneity in customer tastes, with the constituent MNL models possessing above MNL characteristic. Thus, the robust approach can be confidently used whenever these intuitive conditions are met.

We now provide the precise details. In order to derive error bounds for robust estimates, we first derive a bound for the error incurred when we approximate $\mathbb{P}(j \mid \mathcal{M} \cup \{0\})$ by $\mathbb{P}(j \mid \mathcal{M}' \cup \{0\})$, for any assortment $\mathcal{M}' \subset \mathcal{M}$ such that $j \in \mathcal{M}'$. For that, as noted above, the demand (purchase probability) observed for a product in an assortment can be thought of as a combination of its primary demand and the cannibalization effects of all other products present in the assortment. Thus, when we approximate $\mathbb{P}(j \mid \mathcal{M} \cup \{0\})$ by $\mathbb{P}(j \mid \mathcal{M}' \cup \{0\})$, we are ignoring the cannibalization effects of all the products in $\mathcal{M} \setminus \mathcal{M}'$. Let $B_j(\mathcal{M}'; i) \stackrel{\text{def}}{=} \mathbb{P}(j \mid \mathcal{M}' \cup \{0\}) - \mathbb{P}(j \mid \mathcal{M}' \cup \{i, 0\})$, which we call the *incremental cannibalization effect* on j of adding i to $\mathcal{M}'$, and $B_j(\mathcal{M}'; \mathcal{M}) \stackrel{\text{def}}{=} \sum_{i \in \mathcal{M} \setminus \mathcal{M}'} B_j(\mathcal{M}'; i)$. Then, we can prove that $\mathbb{P}(j \mid \mathcal{M}') - B_j(\mathcal{M}'; \mathcal{M}) \leq \mathbb{P}(j \mid \mathcal{M}) \leq \mathbb{P}(j \mid \mathcal{M}')$ (see Lemma 1); the right inequality is straightforward and the left inequality requires a union bound argument. Since we have access to k^{th}-order partial data, we can readily evaluate the error bound $B_j(\mathcal{M}'; \mathcal{M})$ if $|\mathcal{M}'| \leq k - 1$. Letting $S(j, \mathcal{M}) \stackrel{\text{def}}{=} \{\mathcal{M}' \subset \mathcal{M} \colon j \in \mathcal{M}',\ |\mathcal{M}'| \leq k - 1\}$, it now follows that we can readily compute the bound for $\mathbb{P}(j \mid \mathcal{M}') - \mathbb{P}(j \mid \mathcal{M})$ whenever $\mathcal{M}' \in S(j, \mathcal{M})$.

With the above approximation bounds for purchase probabilities, we can now derive error bounds for revenue estimates. First, since we are at liberty to choose any assortment $\mathcal{M}' \in S(j, \mathcal{M})$ to approximate $\mathbb{P}(j \mid \mathcal{M})$, we choose the one the minimizes the relative error. In particular, we choose $\mathcal{M}_j \stackrel{\text{def}}{=} \operatorname{argmin}_{\mathcal{M}' \in S(j,\mathcal{M})} \frac{B_j(\mathcal{M}'; \mathcal{M})}{\mathbb{P}(j \mid \mathcal{M}')}$. Now, letting $\delta_k(j, \mathcal{M}) \stackrel{\text{def}}{=} \frac{B_j(\mathcal{M}_j; \mathcal{M})}{\mathbb{P}(j \mid \mathcal{M}_j)}$, it is easy to see that $(1 - \delta_k(j, \mathcal{M}))\mathbb{P}(j \mid \mathcal{M}_j) \leq \mathbb{P}(j \mid \mathcal{M}) \leq \mathbb{P}(j \mid \mathcal{M}_j)$. Now, let $R_k(\mathcal{M})$ denote $\sum_{j \in \mathcal{M}} p_j \mathbb{P}(j \mid \mathcal{M}_j)$ and $\delta_k(\mathcal{M}) \stackrel{\text{def}}{=} \max_{j \in \mathcal{M}} \delta_k(j, \mathcal{M})$. Then, we can write $(1 - \delta_k(\mathcal{M}))R_k(\mathcal{M}) \leq R(\mathcal{M}) \leq R_k(\mathcal{M})$. Note that both $\delta_k(\mathcal{M})$ and $R_k(\mathcal{M})$ are completely determined by the k^{th}-order partial data that is available. All the bounds we have derived until now are valid for any distribution over permutations that is consistent with the given k^{th}-order partial data. Therefore, we can write $(1 - \delta_k(\mathcal{M}))R_k(\mathcal{M}) \leq R^{\min}(\mathcal{M}) \leq R^{\text{true}}(\mathcal{M}) \leq R_k(\mathcal{M})$,

which immediately gives $R^{\min}(\mathcal{M})/R^{\text{true}}(\mathcal{M}) \geq 1 - \delta_k(\mathcal{M})$. This discussion can now be summarized as the following theorem.

Theorem 1. *Given an assortment $\mathcal{M}$ and product j, suppose we have access to complete k^{th}-order partial information. Then, $\frac{R^{\min}(\mathcal{M})}{R^{\text{true}}(\mathcal{M})} \geq 1 - \delta_k(\mathcal{M})$, where $\delta_k(\mathcal{M}) = \max_{j \in \mathcal{M}} \delta_k(j|\mathcal{M})$.*

The following is now evident from Theorem 1. The relative error in revenue estimation through our robust approach is small if $\delta_k(\mathcal{M})$ is small, which is in turn small if the best relative error in approximating $\mathbb{P}(j \mid \mathcal{M})$ using sets of size at most $k-1$ is small. Therefore, for any given k, the robust approach provides accurate estimates if most of the cannibalization effects can be accounted for by at most $k-2$ products; equivalently, depending on the application at hand, k can be chosen so that most of the cannibalization effects can be accounted for by at most $k-2$ products. This interpretation provides a convenient way to choose k, or equivalently predict the performance of the robust approach for a given k in practical applications.

Specializing $\delta_k(\mathcal{M})$ in Theorem 1 for different parametric models yields guarantees for different models. Specifically, we have the following theorem for the MNL model:

Theorem 2. *Suppose the underlying choice model is the MNL model and we are given kth-order partial information. Without loss of generality, suppose $\mathcal{M} = \{1, 2, \ldots, C\}$ is the assortment for which we want to predict revenues and the weights are such that $w_1 \geq w_2 \geq \cdots \geq w_C$. Then, we have*

$$1 - \frac{R^{\min}(\mathcal{M})}{R^{\text{true}}(\mathcal{M})} \leq \delta_k(\mathcal{M}),$$

where $R^{\min}(\mathcal{M})$ is the estimate produced by the robust approach, $R^{\text{true}}(\mathcal{M})$ is the true revenue, and

$$\delta_k(\mathcal{M}) \leq \frac{w_k + \cdots + w_C}{1 + w_1 + w_2 + \cdots + w_{k-1}}.$$

Thus, for an MNL model cannibalization effects are completely captured by the weights and error is small if the top k products constitute most of the sum of the weights of the product in $\mathcal{M}$. For $k = 2$ and $\mathcal{M} = \mathcal{N}$, this bound evaluates to 0.16 for

the Amazon.com data set we considered in the experiments above. Thus, the bound we derived is tight for this dataset.

Now we consider the MMNL family. Since it is a mixture of MNL models, we express its error bound in terms of error bounds of MNL models. Particularly, we have the following theorem:

Theorem 3. *Suppose the mixing distribution $G(\cdot)$ is multivariate Gaussian with mean vector μ and covariance matrix $\Sigma = \sigma^2 I_D$, where I_D is $D \times D$ identity matrix. Letting $\delta_k^\beta(j, \mathcal{M})$ denote the relative error for MNL model corresponding to β, we can write*

$$\delta_k(j, \mathcal{M}) \leq \frac{\int \delta_k^\beta(j, \mathcal{M}) L_\beta\left(j \mid \mathcal{M}_j^\beta\right) G(d\beta)}{\mathbb{P}(j \mid \mathcal{M}_j)} + \frac{\varepsilon_r}{\mathbb{P}(j \mid \mathcal{M}_j)},$$

where

$$\varepsilon_r = 2kC(D)\frac{\exp(-r^2/(2\sigma^2))}{r}$$

with $C(D)$ a constant that depends only on dimension D of attribute vectors, $r = \min_{i,\ell \in \mathcal{M}} \frac{|\langle \mu, x_i - x_\ell \rangle|}{\|x_i - x_\ell\|}$, and $\mathcal{M}_j, \mathcal{M}_j^\beta \in S_j(\mathcal{M})$ are assortments where the minimization is attained in the expression for $\delta_k^\beta(j, \mathcal{M})$ and $\delta_k(j, \mathcal{M})$ respectively. Here $L_\beta(i \mid \mathcal{M})$ is the choice probability under the MNL model with weight of product i equal to $e^{\langle \beta, x_i \rangle}$, where x_i is vector of features of product i.

In the error bound in the above theorem, the first term corresponds to the error made in individual MNL models and essentially is the best error possible. The second term corresponds to the additional error due to heterogeneity in consumer tastes. This term is small for large r and small σ. Variance σ^2 can be treated as a measure heterogeneity. Further, r will be bigger if the mean utilities of products $\langle \mu, a_i \rangle$ at the mean vector are sufficiently far apart. Therefore, additional error due to heterogeneity will be small if the amount of heterogeneity σ is small, or if the products are sufficiently differentiated in terms of their mean utilities along the mean vector μ. Hence, robust estimates are accurate if the tastes of the population are not too heterogeneous and if the products are sufficiently differentiated along the mean 'taste vector' μ.

The proofs of Theorems 2 and 3 are given in Appendix A.4. Before we con-

clude the section, we formally state Lemma 1 referred to above; its proof is given in Appendix A.4.

Lemma 1. *For any two assortments $\mathcal{M}$, $\mathcal{M}'$ and product j such that $j \in \mathcal{M}'$ and $\mathcal{M}' \subset \mathcal{M}$, we have*

$$\mathbb{P}(j \mid \mathcal{M}') - B_j(\mathcal{M}'; \mathcal{M}) \leq \mathbb{P}(j \mid \mathcal{M}) \leq \mathbb{P}(j \mid \mathcal{M}'),$$

where $B_j(\mathcal{M}; \mathcal{M}') \stackrel{\text{def}}{=} \sum_{i \in \mathcal{M} \setminus \mathcal{M}'} B_j(\mathcal{M}'; i)$ and

$$B_j(\mathcal{M}'; i) \stackrel{\text{def}}{=} \mathbb{P}(j \mid \mathcal{M}' \cup \{0\}) - \mathbb{P}(j \mid \mathcal{M}' \cup \{0, i\}).$$

3.7 Revenue predictions and sparse choice models

We conclude the discussion of our theoretical analysis by establishing a relationship between the "amount" of data and the "complexity" of the choice model used for revenue predictions. As noted in the introduction, the main advantage of the non-parametric approach is that it scales the model complexity with the "amount" of data. This section is devoted to formalize this fact. Specifically, we use *sparsity* or the support size of the distribution over preference lists as a measure of the complexity of the choice model and the dimension of the data vector y as a measure of the "amount" of data. Sparsity is a very appealing property of choice models. Particularly, sparse models allow us to explain the observed substitution behavior using as small a number of customer preference lists as possible; in addition, such a description also provides a great deal of tractability in multiple applications (see, for example van Ryzin and Vulcano [2008]). With these metrics, we show that the sparsity of the choice model used by the robust method for revenue prediction scales with the dimension of the data vector. Sparsity is a rich notion and it serves as a criterion to select a choice model from the family of consistent ones. We devote Chapter 5 to the discussion of sparse models.

In order to show that the complexity of the choice model scales with the "amount"

of data, we obtain a characterization of the choice models implicitly used by the robust approach in terms of their sparsity. More precisely, we first establish that the choice model implicitly used by the robust approach has sparsity within at most one of the sparsity of the sparsest model consistent with the data. Next, we show that the sparsity of the choice model used by the robust approach scales with the dimension of the data vector y thereby establishing that the complexity of the model used by the robust approach scales with the "amount" of data available. This provides a potential explanation for the immunity of the robust approach to over/under fitting issues, as evidenced in our case study.

To state our result formally, define the set $\mathcal{Y}$ as the set of all possible data vectors, namely the convex hull of the columns of the matrix A. For some $y \in \mathcal{Y}$ and an arbitrary offer set, $\mathcal{M}$, let $\lambda^{\min}(y)$ be an optimal *basic feasible* solution to the program used in our revenue estimation procedure, namely, (3.2). Moreover, let, $\lambda^{\text{sparse}}(y)$ be the *sparsest* choice model consistent with the data vector y; i.e. $\lambda^{\text{sparse}}(y)$ is an optimal solution to the program

$$\begin{aligned} \underset{\lambda}{\text{minimize}} \quad & \|\lambda\|_0 \\ \text{subject to} \quad & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{aligned} \tag{3.9}$$

We then have that with probability one, the sparsity (i.e. the number of rank lists with positive mass) under $\lambda^{\min}(y)$ is close to that of $\lambda^{\text{sparse}}(y)$. In particular, we have:

Theorem 4. *For any distribution over $\mathcal{Y}$ that is absolutely continuous with respect to Lebesgue measure on $\mathcal{Y}$, we have with probability 1, that:*

$$0 \leq \|\lambda^{\min}(y)\|_0 - \|\lambda^{\text{sparse}}(y)\|_0 \leq 1$$

Theorem 4 (proved in Appendix A.1) establishes that if K were the support size of the sparsest distribution consistent with y, the sparsity of the choice model used by our revenue estimation procedure is either K or $K+1$ for "almost all" data vectors y. As such, this establishes that the choice model implicitly employed by the robust

procedure is essentially also the sparsest model consistent with the observed data. In addition the proof of the theorem reveals that the sparsity of the robust choice model consistent with the observed data is either[4] m or $m+1$ for almost all data vectors y of dimension m. This yields yet another valuable insight into the choice models implicit in our revenue predictions – the complexity of these models, as measured by their sparsity, grows with the amount of observed data. As such, we see that the complexity of the choice model implicitly employed by the robust procedure scales automatically with the amount of available data, as one would desire from a non-parametric scheme. This provides a potential explanation for the robust procedures lack of susceptibility to the over-fitting observed for the MMNL model in our empirical study.

3.8 Chapter summary and discussion

This chapter considered the application of choice models to making decisions. We focused on the type of applications that are important to the areas of OM and RM. The central decision problem in this context is the *static assortment optimization* problem in which the goal is to find the optimal assortment: the assortment of products with the maximum revenue subject to a constraint on the size of the assortment. Solving the decision problem requires two components: (a) a subroutine that uses historical sales transaction data to predict the expected revenues from offering each assortment of products, and (b) an optimization algorithm that uses the subroutine to find the optimal assortment. Clearly, both components are important in their own right. Specifically, solutions to these two problems lead to a solution to the *single-leg, multiple fare-class yield management problem*, which deals with the allocation of aircraft seat capacity to multiple fare classes when customers exhibit choice behavior.

This chapter focused on designing a nonparametric revenue prediction subroutine. Most of the existing approaches are parametric in nature. However, parametric approaches are limited because the complexity of the choice model used for predictions does not scale with the "amount" of data, making the the model prone to over-fitting

[4]Here, we assume that matrix A has full row rank.

and under-fitting issues. Thus, we considered a nonparametric approach, which overcomes this issue. Specifically, given historical transaction data, we identified a family of distributions that are consistent with the data. Now, given an assortment of products, we predicted revenues by computing the worst-case revenue of all consistent choice models. We addressed the computational issues, and demonstrated the accuracy of our revenue predictions through empirical studies. In particular, in our case-study with transaction data from a major US automaker, our approach succeeded in obtaining around 20% improvement in the accuracy of revenue predictions when compared to popular existing approaches. We also provided theoretical guarantees for the relative errors. The theoretical guarantees confirm the intuition that the error depends on the "complexity" of the underlying choice structure and the "amount" of data that is available.

Although we proposed a nonparametric approach to choice modeling that can be successfully applied in practice, our work is no panacea for all choice modeling problems. In particular, one merit of a structural/ parametric modeling approach to modeling choice is the ability to extrapolate. That is to say, a nonparametric approach such as ours can start making useful predictions about the interactions of a particular product with other products only once *some* data related to that product is observed. With a structural model, one can hope to say useful things about products never seen before. The decision of whether a structural modeling approach is relevant to the problem at hand or whether the approach we offer is a viable alternative thus merits a careful consideration of the context. Of course, as we have discussed earlier, resorting to a parametric approach will typically require expert input on underlying product features that 'matter', and is thus difficult to automate on a large scale.

Chapter 4

Decision problems: assortment optimization

This chapter continues our discussion on using choice models to make decisions. The decision problem of our focus is the *static assortment optimization* problem in which the goal is to find the optimal assortment: the assortment of products with the maximum revenue subject to a constraint on the size of the assortment. Solving the decision problem requires two components: (a) a subroutine that uses historical sales transaction data to predict the expected revenues from offering each assortment of products, and (b) an optimization algorithm that uses the subroutine to find the optimal assortment. The previous chapter focused on using a nonparametric choice modeling approach to make revenue predictions. This chapter focuses on an efficient optimization algorithm to find an approximation of the optimal assortment. As explained in the previous chapter, the revenue prediction subroutine described in the previous chapter along with the assortment optimization algorithm described in this chapter yield a solution to the *single-leg, multiple fare-class yield management problem.*

Given the subroutine to predict revenues as described in the previous chapter, we need an efficient algorithm to search for the optimal assortment. In particular, we

are interested in solving

$$\underset{|\mathcal{M}| \leq C}{\text{argmax}} \quad R(\mathcal{M}),$$

where $R(\mathcal{M})$ is the expected revenue from offering assortment $\mathcal{M}$. In this chapter, we assume access to a subroutine that can efficiently generate revenue predictions for each assortment $\mathcal{M}$, and our goal is to design an optimization algorithm that minimizes the number of calls to the subroutine. The revenue predictions can themselves be generated either using a specific parametric choice model or using the nonparametric approach described in the previous chapter. Assuming there are N products and a constraint of C on the size of the optimal assortment, exhaustive search would require $O(N^C)$ calls to the revenue subroutine. Such an exhaustive search is prohibitive in practice whenever N or C is large. Therefore, our goal is to propose an algorithm that can produce a "good" approximation to the optimal assortment with only a "few" calls to the revenue subroutine. Existing approaches focus on exploiting specific parametric structures of choice models to solve the decision problem efficiently. In this context, Rusmevichientong et al. [2010a] have proposed an efficient algorithm to find the optimal assortment in $O(NC)$ operations whenever the underlying model is the MNL model. Unfortunately, beyond the simple case of the MNL model, the optimization problem or its variants are provably hard (like the NL and MMNL models; see Rusmevichientong et al. [2009] and Rusmevichientong et al. [2010b]). In addition, the algorithms proposed in the literature (both exact and approximate) heavily exploit the structure of the assumed choice model; consequently, the existing algorithms – even without any guarantees – cannot be used with other choice models like the probit model or the mixture of MNL models with a continuous mixture. Given these issues, our goal is to design a general optimization scheme that is (a) not tailored to specific parametric structures and (b) requires only a subroutine that gives revenue estimates for assortments.

Overview of our approach. We propose a general set-function optimization algorithm, which given a general function defined over sets, finds an estimate of the set (or assortment) where the function is maximized. This set-function optimization algo-

rithm clearly applies to the static assortment optimization problem, thereby yielding the optimization scheme with the desired properties. Note that since we are considering a very general setup, there is not much structure to exploit. Hence, we adopt the greedy method – the general technique for designing heuristics for optimization problems. However, a naive greedy implementation algorithm fails even in the simple case of the MNL model. Specifically, consider the simpler un-capacitated decision problem. Here, a naive greedy implementation would start with the empty set and incrementally build the solution set by adding at each stage a product that results in the maximum increase in revenue; this process would terminate when addition of a product no longer results in an increase in revenue. It is easy to see that the naive implementation would succeed in solving the decision problem only if the optimal assortments exhibit a nesting property: the optimal assortment of size C_1 is a subset of the optimal assortment of size C_2 whenever $C_1 < C_2$. Unfortunately, the nesting property does not hold even in the case of the MNL model. In order to overcome this issue, we allow for greedy "exchanges" in addition to greedy "additions." Particularly, at every stage, we allow a new product to be either added (which we call an "addition") to the solution set or replace an existing product (which we call an "exchange") in the solution set; the operation at each stage is chosen greedily. The termination condition now becomes an interesting question. As in the naive implementation, we could terminate the process when addition or exchange no longer results in an increase in revenue. However, since we never run out of products for exchanges, the algorithm may take an exponential (in the number of products) number of steps to terminate. We overcome this issue by introducing a control parameter that caps the number of times a product may be involved in exchanges. Calling that parameter b, we show that the algorithms calls the revenue subroutine $O(N^2bC^2)$ times for the capacitated problem. We thus obtain a general algorithm with the desired properties to solve the static assortment optimization problem.

Guarantees for our algorithm. We derive guarantees to establish the usefulness of our optimization procedure. For that, we first consider the case of the MNL model, where the decision problem is well-understood. Specifically, we assume that

the underlying choice model is an instance of the MNL family and the revenue subroutine yields revenue estimates for assortments under the specific instance. We can show that the the algorithm we propose, when run with $b \geq C$, succeeds in finding the optimal assortment with $O(N^2C^3)$ calls to the revenue subroutine. Therefore, in the special case when the underlying choice model is the MNL model, our algorithm captures what is already known. It also provides a simpler alternative to the more complicated algorithm proposed by Rusmevichientong et al. [2010a]. We also consider the case when noise corrupts the available revenue estimates – a common practical issue. In this case, we show that our algorithm is robust to errors in the revenue estimates produced by the subroutine. Particularly, if the underlying choice model is the MNL model and the revenue estimate produced by the subroutine may not be exact but within a factor $1 - \varepsilon$ of the true value, then we can show that our algorithm finds an estimate of the optimal assortment with revenue that is within $1 - f(\varepsilon)$ of the optimal value; here $f(\varepsilon)$ goes to zero with ε and also depends on C and the parameters of the underlying model. In summary, our theoretical analysis shows that our algorithm finds the exact optimal solution in the noiseless case or a solution with provable guarantees in the noisy case, whenever the underlying choice model is the MNL model. In this sense, our results subsume what is already known in the context of the MNL model.

In the context of the more complicated models like the nested logit (NL) and the mixtures of MNL models, the decision problem is provably hard. As discussed above, even obtaining a PTAS can be very complicated and requires careful exploitation of the structure. We however believe that it is possible to obtain "good" approximations to the optimal assortments in practice.

Organization. Next, we describe in detail the optimization algorithm we propose and the guarantees we can provide. The rest of the chapter is organized as follows. The optimization algorithm, which we call GREEDYOPT is described in Section 4.1. We then describe the precise guarantees we can provide on the algorithm in Section 4.2. Finally, we present the proofs of our results in Section 4.3 before concluding in Section 4.4.

4.1 Description of GREEDYOPT

We now provide the detailed description of our optimization algorithm GREEDYOPT. As noted above, most of the algorithms proposed in the literature – both exact and approximate – are based on heavily exploiting the structure of the assumed choice model. Unfortunately, since we are considering a very general setup, there is not much structure to exploit. Hence, we adopt the greedy method – the general technique for designing heuristics for optimization problems.

A naive greedy implementation however fails even in the simple case of the MNL model. Specifically, consider the simpler un-capacitated decision problem. Here, a naive greedy implementation would start with the empty set and incrementally build the solution set by adding at each stage a product that results in the maximum increase in revenue; this process would terminate when addition of a product no longer results in an increase in revenue. It is easy to see that the naive implementation would succeed in solving the decision problem only if the optimal assortments exhibit a nesting property: the optimal assortment of size C_1 is a subset of the optimal assortment of size C_2 whenever $C_1 < C_2$. Unfortunately, the nesting property does not hold even in the case of the MNL model.

In order to overcome this issues associated with the naive greedy implementation, we allow for greedy "exchanges" in addition to greedy "additions." Particularly, at every stage, we allow a new product to be either added (which we call an "addition") to the solution set or replace an existing product (which we call an "exchange") in the solution set; the operation at each stage is chosen greedily. The termination condition now becomes an interesting question. As in the naive implementation, we could terminate the process when addition or exchange no longer results in an increase in revenue. However, since we never run of products for exchanges, the algorithm may take an exponential (in the number of products) number of steps to terminate. We overcome this issue by introducing a control parameter that caps the number of times a product may be involved in exchanges. Calling that parameter b, we show that the algorithms calls the revenue subroutine $O(N^2bC^2)$ times for the capacitated problem.

We thus obtain a general algorithm with the desired properties to solve the static assortment optimization problem.

The formal description of the algorithm is provided in Figures 4-1 and 4-2. For convenience, whenever an exchange takes place, we call the product that is removed as the product that is *exchanged-out* and the product that is introduced as the product that is *exchanged-in*. Now, the algorithm takes as inputs the capacity C, the initial assortment size S, and a bound b on the number of exchange-outs. The algorithm incrementally builds the solution assortment. Specifically, it searches over all assortments of size S. For each such assortment, the algorithm calls the subroutine GREEDYADD-EXCHANGE (formally described in Figure 4-2) at most $C-S$ times to construct an assortment of size at most C. Of all such constructed assortments, the algorithm returns the one with the maximum revenue.

Figure 4-1: GREEDYOPT

Input: Initial size S, capacity constraint C such that $1 \leq S \leq C \leq N$, and revenue function $R(\cdot)$.

Output: Estimate of optimal assortment $\hat{M}^{\text{OPT}}$ of size $|\hat{M}^{\text{OPT}}| \leq C$

Algorithm:
Initialization: $\hat{M}^{\text{OPT}} \leftarrow \emptyset$
for each $\mathcal{M} \subset \mathcal{N}$ such that $|\mathcal{M}| = S$
  $\mathcal{M}_S \leftarrow \mathcal{M}$
  for $S+1 \leq i \leq C$
    $\mathcal{M}_i \leftarrow$GREEDYADD-EXCHANGE$(\mathcal{M}_{i-1}, \mathcal{N}, b, R(\cdot))$
  end for
  if $R(\hat{M}^{\text{OPT}}) < R(\mathcal{M}_C)$
    $\hat{M}^{\text{OPT}} \leftarrow \mathcal{M}_C$
  end if
end for
Output: $\hat{M}^{\text{OPT}}$

Running-time complexity: It is easy to see that the number of times GREEDY-OPT calls the revenue function $R(\cdot)$ is equal to $(C-S)\binom{N}{S}$ times the number of times GREEDYADD-EXCHANGE calls the revenue function. In order to count the

Figure 4-2: GREEDYADD-EXCHANGE

Input: assortment $\mathcal{M}$, product universe $\mathcal{N}$, revenue function $R(\cdot)$, maximum number of exhange-outs b

Output: Estimate of optimal assortment of size at most $|\mathcal{M}| + 1$

Algorithm:
Initialization: $\mathcal{M}_1 \leftarrow \mathcal{M}$, $\mathcal{N}_1 \leftarrow \mathcal{N}$, $t = 1$, exchange-outs$(i) = 0$ for each $i \in \mathcal{N}$
while $\mathcal{N}_t \setminus \mathcal{M}_t \neq \emptyset$
  //try exchanging products
  $i^*, j^* = \text{argmax}_{i \in \mathcal{M}_t, j \in \mathcal{N}_t \setminus \mathcal{M}_t} R((\mathcal{M}_t \setminus \{i\}) \cup \{j\})$
  $\tilde{\mathcal{M}}_{\text{exchange}} \leftarrow (\mathcal{M}_t \setminus \{i^*\}) \cup \{j^*\}$

  // try adding a product
  $k^* = \text{argmax}_{k \in \mathcal{N}_t \setminus \mathcal{M}_t} R(\mathcal{M}_t \cup \{k\})$
  $\tilde{\mathcal{M}}_{\text{add}} \leftarrow \mathcal{M}_t \cup \{k^*\}$

  if $|\mathcal{M}_t| < |\mathcal{M}| + 1$ **and** $R(\tilde{\mathcal{M}}_{\text{add}}) > R(\mathcal{M}_t)$ **and** $R(\tilde{\mathcal{M}}_{\text{add}}) > R(\tilde{\mathcal{M}}_{\text{exchange}})$
    // add the product k^*
    $\mathcal{M}_{t+1} \leftarrow \tilde{\mathcal{M}}_{\text{add}}$
  else if $R(\tilde{\mathcal{M}}_{\text{exchange}}) > R(\mathcal{M}_t)$
    // exchange products i^* and j^*
    $\mathcal{M}_{t+1} \leftarrow \tilde{\mathcal{M}}_{\text{exchange}}$
    exchange-outs$(i^*) \leftarrow$ exchange-outs$(i^*) + 1$
    if exchange-outs$(i^*) \geq b$
      $\mathcal{N}_{t+1} \leftarrow \mathcal{N}_t \setminus \{i^*\}$
    else
      $\mathcal{N}_{t+1} \leftarrow \mathcal{N}_t$
  else
    break from while
  end if
end while
Output: $\mathcal{M}_t$

number of times GREEDYADD-EXCHANGE calls the revenue function $R(\cdot)$, we first count the number of times the while loop in GREEDYADD-EXCHANGE is executed. The number of times the while loop runs is bounded above by the maximum number of iterations before the set $\mathcal{N}_t \setminus \mathcal{M}_t$ becomes empty. In each iteration either an addition or an exchange takes place. Since there is at most one addition that

can take place and $|\mathcal{N}_t \setminus \mathcal{M}_t|$ decreases by 1 whenever exchange-outs(i) of a product i reaches b, it follows that the while loop runs for at most $Nb+1$ iterations. In each iteration of the while loop, the revenue function is called at most $O(CN)$ times. Thus, GREEDYADD-EXCHANGE calls the revenue function at most $O(CbN^2)$ times. Since $\binom{N}{S} = O(N^S)$, we can now conclude that GREEDYOPT calls the revenue function $O(C^2bN^{S+2})$. The choice of S will depend on the accuracy of revenue estimates we have access to. Next, we provide guarantees on GREEDYOPT, which provide guidance on the choice of S.

4.2 Theoretical guarantees for GREEDYOPT

We now give a precise description of the main results we can establish for the GREEDYOPT algorithm. Specifically, suppose that the underlying choice model is an MNL model with weights $w_0 = 1$ for product 0 and w_i for product $i \in \mathcal{N}$; recall that the choice probabilities are given by

$$\mathbb{P}(i|\mathcal{M}) = \frac{w_i}{1 + \sum_{j \in \mathcal{M}} w_j}.$$

Note that 1 appears in the denominator because of the no-purchase option. In particular, the probability that an arriving customer leaves without purchasing anything when assortment $\mathcal{M}$ is on offer is given by

$$\mathbb{P}(0|\mathcal{M}) = \frac{1}{1 + \sum_{i \in \mathcal{M}} w_i}.$$

Let $R(\mathcal{M})$ denote the expected revenue from assortment $\mathcal{M}$. Under the MNL model, we have

$$R(\mathcal{M}) = \frac{\sum\limits_{i \in \mathcal{M}} p_i w_i}{1 + \sum\limits_{i \in \mathcal{M}} w_i},$$

where p_i is the price or the revenue obtained from the sale of product i.

We now have the following theorem when the revenue subroutine provides exact revenues:

Theorem 5. *Suppose the underlying model is the MNL model with weights $w_1, w_2, \ldots, w_N$ and the revenue subroutine provides exact revenues. Then, for any $S \geq 0$ and $b \geq C + 1$, the* GREEDYOPT *algorithm finds the optimal solution to* CAPACITATED OPT *problem.*

Therefore, taking $S = 0$ and $b = C + 1$, GREEDYOPT finds the optimal assortment of size at most C by calling the revenue function $O(N^2C^3)$. Thus, our algorithm provides a simpler alternative to the more complicated algorithm proposed by Rusmevichientong et al. [2010a].

We next show that the GREEDYOPT algorithm is robust to errors in the available revenue estimates. Specifically, we consider the more realistic setting where one has access to only approximate estimates of revenues i.e., we assume access to a function $\tilde{R}(\cdot)$ such that for any assortment $\mathcal{M}$ we have

$$(1 - \varepsilon(\mathcal{M}))R(\mathcal{M}) \leq \tilde{R}(\mathcal{M}) \leq R(\mathcal{M})$$

for some parameter $0 < \varepsilon(\mathcal{M}) < 1$. Naturally, the parameter $\varepsilon(\mathcal{M})$ determines the quality of revenue estimates we have available. Assuming that we have access to only approximate revenues, we find the optimal assortment by running GREEDYOPT with approximate revenues. In order to describe the result, we need some notation. For any assortment $\mathcal{M}$, let $w(\mathcal{M})$ denote $1 + \sum_{i \in \mathcal{M}} w_i$. Further, let

$$\varepsilon_{\max} \stackrel{\text{def}}{=} \max_{\mathcal{M}:\, |\mathcal{M}| \leq C} \varepsilon(\mathcal{M}) \quad \text{and} \quad W_C^{\max} \stackrel{\text{def}}{=} \max_{\mathcal{M}:\, |\mathcal{M}| \leq C} w(\mathcal{M}).$$

Finally, we defer to the next section the precise definitions of two quantities $\bar{C}(\delta_C)$ and δ_C that we need to describe the theorem; it suffices to say that as $\varepsilon_{\max} \to 0$, we have $\delta_C \to 0$ and $\bar{C}(\delta_C) \to C$.

With these definitions, we can now state our result.

Theorem 6. *Let M_C^{OPT} denote the optimal assortment of size at most C and $\hat{M}_C^{\text{OPT}}$ denote the estimate of the optimal assortment produced by* GREEDYOPT *when run*

with inputs $S \geq 0$ *and* $b \geq \bar{C}(2\delta_C) + 1$. *Then, we must have*

$$\frac{R(M_C^{\mathrm{OPT}}) - R(\hat{M}_C^{\mathrm{OPT}})}{R(M_C^{\mathrm{OPT}})} \leq f(w, \varepsilon_{\max}),$$

where w *denotes the vector of weights* $(w_1, w_2, \ldots, w_N)$ *and*

$$f(w, \varepsilon_{\max}) \stackrel{\text{def}}{=} \frac{W_C^{\max}}{w(M_C^{\mathrm{OPT}})} \eta(\varepsilon_{\max})$$

with $\eta(\varepsilon_{\max}) \stackrel{\text{def}}{=} 4C\varepsilon_{\max}/(1 - \varepsilon_{\max})$.

It is easy to see that the algorithm calls the revenue function $O(N^2C^2\bar{C}(2\delta_C))$ times. Note that as $\varepsilon_{\max} \to 0$, $\eta(\varepsilon_{\max})$ and hence $f(w, \varepsilon_{\max})$ go to zero. In addition, it follows from our definitions that as $\varepsilon_{\max} \to 0$, $\bar{C}(2\delta_C) \to C$. Consequently, taking the error in revenues $\varepsilon_{\max} = 0$ yields in Theorem 6 yields the result of Theorem 5 as the special result. Therefore, we only prove Theorem 6 in the next section.

4.3 Proofs of the main results

In this section we prove Theorem 6; specifically, we establish that the revenues of the optimal assortment and the estimate of the optimal assortment produced by GREEDYOPT are "close". In order to establish this result, for the rest of the section, fix a capacity C. Let M^{OPT} and $\hat{M}^{\mathrm{OPT}}$ respectively denote the optimal assortment and the estimate of the optimal assortment produced by GREEDYOPT. Then, our goal is to show that $R(M^{\mathrm{OPT}})$ and $R(\hat{M}^{\mathrm{OPT}})$ are "close" to each other. We assume that the underlying choice model is the MNL model with parameters $w_1, w_2, \ldots, w_N$. Recall that for any assortment $\mathcal{M}$,

$$R(\mathcal{M}) = \frac{\sum_{i \in \mathcal{M}} p_i w_i}{1 + \sum_{i \in \mathcal{M}} w_i},$$

where p_i is the price of product i. The term in the denominator makes comparison of the revenues of two different assortment difficult. Therefore, instead of dealing with

the revenues of the assortment directly, borrowing ideas from Rusmevichientong et al. [2010a], we consider the following transformation of the revenues of assortments: for any assortment $\mathcal{M}$ and number $u \in \mathbb{R}$,

$$R(\mathcal{M}) - u = \frac{\sum\limits_{i \in \mathcal{M}} p_i w_i}{1 + \sum\limits_{i \in \mathcal{M}} w_i} - u = \frac{\left(\sum\limits_{i \in \mathcal{M}} (p_i - u) w_i\right) - u}{1 + \sum\limits_{i \in \mathcal{M}} w_i} = \frac{H_{\mathcal{M}}(u) - u}{w(\mathcal{M})},$$

where $H_{\mathcal{M}} \colon \mathbb{R} \to \mathbb{R}$ is a function defined as $H_{\mathcal{M}}(u) = \sum_{i \in \mathcal{M}} (p_i - u) w_i$ and $w(\mathcal{M}) \stackrel{\text{def}}{=} 1 + \sum_{i \in \mathcal{M}} w_i$. We can now write

$$H_{\mathcal{M}}(u) = u + w(\mathcal{M})(R(\mathcal{M}) - u). \tag{4.1}$$

It is clear that $H_{\mathcal{M}}(\cdot)$ is directly related to the revenue $R(\mathcal{M})$. Moreover, as will become apparent soon, it is easier to compare the transformations $H_{\mathcal{M}_1}(\cdot)$ and $H_{\mathcal{M}_2}(\cdot)$ of two assortments $\mathcal{M}_1$ and $\mathcal{M}_2$ than their revenues $R(\mathcal{M}_1)$ and $R(\mathcal{M}_2)$. Specifically, we can establish the properties stated in the following proposition.

Proposition 1. *For any two assortments $\mathcal{M}_1$ and $\mathcal{M}_2$, let $H_1(\cdot)$ and $H_2(\cdot)$ respectively denote the functions $H_{\mathcal{M}_1}(\cdot)$ and $H_{\mathcal{M}_2}(\cdot)$. Further, let u_1 and u_2 denote the revenues $R(\mathcal{M}_1)$ and $R(\mathcal{M}_2)$ respectively. We then have*

1. $H_1(u_2) \geq H_2(u_2) \iff R(\mathcal{M}_1) \geq R(\mathcal{M}_2)$.

2. $H_1(u_2) \geq (1 + \delta(\mathcal{M}_1)) H_2(u_2) \implies \tilde{R}(\mathcal{M}_1) \geq \tilde{R}(\mathcal{M}_2)$,

where $\delta(\mathcal{M}_1) \stackrel{\text{def}}{=} \varepsilon(\mathcal{M}_1) w(\mathcal{M}_1) / (1 - \varepsilon(\mathcal{M}_1))$.

Proof. We prove each of the properties in turn. First note that for any assortment $\mathcal{M}$ with revenue $R(\mathcal{M}) = u$, it immediately follows from our definitions that $H_{\mathcal{M}}(u) = u + w(\mathcal{M})(R(\mathcal{M}) - u) = u$. The first property now follows from a straightforward

expansion of the terms involved:

$$\begin{aligned} H_1(u_2) \geq H_2(u_2) &\iff u_2 + w(\mathcal{M}_1)(u_1 - u_2) \geq u_2 \\ &\iff u_1 \geq u_2 \\ &\iff R(\mathcal{M}_1) \geq R(\mathcal{M}_2), \end{aligned}$$

where the second equivalence follows from the fact that $w(\mathcal{M}_1) > 0$. The second property can also be obtained through a similar straightforward expansion of the terms. In particular,

$$\begin{aligned} H_1(u_2) \geq (1+\varepsilon(\mathcal{M}_1))H_2(u_2) &\iff u_2 + w(\mathcal{M}_1)(u_1 - u_2) \geq (1+\delta(\mathcal{M}_1))u_2 \\ &\iff u_1 \geq \left(1 + \frac{\delta(\mathcal{M}_1)}{w(\mathcal{M}_1)}\right) u_2 \\ &\iff u_1 \geq \left(1 + \frac{\varepsilon(\mathcal{M}_1)}{1-\varepsilon(\mathcal{M}_1)}\right) u_2 \\ &\iff (1-\varepsilon(\mathcal{M}_1))u_1 \geq u_2, \end{aligned} \tag{4.2}$$

where the second equivalence follows from the definition of $\delta(\mathcal{M}_1)$. Moreover, it follows from our definitions that $\tilde{R}(\mathcal{M}_1) \geq (1-\varepsilon(\mathcal{M}_1))u_1$ and $u_2 \geq \tilde{R}(\mathcal{M}_2)$. We now conclude from (4.2) that

$$\tilde{R}(\mathcal{M}_1) \geq (1-\varepsilon(\mathcal{M}_1))u_1 \geq u_2 \geq \tilde{R}(\mathcal{M}_2).$$

The result of the proposition now follows. □

The above proposition establishes that if the transformation $H_{\mathcal{M}}(\cdot)$ of one assortment is "sufficiently" larger than the other, then it follows that the revenues of one assortment should be larger than the revenues of the other. Therefore, instead of keeping track of the revenues of the assortments in our algorithm, we keep track of their respective transformations $H_{\mathcal{M}}(\cdot)$.

Next, we establish a loop-invariance property that arises due to greedy additions and exchanges in our algorithms. We make use of this property to prove our theorems.

In order to state the proposition, we introduce the following notation:

$$\delta_C \stackrel{\text{def}}{=} \max_{\mathcal{M}:\,|\mathcal{M}|} \delta(\mathcal{M}) = \max_{\mathcal{M}:\,|\mathcal{M}|} w(\mathcal{M}) \frac{\varepsilon(\mathcal{M})}{1-\varepsilon(\mathcal{M})}.$$

We then have

Proposition 2. *Consider an iteration t of the while loop of the* GREEDYADD-EXCHANGE *algorithm. Let $\mathcal{M}_t$ and $\mathcal{M}_{t+1}$ denote the estimates of the optimal assortments at the beginning and the end of iteration t. Let $\mathcal{N}_t$ denote the universe of products at the beginning of iteration t. Then,*

1. *if a greedy exchange takes place i.e., $\mathcal{M}_{t+1} = (\mathcal{M}_t \setminus \{i^*\}) \cup \{j^*\}$, then for $u = R(\mathcal{M}_{t+1})$, we must have*

$$\begin{aligned} h_{i^*}(u) &\le h_i(u) + \delta_C u, && \text{for all } i \in \mathcal{M}_t \\ h_{j^*}(u) &\ge h_j(u) - \delta_C u, && \text{for all } j \in \mathcal{N}_t \setminus \mathcal{M}_t; \end{aligned}$$

2. *if an addition takes place i.e., $\mathcal{M}_{t+1} = \mathcal{M}_t \cup \{j^*\}$, then for $u = R(\mathcal{M}_{t+1})$ we must have*

$$h_{j^*}(u) \ge h_j(u) - \delta_C u, \quad \text{for all } j \in \mathcal{N}_t \setminus \mathcal{M}_t.$$

Proof. We prove this proposition by contradiction. First consider the case when exchange happens i.e., $\mathcal{M}_{t+1} = (\mathcal{M}_t \setminus \{i^*\}) \cup \{j^*\}$. Note that for any assortment $\mathcal{M} = (\mathcal{M}_t \setminus \{i\}) \cup \{j\}$ with $i \in \mathcal{M}_t$ and $j \in \mathcal{N}_t \setminus \mathcal{M}_t$, letting u denote $R(\mathcal{M}_{t+1})$, we can write

$$H_{\mathcal{M}}(u) - H_{\mathcal{M}_{t+1}}(u) = h_j(u) - h_{j^*}(u) + h_{i^*}(u) - h_i(u). \tag{4.3}$$

Now, if the hypothesis of the proposition pertaining to exchange is false, then at least one of the following should be true: either (1) there exists a product $i \in \mathcal{M}_t$ and $i \neq i^*$ such that $h_{i^*}(u) > h_i(u) + \delta_C u$, or (2) there exists a product $j \in \mathcal{N}_t \setminus \mathcal{M}_t$ and $j \neq j^*$ such that $h_{j^*}(u) < h_j(u) + \delta_C u$. In the first case when $h_{i^*}(u) > h_i(u) + \delta_C u$, by

taking $j = j^*$, we can write from (4.3) that $H_{\mathcal{M}}(u) - H_{\mathcal{M}_{t+1}}(u) > \delta_C u$. Similarly, in the second case when $h_{j^*}(u) < h_j(u) + \delta_C u$, by taking $i = i^*$, we can write from (4.3) that $H_{\mathcal{M}}(u) - H_{\mathcal{M}_{t+1}}(u) > \delta_C u$. Therefore, in both the cases, we have exhibited an assortment $\mathcal{M}$ distinct from $\mathcal{M}_{t+1}$ that can be obtained from $\mathcal{M}_t$ through an exchange and has the property that $H_{\mathcal{M}}(u) - H_{\mathcal{M}_{t+1}}(u) > \delta_C u$. We can now write

$$
\begin{aligned}
& H_{\mathcal{M}}(u) > H_{\mathcal{M}_{t+1}}(u) + \delta_C u && && \text{(4.4a)} \\
\Longrightarrow \quad & H_{\mathcal{M}}(u) > H_{\mathcal{M}_{t+1}}(u) + \delta(\mathcal{M}) u && \text{since } \delta_C \geq \delta(\mathcal{M}) \text{ by definition} && \text{(4.4b)} \\
\Longrightarrow \quad & H_{\mathcal{M}}(u) > (1 + \delta(\mathcal{M})) H_{\mathcal{M}_{t+1}}(u) && \text{since } H_{\mathcal{M}_{t+1}}(u) = u \text{ by definition} && \text{(4.4c)} \\
\Longrightarrow \quad & \tilde{R}(\mathcal{M}) > \tilde{R}(\mathcal{M}_{t+1}) && \text{by Proposition 1.} && \text{(4.4d)}
\end{aligned}
$$

This clearly contradicts the fact that $\mathcal{M}_{t+1}$ is chosen greedily.

The case when addition happens can be proved in the exact similar way. Particularly, suppose there exists a product $j \in \mathcal{N}_t \setminus \mathcal{M}_t$ and $j \neq j^*$ such that $h_{j^*}(u) < h_j(u) - \delta_C u$, where $u = R(\mathcal{M}_{t+1})$ with $\mathcal{M}_{t+1} = \mathcal{M}_t \cup \{j\}$. Letting $\mathcal{M}$ denote the set $\mathcal{M}_t \cup \{j\}$, we can then write

$$H_{\mathcal{M}}(u) - H_{\mathcal{M}_{t+1}}(u) = h_j(u) - h_{j^*}(u) > \delta_C u.$$

This implies – following the sequence of arguments in (4.4) – that $\tilde{R}(\mathcal{M}) > \tilde{R}(\mathcal{M}_t)$, contradicting the fact that $\mathcal{M}_{t+1}$ is chosen greedily.

The result of the proposition now follows. □

The above proposition establishes a key loop-invariance property that results from greedy additions and exchanges. Specifically, let u denote the revenue of the estimate of the optimal assortment obtained at the end of an iteration of the while loop in GREEDYADD-EXCHANGE. Then, the proposition establishes that whenever a product j^* is introduced (either through addition or an exchange-in) greedily, it must be that $h_{j^*}(u)$ is "close" to the maximum $h_j(u)$ of all products j that have been considered for an addition or exchange-in. Similarly, the product i^* that is greedily exchanged-out must be such that $h_{i^*}(u)$ is "close" to the minimum $h_i(u)$ of all products

i that have been considered for an exchange-out.

Using the propositions above, we can establish a key property of the subroutine GREEDYADD-EXCHANGE. For that, we need the following notation. For any u, define

$$B_S(u) \stackrel{\text{def}}{=} \arg \max_{\mathcal{M}:\, |\mathcal{M}| \le S} H_{\mathcal{M}}(u) = \arg \max_{\mathcal{M}:\, |\mathcal{M}| \le S} \sum_{i \in \mathcal{M}} h_i(u).$$

It is easy to see from the above definition that $B_S(u)$ consists of the top at most C products according to $h_i(u)$ such that $h_i(u) > 0$. Since $h_i(\cdot)$ is monotonically decreasing, it is easy to see that

$$|B_S(u_1)| \ge |B_S(u_2)|, \quad \text{whenever } u_1 \le u_2. \tag{4.5}$$

Under appropriate technical assumptions, Rusmevichientong et al. [2010a] showed that for any $1 \le S \le N$, the optimal assortment of size at most S under the MNL model is one of the assortments in the collection $\mathcal{B}_S \stackrel{\text{def}}{=} \{B_S(u)\colon u \in \mathbb{R}\}$. In fact the authors show that if u_S denotes the optimal revenue, then $B_S(u_S)$ is the optimal assortment. An immediate consequence of this result and (4.5) is that for any $u \le u_S$

$$S \ge |B_S(u)| \ge |M_S^{\text{OPT}}|. \tag{4.6}$$

It has been established by Rusmevichientong et al. [2010a] that there can be at most $O(NC)$ distinct assortments in the collection $\mathcal{B}_S$ allowing one to find the optimal assortment by restricting one's search to $O(NC)$ assortments. The following lemma shows that the assortment found by the subroutine GREEDYADD-EXCHANGE is "close" to one of the assortments in $\mathcal{B}_S$. Before we describe the lemma, we need the following notation. For any $\delta > 0$ and $u \in \mathbb{R}$, let

$$i_S(u) \stackrel{\text{def}}{=} \min_{i \in B_S(u)} h_i(u).$$

Moreover, let

$$\bar{B}_S(\delta, u) \stackrel{\text{def}}{=} B_S(u) \cup \left\{ j \in \mathcal{N} \setminus B_S(u)\colon h_{i_S(u)}(u) - h_j(u) \le \delta u \right\},$$

Also, let

$$\bar{C}(\delta) \stackrel{\text{def}}{=} \max_{u \in \mathbb{R}_+} |\bar{B}_S(\delta, u)|,$$

We then have

Lemma 2. *Suppose* GREEDYADD-EXCHANGE *is run with some input assortment* $\mathcal{M}$ *and* $b \geq \bar{C}(\delta_C)+2$, *where* $C \geq S+1$. *Further, suppose that* $|M_{S+1}^{\text{OPT}}| = S+1$. *Then, there exists an iteration* t^* *of the while loop such that if* $\mathcal{M}^*$ *denotes the assortment* $\mathcal{M}_{t^*+1}$ *and* u^* *denotes* $R(\mathcal{M}^*)$, *then*

$$H_{B(u^*)}(u^*) - H_{\mathcal{M}^*}(u^*) \leq 2\tilde{C}_{u^*}\delta_C u^*,$$

where $B(u^*)$ *denotes the assortment* $B_{S+1}(u^*)$ *and* $\tilde{C}^*$ *is a constant denoting* $1 + |B(u^*) \setminus \mathcal{M}^*|$.

We defer the proof of Lemma 2 to the end of the section. We now present the proof of Theorem 6.

4.3.1 Proof of Theorem 6

Let M_C^{OPT} denote the true optimal assortment, and $\hat{M}_C^{\text{OPT}}$ denote the estimate of the optimal assortment produced by GREEDYOPT. Furthermore, let $C^* \leq C$ denote the size of M_C^{OPT}. It follows from Lemma 2 that in the C^*th invocation of the subroutine GREEDYADD-EXCHANGE, there exists an assortment $\mathcal{M}^*$ such that $\tilde{R}(\hat{M}_C^{\text{OPT}}) > \tilde{R}(\mathcal{M}^*)$ and $\mathcal{M}^*$ is such that

$$H_{B(u^*)}(u^*) - H_{\mathcal{M}^*}(u^*) \leq 2\tilde{C}_{u^*}\delta_C u^*,$$

where $\tilde{C}^*$ denotes $|B(u^*) \setminus \mathcal{M}^*| + 1$ and $B(u^*)$ denotes the set $B_{C^*}(u^*)$. It follows by the definition of $B(u^*)$ that $H_{B(u^*)}(u^*) \geq H_{M_C^{\text{OPT}}}(u^*)$. Thus, we can write

$$H_{M_C^{\text{OPT}}}(u^*) - H_{\mathcal{M}^*}(u^*) \leq 2\tilde{C}_{u^*}\delta_C u^* \leq 2C\delta_C u^*. \tag{4.7}$$

Let u_C denote $R(M_C^{\mathrm{OPT}})$. Then, it follows by definition that $H_{M_C^{\mathrm{OPT}}}(u_C) = u_C$. Thus,

$$H_{M_C^{\mathrm{OPT}}}(u_C) - H_{M_C^{\mathrm{OPT}}}(u^*) = \sum_{j \in M_C^{\mathrm{OPT}}} w_j(u^* - u_C) = (u^* - u_C)(w(M_C^{\mathrm{OPT}}) - 1).$$

Since $H_{M_C^{\mathrm{OPT}}}(u_C) = u_C$, we can write

$$H_{M_C^{\mathrm{OPT}}}(u^*) = u_C + (u_C - u^*)(w(M_C^{\mathrm{OPT}}) - 1). \tag{4.8}$$

Since $H_{\mathcal{M}^*}(u^*) = u^*$, it now follows from (4.7) and (4.8) that

$$\begin{aligned} & (u_C - u^*)(w(M_C^{\mathrm{OPT}}) - 1) + u_C - u^* \leq 2C\delta_C u^* \\ \Longrightarrow \quad & (u_C - u^*)w(M_C^{\mathrm{OPT}}) \leq 2C\delta_C u^* \\ \Longrightarrow \quad & u_C \leq (1 + \tilde{\varepsilon})u^*, \end{aligned} \tag{4.9}$$

where $\tilde{\varepsilon} \stackrel{\mathrm{def}}{=} 2C\delta_C / w(M_C^{\mathrm{OPT}})$. Now since $\tilde{R}(\hat{M}_C^{\mathrm{OPT}}) > \tilde{R}(\mathcal{M}^*)$, it follows that

$$(1 - \varepsilon(\mathcal{M}^*))u^* \leq \tilde{R}(\mathcal{M}^*) < \tilde{R}(\hat{M}^{\mathrm{OPT}}) \leq \hat{u}_C,$$

where $\hat{u}_C$ denotes $R(\hat{M}_C^{\mathrm{OPT}})$. It now follows from (4.9) that

$$u_C \leq (1 + \tilde{\varepsilon})u^* \leq \frac{1 + \tilde{\varepsilon}}{1 - \varepsilon(\mathcal{M}^*)}\hat{u}_C.$$

Now since

$$\delta_C = \max_{\mathcal{M}:\, |\mathcal{M}| \leq C} \frac{\varepsilon(\mathcal{M})}{1 - \varepsilon(\mathcal{M})} w(\mathcal{M}),$$

by letting $\varepsilon_{\max} = \max_{\mathcal{M}:\, |\mathcal{M}| \leq C} \varepsilon(\mathcal{M})$ and $W_C^{\max} = \max_{\mathcal{M}:\, |\mathcal{M}| \leq C} w(\mathcal{M})$, we have

$$\delta_C \leq \frac{\varepsilon_{\max}}{1 - \varepsilon_{\max}} W_C^{\max}.$$

Thus,

$$\tilde{\varepsilon} = \frac{2C}{w(M_C^{\mathrm{OPT}})}\delta_C \leq \frac{2C}{w(M_C^{\mathrm{OPT}})} \frac{\varepsilon_{\max}}{1 - \varepsilon_{\max}} W_C^{\max} \stackrel{\mathrm{def}}{=} f(w, \varepsilon_{\max})/2.$$

With these definitions, it is easy to see that $\varepsilon(\mathcal{M}^*) \leq \varepsilon_{\max} \leq f(w, \varepsilon_{\max})/2$. It now follows that

$$\frac{u_C - \hat{u}_C}{u_C} \leq 1 - \frac{1 - \varepsilon(\mathcal{M}^*)}{1 + \tilde{\varepsilon}} \leq \frac{\tilde{\varepsilon} + \varepsilon(\mathcal{M}^*)}{1 + \tilde{\varepsilon}} \leq \varepsilon(\mathcal{M}^*) + \tilde{\varepsilon} \leq f(w, \varepsilon_{\max}).$$

This establishes the result of the theorem.

4.3.2 Proof of Lemma 2

Suppose the while loop in the subroutine terminates at the end of iteration T. Then, it follows from the description of the subroutine that at least one of the following conditions holds at the end of iteration T:

1. The set of products $\mathcal{N}_{T+1} \setminus \mathcal{M}_{T+1}$ available for additions or exchanges is empty.
2. No further additions or exchanges can increase the revenues.

Our goal is to prove the existence of an iteration $t^* \leq T$ such that

$$H_{B(u^*)}(u^*) - H_{\mathcal{M}^*}(u^*) \leq 2\tilde{C}_{u^*}\delta_C u^*,$$

where $\mathcal{M}^*$ denotes the assortment $\mathcal{M}_{t^*+1}$ and u^* denotes $R(\mathcal{M}^*)$. We prove this by considering two cases corresponding to each of the two ways in which the subroutine terminates. Note that in order to simplify the notation, we have dropped the subscript from the notation of $B_{S+1}(\cdot)$.

Case 1: Subroutine terminates with $\mathcal{N}_{T+1} = \mathcal{M}_{T+1}$. We first consider the case when the subroutine terminates when the set of products $\mathcal{N}_{T+1} \setminus \mathcal{M}_{T+1}$ becomes empty. In this case, we prove the existence of an iteration $t^* \leq T$ that satisfies the condition stated in the hypothesis of the lemma. In fact, we prove something stronger; we shall show that the iteration $t^* \leq T^*$, where $T^* \leq T$ is the first iteration such that $\mathcal{N}_{T^*} \subset \mathcal{N}$ (recall that $\mathcal{N}_1 = \mathcal{N}$). We prove this result by contradiction. In particular, suppose that after every iteration $t \leq T^*$ of the while loop, we have

$$H_{B(u)}(u) - H_{\mathcal{M}_{t+1}}(u) > 2\tilde{C}_u\delta_C u, \tag{4.10}$$

where u denotes the revenue $R(\mathcal{M}_{t+1})$ and $\tilde{C}_u$ denotes the constant $1+|B(u)\setminus\mathcal{M}_{t+1}|$. Note that a product i would be removed from the universe $\mathcal{N}_t$ at the end of some iteration t only if it has been exchanged-out b times. Since $b \geq \bar{C}(\delta_C)$, it is easy to see that we arrive at a contradiction if we show that as long (4.10) is satisfied at the end of each iteration, each product i can be exchanged-out at most $\bar{C}(\delta_C)+2$ times.

In order to bound the number of times a product can be exchanged-out, we establish a special property that should be satisfied whenever an exchange happens. Specifically, suppose an exchange happens during iteration t i.e., $\mathcal{M}_{t+1} = (\mathcal{M}_t \setminus \{i^*\}) \cup \{j^*\}$. In addition, let u denote the revenue $R(\mathcal{M}_{t+1})$, and let product $k^* \in \mathcal{N}_t \setminus \mathcal{M}_t$ denote the product such that $h_{k^*}(u) \geq h_k(u)$ for all products $k \in \mathcal{N}_t \setminus \mathcal{M}_t$. Then, we claim that

$$h_{j^*}(u) \geq h_{k^*}(u) - \delta_C u \tag{4.11a}$$

$$h_{i^*}(u) \leq h_{k^*}(u) - \delta_C u. \tag{4.11b}$$

We prove this claim as follows. Since $k^* \in \mathcal{N}_t \setminus \mathcal{M}_t$, (4.11a) follows directly from Proposition 2. We now argue that $h_{i^*}(u) \leq h_{k^*}(u) - \delta_C u$. For that, we first note that

$$h_{i^*}(u) - h_{j^*}(u) \leq 2\delta_C u. \tag{4.12}$$

To see why, note that since an exchange has happened, it must be that $\tilde{R}(\mathcal{M}_t) \leq \tilde{R}(\mathcal{M}_{t+1})$. This implies by Proposition 1 that $H_{\mathcal{M}_t}(u) \leq (1+\delta(\mathcal{M}_t))H_{\mathcal{M}_{t+1}}(u)$. Since $\delta(\mathcal{M}_1) \leq \delta_C$ and $H_{\mathcal{M}_{t+1}}(u) = u$ by definition, we can write

$$\begin{aligned} H_{\mathcal{M}_t}(u) \leq (1+\delta(\mathcal{M}_t))H_{\mathcal{M}_{t+1}}(u) &\implies H_{\mathcal{M}_t}(u) - H_{\mathcal{M}_{t+1}}(u) \leq \delta_C u \\ &\implies h_{i^*}(u) - h_{j^*}(u) \leq \delta_C u < 2\delta_C u. \end{aligned}$$

Now, consider

$$H_{B(u)}(u) - H_{\mathcal{M}_{t+1}}(u) = H_{B(u)}(u) - H_{\mathcal{M}_t}(u) + H_{\mathcal{M}_t}(u) - H_{\mathcal{M}_{t+1}}(u)$$
$$= \sum_{j \in B(u) \setminus \mathcal{M}_t} h_j(u) - \sum_{i \in \mathcal{M}_t \setminus B(u)} h_i(u) + (h_{i^*}(u) - h_{j^*}(u)) .$$

We now collect terms in the above expression as follows. Let $\mathcal{M}_1$ denote the set $\mathcal{M}_t \setminus B(u)$. Further, partition the set $B(u) \setminus \mathcal{M}_t$ into $Mscr_2 \cup \mathcal{M}_3$ such that $\mathcal{M}_2 \cap \mathcal{M}_3 = \emptyset$ and $|\mathcal{M}_2| = |\mathcal{M}_1|$; note that such a partitioning is possible because $|B(u)| = S+1$ (which follows from (4.6) and the hypothesis that $|M_{S+1}^{\text{OPT}} = S+1|$) and $|\mathcal{M}_t| \leq S+1$. Also note that $\mathcal{M}_3 \neq \emptyset$ if and only if $|\mathcal{M}_t| < S+1$. With this partitioning, we can now write

$$H_{B(u)}(u) - H_{\mathcal{M}_{t+1}}(u) = \sum_{i \in \mathcal{M}_1, j \in \mathcal{M}_2} (h_j(u) - h_i(u)) + \sum_{j \in \mathcal{M}_3} h_j(u) + (h_{i^*}(u) - h_{j^*}(u)) .$$

We now claim that at least on of the following must be true: either (1) there exists a pair of products $i \in \mathcal{M}_1$ and $j \in \mathcal{M}_2$ such that $h_j(u) - h_i(u) > 2\delta_C u$, or (2) if $\mathcal{M}_3 \neq \emptyset$, then there exists a product $k \in \mathcal{M}_3$ such that $h_3(u) > 2\delta_C u$. Otherwise, it is easy to see from (4.12) that $H_{B(u)}(u) - H_{\mathcal{M}_{t+1}}(u) \leq 2\tilde{C}_u \delta_C u$, where $\tilde{C}_u = |B(u) \setminus \mathcal{M}_{t+1}| + 1$, contradicting (4.10). We now consider each of the cases in turn.

First suppose that $h_j(u) - h_i(u) > 2\delta_C u$ for some $i \in \mathcal{M}_1$ and $j \in \mathcal{M}_2$. It follows from Proposition 2 that $h_{i^*}(u) \leq h_i(u) + \delta_C u$. Thus, we can write

$$h_{i^*}(u) \leq h_i(u) + \delta_C u < h_j(u) - 2\delta_C u + \delta_C u \leq h_{k^*}(u) - \delta_C u,$$

where the last inequality follows from the definition of k^* and the fact that $j \in \mathcal{M}_2 \subset \mathcal{N} \setminus \mathcal{M}_t$. Thus, for this case, we have established (4.11b).

Now suppose that $\mathcal{M}_3 \neq \emptyset$ and $h_k(u) > 2\delta_C u$ for some $k \in \mathcal{M}_3$. As noted above, in this case, we should have $|\mathcal{M}_{t+1}| < S+1$. This means that an exchange has happened instead of addition, which in turn implies that $\tilde{R}(\tilde{\mathcal{M}}) \leq \tilde{R}(\mathcal{M}_{t+1})$, where

$\tilde{\mathcal{M}}$ denotes the set $\mathcal{M}_t \cup \{k\}$. Thus, by Proposition 1, we should have

$$
\begin{aligned}
& H_{\tilde{\mathcal{M}}}(u) \leq (1+\delta(\tilde{\mathcal{M}}))H_{\mathcal{M}_{t+1}}(u) \\
\implies \quad & H_{\tilde{\mathcal{M}}}(u) - H_{\mathcal{M}_{t+1}}(u) \leq \delta(\tilde{\mathcal{M}})H_{\mathcal{M}_{t+1}}(u) \\
\implies \quad & h_k(u) + h_{i^*}(u) - h_{j^*}(u) \leq \delta_C u \quad \text{as } H_{\mathcal{M}_{t+1}}(u) = u, \delta(\tilde{\mathcal{M}}) \leq \delta_C \\
\implies \quad & h_{i^*}(u) \leq h_{j^*}(u) - h_k(u) + \delta_C u \\
\implies \quad & h_{i^*}(u) \leq h_{j^*}(u) - 2\delta_C u + \delta_C u \qquad \text{since } h_k(u) > 2\delta_C u \\
\implies \quad & h_{i^*}(u) \leq h_{k^*}(u) - \delta_C u \qquad \text{since } h_{j^*}(u) \leq h_{k^*}(u).
\end{aligned}
$$

We have thus established that $h_{i^*}(u) \leq h_{k^*}(u) - \delta_C u$ for both the cases.

We now use (4.11) to bound the number of exchange-outs that can happen for each product. Specifically, as mentioned above, we arrive at a contradiction by showing that each product can be exchanged-out at most $\bar{C}(\delta_C) + 2$ times. For that, for any iteration $t \leq T^*$, let k_t denote the product such that $k_t \in \mathcal{N}_t \setminus \mathcal{M}_t$ and $h_{k_t}(u_{t+1}) \geq h_j(u_{t+1})$ for all products $j \in \mathcal{N}_t \backslash Mscr_t$ and $u_{t+1} = R(\mathcal{M}_{t+1})$. Now define the function

$$
g(u) = \begin{cases} h_{k_t}(u) - \delta_C u & \text{for } u_t < u \leq u_{t+1}, t \leq T^*, \\ h_{k_1}(u_1) - \delta_C u_1 & \text{for } u = u_1. \end{cases}
$$

Note that for the above definition to be meaningful, for any $t \leq T^*$, we need to show that $u_t \leq u_{t+1}$. This should be true because by (4.11), it follows that for $u = R(\mathcal{M}_{t+1})$, we have $h_{i^*}(u) \leq h_{j^*}(u)$; this in turn implies that $H_{\mathcal{M}_t}(u) \leq H_{\mathcal{M}_{t+1}}(u)$, which implies by Proposition 1 that $u_t = R(\mathcal{M}_t) \leq R(\mathcal{M}_{t+1}) = u_{t+1}$. It is easy to see that the function $g(\cdot)$ is piecewise linear. However, note that it may not be continuous.

Now fix a product i, and for this product we argue that it can be exchanged at most $\bar{C}(\delta_C)$ times. For that let t_1 be an iteration in which i is exchanged-out and t_2 be the first iteration after t_1 when i is exchanged-in. Let u_1, u_2 denote $R(\mathcal{M}_{t_1+1})$ and $R(\mathcal{M}_{t_2+1})$ respectively. Furthermore, let k_1 and k_2 respectively denote the products

k_{t_1} and k_{t_2}. It now follows from (4.11) that

$$h_i(u_1) \leq h_{k_1}(u_1) - \delta_C u_1 = g(u_1)$$
$$h_i(u_2) \geq h_{k_2}(u_2) - \delta_C u_2 = g(u_2).$$

This implies that the line $h_i(\cdot)$ is below $g(\cdot)$ at u_1 and above $g(\cdot)$ at u_2. We now argue that $h_i(\cdot)$ intersects $g(\cdot)$ at some $u_1 \leq u \leq u_2$ i.e., $h_i(u) = g(u)$. If $g(\cdot)$ were continuous, this assertion would immediately follow from the intermediate value theorem. However, the way we have defined $g(\cdot)$, it may be discontinuous at some u_t with $t_1 < t \leq t_2$. Now the only way $h_i(\cdot)$ and $g(\cdot)$ do not intersect is if for some $t_1 < t \leq t_2$,

$$g(u_t^-) < h_i(u_t) < g(u_t^+) \quad \text{and} \quad h_i(u) > g(u) \text{ for } u_t \leq u \leq u_2.$$

We argue that this cannot happen. For that consider iteration t. By definition $i \notin \mathcal{M}_t$. Since $\mathcal{N}_t = \mathcal{N}$, it follows by our definition that $h_{k_t}(u_{t+1}) \geq h_i(u_{t+1})$, which in turn implies that $g(u_{t+1}) \geq h_i(u_{t+1})$ resulting in a contradiction. Thus, $h_i(\cdot)$ intersects $g(\cdot)$ from below at some u such that $u_1 \leq u \leq u_2$.

Hence, we can correspond each exchange-out with an intersection point corresponding to $h_i(\cdot)$ intersecting $g(\cdot)$ from below. This implies that the total number of exchage-outs can be bounded above by one plus the number of times $h_i(\cdot)$ intersects $g(\cdot)$ from below beyond u_i, where u_i is the revenue of the assortment $\mathcal{M}_t$ immediately after i is added to it (either through an exchange-in or addition). Note that $h_i(\cdot)$ intersects $g(\cdot)$ at $u \geq u_i$ if and only if $w_i \leq w_{k(u)}$ and $h_{k(u)}(u_i) \geq h_i(u_i)$, where $k(u)$ is the product such that $k(u) = k_t$, where $u_t < u \leq u_{t+1}$. Thus, the number of intersection points can be bounded above by the number of products k such that $h_k(u_i) \geq h_i(u_i)$. We now argue that $i \in \bar{B}_{S+1}(\delta_C, u_i)$. If this is true, then it implies that there can be at most $|\bar{B}_{S+1}(\delta_C, u_i)| \leq \bar{C}(\delta_C)$ intersection points, which immediately implies that there can be at most $1 + \bar{C}(\delta_C)$ exchange-outs.

The only thing we are left with is to argue that $i \in \bar{B}_{S+1}(\delta_C, u_i)$. To see this, let $\tilde{\mathcal{M}}$ be the assortment obtained after i is added or exhanged-in for the first time. Then, according to our definition, we have that $u_i = R(\tilde{\mathcal{M}})$. Further, since $H_{B(u_i)}(u_i) - H_{\tilde{\mathcal{M}}}(u_i) > 0$, there exists a product $k \in B(u_i) \setminus \tilde{\mathcal{M}}$. It now follows by Proposition 2 that

$$h_i(u_i) \geq h_k(u_i) - \delta_C u_i \geq h_{i_{S+1}(u_i)} - \delta_C u_i,$$

where $i_{S+1}(u_i)$ is as defined above i.e., $i_{S+1}(u_i) \stackrel{\text{def}}{=} \arg\min_{j \in B(u_i)} h_j(u_i)$. It now follows by the definition of $\bar{B}_{S+1}(\delta_C, u_i)$ that $i \in \bar{B}_{S+1}(\delta_C, u_i)$.

Case 2: Subroutine terminates because no further additions or exchanges increase revenue. We now consider the case when subroutine terminates at iteration T because no further additions or exchanges increase the revenue. Now there are two possibilities: either $\mathcal{N}_t = \mathcal{N}$ for all $t \leq T$ or not. In the latter case let T^* be the first iteration t when $\mathcal{N}_t \subset \mathcal{N}$. It then follows from our arguments for the above case that there exists an iteration $t^* \leq T^*$ that satisfies the properties of the lemma. Thus, we consider the case when $\mathcal{N}_t = \mathcal{N}$ for all $t \leq T^*$. Assuming this, we prove the result by contradiction. In particular, suppose at the end of iteration T we have

$$H_{B(u)}(u) - H_{\mathcal{M}_{T+1}}(u) \geq 2\tilde{C}_u \delta_C u, \tag{4.13}$$

Now consider

$$H_{B(u)}(u) - H_{\mathcal{M}_{T+1}}(u) = \sum_{k \in \mathcal{M}_3} h_j(u) + \sum_{i \in \mathcal{M}_1, j \in \mathcal{M}_2} (h_j(u) - h_i(u)),$$

where as above, $\mathcal{M}_1$ denotes the assortment $\mathcal{M}_{T+1} \setminus B(u)$ and the set $B(u) \setminus \mathcal{M}_{T+1}$ is partitioned into $\mathcal{M}_2 \cup \mathcal{M}_3$ such that $\mathcal{M}_2 \cap \mathcal{M}_3 = \emptyset$ and $|\mathcal{M}_2| = |\mathcal{M}_1|$; such a partitioning is possible since $|B(u)| = S + 1$ (which follows from (4.6) and the hypothesis that $|M_{S+1}^{\text{OPT}} = S + 1|$) and $|\mathcal{M}_{T+1}| \leq S + 1$. It now follows that one of the following conditions should hold: either (1) there exists a pair of products $i \in \mathcal{M}_1$ and $j \in \mathcal{M}_2$ such that $h_j(u) - h_i(u) > 2\delta_C u$, or (2) if $\mathcal{M}_3 \neq \emptyset$, then there exists a product $k \in \mathcal{M}_3$ such that $h_3(u) > 2\delta_C u$. Otherwise, it is easy to see that

$H_{B(u)}(u) - H_{\mathcal{M}_{T+1}}(u) \leq 2\tilde{C}_u \delta_C u$, where $\tilde{C}_u = |B(u) \setminus \mathcal{M}_{T+1}| + 1$, contradicting (4.13). We consider each of the cases in turn.

First, suppose that there exist a pair of products $i \in \mathcal{M}_1$ and $j \in \mathcal{M}_2$ such that $h_j(u) - h_i(u) > 2\delta_C u$. Let $\tilde{\mathcal{M}}$ denote the assortment $(\mathcal{M}_{T+1} \setminus \{i\}) \cup \{j\}$. We can then write

$$H_{\tilde{\mathcal{M}}}(u) - H_{\mathcal{M}_{T+1}}(u) = h_j(u) - h_i(u) > 2\delta_C u.$$

Since $H_{\mathcal{M}_{T+1}}(u) = u$ and $\delta_C \geq \delta(\tilde{\mathcal{M}})$, it follows that by Proposition 1 that $\tilde{R}\tilde{\mathcal{M}} > \tilde{R}\mathcal{M}_{T+1}$. This contradicts the assumption that the subroutine terminates with $\mathcal{M}_{T+1}$ because no further additions or exchanges result in an increase of revenue.

Next, suppose $\mathcal{M}_3 \neq \emptyset$ and $h_k(u) > 2\delta_C u$ for some $k \in \mathcal{M}_3$. Now let $\tilde{\mathcal{M}} = \mathcal{M}_{T+1} \cup \{k\}$; note that since $\mathcal{M}_3 \neq \emptyset$, it must be that $|\mathcal{M}_{T+1}| = S$. We can now write

$$H_{\tilde{\mathcal{M}}}(u) - H_{\mathcal{M}_{T+1}}(u) = h_k(u) > 2\delta_C u.$$

Since $H_{\mathcal{M}_{T+1}}(u) = u$ and $\delta_C \geq \delta(\tilde{\mathcal{M}})$, it follows that by Proposition 1 that $\tilde{R}\tilde{\mathcal{M}} > \tilde{R}\mathcal{M}_{T+1}$. This contradicts the assumption that the subroutine terminates with $\mathcal{M}_{T+1}$ because no further additions or exchanges result in an increase of revenue. This finishes the proof of this case.

The proof of the lemma now follows.

4.4 Chapter summary and discussion

This chapter continued the discussion of using choice models to make decisions. Assuming that we have access to a revenue prediction subroutine, we designed an algorithm to find an approximation of the optimal assortment with as few calls to the revenue subroutine as possible.

We designed a general algorithm for the optimization of set-functions to solve the static assortment optimization algorithms. Most existing algorithms (both exact and approximate) heavily exploit the structure of the assumed choice model; consequently, the existing algorithms – even without any guarantees – cannot be used with other

choice models like the probit model or the mixture of MNL models with a continuous mixture. Given these issues, we designed an algorithm that is (a) not tailored to specific parametric structures and (b) requires only a subroutine that gives revenue estimates for assortments. Our algorithm is a sophisticated form of greedy algorithm, where the solution is constructed from a smaller assortment through greedy additions and exchanges. The algorithm is proved to find the optimal assortment exactly when the underlying choice model is the MNL model. We also showed that the algorithm is robust to errors in the revenue estimates provided by the revenue subroutine, as long as the underlying choice model is the MNL model.

Note that the focus of the current and previous chapters has been decision making using choice models. The rationale here is that such decision problems appear in several important practical applications, and the performance of the decisions can have a huge impact on the revenues. In this context, we have attempted to solve the decision problem directly, avoiding the unnecessary step of actually learning the underlying choice model. Interestingly, a question we haven't explicitly addressed here identification of a particular choice model itself. Given that the data allows us to only identify a family of consistent distributions, we need a criterion to pick one of the distributions. Ideally, we want a criterion that is independent of the decision context. A criterion we briefly broached in the context of characterizing the choice model used for revenue predictions is that of sparsity. Sparsity is an appealing notion and has recently found a lot of success in the area of high-dimensional statistics Donoho [2006]. The next chapter deals with the sparsity as a criterion to pick choice models and the rich theory associated with it.

Chapter 5

Learning choice models

In the problems discussed in the previous chapter so far, our ultimate goal was to use a choice model to make a decision. As a result, we avoided the unnecessary indirection of actually learning the underlying choice model and developed methods to directly solve the decision problem. However, as elaborated below, explicit learning of choice models is required in many important applications.

For instance, consider the problem of 'customer segmentation.' It is central to several important applications and involves segmenting the customer population into groups of individuals that have similar preferences. A statement to the extent "your customers are mainly of 3 types, and their preferences are described by these preference lists" is of high value in these contexts. One way to usefully segment customers is to learn a choice model over K preference lists and then segment the customer population into K classes with the preferences of each class described by one of the K learned preference lists. Such segmentation is especially crucial to applications that need effective targeting of resources. The classical application that heavily uses customer segmentation is marketing, where it has long since been known that the marketing strategy needs to be effectively targeted to the specific customer type. Interestingly, a non-traditional application area that also relies on customer segmentation is the area of the recently popular recommendation/discovery systems, where the recommendations (be it movies, books, or news stories) need to be tailored in a useful way to the specific customer types.

In addition to applications related to OM/RM, another broad class of problems where learning distributions over preference lists becomes important is 'rank aggregation'. As mentioned above, this is an important problem that arises in various contexts like web-search, polling, betting, and elections, in which the goal essentially is to come up with a final ranking given some partial preference information. The rank-aggregation problem has been studied extensively in the area of social choice theory, where extensive work has been done to determine the "right" final ranking given access to the entire distribution over rankings. Of course, in most practical applications, such a distribution is not readily available. What is readily available though is partial information about the distribution: for instance, in the context of web-search, clicks give information about which document is preferred from a set of documents that were shown; similarly, in the context of polling, one may have access to pairwise comparison information (see was and wfn). Given this, a reasonable approach to aggregating rankings is to learn a distribution over rankings that captures the underlying choice structure from marginal preference information and use any of the several methods developed in the social choice literature for aggregation.

Finally, there is a host of other applications in which distributions over rankings are compressed in order to store efficiently[1] by retaining only partial information (typically in the form of a subset of Fourier coefficients). For instance, an important application, which has received a lot of attention recently, is the *Identity Management Problem* or the *Multi-object tracking problem.* This problem is motivated by applications in air traffic control and sensor networks, where the goal is to track the identities of N objects from noisy measurements of identities and positions. Specifically, consider an area with sensors deployed that can identify the unique signature and the position associated with each object when it passes close to it. Let the objects be labeled $1, 2, \ldots, N$ and let $x(t) = (x_1(t), x_2(t), \ldots, x_N(t))$ denote the vector of positions of the N objects at time t. Whenever a sensor registers the signature of an object the vector $x(t)$ is updated. A problem, however, arises when two objects, say

[1]Such compression is indeed a necessity given that distributions over rankings have a factorial (in N) blow-up.

i, j, pass close to a sensor simultaneously. Because the sensors are inexpensive, they tend to confuse the signatures of the two objects; thus, after the two objects pass, the sensor has information about the positions of the objects, but it only has beliefs about which position belongs to which object. This problem is typically modeled as a probability distribution over permutations, where, given a position vector $x(t)$, a permutation σ of $1, 2, \ldots, N$ describes the assignment of the positions to objects. Because the measurements are noisy, to each position vector $x(t)$, we assign, not a single permutation, but a distribution over permutations. Since we now have a distribution over permutations, the factorial blow-up makes it challenging to maintain it. Thus, it is often approximated using a partial set of Fourier coefficients. Recent work by Huang et al. [2008], Kondor et al. [2007] deals with updating the distribution with new observations in the Fourier domain. In order to obtain the final beliefs one has to recover the distribution over permutations from a partial set of Fourier coefficients.

In summary, there is a wide-range of important applications that need learning of the underlying choice model given marginal information. Now, as explained in the previous chapters, given marginal information, we can identify a family of choice models that are consistent with the given information; the family is almost certainly not a singleton because the data is insufficient to specify a unique distribution. Therefore, the problem of learning the choice model reduces to the problem of finding an appropriate criterion to select one of the models consistent with the available data. Now, a popular statistical criterion for model selection that has been extensively used in many contexts is the criterion of *parsimony*, which encourages the selection of the 'most parsimonious' model that is consistent with the data.

The criterion of parsimony is justified in many ways. Philosophically speaking, this criterion is consistent with the *Occam's razor* philosophy, which roughly stated, suggests that under the absence of additional information, one should tend toward 'simpler' theories. Statistically speaking, parsimony is born out of the need not to over-fit. Finally, operationally speaking, parsimony is desired because parsimonious models are easier to handle in practice – both computationally and otherwise. Of course, parsimony is nuanced idea and it is not straightforward to operationalize the

criterion. In parametric models, parsimony has often been translated into parsimony of parameters; for instance, an MNL model with N parameters can be considered 'more parsimonious' than an exponential family with N^2 parameters. In the nonparametric case, however, the sparsity (or the support size) of the distribution becomes a natural candidate for a measure of parsimony of the model. In addition, as described above in the context of choice models used for revenue predictions, sparsity is also naturally born out of the fact that only marginal information available; more precisely, a distribution of sparsity no more than[2] $(N-1)^2+1$ is needed to describe the first-order information, which captures the probability that i is ranked at position r for all i and r. Finally, sparse models have found immense success (in both theory and practice) in the area of *compressive sensing*, which has gained recent popularity in the areas of signal processing, coding theory, and streaming algorithms (see Donoho [2006], Candes et al. [2006b,a]).

Given the considerations above, we propose to recover the underlying choice model by identifying the sparsest distribution that is consistent with the available data. From an operational perspective, two main questions arise at this point: (1) how does one find the sparsest consistent distribution in an efficient manner? and (2) how "good" are sparse models in practice? In addition, from a theoretical standpoint, the question about the discriminative power of the sparsest-fit criterion arises. More precisely, it is useful to obtain a description of the the family of models that can be *identified* as the unique sparsest models consistent with the marginal information. Intuitively, we expect a characterization of the form: if the underlying model is "sparse enough", then it can be identified by as the unique sparsest solution consistent with the marginal information; the sparsity bound of course should depend on the dimension of the data, which can be treated as a measure of the "amount" of information that is available. Next, we describe the contributions we make to answering these questions.

The rest of the chapter is organized as follows. Section 5.1 gives a brief overview

[2]This statement follows from Caratheodory's theorem that states that every point in a convex polytope of dimension d can be decomposed into a convex combination of at most $d+1$ extreme points, and the fact that doubly stochastic matrices have a dimension of $(N-1)^2$.

of the related work on learning sparse models from marginal information. We give a formal description of the problems in Section 5.2. Sections 5.3, 5.4, and 5.5 are devoted to the noiseless setting: Section 5.3 gives precise descriptions of the problems we consider; Section 5.4 describes our main results; and finally, Section 5.5 describes the algorithm we propose for efficient determination of the sparsest model from the given marginal information. Similarly, Sections 5.6, 5.7, 5.8, and 5.9 are devoted to the noisy setting: Section 5.6 gives precise descriptions of the problems we consider; Section 5.7 describes our main results; Section 5.8 describes the algorithm we propose for efficient recovery of the sparsest model in the noisy setting; and finally, Section 5.9 presents the proofs for our main results for the noisy setting. We describe the results from our empirical studies in Section 5.11 before concluding with the chapter summary and discussion in Section 5.12.

5.1 Relevant work

As described in detail in Chapter 2, there is large body of work done on learning structurally simple parametric choice models from partial information. Our goal however is to fit sparse (as measured by the support size) models to data. Fitting sparse models to observed data has been a classical approach used in statistics for model recovery and is inspired by the philosophy of *Occam's Razor*. Motivated by this, sufficient conditions based on sparsity for learnability have been of great interest over years in the context of communication, signal processing and statistics, cf. Shannon [1949], Nyquist [2002]. In recent years, this approach has become of particular interest due to exciting developments and wide ranging applications including:

- In signal processing (see Candes and Tao [2005], Candes et al. [2006b], Candes and Romberg [2006], Candes et al. [2006a], Donoho [2006]) where the goal is to estimate a 'signal' by means of minimal number of measurements. This is referred to as compressive sensing.

- In coding theory through the design of low-density parity check codes Gallager

[1962], Sipser and Spielman [1996], Luby et al. [2001] or in the design Reed Solomon codes Reed and Solomon [1960] where the aim is to design a coding scheme with maximal communication rate.

- In the context of streaming algorithms through the design of 'sketches' (see Tropp [2006, 2004], Berinde et al. [2008], Cormode and Muthukrishnan [2006], Gilbert et al. [2007]) for the purpose of maintaining a minimal 'memory state' for the streaming algorithm's operation.

In all of the above work, the basic question (see Muthukrishnan [2005]) pertains to the design of an $m \times p$ "measurement" matrix A so that x can be recovered efficiently from measurements $y = Ax$ (or its noisy version) using the "fewest" possible number measurements m. The setup of interest is when x is sparse and when $m < p$ or $m \ll p$. The type of interesting results (such as those cited above) pertain to characterization of the sparsity K of x that can be recovered for a given number of measurements m. The usual tension is between the ability to recover x with large sparsity K as possible and using a sensing matrix A with as few measurements m as possible.

The sparsest recovery approach of this paper is similar (in flavor) to the above stated work; in fact, as has been shown in Chapter 3, the partial information we consider can be written as a linear transform of the the model λ. However, the methods or approaches of the prior work do not apply. Specifically, the work considers finding the sparsest function consistent with the given partial information by solving the corresponding ℓ_1 relaxation problem. The work derives a necessary and sufficient condition, called the *Restricted Nullspace Property*, on the structure of the matrix A that guarantees that the solutions to the ℓ_0 and ℓ_1 relaxation problems are the same (see Candes et al. [2006b], Berinde et al. [2008]). However, such sufficient conditions trivially fail in our setup (see Jagabathula and Shah [2008]). Therefore, our work provides an alternate set of conditions that guarantee efficient recovery of the sparsest model.

5.2 Setup and data

In this section, we establish the relevant notation and give precise statements of the problems we are interested in solving. As before, we assume there are N alternatives and λ denotes the underlying choice model. Let S_N denote the space of the $N!$ rankings of the N alternatives. Our interest is in learning the underlying model λ given marginal information. We assume that marginal information is available to us in the form of $y = A\lambda + \eta$, where we have abused notation and think of λ as an $N! \times 1$ vector, y is an $m \times 1$ data vector, A is an $m \times N!$ matrix with entries in the set $\{0, 1\}$, and η is the $m \times 1$ noise vector; the magnitude of noise is assumed to be bounded i.e., $\|\eta\| < \varepsilon$ for some $\varepsilon > 0$. As discussed in Section 3.2.2 of Chapter 3, the sales transaction data that is available in practice can be readily cast into the form $y = A\lambda + \eta$.

In addition to the sales transaction data, there is another rich class of marginal data that can be cast into the form $y = A\lambda + \eta$; this class comprises different types of data that can be constructed from the Fourier transform of λ. The simplest type of data belonging to this class is what is known as first-order marginal information; we represent it by an $N \times N$ doubly stochastic matrix $M^{(N-1,1)}(\lambda)$ (the reason for the superscript in the notation will become clear shortly) in which the element in the rth row and ith column is the probability under λ that i is ranked at position r. More precisely,

$$M_{ri}^{(N-1,1)}(\lambda) = \sum_{\sigma \in S_N} \lambda(\sigma) \mathbb{1}_{\{\sigma(i)=r\}}.$$

We can generalize this idea to consider other types of partial information. An immediate generalization is to consider what is called second-order partial information, which we represent by an $N^2 \times N^2$ matrix $M^{(N-2,1,1)}(\lambda)$ in which each row corresponds to a pair of positions and each column corresponds to a pair of alternatives. Specifically, if the rth row corresponds to the pair of positions (r_1, r_2) and the ith column corresponds to the pair of alternatives (i_1, i_2), then

$$M_{ri}^{(N-2,1,1)}(\lambda) = \sum_{\sigma \in S_N} \lambda(\sigma) \mathbb{1}_{\{\sigma(i_1)=r_1, \sigma(i_2)=r_2\}}.$$

It is easy to see that one can obtain other generalizations by considering triples (i_1, i_2, i_3) of alternatives or unordered pairs $\{i_1, i_2\}$ of alternatives. The rigorous way to construct various generalizations is to consider various partitions of N. More precisely, define a partition ρ of N as a tuple $\rho = (\rho_1, \rho_2, \ldots, \rho_s)$ such that $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_s$ and $\rho_1 + \rho_2 + \cdots + \rho_s = N$. For instance, $\rho = (N-1, 1)$ and $\rho = (N-2, 1, 1)$ are partitions of N. For any partition ρ of N, we can consider different ρ-partitions of the set $\mathcal{N} = \{1, 2, \ldots, N\}$ in which we partition the alternatives in $\mathcal{N}$ into s partitions with partition i containing ρ_i elements. It is easy to see that the total number of distinct ρ partitions of $\mathcal{N}$ is given by

$$D_\rho = \frac{N!}{\prod_{i=1}^{s} \rho_i!}.$$

Let the distinct ρ partitions of $\mathcal{N}$ be denoted by t_i, $1 \leq i \leq D_\rho$[3]. For example, for $\rho = (N-1, 1)$, there are $D_\rho = N!/(N-1)! = N$ distinct ρ partitions given by

$$t_i \equiv \{1, \ldots, i-1, i+1, \ldots, N\}\{i\}, \ 1 \leq i \leq N.$$

Given a permutation $\sigma \in S_N$, its action on t_i is defined through its action on the N elements of t_i, resulting in a ρ partition with the N elements permuted. For instance, in the above example with $\rho = (N-1, 1)$, σ acts on t_i to give the ρ-partition $t_{\sigma(i)}$, where

$$t_{\sigma(i)} \equiv \{1, \ldots, \sigma(i)-1, \sigma(i)+1, \ldots, n\}\{\sigma(i)\}.$$

With these definitions, we can now define what we call ρ-order marginal information, which is denoted by the $D_\rho \times D_\rho$ matrix $M^\rho(\lambda)$ where the i,jth element expressed as

$$M^\rho_{ij}(\lambda) = \sum_{\sigma \in S_N} \lambda(\sigma) \mathbb{1}_{\{\sigma(t_j) = t_i\}}.$$

It is easy to see that the first-order and second-order marginal information defined

[3]To keep the notation simple, we use t_i instead of t_i^ρ that takes explicit dependence on ρ into account.

above can be obtained as special cases of the general ρ-order partial information by setting $\rho = (N-1,1)$ and $\rho = (N-2,1,1)$. This now clarifies our choice of notation for the first and second order marginal information above.

It is not difficult to see that the ρ-order partial information is related to the underlying choice model λ through a linear transform. Specifically, we can write

$$M^\rho(\lambda) = \sum_{\sigma \in S_N} \lambda(\sigma) A^\rho(\sigma),$$

where $A^\rho(\sigma)$ is a $D_\rho \times D_\rho$ matrix with the i,jth element defined as

$$A^\rho_{ij}(\sigma) = \begin{cases} 1, & \text{if } \sigma(t_j) = t_i, \\ 0, & \text{otherwise.} \end{cases}$$

Thinking of the $D_\rho \times D_\rho$ matrix $M^\rho(\lambda)$ as a $D_\rho^2 \times 1$ vector y and each of the matrices $A^\rho(\sigma)$ as a $D_\rho^2 \times 1$ column of the A matrix, we can cast the ρ-order partial information in the form $y = A\lambda$.

We note here that the ρ-order partial information we have defined here is related to the group theoretic Fourier transform of the underlying choice model λ. Specifically, it can be shown that the ρ-order partial information conveniently captures the information contained in a subset of lower-order Fourier coefficients up to the partition ρ (see Diaconis [1988] for more details). The motivation for considering group theoretic Fourier coefficients is two fold: first, they provide a rigorous way to compress the high-dimensional function λ (as used in Huang et al. [2008], Kondor et al. [2007]), and second, Fourier coefficients at different representations have natural interpretations, which makes it easy to gather in practice.

Given the above definitions, we consider the problem of learning the choice model given marginal information that is either of the form of sales transaction data or of the form of ρ-order partial information. We consider two different cases – the noiseless case and the noisy case – and describe the methods we propose and the main results we obtain for each of the cases in turn.

5.3 Noiseless case: problem formulations

In the noiseless case, we assume that the underlying choice model is λ, and we have access to the exact marginal information $y = A\lambda$. Given this, our goal is to determine the sparsest distribution that is consistent with the given marginal information y. In other words, our interest is in solving the problem:

$$\begin{aligned} \underset{\mu}{\text{minimize}} \quad & \|\mu\|_0 \\ \text{subject to} \quad & A\mu = y, \\ & \mathbf{1}^\top \mu = 1, \\ & \mu \geq 0. \end{aligned} \tag{5.1}$$

In this section, we discuss in detail the important questions that arise in this context. For each question, we discuss its importance and the type of answer we should expect.

The program in (5.1) first raises the following two important questions:

Question one. From a theoretical standpoint, can we obtain a description of the family of models that can be *identified* from marginal information via the sparsest-fit criterion?

Question two. From an operational standpoint, is there a reasonably large family of models for which it is possible to find the sparsest consistent distribution in an efficient manner?

We answer both questions by identifying a family of models $\mathcal{F}$, which can be *efficiently* recovered from marginal information as the *unique optimal solutions* to the program in (1.1); the family of models $\mathcal{F}$ is described in the next section.

Before we move to the next question, we try and understand the nature of conditions we expect to be imposed. Now, a condition that first comes to mind is the linear independence of the columns of A corresponding to the permutations. This condition is of course necessary; unfortunately it is not sufficient. Moreover, unlike in the popular literature (cf. compressed sensing), the sufficient conditions cannot be based on sparsity alone. In order to see why linear independence and sparsity conditions are not sufficient and gain additional intuition into the conditions that need to

be imposed, we consider the following example:

Example 1. *For any $N \geq 4$, consider the four permutations $\sigma_1 = (1,2)$, $\sigma_2 = (3,4)$, $\sigma_3 = (1,2)(3,4)$ and $\sigma_4 = \mathrm{id}$, where* id *is the identity permutation; here, we are representing permutations using the cycle notation. In addition, consider the partition $\rho = (N-1,1)$. It is then easy to see that*

$$A^\rho(\sigma_1) + A^\rho(\sigma_2) = A^\rho(\sigma_3) + A^\rho(\sigma_4).$$

In order to illustrate the issues that arise with straightforward conditions, we consider three cases:

1. *This example shows that a sparsity bound (even 4) on λ is not sufficient to guarantee that λ will indeed be the sparsest solution. Specifically, suppose that $\lambda(\sigma_i) = p_i$, where $p_i \in \mathbb{R}_+$ for $1 \leq i \leq 4$, and $\lambda(\sigma) = 0$ for all other $\sigma \in S_N$. Without loss of generality, let $p_1 \leq p_2$. Then,*

$$\begin{aligned} & M^\rho(\lambda) \\ =& p_1 A^\rho(\sigma_1) + p_2 A^\rho(\sigma_2) + p_3 A^\rho(\sigma_3) + p_4 A^\rho(\sigma_4) \\ =& (p_2 - p_1) A^\rho(\sigma_2) + (p_3 + p_1) A^\rho(\sigma_3) \\ & + (p_4 + p_1) A^\rho(\sigma_4). \end{aligned}$$

   Thus, the distribution $\tilde{\lambda}$ with $\tilde{\lambda}(\sigma_2) = p_2 - p_1$, $\tilde{\lambda}(\sigma_3) = p_3 + p_1$, $\tilde{\lambda}(\sigma_4) = p_4 + p_1$ and $\tilde{\lambda}(\sigma) = 0$ for all other $\sigma \in S_N$ is such that $M^\rho(\tilde{\lambda}) = M^\rho(\lambda)$ but $\|\tilde{\lambda}\|_0 = 3 < 4 = \|\lambda\|_0$. That is, λ can not be recovered as the solution of ℓ_0 optimization problem (5.1) *even when support of λ is only* 4.

2. *This example shows that there are cases where λ is the sparsest solution, but is not the* unique *sparsest solution. Specifically, suppose that $\lambda(\sigma_1) = \lambda(\sigma_2) = p$ and $\lambda(\sigma) = 0$ for all other $\sigma \in S$. Then, $M^\rho(\lambda) = pA^\rho(\sigma_1) + pA^\rho(\sigma_2) = pA^\rho(\sigma_3) + pA^\rho(\sigma_4)$. Thus, the sparsest solution is not unique.*

3. *Finally, this example shows that even though the support of λ corresponds to*

a linearly independent set of columns of A, the sparsest solution may not be unique. Specifically, suppose that $\lambda(\sigma_i) = p_i$, where $p_i \in \mathbb{R}_+$ for $1 \leq i \leq 3$, and $\lambda(\sigma) = 0$ for all other $\sigma \in S_N$. Without loss of generality, let $p_1 \leq p_2$. Then,

$$\begin{aligned} & M^\rho(\lambda) \\ =& p_1 A^\rho(\sigma_1) + p_2 A^\rho(\sigma_2) + p_3 A^\rho(\sigma_3) \\ =& (p_2 - p_1) A^\rho(\sigma_2) + (p_3 + p_1) A^\rho(\sigma_3) + p_1 A^\rho(\sigma_4). \end{aligned}$$

Here, note that $\{A^\rho(\sigma_1), A^\rho(\sigma_2), A^\rho(\sigma_3)\}$ is linearly independent, yet the sparsest solution is not unique.

Given the above example, it is clear that a simple sparsity bound or linear independence of the columns of A corresponding to the support of λ are not sufficient. Nevertheless, we manage to identify two sufficient conditions, as detailed in the next section.

Once we identify the conditions that define the family of models $\mathcal{F}$, we face the following natural question:

Question three. Given the conditions that define the family of models $\mathcal{F}$, how restrictive are these conditions?

We expect the conditions imposed on $\mathcal{F}$ to translate into a condition on the sparsity of the models; particularly, it is natural to expect a characterization of the form: if the model is "sparse enough", then it can be identified from marginal information via the sparsest-fit criterion. Such a characterization would quantify the relationship between the complexity (as measured by the sparsity) of the models and the amount of information (as measured by the dimension of the marginal information) available. Unfortunately, as discussed above, the sufficient conditions cannot translate into a simple sparsity bound on the models. In that case, can we find a sparsity bound such that "most," if not all, distributions that satisfy the sparsity bound can be recovered from marginal information via the sparsest-fit criterion? It turns out we can, as described in the next section. We make the notion of "most" distributions precise by

proposing a natural random generative model for functions with a given sparsity:

Definition 1 (Random Model). *Given $K \in \mathbb{Z}_+$ and an interval $\mathscr{C} = [a, b]$, $0 < a < b$, a random distribution λ with sparsity K and values in $\mathscr{C}$ is generated as follows: choose K permutations from S_N independently and uniformly at random* [4]*, say $\sigma_1, \ldots, \sigma_K$; select K values from $\mathscr{C}$ uniformly at random, say $p_1, \ldots, p_K$; then the distribution λ is defined as*

$$\lambda(\sigma) = \begin{cases} p_i & \text{if} \quad \sigma = \sigma_i, \ 1 \leq i \leq K \\ 0 & \text{otherwise.} \end{cases}$$

We denote this model by $R(K, \mathscr{C})$.

Given this model, we expect to derive a sparsity bound K^* (which depends on the type of marginal information) such that whenever $K \leq K^*$, a random distribution λ of sparsity K generated according to the random model $R(K, \mathscr{C})$ belongs to the family $\mathcal{F}$ with a high probability. Such a statement essentially establishes that the type of situations illustrated in Example 1 occur with a vanishing probability.

Finally, we would like to understand the limitations on the type of models that could be recovered form marginal information. Specifically, since the marginal information is limited, we expect not to be able to recover models that are very complex. In other words, we raise the following question:

Question four. Can we characterize the limitation on the ability of *any* algorithm to recover λ from only marginal information y?

In the next section, we describe the answers we propose for each of the questions raised above.

5.4 Noiseless case: main results

As the main results of this paper, we provide answers to the four questions raised in stated in the previous section.

[4] Throughout, we assume that the random selection is done *with* replacement.

Answers one & two. To answer the first two questions, we need to find sufficiency conditions that guarantee that the model λ can be recovered from marginal information via the sparsest-fit criterion and design a simple algorithm to find the sparsest model consistent with the available information efficiently. For that, we first try to gain a qualitative understanding of the conditions that the model λ should satisfy. Note that a necessary condition for identification of λ is that the program in (5.1) must have a *unique* solution; otherwise, without any additional information, we wouldn't know which of the multiple solutions is the true solution. It is clear that a necessary condition for (5.1) to have a unique optimal solution is that columns of A corresponding to the permutations in the support of λ must be linearly independent. However, as illustrated above, this linear independence condition is, in general, not sufficient to guarantee a unique solution; in particular, even if the columns of A corresponding to the permutations in the support S_1 of λ are linearly independent, there could exist a set of permutations S_2 such that $|S_2| \leq |S_1|$ and $A\lambda = A\lambda'$, where S_2 is the support of λ'; Example 1 illustrates exactly such a scenario. Thus, a sufficient condition for the distribution λ to be the unique sparsest solution is that not only is the set of columns of A corresponding to the set S_1 of permutations linearly independent, but the set of columns of corresponding to the permutations in the set $S_1 \cup S_2$ are linearly independent for all set of permutations S_2 such that $|S_2| \leq |S_1|$; in other words, not only we want the columns corresponding to the support to be linearly independent, but we want them to remain linearly independent even after the addition of at most K permutations to the support of λ. Note that this condition is similar to the Restricted Isometry Property (RIP) introduced in Candes and Tao [2005], which roughly translates to the condition that ℓ_0 optimization recovers x of sparsity K from $z = Bx$ provided every subset of $2K$ columns of B is linearly independent. Motivated by this, we impose the following conditions on λ.

Condition 1 (Sufficiency Conditions)**.** We require the model λ with support $\text{supp}\,(\lambda) = \{\sigma_1, \sigma_2, \ldots, \sigma_K\}$ to satisfy the following conditions:

- *Signature condition*: for any $\sigma \in \text{supp}\,(\lambda)$, there exists a row $1 \leq i \leq m$ of the

matrix A such that $A_{i\sigma} = 1$ and $A_{i\sigma'} = 0$ for all other permutations σ' such that $\sigma' \in \text{supp}(\lambda)$.

- *Linear Independence*: for any collection of integers $c_1, \ldots, c_K$ taking values in $\{-K, \ldots, K\}$, $\sum_{k=1}^{K} c_k \lambda(\sigma_k) \neq 0$, unless $c_1 = \cdots = c_K = 0$.

The discussion above motivates the 'signature' condition; indeed, whenever λ satisfies the signature condition, it is easy to see that the columns of A corresponding to the $\text{supp}(\lambda)$ are linearly independent. In addition, as shown in the proof of Theorem 7, the *linear independence* condition is required to establish the uniqueness of the sparsest solution.

Now we state the formal result that establishes Condition 1 as sufficient for recovery of λ as the unique sparsest solution consistent with the marginal information. Further, the conditions allow for a simple, iterative recovery algorithm. Thus, Theorem 7 provides answers to *Question one* and *Question two* from the previous section.

Theorem 7. *Given $y = A\lambda$, suppose λ satisfies Condition 1. Then, λ can be recovered from y as the unique optimal solution to the ℓ_0 optimization problem in* (5.1). *Further, a simple, iterative algorithm called the sparsest-fit algorithm, described in Section 5.5, recovers λ.*

Linear programs don't work. Theorem 7 states that whenever λ satisfies Condition 1, it can be recovered as the unique sparsest solution consistent with the available information. In order to solve the ℓ_0 optimization problem efficiently, it is natural to consider its convex relaxation and solve the following ℓ_1 optimization problem:

$$\begin{aligned} \underset{\lambda}{\text{minimize}} \quad & \|\lambda\|_1 \\ \text{subject to} \quad & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0, \end{aligned} \tag{5.2}$$

This approach has found a lot of success in the work done in the recently popular compressive sensing literature. (cf. Candes et al. [2006b,a], Donoho [2006], Berinde

et al. [2008]). Unfortunately, such convex relaxations have no bite in our setting. For instance, it is clear from the constraints in (1.2) that all feasible models have the same ℓ_1 norm (equal to 1). Moreover, it is hardly the case that the set of consistent models is a singleton as illustrated in Example 1. In fact, we can prove that for a marginal data vector y that comes from a distribution generated as per the random model $R(K, \mathscr{C})$ with $K \geq 2$, the set of consistent distributions is not a singleton. More precisely, we can prove the following theorem.

Theorem 8. *Consider a distribution λ generated according to the random model $R(K, \mathscr{C})$ described in Definition 1 with sparsity $K \geq 2$. Then, as long as ρ is not the partition $(1, 1, \ldots, 1)$ (N times), with probability $1 - o(1)$, there exists a distribution λ' distinct from λ such that $M^{\rho}(\lambda) = M^{\rho}(\lambda')$ and $\|\lambda'\|_1 = \|\lambda\|_1$.*

Thus, the ℓ_1 criterion is not reducing the family of consistent distributions at all, and we cannot guarantee the optimality of an arbitrarily selected model from the set of feasible models. In fact, as described in the the noisy setting (see discussion after Theorem 14), for the first-oder marginal information, selecting an arbitrary basic feasible solution that minimizes the ℓ_1 norm can have a sparsity of $O(N^2)$, whereas the sparsest distribution can be shown to have a sparsity of $O(N)$.

Answer three. Next, we turn to the third question. Specifically, we study the conditions for high probability recoverability of a random distribution λ in terms of its sparsity. In what follows, we spell out our result starting with few specific cases so as to better explain the dependency of the sparsity bound on the type of marginal information.

Case 1: We first consider the three of the simplest types of marginal information: comparison information, top-set information, and first-order marginal information. All three are described in detail in Section 3.2.2 of Chapter 3. First-order information is also described above and corresponds to the ρ-order marginal information with $\rho = (N - 1, 1)$. In the context of these three types of marginal information, we can establish the following theorem:

Theorem 9. *Suppose λ is a choice model of support size K drawn from the generative model $R(K, \mathscr{C})$. Then, λ satisfies the signature and linear independence conditions with probability $1 - o(1)$ as $N \to \infty$ provided $K = o(\log N)$ for comparison information, $K = o(\sqrt{N})$ for the top-set information, and $K = (1 - \varepsilon) N \log N$ for any fixed $\varepsilon > 0$.*

Case 2: We now consider the ρ-order marginal information for $\rho = (N - v, v)$ with $1 < v = O(1)$. Here $D_\rho = \Theta(N^v)$ and $M^\rho(\lambda)$ provides the *vth order* marginal information. As stated next, for this case we find that the sparsity bound scales at least as $N^v \log N$.

Theorem 10. *Suppose λ is a choice model of support size K drawn from the generative model $R(K, \mathscr{C})$. Then, λ satisfies the signature and linear independence conditions with probability $1 - o(1)$ as $N \to \infty$ provided $K \leq \frac{1-\varepsilon}{v!} N^v \log N$ for any fixed $\varepsilon > 0$ and ρ-order marginal information with $\rho = (N - v, v)$, $v = O(1)$.*

In general, for any ρ with $\rho_1 = N - v$ and $v = O(1)$, the arguments of Theorem 10 can be adapted to show that the sparsity bound scales as $N^v \log N$. Theorems 9 and 10 suggest that the sparsity bound scales as $D_\rho \log D_\rho$ for $\rho = (\rho_1, \ldots, \rho_s)$ with $\rho_1 = N - v$ for $v = O(1)$. Next, we consider the case of more general ρ.

Case 3: $\rho = (\rho_1, \ldots, \rho_s)$ with $\rho_1 = N - O\left(N^{\frac{2}{9}-\delta}\right)$ for any $\delta > 0$. As stated next, for this case, the sparsity bound scales at least as $D_\rho \log \log D_\rho$.

Theorem 11. *Suppose λ is a choice model of support size K drawn from the generative model $R(K, \mathscr{C})$. Then, λ satisfies the signature and linear independence conditions with probability $1 - o(1)$ as $N \to \infty$ provided $K \leq (1 - \varepsilon) D_\rho \log \log D_\rho$ for any fixed $\varepsilon > 0$ and ρ-order marginal information with $\rho = (\rho_1, \ldots, \rho_s)$, where $\rho_1 = N - N^{\frac{2}{9}-\delta}$ for any $\delta > 0$.*

We next consider the most general case of ρ-order marginal information.

Case 4: Any $\rho = (\rho_1, \ldots, \rho_r)$. The results stated thus far suggest that the threshold is essentially D_ρ, ignoring the logarithm term. For general ρ, we establish a

sparsity bound as stated in Theorem 12 below. Before stating the result, we introduce some notation. For given ρ, define $\alpha = (\alpha_1, \ldots, \alpha_s)$ with $\alpha_i = \rho_i/N$, $1 \leq i \leq s$. Let

$$H(\alpha) = -\sum_{i=1}^{s} \alpha_i \log \alpha_i, \quad \text{and} \quad H'(\alpha) = -\sum_{i=2}^{s} \alpha_i \log \alpha_i.$$

Theorem 12. *Suppose λ is a choice model of support size K drawn from the generative model $R(K, \mathscr{C})$. Then, λ satisfies the signature and linear independence conditions for ρ-order marginal information with probability $1 - o(1)$ as $N \to \infty$ provided*

$$K \leq C D_\rho^{\gamma(\alpha)}, \tag{5.3}$$

where

$$\gamma(\alpha) = \frac{T}{T+1}\left[1 - C'\frac{H(\alpha) - H'(\alpha)}{H(\alpha)}\right],$$
$$\text{with } T = \left\lfloor \frac{1}{1-\alpha_1} \right\rfloor,$$

$0 < C, C' < \infty$ are constants.

At a first glance, the above result in Theorem 12 seems very different from the crisp formulas of Theorems 9-11. To understand if that is indeed the case, consider the following special cases. First, observe that as $\alpha_1 \uparrow 1$, $T/(T+1) \to 1$. Further, as stated in Lemma 3, $H'(\alpha)/H(\alpha) \to 1$. Thus, we find that the bound on sparsity essentially scales as D_ρ for these special cases. Note that the cases 1, 2 and 3 fall squarely under this scenario since $\alpha_1 = \lambda_1/n = 1 - o(1)$. Thus, this general result contains the results of Theorems 9-11 (ignoring the logarithm terms). Next, consider the other extreme of $\alpha_1 \downarrow 0$. Then, $T \to 1$ and again by Lemma 3, $H'(\alpha)/H(\alpha) \to 1$. Therefore, the bound on sparsity scales as $\sqrt{D_\rho}$. This ought to be the case because for $\rho = (1, \ldots, 1)$ we have $\alpha_1 = 1/N \to 0$, $D_\lambda = N!$, and signature condition holds only up to $o(\sqrt{D_\rho}) = o(\sqrt{N!})$ due to the standard Birthday paradox.

In summary, the result in Theorem 12 appears reasonably tight for the general form of partial information ρ. We now state the Lemma 3 used above.

Lemma 3. *Consider any* $\alpha = (\alpha_1, \ldots, \alpha_s)$ *with* $1 \geq \alpha_1 \geq \cdots \geq \alpha_s \geq 0$ *and* $\sum_{i=1}^{s} \alpha_i = 1$. *Then,*

$$\lim_{\alpha_1 \uparrow 1} \frac{H'(\alpha)}{H(\alpha)} = 1,$$

and

$$\lim_{\alpha_1 \downarrow 0} \frac{H'(\alpha)}{H(\alpha)} = 1.$$

Answer four. Finally, we wish to understand the fundamental limitation on the ability to recover λ from the marginal information by *any* algorithm. To obtain a meaningful bound (cf. Example 1), we examine this question under an appropriate information theoretic setup.

To this end, suppose the model λ with sparsity K is drawn from the random model $R(K, \mathscr{C})$. However, for technical reasons (or limitations), we assume that the values p_is are chosen from a discrete set. Specifically, let each p_i be chosen from integers $\{1, \ldots, L\}$ instead of compact set $\mathscr{C}$. We denote this random model by $R(K, L)$.

Now, consider any algorithm that attempts to recover λ from $M^{\rho}(\lambda)$. Let $\hat{\lambda}$ be the estimate produced by the algorithm. Define the probability of error of the algorithm as

$$p_{\text{err}} = \mathbb{P}\left(\hat{\lambda} \neq \lambda\right).$$

Then, we can prove the following result.

Theorem 13. *Suppose the choice model* λ *of sparsity* K *is drawn from the random model* $R(K, L)$. *Let* $\hat{\lambda}$ *be the estimate of* λ *produced by some algorithm from access to only* $M^{\rho}(\lambda)$. *Then, the probability of error is uniformly bounded away from* 0 *for all* N *large enough and any* ρ, *if*

$$K \geq \frac{3D_{\rho}^2}{N \log N} \left[\log \left(\frac{D_{\rho}^2}{N \log N} \vee L \right) \right],$$

where for any two numbers x *and* y, $x \vee y$ *denotes* $\max\{x, y\}$.

The proofs of all the theorems are given in the appendix. Next, we describe the sparsest-fit algorithm.

5.5 Noiseless case: sparsest-fit algorithm

As mentioned above, finding the sparsest distribution that is consistent with the given partial information is in general a computationally hard problem. In this section, we propose an efficient algorithm to fit the sparsest distribution to the given marginal information $y = A\lambda$ as long as the underlying model λ satisfies the signature and linear independence conditions. Specifically, when run with marginal information y, the sparsest-fit algorithm terminates with either (a) the sparsest distribution if the data vector y is generated by a model satisfying Condition 1 or (b) a certificate that y is not generated by a model satisfying Condition 1. It follows from Theorem 7 that whenever the sparsest-fit algorithm succeeds in finding the sparsest distribution, it find a solution satisfying the signature and linear independence conditions.

The algorithm takes the data vector y as an explicit input with the prior knowledge of the structure of A as an auxiliary input. It's aim is to produce λ. In particular, the algorithm outputs the sparsity of λ, $K = \|\lambda\|_0$, permutations $\sigma_1, \ldots, \sigma_K$ so that $\lambda(\sigma_i) \neq 0$, $1 \leq i \leq K$ and the values $\lambda(\sigma_i)$, $1 \leq i \leq K$. Without loss of generality, assume that the values $y_1, \ldots, y_m$ are sorted with $y_1 \leq \cdots \leq y_m$ and further that $\lambda(\sigma_1) \leq \lambda(\sigma_2) \leq \ldots \lambda(\sigma_K)$.

Before we describe the algorithm, we observe the implication of the two signature and linear independence conditions. An implication of the linear independence condition is that for any two non-empty distinct subsets $S, S' \subset \{1, \ldots, K\}$, $S \neq S'$,

$$\sum_{i \in S} \lambda(\sigma_i) \neq \sum_{j \in S'} \lambda(\sigma_j).$$

This means that if we know all $\lambda(\sigma_i)$, $1 \leq i \leq K$ and since we know $y_d, 1 \leq d \leq m$, then we can recover $A(\sigma_i)_d, i = 1, 2, \ldots, K$ as the unique solution to $y_d = \sum_{i=1}^{K} A(\sigma_i)_d \lambda(\sigma_i)$ in $\{0,1\}^K$.

Therefore, the non-triviality lies in finding K and $\lambda(\sigma_i)$, $1 \leq i \leq K$. This issue is resolved by use of the signature condition in conjunction with the above described properties in an appropriate recursive manner. Specifically, recall that the signature condition implies that for each σ_i for which $\lambda(\sigma_i) \neq 0$, there exists d such that

$y_d = \lambda(\sigma_i)$. By linear independence, it follows that all $\lambda(\sigma_i)$s are distinct and it follows from our assumption above that

$$\lambda(\sigma_1) < \lambda(\sigma_2) < \cdots < \lambda(\sigma_K).$$

Therefore, it must be that the smallest value, y_1 equals $\lambda(\sigma_1)$. Moreover, $A(\sigma_1)_1 = 1$ and $A(\sigma_i)_1 = 0$ for all $i \neq 1$. Next, if $y_2 = y_1$ then it must be that $A(\sigma_1)_2 = 1$ and $A(\sigma_i)_2 = 0$ for all $i \neq 1$. We continue in this fashion until we reach a d' such that $y_{d'-1} = y_1$ but $y_{d'} > y_1$. Using similar reasoning it can be argued that $y_{d'} = \lambda(\sigma_2)$, $A(\sigma_2)_{d'} = 1$ and $A(\sigma_i)_{d'} = 0$ for all $i \neq 2$. Continuing in this fashion and repeating essentially the above argument with appropriate modifications leads to recovery of the sparsity K, the corresponding $\lambda(\sigma_i)$ and $A(\sigma_i)$ for $1 \leq i \leq K$. The complete procedural description of the algorithm is given below.

Sparsest Fit Algorithm:

Initialization: $k(1) = 1$, $d = 1$, $\lambda(\sigma_1) = y_1$ and $A(\sigma_1)_1 = 1$, $A(\sigma_1)_\ell = 0$, $2 \leq \ell \leq m$.

for $d = 2$ to m

  if $y_d = \sum_{i \in T} \lambda(\sigma_i)$ for some $T \subseteq \{1, \ldots, k(d-1)\}$

    $k(d) = k(d-1)$

    $A(\sigma_i)_d = 1 \quad \forall \quad i \in T$

  else

    $k(d) = k(d-1) + 1$

    $\lambda(\sigma_{k(d)}) = y_d$

    $A(\sigma_{k(d)})_d = 1$ and $A(\sigma_{k(d)})_\ell = 0$, for $1 \leq \ell \leq m, \ell \neq d$

  end if

end for

Output $K = k(m)$ and $(\lambda(\sigma_i), A(\sigma_i)), 1 \leq i \leq K$.

The correctness of this algorithm is established in the proof of Theorem 7.

Complexity of the algorithm. Initially, we sort at most m elements of the vector y. This has a complexity of $O(m^2 \log m)$. Further, note that the **for** loop in the algorithm iterates for at most m times. In each iteration, we are solving a subset-sum problem. Since there are at most K elements, the worst-case complexity of subset-sum in each iteration is $O(2^K)$. Thus, the worst-case complexity of the algorithm is $O(m \log m + m2^K)$. The average case complexity can be shown to be much smaller. Specifically, suppose the marginal data vector y corresponds to ρ-order partial information and the underlying model λ is drawn from the random model $R(K, \mathscr{C})$. Then, using the standard balls and bins argument, we can prove that for $K = O(D_\rho \log D_\rho)$, with a high probability, there are at most $O(\log D_\rho)$ elements in each subset-sum problem. Thus, the complexity would then be $O\left(\exp(\log^2 D_\rho)\right)$ with a high probability.

5.6 Noisy case: problem formulations

We now consider the more practical scenario when the data vector y might be corrupted by noise. This section will be completely focused on the first-order marginal information, with the aim to simplify exposition. The methods we propose in this section do extend to higher-order marginal information (for ρ different from $(N-1,1)$) and the 'transaction-type' marginal information introduced above. However, some of the guarantees we provide for the running time complexity of the methods do not readily extend. We discuss these extensions in detail in the discussion section (Section 5.12).

Precisely, the setup we consider is as follows. We assume we are given first-order marginal information about a choice model λ that may be corrupted by noise; that is, we are given an $N \times N$ observation matrix Y that is related to the underlying choice model λ as $Y = M^\rho(\lambda) + \eta$, where, as defined above, $M^\rho(\lambda)$ is the $N \times N$ first-order marginal information matrix of λ and η is an $N \times N$ matrix capturing the noise. Note that we have deviated from the notation above, where we have denoted the observations (noisy or noiseless) by a vector y. We deviate in order to explicitly

account for the matrix structure of the first-order information. Moreover, since we fix $\rho = (N-1, 1)$ for this section, we drop the ρ and write $Y = M(\lambda) + \eta$; thus, unless mentioned otherwise, $M(\lambda)$ denotes the first-order marginal information in this section. We also assume that the noise can be bounded above as $\|\eta\|_2 \leq \delta$ for some $\delta > 0$, where

$$\|\eta\|_2^2 \stackrel{\text{def}}{=} \sum_{i,j=1}^{N} \eta_{i,j}^2.$$

Given the data Y and an approximation error $\varepsilon > 0$, our goal is to find the sparsest distribution $\hat{\lambda}$ such that $\|Y - M(\hat{\lambda})\|_2 \leq \varepsilon$; for brevity, we call such a model $\hat{\lambda}$ an ε-fit to the observations Y. In particular, our goal is to solve

$$\begin{aligned} \underset{\mu}{\text{minimize}} \quad & \|\mu\|_0 \\ \text{subject to} \quad & \|Y - M(\mu)\|_2 \leq \varepsilon, \\ & 1^\top \mu = 1, \\ & \mu \geq 0. \end{aligned} \tag{5.4}$$

Given the program in (5.4), a natural question that arises is whether we can solve it efficiently. Interestingly, a more fundamental question, which informs the question of efficient solvability of (5.4), is "how sparse can the sparsest solution be?" To elaborate further, first observe that for any choice model μ, the first-order marginal matrix $M(\mu)$ is doubly stochastic. Thus, it follows from $Y = M(\lambda) + \eta$ and $\|\eta\|_2 \leq \delta$ that solving the program in (5.4) is essentially equivalent to determining the convex decompositions of all doubly stochastic matrices that are within a ball of radius $\delta + \varepsilon$ of $M(\lambda)$ and choosing the sparsest convex decomposition. Now, it follows from Birkhoff-von Neumann's celebrated result (see Birkhoff [1946] and von Neumann [1953]) that a doubly stochastic matrix belongs to an $(N-1)^2$ dimensional polytope with the permutation matrices as the extreme points. Therefore Caratheodory's theorem tells us that it is possible to find a convex decomposition of any doubly stochastic matrix with at most $(N-1)^2 + 1$ extreme points, which in turn implies that the sparsest model consistent with the observations has a support of at most $(N-1)^2+1 = \Theta(N^2)$. We raise the following natural question at this point:

Question one. Given *any* doubly stochastic matrix M, does there exist a choice model λ with sparsity significantly smaller than $\Theta(N^2)$ such that $\|M(\lambda) - M\|_2 \leq \varepsilon$.

Geometrically speaking, the question above translates to: given a ball of radius ε around M, is there a subspace spanned by K extreme points that intersects the ball, for *any* double stochastic matrix M and some K that is significantly smaller than $\Theta(N^2)$? Note that the answer to this question can give us an indication of whether a straightforward approach (such as convex relaxation) can produce good approximations to the sparsest solution. In particular, it is possible that for a general doubly stochastic matrix M, there do not exist models λ with sparsity significantly smaller than $\Theta(N^2)$ such that $\|M(\lambda) - M\|_2 \leq \varepsilon$. In other words, it is possible that for a general M, there is no subspace of K extreme points that intersects the ε ball around M for $K \ll N^2$. If such is the case, then convex relaxations – which result in a models of sparsity $O(N^2)$ – can produce solutions close to the optimal, at least for general M. In that case, we could attempt to characterize the class of matrices M for which we can solve (5.4) through convex relaxation. If, on the other hand, we can prove that for a general M, the sparsest model can have sparsity significantly smaller than $\Theta(N^2)$, then using straightforward approaches like convex relaxations can result in a highly suboptimal solution. As explained in the next section, it happens that we can show that the sparsest solution indeed can have a sparsity that is significantly smaller than $\Theta(N^2)$. Thus, we expect to go beyond convex relaxations in order to efficiently recover the sparsest solution, leading us to our next question about efficient recovery of the sparsest solution.

As mentioned above, if the sparsest solution has sparsity K, then the brute-force approach would require searching over $\binom{N!}{K} \approx \exp(\Theta(KN \log N))$ options. Thus, we ask:

Question two. Is it possible to solve (5.4) with a running time complexity that is far better than $O(\exp(KN \log N))$, at least for a reasonable large class of observations Y?

We obtain a faster algorithm by restricting our search to models that belong to the signature family. The structure of the family allows efficient search. In addition, we

can establish that the signature family is appropriately "dense", thereby by restricting our search to the signature family, we are not losing much. Before we describe our answers to the questions above, we quickly recall the definition of the signature family specialized to first-order information:

> **Signature family.** A distribution (choice model) λ is said to belong the signature family if for each permutation σ that is in the support (i.e., $\lambda(\sigma) > 0$) there exist an pair i, j such that $\sigma(i) = j$ and $\sigma'(i) \neq j$ for any permutation σ' in the support. Equivalently, for every permutation σ in the support of λ, there exists a pair i, j such that σ ranks i at position j, but no other permutation in the support ranks i at position r.

In the next section, we provide our answers to the above questions.

5.7 Noisy case: main results

As main results for the noisy case, we provide answers to the two questions raised above. We provide the answers to each of the questions in turn.

Question one: sparse approximation. As the first result, we establish that given *any* doubly stochastic matrix M and $\varepsilon > 0$, there exists a model λ with sparsity $O(N/\varepsilon^2)$ such that $\|M(\lambda) - M\|_2 \leq \varepsilon$. Thus, we show that by allowing a "small" error of ε, one can obtain a significant reduction from $\Theta(N^2)$ to $O(N/\varepsilon^2)$ in the sparsity of the model that is needed to explain the observations. Precisely, we have the following theorem.

Theorem 14. *For any doubly stochastic matrix M and $\varepsilon \in (0, 1)$, there exists a choice model λ such that $\|\lambda\|_0 = O(N/\varepsilon^2)$ and $\|M(\lambda) - M\|_2 \leq \varepsilon$.*

We emphasize here that this result holds for *any* doubly stochastic matrix M. In such generality, this result is in fact tight in terms of the dependence on N of the required sparsity. To see that consider the uniform doubly stochastic matrix M with all of its entries equal to $1/N$. Then, any choice model λ with $o(N)$ support can have

at most $N * o(N) = o(N^2)$ non-zero entries, which in turn means that the ℓ_2. Thus, the ℓ_2 error $\|M(\lambda) - M\|_2$ is at least $\sqrt{(N^2 - o(N^2))/N^2} \approx 1$ for large N.

The result of Theorem 14 also justifies why convex relaxations don't have any bite in our setting. Specifically, suppose we are given a doubly stochastic matrix M and a tolerance parameter $\varepsilon > 0$. Then, all the consistent choice models λ, which satisfy $\|M(\lambda) - M\|_0 \leq \varepsilon$, have the same ℓ_1 norm. We claim that "most" of such consistent models λ have sparsity $\Theta(N^2)$. More precisely, following the arguments presented in the proof of Theorem 4, we can show that the set of doubly stochastic matrices $\tilde{M}$ such that $\|\tilde{M} - M\|_2 \leq \varepsilon$ and can be written as $M(\lambda) = \tilde{M}$ for some model λ with sparsity $K < (N-1)^2$ has an $(N-1)^2$ dimensional volume of zero. It thus follows that picking an arbitrary consistent model λ will most certainly yield a model with sparsity $\Theta(N^2)$; this is a factor N off from the sparsest solution, which has a sparsity of $O(N)$ (ignoring the ε dependence).

Question two: efficient algorithm to solve (5.4). We now consider the question of efficiently solving the program in (5.4). As explained above, a brute-force search for a model of sparsity K that is consistent with the data requires searching over $\exp(\Theta(KN \log N))$ options. We now show that by restricting ourselves to a reasonably large class of choice models, we can improve the running time complexity to $O(\exp(\Theta(K \log N)))$ – effectively shaving off a factor of N from the exponent. More precisely, we can establish the following result.

Theorem 15. *Given a noisy observation Y and $\varepsilon \in (0, 1/2)$, suppose there exists a choice model λ in the signature family such that $\|\lambda\|_0 = K$ and $\|Y - M(\lambda)\|_2 \leq \varepsilon$. Then, with a running time complexity of $\exp\big(\Theta(K \log N)\big)$, we can find a choice model $\hat{\lambda}$ such that $\|\hat{\lambda}\|_0 = O(\varepsilon^{-2} K \log N)$ and $\|M(\hat{\lambda}) - D\|_2 \leq 2\varepsilon$.*

Several remarks are in order. We propose a method in Section 5.8 to find sparse models consistent with the data efficiently. The result of Theorem 15 establishes guarantees for this method. The result of Theorem 15 essentially establishes that as long as there is a sparse choice model of sparsity K in the signature family that is an ε-fit to the observations Y, we can shave off a factor of N in the exponent

from the running time complexity at the cost of finding a model with sparsity that is essentially within a factor of $\log N$ of K. In other words, we can obtain an exponential reduction in the running time complexity at the cost of introducing a factor of $\log N$ in the sparsity.

It is worth pausing here to understand how good (or bad) the computation cost of $\exp\left(\Theta(K \log N)\right)$ is. As discussed below (in Theorem 16), for a large class of choice models, the sparsity K scales as $O(\varepsilon^{-2}N)$, which implies that the computation cost scales as $\exp\left(\Theta(N \log N)\right)$ (ignoring ε to focus on dependence on N). That is, the computation cost is polynominal in $N! = \exp\left(\Theta(N \log N)\right)$, or equivalently, polynomial in the dimension of the ambient space. To put this in perspective, the scaling we obtain is very similar to the scaling obtained in the recently popular compressive sensing literature, where sparse models are recovered by solving linear or convex programs, which result in a computational complexity that is polynomial in the ambient dimension.

Finally, the guarantee of Theorem 15 is conditional on the existence of a sparse choice model in the signature family that is an ε-fit to the data. It is natural to wonder if such a requirement is restrictive. Specifically, given any doubly stochastic matrix M, there are two possibilities. Firstly it may be the case that there is no model in the signature family that is an ε-fit to the data; in such a case, we may have to lose precision by increasing ε in order to find a model in the signature family. Secondly, even if there did exist such a model, it may not be "sparse enough"; in other words, we may end up with a solution in the signature family whose sparsity scales like $\Theta(N^2)$. Our next result shows that both scenarios described above do not happen; essentially, it establishes that the signature family of models is "dense" enough that for a "large" class of data vectors, we can find a "sparse enough" model in the signature family that is an ε-fit to the data. More specifically, we can establish that the signature family is "dense" as long as the observations are generated by an MNL model or a max-ent (maximum-entropy) distribution. Recall the descriptions of the MNL model and the max-ent distributions (refer to Chapter 2 for a detailed discussion):

Multinomial Logit (MNL) model. The MNL model is a parametric model with N positive valued parameters – one for each of the N alternatives. Let $w_i > 0$ be parameter associated with alternative i. Then, the probability mass assigned to permutation $\sigma \in S_N$ is given by (for example, see Marden [1995])

$$\mathbb{P}_w(\sigma) = \prod_{j=1}^{N} \frac{w_{\sigma^{-1}(j)}}{w_{\sigma^{-1}(j)} + w_{\sigma^{-1}(j+1)} + \cdots + w_{\sigma^{-1}(N)}}, \tag{5.5}$$

where $\sigma^{-1}(j) = i$ if $\sigma(i) = j$.

Max-ent distribution. The max-ent distribution is a member of the exponential family of models. It is the distribution with the maximum Shannon entropy of all distributions that are consistent with given first-order marginal information. It is parametrized by N^2 parameters θ_{ij} for $1 \leq i, j \leq N$. Given such a vector of parameters θ, the probability mass assigned to permutation σ is given by

$$\begin{aligned} \mathbb{P}_\theta(\sigma) &\propto \exp\left(\sum_{1\leq i,j\leq N} \theta_{ij}\sigma_{ij}\right), \\ &= \frac{1}{Z(\theta)} \exp\left(\sum_{1\leq i,j\leq N} \theta_{ij}\sigma_{ij}\right), \end{aligned} \tag{5.6}$$

where $Z(\theta) = \sum_{\sigma\in S_N} \exp\left(\sum_{1\leq i,j\leq N} \theta_{ij}\sigma_{ij}\right)$; $\sigma_{ij} = 1$ iff $\sigma(i) = j$ and $\sigma_{ij} = 0$ otherwise. It is well known that with respect to the space of all first-order marginal distributions, the above exponential family is dense. Specifically, for any doubly stochastic matrix (the first-order marginals) $M = [M_{ij}]$ with $M_{ij} > 0$ for all i, j, there exists $\theta \in \mathbb{R}^{N\times N}$ so that the first-order marginal induced by the corresponding exponential family is precisely M. An interested reader is referred to, for example, monograph Wainwright and Jordan [2008] for details on this correspondance between parameters of exponential family and its marginals.

We can establish the following result about how dense the signature family is.

Theorem 16. *Suppose Y is a noisy observation of first-order marginal $M(\lambda)$ with $\|Y - M(\lambda)\|_2 \leq \varepsilon$ for some $\varepsilon \in (0, 1/2)$ and choice model λ such that*

1. *either, λ is from MNL model with parameters $w_1, \ldots, w_N$ (and without loss of*

generality $w_1 < w_2 < \ldots < w_N$*) such that*

$$\frac{w_N}{\sum_{k=1}^{N-L} w_k} \leq \frac{\sqrt{\log N}}{N}, \tag{5.7}$$

for $L = N^\delta$ *for some* $\delta \in (0,1)$*;*

2. *or,* λ *is from the max-ent exponential family with parameters* θ *such that for any set of four distinct tuples of integers* (i_1, j_1),$(i_{2,}, j_2)$, (i_3, j_3), *and* (i_4, j_4) *(with* $1 \leq i_k, j_k \leq N$ *for* $1 \leq k \leq 4$*)*

$$\frac{\exp(\theta_{i_1 j_1} + \theta_{i_2 j_2})}{\exp(\theta_{i_3 j_3} + \theta_{i_4 j_4})} \leq \sqrt{\log N}. \tag{5.8}$$

Then, there exists a $\hat{\lambda}$ *in the signature family such that:* $\|Y - \hat{\lambda}\|_2 \leq 2\varepsilon$ *and* $\|\hat{\lambda}\|_0 = O\left(N/\varepsilon^2\right)$.

Remark. The conditions (5.7) and (5.8) can be further relaxed by replacing $\sqrt{\log N}$ (in both of them) by $C \log N/\varepsilon^2$ for an appropriately chosen (small enough) constant $C > 0$. For the clarity of the exposition, we have chosen a somewhat weaker condition.

We have established in Theorem 16 that (under appropriate conditions) the rich families of MNL and max-ent exponential families can be approximated by sparse models in signature families as far as first-order marginals are concerned. Note that both families induce distributions that are full support. Thus, if the only thing we care about are first-order marginals, then we can just use sparse models in the signature family with sparsity only $O(N)$ (ignoring ε dependence) rather than distributions that have full support. It is also interesting to note that in Theorem 14, we establish the existence of a sparse model of $O(N/\varepsilon^2)$ that is an ε-fit to the observations. The result of Theorem 16 establishes that by restricting to the signature family, the sparsity scaling is still $O(N/\varepsilon^2)$ implying that we are not losing much in terms of sparsity by the restriction to the signature family.

In the next section, we describe the algorithm we propose to solve the program in (5.4) efficiently and also prove Theorem 15. We present the proofs of Theorems 14

and 16 subsequently.

5.8 Noisy case: efficient recovery of sparse models

In this section, we describe the algorithm we propose for efficient recovery of sparse models. In the process, we also prove Theorem 15.

The setup for the algorithm is as follows. We are a given first-order observation matrix Y. Suppose there exists a choice model μ in the signature family such that $\|\mu\|_0 = K$ and $\|M(\mu) - Y\|_2 \leq \varepsilon$. Then, the algorithm we describe below finds a choice model $\hat{\lambda}$ such that $\|\hat{\lambda}\|_0 = O(\varepsilon^{-2} K \log N)$ and $\|M(\hat{\lambda}) - Y\|_\infty \leq 2\varepsilon$ with a running-time complexity of $\exp(\Theta(K \log N))$. The algorithm requires effectively searching over space of choice models from signature family. Before we can describe the algorithm, we introduce a representation of the models in the signature family, which allows us reduce the problem into solving a collection of LPs.

Representation of signature family. We start by developing a representation of choice models from the signature family that is based on their first order marginal information. All the relevant variables are represented by vectors in N^2 dimension. For instance, the observation matrix $Y = [Y_{ij}]$ is represented as an N^2 dimensional vector with its (i,j)th component Y_{ij} indexed by the tuple (i,j). Further, the components are ordered according to the lexicographic ordering of the tuples: $(i,j) < (i',j')$ iff $i < i'$ or $i = i'$ and $j < j'$. Thus, the observation matrix Y is represented in a column vector form as

$$Y = [Y_{(1,1)}\ Y_{(1,2)} \ldots Y_{(1,N)}\ Y_{(2,1)} \ldots Y_{(N,N)}]^T.$$

In a similar manner, we represent a permutation $\sigma \in S_N$ as a 0-1 valued N^2 dimensional vector $\sigma = [\sigma_{(i,j)}]$ with $\sigma_{(i,j)} = 1$ if $\sigma(i) = j$ and 0 otherwise.

Now consider a choice model in the signature family with support K. Suppose it has the support $\sigma^1, \ldots, \sigma^K$ with respective probabilities $p_1, \ldots, p_K$. Since the model belongs to the signature family, the K permutations have distinct *signature* components. Specifically, for each k, $1 \leq k \leq K$, let (i_k, j_k) be the signature component of

permutation σ^k so that $\sigma^k(i_k) = j_k$ (i.e. $\sigma^k_{(i_k,j_k)} = 1$) but $\sigma^{k'}(i_k) \neq j_k$ (i.e. $\sigma^{k'}_{(i_k,j_k)} = 0$) for all $k' \neq k$, $1 \leq k' \leq K$. Now let $M = [M_{(i,j)}]$ be first order marginal information of this choice model. Then, it is clear from our notation that $M_{(i_k,j_k)} = p_k$ for $1 \leq k \leq K$ and $M_{(i,j)}$ is a summation of a subset of the K values $p_1, \ldots, p_K$, for any other (i,j), $1 \leq i,j \leq N$,

The above discussion leads to the following representation of a choice model from signature family. Each choice model is represented by an $N^2 \times N^2$ matrix with 0-1 entries, say $Z = [Z_{(i,j)(i',j')}]$ for $1 \leq i,j,i',j' \leq N$: in $Z_{(i,j)(i',j')}$, (i,j) represents a row index while (i',j') represents a column index. The choice model with support K is identified with its K signature components (i_k,j_k), $1 \leq k \leq K$. The corresponding matrix Z has all of the $N^2 - K$ columns corresponding to indices other than these K tuples equal to 0. The columns corresponding to (i_k,j_k), $1 \leq k \leq K$, indices are non-zero with each representing a permutation consistent with signature condition. In particular, the Z matrix satisfies the following constraints: for each (i_k,j_k), $1 \leq k \leq K$,

$$Z_{(i,j)(i_k,j_k)} \in \{0,1\}, \quad \text{for all } 1 \leq i,j \leq N, \tag{5.9}$$

$$Z_{(i_k,j_k)(i_k,j_k)} = 1, \tag{5.10}$$

$$Z_{(i,j)(i_k,j_k)} = 0, \quad \text{if } (i,j) \in \{(i_{k'},j_{k'}) : 1 \leq k' \leq K,\ k' \neq k\}, \tag{5.11}$$

$$\sum_{\ell=1}^{N} Z_{(i,\ell)(i_k,j_k)} = 1, \quad \sum_{\ell=1}^{N} Z_{(\ell,j)(i_k,j_k)} = 1, \quad \text{for all } 1 \leq i,j \leq N. \tag{5.12}$$

Observe that (5.10)-(5.11) enforce the signature condition while (5.12) enforces the permutation structure. In summary, given a set of K distinct pairs of indices, (i_k,j_k), $1 \leq k \leq K$ with $1 \leq i_k,j_k \leq N$, (5.9)-(5.12) represent the set of all possible signature family with these indices as their signature components.

Efficient representation of signature family. Indeed, a signature family choice model with support K can, in principle, have any K of the N^2 possible tuples as its signature components. Therefore, one way to search the signature family choice model with support K is to first choose one of the $\binom{N^2}{K}$ tuples and then for those

particular K tuples, search among all Zs satisfying (5.9)-(5.12). As we shall see, the basic searching complexity for determining a sparse choice model arises from going over each of the distinct $\binom{N^2}{K}$ tuples. This is because the the points described by (5.9)-(5.12) form the extreme points of the following relaxation: for each (i_k, j_k), $1 \leq k \leq K$,

$$Z_{(i,j)(i_k,j_k)} \in [0,1], \quad \text{for all } 1 \leq i,j \leq N, \tag{5.13}$$

$$Z_{(i_k,j_k)(i_k,j_k)} = 1, \tag{5.14}$$

$$Z_{(i,j)(i_k,j_k)} = 0, \quad \text{if } (i,j) \in \{(i_{k'}, j_{k'}) : 1 \leq k' \leq K,\ k' \neq k\}, \tag{5.15}$$

$$\sum_{\ell=1}^{N} Z_{(i,\ell)(i_k,j_k)} = 1, \quad \sum_{\ell=1}^{N} Z_{(\ell,j)(i_k,j_k)} = 1, \quad \text{for all } 1 \leq i,j \leq N. \tag{5.16}$$

It is easy to see that the points described by the set of equations (5.9)-(5.12) are contained in the polytope above described by equations (5.13)-(5.16). Thus, in order to justify the claim that the polytope above is the convex hull of the points described by (5.9)-(5.12), it is sufficient to argue that the extreme points of the polytope are integral. For that, we invoke the Birkhoff and Von Neumann Birkhoff [1946], von Neumann [1953] result which states that permutation matrices are extreme points of the polytope describing the doubly stochastic matrices. Now, it is easy to see that the polytope above is the polytope of double stochastic matrices with some of the coordinates set to 0s and 1s. Fortunately, such constraints introduce hyperplanes that do not 'cut through' the doubly stochastic polytope and hence do not introduce any new extreme points. Thus, the resulting polytope has integral extreme points. An important consequence of this result is that we can now solve an IP (integer program) over the constraints described by (5.9)-(5.12) efficiently (polynomial in N time) by relaxing it to an LP described by (5.13)-(5.16).

Feasibility of signature components. It is important to check whether for a given set of K tuples, (i_k, j_k) for $1 \leq k \leq K$, there exists a choice model in the signature family with the given K tuples as signature components. Equivalently, we wish to check the feasibility of the set of constraints (5.9)-(5.12) (or its relaxation).

An efficient way to check the feasibility is to reduce this problem to finding maximum size matchings in K distinct $N \times N$ bipartite graphs, as described next. Given K signature components (i_k, j_k), $1 \leq k \leq K$, we construct K bipartite $N \times N$ graphs as follows: for each $1 \leq k \leq K$, we start with a complete $N \times N$ bipartite graph and remove all the edges (i, j) between the left vertex i and right vertex j such that $i = i_{k'}$ and $j = j_{k'}$ for some $1 \leq k' \leq K$ and $k' \neq k$. Note that a perfect matching in any of the bipartite graphs corresponds to a permutation. Thus, once we construct the K bipartite graphs, we can check the feasibility of the constraints (5.10)-(5.11) by checking if there exist perfect matchings in all of the K bipartite graphs. Equivalently, we need to check if the size of the maximum size matchings in each of the bipartite graphs is equal to N. Finding the maximum size matching (and its size) in a bipartite graph has a computational complexity of $O(N^{2.5})$ (cf. Micali and Vazirani [1980], also see Edmonds and Karp [1972]). Thus, given a subset of K tuples, (i_k, j_k) for $1 \leq k \leq K$, it can be verified in $O(KN^{2.5})$ time if it is feasible to have a choice model with these K as signature components.

Searching in signature family. We now describe the main algorithm that will establish the result of Theorem 15. The algorithm succeeds in finding a choice model $\hat{\lambda}$ with sparsity $\|\hat{\lambda}\|_0 = O(\varepsilon^{-2} K \log N)$ and error $\|M(\hat{\lambda}) - Y\|_\infty \leq 2\varepsilon$ if there exists a choice model μ in signature family with sparsity K that is near consistent with Y in the sense that $\|M(\mu) - Y\|_\infty \leq \varepsilon$ (note that $\|\cdot\|_2 \leq \|\cdot\|_\infty$). The computation cost scales as $\exp\big(\Theta(K \log N)\big)$. Our algorithm uses the so called *Multiplicative Weight* algorithm utilized in the context of the framework developed by Plotkin-Shmoys-Tardos Plotkin et al. [1991] for fractional packing (also see Arora et al. [2005]).

The algorithm starts by going over all possible $\binom{N^2}{K}$ subset of possible signature components in any order till desired choice model $\hat{\lambda}$ is found or all combinations are exhausted – in the latter case, we declare that there exists no choice model of sparsity K in the signature family that is an ε-fit to the data Y. Now consider any such set of K signature components, (i_k, j_k) with $1 \leq k \leq K$. By the definition of the signature family, the $Y_{(i_k, j_k)}$ for $1 \leq k \leq K$ are the probabilities of the K permutations in the support. Therefore, we check if $1 - \varepsilon \leq \sum_{k=1}^{K} Y_{(i_k, j_k)} \leq 1 + \varepsilon$. If no, we reject

this set of K tuples as signature components and move to the next set. If yes, we continue towards finding a choice model with these K as signature components and the corresponding probabilities.

The first step is to verify if there exists a choice model with the chosen K components as signature components. As discussed above, we can verify that in $O(KN^{2.5})$ time. If such a model indeed exists, we search for the appropriate model. The choice model of our interest, represented by Z satisfying (5.9)-(5.12), should be such that $Y \approx ZY$, where Y is viewed as an N^2 dimensional vector and Z as $N^2 \times N^2$ matrix. Putting it other way, we are interested in finding Z so that

$$Y_{(i,j)} - \varepsilon \leq \sum_{k=1}^{K} Z_{(i,j)(i_k,j_k)} Y_{(i_k,j_k)} \leq Y_{(i,j)} + \varepsilon, \quad \text{for all } 1 \leq i,j \leq N^2 \tag{5.17}$$

$$Z \text{ satisfies } (5.9) - (5.12). \tag{5.18}$$

This is precisely the setting considered by Plotkin-Shmoys-Tardos Plotkin et al. [1991]: Z is required to satisfy a certain collection of 'difficult' linear inequalities (5.17) and a certain other collection of 'easy' convex constraints (5.18) (due to its exact linear relaxation (5.13)-(5.16) as discussed earlier). If there is a feasible solution satisfying (5.17)-(5.18), then Plotkin et al. [1991] finds a Z that approximately satisfies (5.17) and exactly (5.18). Or else, it discovers non-feasibility of the above optimization problem. We describe this precise algorithm next.

For ease of notation, we denote the choice model matrix Z of dimension $N^2 \times N^2$ (effectively $N^2 \times K$) by a vector z of KN^2 dimension; we think of (5.17) as $2N^2$ inequalities denoted by $Bz \geq b$ with B being $2N^2 \times KN^2$ matrix and b being $2N^2$ dimensional vector; finally, (5.18), through the linear relaxation (5.13)-(5.16), is denoted by $\mathcal{P}$. Thus, we are interested in finding $z \in \mathcal{P}$ such that $Bz \geq b$.

The Plotkin et al. [1991] framework essentially tries to solve the *Lagrangian* relaxation of $Bz \geq b$ over $z \in \mathcal{P}$ in an iterative manner. To that end, let p_ℓ be the Lagrangian variable (or weight) parameter associated with the ℓth constraint $a_\ell^T z \geq b_\ell$ for $1 \leq \ell \leq 2N^2$ (where a_ℓ is the ℓth row of B). We update the weights iteratively: let $t \in \{0, 1, \dots\}$ represent the index of the iteration. Initially, $t = 0$ and $p_\ell(0) = 1$

for all ℓ. Given $p(t) = [p_\ell(t)]$, find z^t by solving linear program

$$\begin{aligned} &\text{maximize } \sum_\ell p_\ell(t)(a_\ell^T z - b_\ell) \\ &\text{over } \ z \in \mathcal{P}. \end{aligned} \tag{5.19}$$

We insist on z^t being an extreme point of $\mathcal{P}$. Thus, even in case there are multiple solutions, z^t will be an integral (its components will be 0 or 1) corresponding to a K-sparse choice model in the signature family. Given such a z^t, the weights $p(t+1)$ are obtained as follows: for $\delta = \min\left(\varepsilon/8, 1/2\right)$

$$p_\ell(t+1) = p_\ell\Big(1 - \delta(a_\ell^T z^t - b_\ell)\Big). \tag{5.20}$$

The above update (5.20) suggests that if the ℓth inequality is non satisfied, we should increase the penalty imposed by $p_\ell(t)$ or else we should decrease. Now $b_\ell \in [0,1]$ since it corresponds to an entry in a non-negative doubly stochastic matrix Y. The $a_\ell^T z^t \in [0, 1+\varepsilon]$ since it corresponds to the summation of a subset of (non-negative) K entries $Y_{(i_k,j_k)}$, $1 \le k \le K$ and by choice, we have made sure that the summation of all of these K entries is at most $1+\varepsilon$. Therefore, $a_\ell^T z^t - b_\ell \in [-2, 2]$. Hence, the multiplicative update to each of the $p_\ell(\cdot)$ is by a factor of at most $(1 \pm 2\delta)$ in one iteration. Such bounded change is necessary in order to guarantee good performance of the eventual algorithm.

Now consider the sequence of z^t produced for $t \le T$ where $T = O(\varepsilon^{-2} \log N)$ (precisely, $T = 64\varepsilon^{-2} \ln(2N^2)$ as per [Arora et al., 2005, Corollary 4] and its utilization in [Arora et al., 2005, Section 3.2]). If for any t, the value of the objective in (5.19) is less than 0, then we declare infeasibility. This is because, if there indeed was a z that is a feasible solution to (5.17)-(5.18), then this optimization problem must have cost of optimal solution ≥ 0. On the other hand, if for all $t \le T$ iterations, the optimal value of (5.19) is ≥ 0, then $\hat{z} = \frac{1}{T}(\sum_{t=0}^{T-1} z^t)$ is such that (see [Arora et al., 2005,

Section 3.2])

$$a_\ell^T \hat{z} \geq b_\ell - \varepsilon, \quad \text{for all } 1 \leq \ell \leq 2N^2. \tag{5.21}$$

Now $\hat{z}$ corresponds to a choice model with support at most $O(KT) = O(\varepsilon^{-2} K \log N)$ permutations since each z^t is a choice model with support over K permutations in a signature family. Further this choice model, say $\hat{\lambda}$ (implied by $\hat{z}$) is such that $\|M(\hat{\lambda}) - Y\|_\infty \leq 2\varepsilon$.

Note that the complexity of the above described algorithm, for given subset of K signature components is polynomial in N. Therefore, the overall computation cost of the above described algorithm is dominated by term $\binom{N^2}{K}$ which is at most N^{2K}. That is, for any $K \geq 1$, the overall computation cost of the algorithm is bounded above by $\exp\big(\Theta(K \log N)\big)$. This also establishes the result of Theorem 15.

Utilizing the algorithm. Indeed, it is not clear a priori if for given observation Y, there exists a signature family of sparsity K within some small error $\varepsilon > 0$ with $\varepsilon \leq \varepsilon_0$ where ε_0 is most error we wish to tolerate. The natural way to adapt the above algorithm is as follows. Search over increasing values of K and for each K search for $\varepsilon = \varepsilon_0$. For the first K for which the algorithm succeeds, it may be worth optimizing over error ε by means of a binary search: $\varepsilon_0/2, \varepsilon_0/4, \ldots$. Clearly such a procedure would require $O(\log 1/\varepsilon)$ additional run of the same algorithm for the given K, where ε is the best precision we can obtain.

5.9 Proofs for Section 5.7

5.9.1 Proof of Theorem 14

In this section, we prove Theorem 14. We prove this theorem using the probabilistic method. Given the double stochastic matrix M, there exists a choice model (by Birkhoff-von Neumann's result) λ such that $M(\lambda) = M$. Suppose we draw T permutations (samples) independently according to the distribution λ. Let $\hat{\lambda}$ denote the empirical distribution based on these T samples. We show that for $T = N/\varepsilon^2$, on

average $\|M(\hat{\lambda}) - M\|_2 \leq \varepsilon$. Therefore, there must exist a choice model with $T = N/\varepsilon^2$ support size whose first-order marginals approximate M within an ℓ_2 error of ε.

To that end, let $\sigma_1, \sigma_2, \ldots, \sigma_T$ denote the T samples of permutations and $\hat{\lambda}$ be the empirical distribution (or choice model) that puts $1/T$ probability mass over each of the sampled permutations. Now consider a pair of indices $1 \leq i, j \leq N$. Let X_{ij}^t denote the indicator variable of the event that $\sigma_t(i) = j$. Since the permutations are drawn independently and in an identically distributed manner, X_{ij}^t are independent and identically distributed (i.i.d.) Bernoulli variables for $1 \leq t \leq T$. Further,

$$\mathbb{P}(X_{ij}^t = 1) \;=\; \mathsf{E}[X_{ij}^t] \;=\; M_{ij}.$$

Therefore, the (i, j) component of the first-order marginal $M(\hat{\lambda})$ of $\hat{\lambda}$ is the empirical mean of a Binomial random variable with parameters T and M_{ij}, denoted by $B(T, M_{ij})$. Therefore, with respect to the randomness of sampling,

$$\begin{aligned}
\mathsf{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T} X_{ij}^t - M_{ij}\right)^2\right] &= \frac{1}{T^2}\operatorname{Var}\Big(B(T, M_{ij})\Big) \\
&= \frac{1}{T^2} T M_{ij}(1 - M_{ij}) \\
&\leq \frac{M_{ij}}{T}, \qquad (5.22)
\end{aligned}$$

where we used the fact that $M_{ij} \in [0, 1]$ for all $1 \leq i, j \leq N$. Therefore,

$$\begin{aligned}
\mathsf{E}\Big[\|M(\hat{\lambda}) - M\|_2^2\Big] &= \mathsf{E}\Big[\sum_{ij}\Big(\frac{1}{T}\sum_{t=1}^{T} X_{ij}^t - M_{ij}\Big)^2\Big] \\
&\leq \sum_{ij} \frac{M_{ij}}{T} \\
&= \frac{N}{T}, \qquad (5.23)
\end{aligned}$$

where the last equality follows from the fact that M is a doubly stochastic matrix and hence its entries sum up to N. From (5.23), it follows that by selecting $T = N/\varepsilon^2$, the error in approximating the first-order marginals, $\|M(\hat{\lambda}) - M\|_2$, is within ε on average. Therefore, the existence of such a choice model follows by the probabilistic

method. This completes the proof of Theorem 14.

5.9.2 Proof of Theorem 16

We prove Theorem 16 using the probabilistic method as well. Suppose Y is a noisy observation of the first-order marginals $M(\lambda)$ of the underlying choice model λ. As per the hypothesis of the theorem 16, we assume that λ is either from the MNL family or the max-ent exponential family with the corresponding regularity conditions on the parameters. For such models λ, we establish the existence of a choice model $\hat{\lambda}$ that belongs to the signature family approximates $M(\lambda)$ (and hence approximates Y) well.

Fix a model λ, and as in the proof of Theorem 14, consider T permutations drawn independently and in an identical manner from the distribution λ. Let $\hat{\lambda}$ be the empirical distribution of these T samples as considered before. Following the arguments there, we obtain (like (5.23)) that

$$\mathsf{E}\Big[\|M(\hat{\lambda}) - M(\lambda)\|_2^2\Big] \leq \frac{N}{T}, \tag{5.24}$$

For the choice of $T = 4N/\varepsilon^2$, using Markov's inequality, it follows that

$$\mathbb{P}\Big(\|M(\hat{\lambda}) - M(\lambda)\|_2^2 \geq \varepsilon^2\Big) \leq \frac{1}{4}. \tag{5.25}$$

Since $\|M(\lambda) - M\|_2 \leq \varepsilon$, it follows that $\|M(\hat{\lambda}) - Y\|_2 \leq 2\varepsilon$ with probability at least 3/4.

Next, we show that $\hat{\lambda}$, thus generated satisfies signature conditions with a high probability (at least 1/2) as well. Therefore, by union bound we can conclude that $\hat{\lambda}$ satisfies the properties claimed by Theorem 16 with probability at least 1/4.

To that end, let $\mathcal{E}_t$ be the event that σ_t satisfies signature condition with respect to set $(\sigma_1, \ldots, \sigma_T)$. Since all $\sigma_1, \ldots, \sigma_T$ are chosen in an i.i.d. manner, the probability of each event is identical. We wish to show that $\mathbb{P}\Big(\cup_{1\leq t\leq T} \mathcal{E}_t^c\Big) \leq 1/2$. This will follow from establishing that $T\mathbb{P}(E_1^c) \leq 1/2$. To establish this, it is sufficient to show that $\mathbb{P}(E_1^c) \leq 1/N^2$ since $T = 4N/\varepsilon^2$.

To that end, suppose σ_1 is such that $\sigma_1(1) = i_1, \ldots, \sigma_1(N) = i_N$. Let $\mathscr{F}_j = \{\sigma_t(j) \neq i_j,\ 2 \leq t \leq T\}$. Then by definition of signature condition, it follows that

$$\mathcal{E}_1 = \cup_{j=1}^{N} \mathscr{F}_j.$$

Therefore,

$$\begin{aligned}
\mathbb{P}\Big(\mathcal{E}_1^c\Big) &= \mathbb{P}\Big(\bigcap_{j=1}^{N} \mathscr{F}_j^c\Big) \\
&\leq \mathbb{P}\Big(\bigcap_{j=1}^{L} \mathscr{F}_j^c\Big) \\
&= \mathbb{P}\Big(\mathscr{F}_1^c\Big) \prod_{j=2}^{L} \mathbb{P}\Big(\mathscr{F}_j^c \Big| \bigcap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\Big). \qquad (5.26)
\end{aligned}$$

We establish that the right hand side of (5.26) is bounded above by $O(1/N^2)$ and hence $T\mathbb{P}(\mathcal{E}_1^c) = O(\varepsilon^{-2}/N) \ll 1/2$ for N large enough as desired. To establish this bound of $O(1/N^2)$ under the two different conditions stated in Theorem 16, we consider in turn the two cases: (i) λ belongs to the MNL family with condition (5.7) satisfied, and (ii) λ belongs to the max-ent exponential family model with the condition (5.8) satisfied.

Bounding (5.26) *under the MNL model with* (5.7). Let $L = N^\delta$ for some $\delta > 0$ as in hypothesis of Theorem 16 under which (5.7) holds. Now

$$\begin{aligned}
\mathbb{P}\Big(\mathscr{F}_1^c\Big) &= 1 - \mathbb{P}\Big(\mathscr{F}_1\Big) \\
&= 1 - \mathbb{P}\Big(\sigma_t(1) \neq i_1;\ 2 \leq t \leq T\Big) \\
&= 1 - \mathbb{P}\Big(\sigma_2(1) \neq i_1\Big)^{T-1} \\
&= 1 - \Big(1 - \frac{w_{i_1}}{\sum_{k=1}^{N} w_k}\Big)^{T-1}. \qquad (5.27)
\end{aligned}$$

For $j \geq 2$, to evaluate $\mathbb{P}\Big(\mathscr{F}_j^c | \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\Big)$, we evaluate $1 - \mathbb{P}\Big(\mathscr{F}_j | \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\Big)$. To evaluate, $\mathbb{P}\Big(\mathscr{F}_j | \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\Big)$, note that the conditioning event $\cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c$ suggets that for each σ_t, $2 \leq t \leq T$, some assignments (ranks) for first $j-1$ items are given and we need

to find the probability that jth item of each of the $\sigma_2, \ldots, \sigma_T$ are not mapped to i_j. Therefore, given $\cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c$, the probability that $\sigma_2(j)$ does map to i_j is $w_j/(\sum_{k\in X} w_k)$, where X is the set of $N-j+1$ elements that does not include the $j-1$ elements to which $\sigma_2(1), \ldots, \sigma_2(j-1)$ are mapped to. Since by assumption (without loss of generality), $w_1 < \ldots < w_N$, it follows that $\sum_{k\in X} w_k \geq \sum_{k=1}^{N-j+1} w_k$. Therefore,

$$\mathbb{P}\Big(\mathscr{F}_j \Big| \bigcap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\Big) \geq \Big(1 - \frac{w_{i_j}}{\sum_{k=1}^{N-j+1} w_k}\Big)^{T-1}. \tag{5.28}$$

Therefore, it follows that

$$\begin{aligned}
\mathbb{P}\Big(\mathcal{E}_1^c\Big) &\leq \prod_{j=1}^{L} \Big[1 - \Big(1 - \frac{w_{i_j}}{\sum_{k=1}^{N-j+1} w_k}\Big)^{T-1}\Big] \\
&\leq \prod_{j=1}^{L} \Big[1 - \Big(1 - \frac{w_N}{\sum_{k=1}^{N-j+1} w_k}\Big)^{T-1}\Big] \\
&\leq \prod_{j=1}^{L} \Big[1 - \Big(1 - \frac{w_N}{\sum_{k=1}^{N-L+1} w_k}\Big)^{T-1}\Big] \\
&= \Big[1 - \Big(1 - \frac{w_N}{\sum_{k=1}^{N-L+1} w_k}\Big)^{T-1}\Big]^L.
\end{aligned} \tag{5.29}$$

Let $W(L,N) = w_N/(\sum_{k=1}^{N-L+1} w_k)$. By hypothesis of Theorem 16, it follows that $W(L,N) \leq \sqrt{\log N}/N$ and $L = N^\delta$. Therefore, from above it follows that

$$\begin{aligned}
\mathbb{P}\Big(\mathcal{E}_1^c\Big) &\leq \Big[1 - \Big(1 - \frac{\sqrt{\log N}}{N}\Big)^{T-1}\Big]^L \\
&\leq \Big[1 - \Theta\Big(\exp\Big(-\frac{T\sqrt{\log N}}{N}\Big)\Big)\Big]^L,
\end{aligned} \tag{5.30}$$

where we have used the fact that $1 - x = \exp(-x)(1 + O(x^2))$ for $x \in [0,1]$ (with $x = \sqrt{\log N}/N$) and since $T = N/\varepsilon$, $(1 + O(\log N/N^2))^T = 1 + o(1) = \Theta(1)$. Now

$$\exp\Big(-\frac{T\sqrt{\log N}}{N}\Big) = \exp\Big(-4\sqrt{\log N}/\varepsilon^2\Big) \ll 1. \tag{5.31}$$

Therefore, using the inequality $1 - x \leq \exp(-x)$ for $x \in [0, 1]$, we have

$$\mathbb{P}\Big(\mathcal{E}_1^c\Big) \leq \exp\Big(- L \exp(-4\sqrt{\log N}/\varepsilon^2)\Big). \tag{5.32}$$

Since $L = N^\delta$ for some $\delta > 0$ and $\exp(-4\sqrt{\log N}/\varepsilon^2) = o(N^{\delta/2})$ for any $\delta > 0$, it follows that

$$\begin{aligned}\mathbb{P}\Big(\mathcal{E}_1^c\Big) &\leq \exp\Big(- \Theta\Big(N^{\delta/2}\Big)\Big) \\ &\leq O(1/N^2).\end{aligned} \tag{5.33}$$

Therefore, it follows that all the T samples satisfy the signature condition with respect to each other with probability at least $O(1/N) \leq 1/4$ for N large enough. Therefore, we have established the existence of desired sparse choice model in signature family. This completes the proof of Theorem 16 under MNL model with condition (5.7).

Bounding (5.26) *under the max-ent exponential family with* (5.8). As before, let $L = N^\delta$ for some $\delta > 0$ (choice of $\delta > 0$ here is arbitrary; for simplicity, we shall think of this δ as being same as that used above). Now

$$\begin{aligned}\mathbb{P}\Big(\mathscr{F}_1^c\Big) &= 1 - \mathbb{P}\Big(\mathscr{F}_1\Big) \\ &= 1 - \mathbb{P}\Big(\sigma_t(1) \neq i_1;\ 2 \leq t \leq T\Big) \\ &= 1 - \mathbb{P}\Big(\sigma_2(1) \neq i_1\Big)^{T-1}.\end{aligned} \tag{5.34}$$

To bound the right hand side of (5.34), we need to carefully understand the implication of (5.8) for the exponential family distribution. For that, first consider the simple case where the parameters θ_{ij} are all equal. In that case, it is easy to see that all permutations have equal $(1/N!)$ probability assigned and hence the probability $\mathbb{P}(\sigma_2(1) \neq i_1)$ equals $1 - 1/N$. However such an evaluation (or bounding) is not straightforward for general parameter values because it involves computation of the 'partition' function. To that end, consider $1 \leq i \neq i' \leq N$. Now by the definition of

exponential family (and σ_2 is chosen as per it),

$$\mathbb{P}\Big(\sigma_2(1)=i\Big) = \frac{1}{Z(\theta)}\Big[\sum_{\sigma\in S_N(1\to i)} \exp\Big(\sum_{kl}\theta_{kl}\sigma_{kl}\Big)\Big]$$
$$= \frac{\exp\big(\theta_{1i}\big)}{Z(\theta)}\Big[\sum_{\sigma\in S_N(1\to i)} \exp\Big(\sum_{k\neq 1,l}\theta_{kl}\sigma_{kl}\Big)\Big]. \tag{5.35}$$

In above $S_N(1\to i)$ denotes the set of all permutations in S_N that map 1 to i:

$$S_N(1\to i) = \Big\{\sigma\in S_N : \sigma(1)=i\Big\}.$$

Given this, it follows that

$$\frac{\mathbb{P}\Big(\sigma_2(1)=i\Big)}{\mathbb{P}\Big(\sigma_2(1)=i'\Big)} = \frac{\exp(\theta_{1i})\Big[\sum_{\sigma\in S_N(1\to i)}\exp\Big(\sum_{k\neq 1,l}\theta_{kl}\sigma_{kl}\Big)\Big]}{\exp(\theta_{1i'})\Big[\sum_{\rho\in S_N(1\to i')}\exp\Big(\sum_{k\neq 1,l}\theta_{kl}\rho_{kl}\Big)\Big]}. \tag{5.36}$$

Next, we consider a one-to-one and onto map from $S_N(1\to i)$ to $S_N(1\to i')$ (which are of the same cardinality). Our goal is to construct a mapping such that if $\sigma\in S_N(1\to i)$ is mapped to $\rho\in S_N(1\to i')$, then we have that

$$\exp\Big(\sum_{kl}\sigma_{kl}\theta_{kl}\Big) \leq \sqrt{\log N}\exp\Big(\sum_{kl}\rho_{kl}\theta_{kl}\Big). \tag{5.37}$$

In that case, (5.36) implies that

$$\frac{\mathbb{P}\Big(\sigma_2(1)=i\Big)}{\mathbb{P}\Big(\sigma_2(1)=i'\Big)} \leq \sqrt{\log N}. \tag{5.38}$$

This in turn implies that for any i, $\mathbb{P}(\sigma_2(1)=i)\leq\sqrt{\log N}/N$, which we can use in bounding (5.34).

To that end,we consider the following mapping from $S_N(1\to i)$ to $S_N(1\to i')$. For any $\sigma\in S_N(1\to i)$, it follows by definition that $\sigma(1)=i$. Let q be such that $\sigma(q)=i'$. Then map σ to $\rho\in S_N(1\to i')$ where $\rho(1)=i'$, $\rho(q)=i$ and $\rho(k)=\sigma(k)$

for $k \neq 1, q$. Then,

$$\frac{\exp\Big(\sum_{kl} \sigma_{kl}\theta_{kl}\Big)}{\exp\Big(\sum_{kl} \rho_{kl}\theta_{kl}\Big)} = \frac{\exp\big(\theta_{1i} + \theta_{qi'}\big)}{\exp\big(\theta_{1i'} + \theta_{qi}\big)}$$
$$\leq \sqrt{\log N}, \tag{5.39}$$

where the last inequality follows from condition (5.8) in the statement of Theorem 16. From the above discussion, we conclude that

$$\mathbb{P}\Big(\mathscr{F}_1^c\Big) = 1 - \mathbb{P}\Big(\sigma_2(1) \neq i_1\Big)^{T-1}$$
$$\leq 1 - \Big(1 - \frac{\sqrt{\log N}}{N}\Big)^{T-1}. \tag{5.40}$$

For $j \geq 2$, in order to evaluate $\mathbb{P}\Big(\mathscr{F}_j^c | \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\Big)$, we evaluate $1 - \mathbb{P}\Big(\mathscr{F}_j | \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\Big)$. To evaluate, $\mathbb{P}\Big(\mathscr{F}_j | \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\Big)$, note that the conditioning event $\cap_{\ell=1}^{j-1}\mathscr{F}_\ell^c$ suggets that for each σ_t, $2 \leq t \leq T$, some assignments (ranks) for first $j-1$ items are given and we need to find the probability that jth item of each of the $\sigma_2, \ldots, \sigma_T$ are not mapped to i_j. Therefore, given $\cap_{\ell=1}^{j-1}\mathscr{F}_\ell^c$, we wish to evaluate (an upper bound on) probability of $\sigma_2(j)$ mapping i_j given that we know assignments of $\sigma_2(1), \ldots, \sigma_2(j-1)$. By the form of the exponential family, conditioning on the assignments $\sigma_2(1), \ldots, \sigma_2(j-1)$, effectively we have an exponential family on the space of permutations of remaining $N - j + 1$ element. And with respect to that, we wish to evaluate bound on the marginal probability of $\sigma_2(j)$ mapping to i_j. By an argument identical to the one used above to show that $\mathbb{P}(\sigma_2(1) = i) \leq \sqrt{\log N}/N$, it follows that

$$\mathbb{P}\Big(\sigma_2(j) = i_j | \bigcap \mathscr{F}_j^c\Big) \leq \frac{\sqrt{\log N}}{N - j + 1}$$
$$\leq \frac{2\sqrt{\log N}}{N}, \tag{5.41}$$

where we have used the fact that $j \leq L = N^\delta \leq N/2$ (for N large enough). Therefore,

it follows that

$$\mathbb{P}\Big(\mathcal{E}_1^c\Big) \leq \Big[1 - \Big(1 - \frac{2\sqrt{\log N}}{N}\Big)^{T-1}\Big]^L. \tag{5.42}$$

From here on, using arguments identical to those used above (under MNL model), we conclude that

$$\begin{aligned}\mathbb{P}\Big(\mathcal{E}_1^c\Big) &\leq \exp\Big(-\Theta\Big(N^{\delta/2}\Big)\Big) \\ &\leq O(1/N^2).\end{aligned} \tag{5.43}$$

This completes the proof for max-ent exponential family with condition (5.8) and hence that of Theorem 16.

5.10 Noisy setting: greedy heuristic

In Section 5.8 we proposed a novel scheme based on the multiplicative weight update idea used by Plotkin et al. [1991]. For the case of first-order marginal information, we established that the running time complexity of the algorithm is $O(\exp(K \log N))$ provided there is a choice model in the signature family with sparsity K that is good fit to the data. As established in Theorem 16, for a large class of models, the sparsity K scales as $O(N)$ (ignoring the ε dependence), resulting in a computational complexity for our algorithm the is polynomial in $N! = \exp(\Theta(N \log N))$. While our algorithm has managed to obtain a significant reduction in complexity from the brute-force approach, anything polynomial in $N!$ is still prohibitive for many common applications in practice. In order to address this situation, we present in this section a greedy heuristic called the *greedy sparsest-fit* algorithm, which is based on the sparsest-fit algorithm proposed above in Section 5.5; specifically, the greedy sparsest-fit algorithm is a generalization of the sparsest-fit algorithm to the case when the observations may be corrupted by noise. The reason for the specific modifier 'greedy' in the name of the algorithm will become clear once we describe the algorithm.

We can guarantee that the sparsest-fit algorithm always outputs a valid distribution (see Theorem 17 described below), and in addition, if the underlying model λ satisfies the signature and linear independence conditions and the marginal information is noise-free, then it recovers λ (see Theorem 18 described below). In this sense, the greedy sparsest-fit algorithm subsumes the sparsest-fit algorithm. Moreover, we can guarantee that the running time of the algorithm is essentially polynomial in N for first-order marginals and the type of data obtained from sales transactions. Unfortunately, we cannot provide any guarantees on the sparsity of the models that will be identified – and hence we call it a heuristic. However, we find that this algorithm has good performance in practice, as discussed in Section 5.11 in the context of our empirical study.

We now describe the greedy sparsest algorithm. The theorems that establish the algorithm's correctness are described subsequently. We describe the algorithm assuming that we have access to (possibly noisy) observations of ρ-order marginal information. In order to keep the exposition simple, we don't describe the extension of the greedy sparsest-fit algorithm to transaction type data; nevertheless, it does extend in a straightforward manner.

Loosely speaking, the greedy sparsest-fit algorithm outputs a valid distribution (see Theorem 17) with "as few permutations in the support as possible" while approximating the given partial information "as closely as possible." Similar to the sparsest-fit algorithm, the greedy sparsest-fit algorithm processes the components of the marginal information matrix Y sequentially and build the permutations in the support incrementally. While processing each element, the greedy algorithm avoids introducing a new permutation into the support by trying to express the element using the permutations already in the support. In this sense, the algorithm makes 'local' greedy decisions. Local greedy decisions can however lead to meaningless distributions if not done carefully because of the complicated dependence structure of permutations. The 'greedy sparsest-fit' algorithm avoids this issue by using a global view obtained by carefully exploiting permutation structure while making local decisions.

More precisely, we assume access to observations matrix Y that is related to the underlying choice model λ as $Y = M^{\rho}(\lambda) + \eta$, where η is the noise vector and ρ is some partition of N. We consider the general case, where λ may not satisfy either the signature or the linear independence conditions. As mentioned above, the algorithm exploits the permutation structure to make local greedy decisions with a global view. In order to reveal this structure, it is useful to represent the $D_\rho \times D_\rho$ matrix Y as a $D_\rho \times D_\rho$ weighted bipartite graph with the nodes on each side corresponding to distinct ρ-partitions of N and the weight of edge (t_i, t_j) equal to Y_{t_i,t_j}. With this representation, it is easy to see that the matrix $A^{\rho}(\sigma)$ of any permutation σ can be thought of as a $D_\rho \times D_\rho$ perfect matching. Thus, Y is a weighted combination of perfect matchings with weights $\lambda(\sigma)$ for the corresponding σ. Note that every $D_\rho \times D_\rho$ perfect matching does not correspond to a valid permutation[5] (except of course when $\rho = (N-1, 1)$). Therefore, we call a perfect matching that corresponds to a permutation a 'valid perfect matching'.

Before we describe the algorithm, we make the following assumptions. We assume that every element of the data matrix Y is non-negative. We let $\{q_1, q_2, \ldots, q_L\}$ denote the non-zero data elements sorted in ascending order $q_1 \leq q_2 \leq \cdots \leq q_L$; each index $1 \leq \ell \leq L$ corresponds to an edge in the weighted bipartite graph, or equivalently, a pair of ρ-partitions t_i, t_j. At each stage, the algorithm maintains partial rankings, which we think of as partial matchings in a $D_\rho \times D_\rho$ bipartite graph that can be completed into valid perfect matchings. The algorithm maintains a partial ranking as a set of indices/edges $B \subset \{1, 2, \ldots, L\}$.

The algorithm also assumes access to what we call a 'complete-order' oracle: given sets of edges $B, S \subset \{1, 2, \ldots, L\}$, the oracle outputs whether B can be completed into a *valid* perfect matching using the edges in set S. More formally, we assume

[5]Of course, there are $D_\rho!$ possible perfect matchings and only $N!$ permutations.

access to a function $\mathsf{oracle}(B, S)$ such that

$$\mathsf{oracle}(B,S) = \begin{cases} \mathsf{true} & \text{if } \exists \text{ valid perfect matching } M \text{ s.t.} \\ & B \subset M \subset B \cup S \\ \mathsf{false} & \text{otherwise.} \end{cases}$$

Since not every perfect matching corresponds to a permutation, this entails more than just determining whether there exists perfect matching that contains edges in B and uses edges in S. Therefore, except in the first-order case, one may not be able to design a computationally efficient 'complete-order' oracle. Going into the details of how one can overcome this issue is beyond the scope of the current work. It suffices to mention here that depending on the situation at hand one, can design efficient randomized oracles that output false whenever there is no valid perfect matching that satisfies $A \subset M \subset A \cup S$, and true with probability $1 - \varepsilon$ whenever there exists such a perfect matching. The error rate of ε is the price one pays for computational efficiency.

Description. The formal description of the algorithm in given in Figure 5.10. Similar to the sparsest-fit algorithm, the algorithm takes as inputs a sequence of positive values sorted in ascending order $q_1 \leq q_2 \leq \cdots \leq q_L$, a tolerance parameter $\varepsilon > 0$, and a mapping of the indices $1 \leq \ell \leq L$ to edges of a $D_\rho \times D_\rho$ bipartite graph. Given these inputs, the algorithm strives to find a distribution that matches the given data q_ℓ to within a relative error of ε while trying to use as few permutations as possible in the support.

The general idea of the algorithm is as follows. The algorithm processes the values $q_1, q_2, \ldots, q_L$ sequentially. Roughly speaking, for each edge, the algorithm makes an attempt to greedily include it into one of the (partial) permutations already in the support; only in cases it cannot, the algorithm introduces a new permutation into the support. When attempting to include the edge into one of the permutations already in the support, the algorithm uses q_ℓ and the values $p_1, p_2, \ldots, p_{k(\ell)}$ of the permutations already in the support as one of the signals. Since greedy local decisions can result in

ending up with invalid or incomplete permutations in the support, the algorithm uses $\mathsf{oracle}(\cdot,\cdot)$ to narrow down the set of permutations in the support where the edge can be added. This issue was avoided in the sparsest-fit algorithm through the 'unique witness' and 'linear independence' conditions.

More formally, in each iteration ℓ, the algorithm maintains the number of (partial) permutations in the support at the end of iteration ℓ as $k(\ell)$. Partial permutations are maintained as sets B_k, which at the end of iteration ℓ are such that $B_k \subset \{1, 2, \ldots, \ell\}$ and correspond to a partial $D_\rho \times D_\rho$ matchings that are subsets of valid perfect matchings. The values corresponding to the partial permutations are maintained as $p_1, p_2, \ldots, p_{k(\ell)}$. Finally, we maintain a set of 'left-over' edges S_ℓ, which at the end of iteration ℓ contains all edges (or indices in $\{1, 2, \ldots, L\}$) except the ones already processed. We define S_ℓ such that edge ℓ is not present in S_ℓ. A the beginning of the algorithm, $k(0)$ is initialized to zero, S_0 is initialized to contain all the L edges, and the sets B_k are initialized to be empty.

At the beginning of iteration ℓ, the algorithm first determines a set of 'feasible' permutations $\mathsf{feas\text{-}perms}$ containing indices k such that after edge ℓ is added to B_k, it is still possible to complete it into a valid perfect matching using only the left over edges in S_ℓ i.e, $\mathsf{feas\text{-}perms} = \{1 \leq k \leq k(\ell) \colon \mathsf{oracle}(B_k \cup \{\ell\}, S_\ell) = \mathsf{true}\}$. In addition, the algorithm determines a set of 'must be added to' permutations $\mathsf{must\text{-}perms} \subset \mathsf{feas\text{-}perms}$ containing indices k such that if edge ℓ is not added to B_k, then it cannot be completed into a valid perfect matching using only the available edges in S_ℓ i.e., $\mathsf{must\text{-}perms} = \mathsf{feas\text{-}perms} \cap \{1 \leq k \leq k(\ell) \colon \mathsf{oracle}(B_k, S_\ell) = \mathsf{false}\}$. It is by determining $\mathsf{feas\text{-}perms}$ and $\mathsf{must\text{-}perms}$ that the algorithm maintains a "global view." Moreover, it is the determination of $\mathsf{feas\text{-}perms}$ and $\mathsf{must\text{-}perms}$ that requires exploitation of the permutation structure through $\mathsf{oracle}(\cdot,\cdot)$. The algorithm then adds edge ℓ to all sets B_k such that $k \in \mathsf{must\text{-}perms}$. It removes the weight added from the edge i.e., updates q_ℓ to $q_\ell - \sum_{k \in \mathsf{must\text{-}perms}} p_k$. If the leftover weight is "too small" i.e., if updated value q_ℓ is less than ε times its original value $q_\ell + \sum_{k \in \mathsf{must\text{-}perms}} p_k$, then the algorithm stops the iteration here and moves to the next iteration.

If the leftover weight q_ℓ is large enough then the algorithm tries to include the

edge in some of the 'feasible' permutations. More precisely, the algorithm determines a subset $T \subset$ feas-perms $\setminus$ must-perms such that the error $|q_\ell - \sum_{k \in T} p_k|$ is small as possible; let's call the subset the gives the smallest error best-set and the corresponding error best-error. If the best error is within εq_ℓ, then edge ℓ is added to the permutations in best-set and the algorithm moves onto the next iteration.

If the best-error is not within εq_ℓ, then the algorithm is forced to introduce a new permutation into the support. Before introducing the new permutation though, the algorithm checks whether it is possible to create a complete permutation using only the edges available in set S_ℓ; in particular, it checks whether oracle$(\{\ell\}, S_\ell)$ is true. If it is indeed possible to add a new permutation to the support, the algorithm increments $k(\ell - 1)$ to $k(\ell) = k(\ell - 1) + 1$ and adds the edge to the set $B_{k(\ell)}$. In case it is not possible to create a new permutation with the existing edges i.e., oracle$(\{\ell\}, S_\ell) =$ false, then the algorithm includes the edge in the best-set although best-error may not be within εq_ℓ. Note that in this case, it is possible for feas-perms to be empty, in which case this edge would not be added to any permutation in the support.

Complexity of the algorithm. The derivation of the complexity of the algorithm is similar to that of the sparsest-fit algorithm. To start with, we sort at most D_ρ^2 elements. This has a complexity of $O(D_\rho^2 \log D_\rho)$. Further, note that the **for** loop in the algorithm iterates for at most D_ρ^2 times. In each iteration, we call the oracle$(\cdot, \cdot)$ at most $2K$ times. In addition, we solve a subset-sum problem, and since there are at most K elements, the worst-case complexity of subset-sum in each iteration is $O(2^K)$. Thus, the worst-case complexity of the algorithm is $O(D_\rho^2 \log D_\rho + D_\rho^2 2^K C(\text{oracle}))$, where $C(\text{oracle})$ denotes the computational complexity of the oracle. For the first-order information, it is known that we can construct such an oracle with a computational complexity of $O(N^{2.5})$ (see Micali and Vazirani [1980], Edmonds and Karp [1972]). Similar to the case of the sparsest-fit algorithm, the average case complexity of the algorithm can be shown to be much smaller. Specifically, suppose the ρ-order partial information comes from an underlying model λ that is drawn from the random model $R(K, \mathscr{C})$. Then, using the standard balls and bins argument, we can prove

Figure 5-1: Greedy sparsest-fit algorithm

Input: Tolerance parameter $\varepsilon > 0$ and non-negative values $\{q_1, q_2, \ldots, q_L\}$ sorted in ascending order i.e., $q_1 \leq q_2 \leq \cdots \leq q_L$ and a mapping of indices $1 \leq \ell \leq L$ to edges in $D_\rho \times D_\rho$ bipartite graph.

Output: Positive numbers $\{p_1, p_2, \ldots, p_K\}$ s.t. $\sum_{k=1}^{K} p_k = 1$, and for $1 \leq k \leq K$ set $B_k \subset \{1, 2, \ldots, L\}$ s.t. B_k corresponds to a valid permutation.

Algorithm:

initialization: $p_0 = 0, k(0) = 0$, $B_k = \emptyset$ for all possible k, $S_0 = \{1, 2, \ldots, L\}$.
for $\ell = 1$ to L
  $S_\ell = S_{\ell-1} \setminus \{\ell\}$
  feas-perms $\leftarrow \left\{1 \leq k' \leq k(\ell-1) : \right.$
    $\left. \text{oracle}(B_{k'} \cup \{\ell\}, S_\ell) = \text{true} \right\}$
  must-perms $\leftarrow$ feas-perms $\cap \left\{1 \leq k' \leq k(\ell-1) : \right.$
    $\left. \text{oracle}(B_{k'}, S_\ell) = \text{false} \right\}$
  $B_{k'} \leftarrow B_{k'} \cup \{\ell\}$ $\forall\, k' \in$ must-perms
  $q_\ell \leftarrow q_\ell - \sum_{k' \in \text{must-perms}} p_{k'}$
  if $q_\ell < \varepsilon(q_\ell + \sum_{k' \in \text{must-perms}} p_{k'})$
    $k(\ell) = k(\ell-1)$
    skip rest of loop
  end if
  cand-perms = feas-perms $\setminus$ must-perms
  best-set $\leftarrow \arg\min_{T \subset \text{cand-perms}} |q_\ell - \sum_{k' \in T} p_{k'}|$
  best-error $\leftarrow |q_\ell - \sum_{k' \in \text{best-set}} p_{k'}|$
  if best-error $< \varepsilon q_\ell$
    $B_{k'} \leftarrow B_{k'} \cup \{\ell\}$ $\forall\, k' \in$ best-set
    $k(\ell) = k(\ell-1)$
  else if oracle$(\{\ell\}, S_\ell) =$ true
    $k(\ell) = k(\ell-1) + 1$
    $p_{k(\ell)} = q_\ell$
    $B_{k(\ell)} \leftarrow B_{k(\ell)} \cup \{\ell\}$
  else
    $B_{k'} \leftarrow B_{k'} \cup \{\ell\}$ $\forall\, k' \in$ best-set
    $k(\ell) = k(\ell-1)$
  end if
end for
$p_{k'} \leftarrow p_{k'} / \sum_{k''=1}^{k(L)} p_{k''}$ for $k' = 1$ to $k(L)$
Output $K = k(L)$ and $(p_{k'}, B_{k'}), 1 \leq k' \leq K$.

that for $K = O(D_\rho \log D_\rho)$, with a high probability, there are at most $O(\log D_\rho)$ elements in each subset-sum problem. Thus, for the first-order marginal information, the complexity would then be $O\left(\exp(\log^2 D_\rho)\right)$ with a high probability.

We now state two theorems about the correctness of the algorithm. The first theorem establishes the correctness of the algorithm, and the second theorem establishes that the greedy sparsest-fit algorithm subsumes the sparsest-fit algorithm.

Theorem 17. *Given any tolerance parameter $\varepsilon > 0$ and non-negative partial information $\{q_1, q_2, \ldots, q_L\}$ corresponding to partition ρ, the 'greedy' heuristic always outputs a valid distribution over permutations, provided there exists at least one valid perfect matching in the edges $\{1, 2, \ldots, L\}$.*

Theorem 18. *Consider the noiseless case where the partial information $\{q_1, q_2, \ldots, q_L\}$ is generated by a distribution λ that satisfies the signature and linear independence conditions. Then, the greedy heuristic with $\varepsilon = 0$ recovers λ exactly.*

Next, we present the proofs of the two theorems.

5.10.1 Proof of Theorem 17

Note that the algorithm normalizes the values so that $\sum_{k=1}^{K} p_k = 1$. Therefore, we only need to prove that the non-empty sets of edges B_k, $1 \leq k \leq K$, correspond to permutations (valid perfect matchings). For that, we claim that the following statement is a loop-invariant: at the end of every iteration $\ell \in \{1, 2, \ldots, L\}$, for every non-empty set B_k, there exists a valid perfect matching $M_{k,\ell}$ such that $B_k \subset M_{k,\ell} \subset B_k \cup S_\ell$. Before we prove the claim, note that the result of the theorem readily follows from this claim. In particular, at the end of iteration L, $S_L = \emptyset$ and hence for every non-empty set B_k, we must have $B_k \subset M_{k,L} \subset B_k$, which implies that every non-empty subset B_k corresponds to a valid permutation. Furthermore, by hypothesis, since there exists at least one valid perfect matching in the set $\{1, 2, \ldots, L\}$, it follows that there exists at least one non-empty set B_k.

We are now only left with proving the above claim, for which we use induction on ℓ. Let $1 \leq \ell_0 < L$ be the first iteration such that B_1 is non-empty at the end

of the iteration. Note that $\ell_0 < L$ by the hypothesis that there exists at least one valid perfect matching in the set $\{1, 2, \ldots, L\}$. By the end of iteration ℓ_0, edge ℓ_0 will have been added to set B_1 resulting in $B_1 = \{\ell_0\}$. This necessarily implies that $\mathsf{oracle}(\{\ell_0\}, S_{\ell_0}) = \mathsf{true}$, and it follows from the definition of $\mathsf{oracle}(\cdot, \cdot)$ that there exists a valid prefect matching M such that

$$\begin{aligned} & \{\ell_0\} \subset M \subset \{\ell_0\} \cup S_{\ell_0} \\ \implies \quad & B_1 \subset M \subset B_1 \cup S_{\ell_0}, \end{aligned}$$

where the implication follows from the fact that $B_1 = \{\ell_0\}$. This proves the base case.

Now assuming that the claim is true for $\ell - 1$, we prove it for $\ell_0 < \ell \leq L$. Let $\mathcal{K}_\ell$ denote the set of ks such that edge ℓ was added to set B_k during iteration ℓ i.e., at the end iteration ℓ, $\ell \in B_k \; \forall \; k \in \mathcal{K}_\ell$. It now follows from the algorithm that for every $k \in \mathcal{K}_\ell$, we must have $\mathsf{oracle}(B_k, S_\ell) = \mathsf{true}$, which implies that for every $k \in \mathcal{K}_\ell$ there exists a valid perfect matching $M_{k,\ell}$ such that

$$B_k \subset M_{k,\ell} \subset B_k \cup S_\ell.$$

Now consider any k such that $k \notin \mathcal{K}_\ell$ and B_k is non-empty. Since edge ℓ was not added to B_k, it must be that $k \notin \mathsf{must\text{-}perms}$. This means that either $\mathsf{oracle}(B_k \cup \{\ell\}, S_\ell) = \mathsf{false}$ or $\mathsf{oracle}(B_k, S_\ell) = \mathsf{true}$. In the first case, since by induction hypothesis there exists a valid perfect matching M such that $B_k \subset M \subset B_k \cup S_\ell \cup \{\ell\}$, it must be the case that M does not contain edge ℓ and hence $M \subset B_k \cup S_\ell$. In the second case, it follows from the definition of $\mathsf{oracle}(\cdot, \cdot)$ that there exists a valid perfect matching such that $B_k \subset M \subset B_k \cup S_\ell$. This establishes the claim for ℓ assuming the claim is true for $\ell - 1$. The claim now follows by induction.

5.10.2 Proof of Theorem 18

Since λ satisfies the signature and linear independence conditions, we can run the sparsest-fit algorithm to obtain the support and values of λ as (B'_k, p'_k) for $1 \leq k \leq K'$. Suppose we also store the number of partial rankings $k'(\ell)$ at the end of each iteration $1 \leq \ell \leq L$ of the sparsest-fit algorithm. For convenience, we let $B'_k = \emptyset$ for $k > K'$. In addition, let B_k be the sets of edges and p_k the values obtained from running the greedy heuristic with $\varepsilon = 0$. We claim that at the end of any iteration ℓ of the greedy heuristic, $B_k \subset B'_k \subset B_k \cup S_\ell$ for all possible k; $p_k = p'_k$ for all $1 \leq k \leq k(\ell)$; and $k(\ell) = k'(\ell)$. Assuming this claim is true, since $S_L = \emptyset$, it immediately follows that the greedy heuristic recovers λ exactly. We are now only left with proving the claim.

We prove the claim using induction on ℓ. As the base case, let's consider the first iteration. Because of the signature condition, it follows that $q_1 = p'_1$; further, $\mathsf{oracle}(\{1\}, S_1) = \mathsf{true}$ because $\{1\} \subset B'_1 \subset \{1\} \cup S_1$. Hence, at the end of first iteration of the greedy heuristic we will have $B_1 = \{1\}$, $p_1 = p'_1$, and $B_k = \emptyset$ for all $k \geq 2$. Since *only* B_1 contains edge 1 because of the signature condition, it follows that $B_k \subset B'_k \subset B_k \cup S_1$ for all possible k, $p_1 = p'_1$ and $k(1) = k'(1) = 1$. This establishes the base case.

Now assuming that the claim is true for all $\ell - 1$, we prove it for $1 < \ell \leq L$. Let $\mathcal{K}'_\ell$ denote the set of ks such that B'_k contains edge ℓ, and consider the beginning of iteration ℓ. Note that it follows by induction hypothesis that at the beginning of iteration ℓ

$$\begin{aligned} B_k \subset B'_k \subset B_k \cup S_\ell, &\quad \text{for any } k \notin \mathcal{K}'_\ell, \\ B_k \cup \{\ell\} \subset B'_k \subset B_k \cup \{\ell\} \cup S_\ell, &\quad \text{for any } k \in \mathcal{K}'_\ell. \end{aligned}$$

Therefore, in order to prove the induction step, it is sufficient to to prove two things:

1. During iteration ℓ, edge ℓ is added to sets B_k for k in $\mathcal{K}'_\ell$ and only in $\mathcal{K}'_\ell$, and
2. If $k(\ell) = k(\ell - 1) + 1$, then $p_{k(\ell)} = p'_{k(\ell)}$ and $k(\ell) = k'(\ell)$.

Before we prove the two things mentioned above, we make the following observa-

tions. For any $k \in \mathcal{K}'_\ell$, it must follow by induction hypothesis that $B_k \cup \{\ell\} \subset B'_k \subset B_k \cup \{\ell\} \cup S_\ell$ at the beginning of iteration ℓ. This means that $\mathsf{oracle}(B_k \cup \{\ell\}, S_\ell) = \mathsf{true}$ for any $k \in \mathcal{K}'_\ell$. This leads us to conclude that

$$\mathcal{K}'_\ell \cap \{k\colon B_k \neq \emptyset\} \subset \textsf{feas-perms at start of iteration } \ell. \tag{5.44}$$

Similarly, it follows from the induction hypothesis that, for any $k \notin \mathcal{K}'_\ell$, $B_k \subset B'_k \subset B_k \cup S_\ell$ at the beginning of iteration ℓ because $\ell \notin B'_k$. Therefore, $\mathsf{oracle}(B_k, S_\ell) = \mathsf{true}$ leading us to conclude that

$$\textsf{must-perms} \subset \mathcal{K}'_\ell \cap \{k\colon B_k \neq \emptyset\} \textsf{ at start of iteration } \ell. \tag{5.45}$$

To prove the induction step, we consider two cases. First, consider the case when we introduce a new permutation into the support i.e., when $K'_\ell = \{k_\ell\}$ and $\textsf{must-perms} = \emptyset$. Note that because of the linear independence condition and the fact that $\varepsilon = 0$, edge ℓ will not be added to any B_k for $k \in \textsf{feas-perms}$. Further, note that $\mathsf{oracle}(\{\ell\}, S_\ell) = \mathsf{true}$ since $\{\ell\} \subset B'_{k_\ell} \subset \{\ell\} \cup S_\ell$. Thus, the algorithm adds edge ℓ to set $B_{k(\ell)}$, where $k(\ell) = k(\ell-1)+1 = k'(\ell-1)+1$, the last equality following from the induction hypothesis that $k(\ell-1) = k'(\ell-1)$. In addition, since must-perms is empty, we will have $p_{k(\ell)} = q_\ell = p'_{k_\ell}$ (because of unique witness condition). The only thing we are now left to prove is that $k(\ell) = k'(\ell) = k_\ell$. To prove this, we argue that ℓ is the lowest index in the set B'_{k_ℓ}. If ℓ is the first edge that is added to set B'_{k_ℓ}, it follows from the property of the sparsest-fit algorithm that $k'(\ell) = k_\ell$ and $k'(\ell-1) = k'(\ell) - 1$. This immediately implies that $k(\ell) = k(\ell-1) + 1 = k'(\ell-1) + 1 = k'(\ell) = k_\ell$. Therefore, in this case, edge ℓ will be added to B_k for k in $\mathcal{K}'_\ell$ and only in $\mathcal{K}'_\ell$, and $k(\ell) = k(\ell-1) + 1 = k'(\ell)$ with $p_{k(\ell)} = p'_{k(\ell)}$. To finish this case, we now argue that ℓ is the first edge added to set B'_{k_ℓ}. First, note that it follows by the definition of must-perms that must-perms is empty only if B_{k_ℓ} is empty at the beginning of iteration ℓ. Therefore, ℓ must be the first edge added to B'_{k_ℓ} because otherwise the induction hypothesis $B_{k_\ell} \subset B'_{k_\ell} \subset B_{k_\ell} \cup S_{\ell-1}$ will be violated as B_{k_ℓ} is empty and $S_{\ell-1}$ does not contain edge ℓ'.

Now consider the case when we don't introduce a new permutation into the support i.e., either $|\mathcal{K}'_\ell| \geq 2$, or $|\mathcal{K}'_\ell = 1|$ and must-perms $\neq \emptyset$. In this case, we argue that at the beginning of iteration ℓ the sets B_k for all $k \in \mathcal{K}'_\ell$ are non-empty. Assuming this is true, it follows from (5.44) and (5.45) that

$$\text{must-perms} \subset \mathcal{K}'_\ell \subset \text{feas-perms}\,.$$

Therefore, the algorithm adds edge ℓ to must-perms and then to B_k for all $k \in \mathcal{K}'_\ell \setminus$ must-perms because of the linear independence condition and the fact that $\varepsilon = 0$. After edge ℓ is added to B_k for $k \in \mathcal{K}'_\ell$, its weight becomes zero and the iteration ends. Therefore, edge ℓ is added to B_k for k in $\mathcal{K}'_\ell$ and only in $\mathcal{K}'_\ell$ and $k(\ell) = k(\ell - 1)$. The only thing we are left with is to argue that B_k is non-empty at the beginning of iteration ℓ for all $k \in \mathcal{K}'_\ell$. We consider two cases. In the case when $|\mathcal{K}'_\ell| \geq 2$, the unique witness edges of all permutations in $\mathcal{K}'_\ell$ should have been encountered before edge ℓ i.e., for each $k \in \mathcal{K}'_\ell$, there exists a unique witness edge $\ell' < \ell$ such that $\ell' \in B'_k$; if B_k were empty, then the induction hypothesis that $B_k \subset B'_k \subset B_k \cup S_{\ell-1}$ would be violated since $S_{\ell-1}$ does not contain ℓ'. Therefore, in the case when $|\mathcal{K}'_\ell| \geq 2$, we must have B_k must be non-empty at the beginning of iteration ℓ for all $k \in \mathcal{K}'_\ell$. Now consider the case when $|\mathcal{K}'_\ell| = 1$ and must-perms $\neq \emptyset$. In this case, since $\mathcal{K}'_\ell$ contains only one element and must-perms $\subset \mathcal{K}'_\ell \cap \{k\colon B_k \neq \emptyset\}$, it must be the case that B_k must be non-empty for $k \in \mathcal{K}'_\ell$. This finishes the proof of the induction step for this case.

The result now follows by induction.

5.11 An empirical study

In this section, we describe an empirical study we conducted to demonstrate that sparse models are effective in capturing the underlying structure of the problem. For the purpose of the study, we used the well-known APA (American Psychological Association) dataset that was first used by Diaconis [1989] in order to demonstrate the

underlying structure one can unearth by studying the appropriate lower-dimensional 'projections' of choice models, which include first and second order marginals.

Specifically, the dataset comprises the ballots collected for electing the president of APA. Each member expresses her/his preferences by rank ordering the candidates contesting the election. In the year under consideration, there were five candidates contesting the election and a total of 5,738 votes that were complete rankings. This information yields a distribution mapping each permutation to the fraction voters who vote for it. Given all the votes, the winning candidate is determined using the *Hare* system (see Fishburn and Brams [1983] for details about the Hare system).

A common issue in such election systems is that it is a difficult cognitive task for voters to rank order all the candidates even if the number of candidates is only five. This, for example, is evidenced by the fact out of more than 15,000 ballots cast in the APA election, only 5,738 of them are complete. The problem only worsens as the number of candidates to rank increases. One way to overcome this issue is to design an election system that collects only partial information from members. The partial information still retains some of the structure of the underlying distribution, and the loss of information is the price one pays for the simplicity of the election process. For example, one can gather first-order partial information i.e., the fraction of people who rank candidate i to position r. As discussed by Diaconis [1989], the first-order marginals retain useful underlying structure like: (1) candidate 3 has a lot of "love" (28% of the first-position vote) and "hate" (23% of the last-position vote) vote; (2) candidate 1 is strong in second position (26% of the vote) and low hate vote (15% of last-position vote); (3) voters seem indifferent about candidate 5 (see Table 5.1).

If we collect only first-order information, then we are faced with the issue of using this partial information to determine the winner. Having complete distribution affords us the flexibility of using any of the several rank aggregation systems out there. In order to retain this flexibility, our approach is to fit a sparse distribution to the partial information and use the distribution with the favorite rank aggregation system to determine the "winning" ranking. Such an approach would be of value if the sparse distribution can capture the underlying structural information of the problem

	Rank				
Candidate	**1**	**2**	**3**	**4**	**5**
1	18	26	23	17	15
2	14	19	25	24	18
3	28	17	14	18	23
4	20	17	19	20	23
5	20	21	20	19	20

Table 5.1: The first-order marginal matrix where the entry corresponding to candidate i and rank j is the percentage of voters who rank candidate i to position j

at hand. Therefore, with the aim to understand the type of structure sparse models can capture, we used the first-order marginal matrix M that is obtained from 5,738 complete rankings and given in Table 5.1 with the heuristic described in Section 5.10 to determine the following sparse model $\hat{\lambda}$:

$$\begin{array}{ll} 24153 & 0.211990 \\ 32541 & 0.202406 \\ 15432 & 0.197331 \\ 43215 & 0.180417 \\ 51324 & 0.145649 \\ 23154 & 0.062206 \end{array}$$

In above description of the model $\hat{\lambda}$, we adopted the notation used in Diaconis [1989] to represent each rank-list by a five-digit number in which each candidate is shown in the position it is ranked i.e., 24153 represents the rank-list in which candidate 2 is ranked at position 1, candidate 4 is ranked at position 2, candidate 1 is ranked at position 3, candidate 5 is ranked at position 4, and candidate 3 is ranked at position 5. Note that the support size of $\hat{\lambda}$ is only 6, which is a significant reduction from the full support size of $5! = 120$ of the underlying distribution. The average relative error in the approximation of M by the first-order marginals $M(\hat{\lambda})$ is less than 0.075,

where the average relative error is defined as

$$\sum_{1 \leq i,j \leq 5} \frac{|M(\hat{\lambda})_{ij} - M_{ij}|}{M_{ij}}.$$

Note that this is a much stronger (since its relative) guarantee than just an additive error guarantee. The main conclusion we can draw from the small relative error we obtained is that the heuristic proposed in Section 5.10 can successfully to find sparse models that are a good fit to the data in interesting practical cases. Now that we managed to obtain a huge reduction in sparsity at the cost of an average relative error of 0.075 in approximating first-order marginals, we next try to understand the type of structure the sparse model is able to capture from just the first-order marginals.

In order understand that, we compared the 'stair-case' curves of the cumulative distribution functions (CDF) of the actual distribution λ and $\hat{\lambda}$ in Figure 5.11. Along the x-axis in the plot, the permutations are ordered such that nearby permutations are "close" to each other in the sense that only a few transpositions (pairwise swaps) are needed to go from one permutation to another. The figure visually represents the how well the sparse model approximates the true CDF.

We next considered structural similarities between λ and $\hat{\lambda}$ beyond the similarities between their CDFs. Particularly, as evidenced by the applications discussed in Chapter 3, in several practical applications, what one is eventually interested in is a functional of the distribution – rather than the distribution itself. In the case of APA dataset, a functional of interest is the one that declares the outcome of the election in the form of a "socially preferred" ranking of the 5 candidates. As mentioned above, the *Hare system* was used to determine the winning ranking. When applied to both the distributions, the winning ranking obtained from the original distribution λ was 23145 and from the sparse distribution $\hat{\lambda}$ was 13245. Given these outcomes, we make the following observations. As is not surprising, the rankings obtained are clearly different, but the sparse model manages to capture the ranking of the candidates 4 and 5. The sparse model declares candidate 1 as the winner, whereas the original ranking declared candidate 2 the winner. We argue that declaring candidate

1 as the winner is not totally unreasonable, and in fact arguably the better choice. Specifically,it was observed in Diaconis [1989] that the data shows that candidate 1 has strongest 'second' position vote and the least "hate" vote or last position vote (see Table 5.1. Moreover, as observed in Diaconis [1989], the APA has two groups of voters with different political inclinations: *academics* and *clinicians*. The authors conclude that candidates $\{2,3\}$ and $\{4,5\}$ belong to different groups with candidate 1 somewhat neutral. From these two observations, one could argue that candidate 1 is a better choice than candidate 2. Furthermore, it is from the winning ranking of $\hat{\lambda}$ the grouping of candidates 2,3 and 4,5 are evident, leading us to conclude that the sparse model $\hat{\lambda}$ is able to capture the partitioning of the APA into two groups.

Finally, it is somewhat tantalizing to relate the support size 6 of the sparse approximation with the structure observed in the dataset by Diaconis [1989]: there are effectively three types (groups) of candidates $\{1\}$, $\{2,3\}$ and $\{4,5\}$ in view of the voters. Therefore, all votes are effectively exhibiting their ordering/preferences over these three groups and therefore effectively the votes are representing $3! = 6$ distinct preferences. And 6 is the suppose of the sparse approximation.

5.12 Chapter summary and discussion

In this chapter, we considered the problem of learning choice models from marginal information about the model. Such problems naturally arise in several important application settings like customer segmentation, rank aggregation, and compression of distributions over permutations. Since the information available is limited, learning the model is equivalent to choosing a criterion for selecting one of the potentially many models that are consistent with the data. The criterion we choose is that of *parsimony* or *sparsity*. Specifically, we propose to select the sparsest model that is consistent with the data.

In the context of learning choice models, this chapter addressed broadly two main questions: (1) how to learn the sparsest model efficiently? and (2) how "good" are sparse models in practice. In order to answer the first question, we considered two

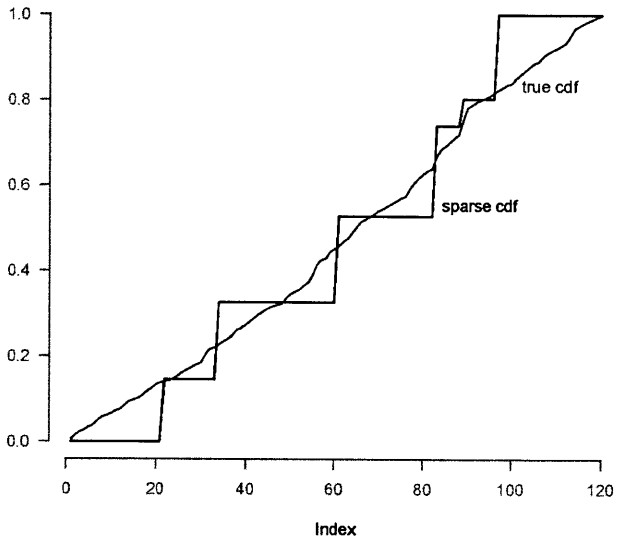


Figure 5-2: Comparison of the CDFs of the true distribution and the sparse approximation we obtain for the APA dataset. The x-axis represents the $5! = 120$ different permutations ordered so that nearby permutations are close to each other with respect to the pairwise transposition distance.

settings: the noiseless and the noisy setting.

For the noiseless setting, given marginal information $y = A\lambda$, we identified a family of models called the 'signature' family $\mathcal{F}$, whose definition depends on the type of marginal information available. We showed that $\mathcal{F}$ possesses the following two appealing properties: given $y = A\lambda$ such that λ belongs to $\mathcal{F}$, then (a) λ is also the sparsest model consistent with y, and (b) λ can be recovered from y in a efficient manner using the sparsest fit algorithm. In addition, we showed that the signature family is not merely a theoretical construct by establishing that a randomly chosen choice model that is sparse enough belongs to the signature family with a high probability.

For the noisy setting on the other hand, we showed that for the first-order information, if we are willing to allow for an ℓ_2 error of ε in approximating a given doubly stochastic matrix, then there exists a choice model with sparsity $O(N/\varepsilon)$ that is an ε-fit to the data. Note that this is a significant reduction from $\Theta(N^2)$ that is guaranteed by the Birhkoff-von Neumann and Caratheodory's theorems. Given that we can expect to find sparse models, we considered the issue of efficient recovery of sparse models. We showed that as long as there is a choice model λ of sparsity K in the signature family that is an ε-fit to the data, we can find a choice model of sparsity $O(K \log N)$ that is a 2ε-fit to the data in time $O(\exp(\Theta(K \log N)))$ as opposed to the brute-force time complexity of $O(\exp(\Theta(KN \log N)))$. Finally, we justified the existence of a model in signature family that is a good fit to the data by showing that the signature family is appropriately "dense" for a large class of models.

Given the results above, we note here their connection to the recently popular area of compressive sensing. As mentioned above, the main goal in compressive sensing is to learn a sparse vector in p dimensions from m linear "measurements", where $m \ll p$. The main result of this work is that a the restricted null space condition on the measurement matrix is necessary and sufficient for linear (convex) programs to be able to find the sparsest solution. As argued above, the restricted null space condition is not met and hence linear (convex) programs fail in our setting. Thus, the signature conditions we derive are another set of sufficient conditions that allow

efficient recovery of sparse models.

Finally, we discuss the applicability of our results in the noisy setting to types of marginal information other than the first-order marginal information. The proof for the result (Theorem 14) on "how sparse the sparse models are" does not exploit the structure of first-order marginals and hence can be extended in a reasonably straightforward manner to other types of marginal information. Similarly, we strongly believe the result (Theorem 16) that we can find good approximations in the signature family for a large subclass of choice models extends to other types of ρ-order marginal information. In particular, for any ρ, one should be able to combine the proof ideas of Theorem 16 and the theorem corresponding to ρ in Section 5.4 for the noiseless setting to obtain a similar result. However, the result about computational efficiency (Theorem 15) strongly relies on the efficient description of the first-order marginal polytope by means of Birkhoff-Von Neumann result and will not readily extend to other types of marginal information. Finally, the heuristic we presented in Section 5.10 clearly does extend to case of general ρ-order marginal information, but the computational complexity depends on the computational complexity of the oracle. Indeed, an important future direction of research is to overcome this issue, possibly developing better computational approximations.

Chapter 6

Conclusions and future work

The thesis studied a nonparametric approach to choice modeling. The main contribution of this thesis is to operationalize such a nonparametric approach to choice modeling. We distinguished two broad setups based on the ultimate goals: (a) to learn the choice model, and (b) to use the choice model to make predictions or decisions. The latter setup is clearly a special case of the former setup. However, since the second setup is simpler, we can obtain stronger results.

We first considered the problem of using choice models to make decisions. We focused on the type of applications that are important to the areas of OM and RM. The central decision problem in this context is the *static assortment optimization* problem in which the goal is to find the optimal assortment: the assortment of products with the maximum revenue subject to a constraint on the size of the assortment. Solving the decision problem requires two components: (a) a subroutine that uses historical sales transaction data to predict the expected revenues from offering each assortment of products, and (b) an optimization algorithm that uses the subroutine to find the optimal assortment. Clearly, both components are important in their own right. Specifically, solutions to these two problems lead to a solution to the *single-leg, multiple fare-class yield management problem*, which deals with the allocation of aircraft seat capacity to multiple fare classes when customers exhibit choice behavior.

In the context of making decisions, we first considered the problem of predicting revenues from offering a particular assortment of products. Most of the existing

approaches are parametric in nature. However, parametric approaches are limited because the complexity of the choice model used for predictions does not scale with the "amount" of data, making the the model prone to over-fitting and under-fitting issues. Thus, we considered a nonparametric approach, which overcomes this issue. Specifically, given historical transaction data, we identified a family of distributions that are consistent with the data. Now, given an assortment of products, we predicted revenues by computing the worst-case revenue of all consistent choice models. We addressed the computational issues, and demonstrated the accuracy of our revenue predictions through empirical studies. In particular, in our case-study with transaction data from a major US automaker, our approach succeeded in obtaining around 20% improvement in the accuracy of revenue predictions when compared to popular existing approaches. We also provided theoretical guarantees for the relative errors. The theoretical guarantees confirm the intuition that the error depends on the "complexity" of the underlying choice structure and the "amount" of data that is available.

We next considered the problem of assortment optimization. Assuming that we have access to a revenue prediction subroutine, we designed an algorithm to find an approximation of the optimal assortment with as few calls to the revenue subroutine as possible. We designed a general algorithm for the optimization of set-functions to solve the static assortment optimization algorithms. Most existing algorithms (both exact and approximate) heavily exploit the structure of the assumed choice model; consequently, the existing algorithms – even without any guarantees – cannot be used with other choice models like the probit model or the mixture of MNL models with a continuous mixture. Given these issues, we designed an algorithm that is (a) not tailored to specific parametric structures and (b) requires only a subroutine that gives revenue estimates for assortments. Our algorithm is a sophisticated form of greedy algorithm, where the solution is constructed from a smaller assortment through greedy additions and exchanges. The algorithm is proved to find the optimal assortment exactly when the underlying choice model is the MNL model. We also showed that the algorithm is robust to errors in the revenue estimates provided by

the revenue subroutine, as long as the underlying choice model is the MNL model.

Finally, we considered the problem of learning choice models from marginal information about the model. Such problems naturally arise in several important application settings like customer segmentation, rank aggregation, and compression of distributions over permutations. Since the information available is limited, learning the model is equivalent to choosing a criterion for selecting one of the potentially many models that are consistent with the data. The criterion we choose is that of *parsimony* or *sparsity*. Specifically, we propose to select the sparsest model that is consistent with the data.

In the context of learning choice models, we addressed broadly two main questions: (1) how to learn the sparsest model efficiently? and (2) how "good" are sparse models in practice. In order to answer the first question, we considered two settings: the noiseless and the noisy setting.

For the noiseless setting, given marginal information $y = A\lambda$, we identified a family of models called the 'signature' family $\mathcal{F}$, whose definition depends on the type of marginal information available. We showed that $\mathcal{F}$ possesses the following two appealing properties: given $y = A\lambda$ such that λ belongs to $\mathcal{F}$, then (a) λ is also the sparsest model consistent with y, and (b) λ can be recovered from y in a efficient manner using the sparsest fit algorithm. In addition, we showed that the signature family is not merely a theoretical construct by establishing that a randomly chosen choice model that is sparse enough belongs to the signature family with a high probability.

For the noisy setting on the other hand, we showed that for the first-order information, if we are willing to allow for an ℓ_2 error of ε in approximating a given doubly stochastic matrix, then there exists a choice model with sparsity $O(N/\varepsilon)$ that is an ε-fit to the data. Note that this is a significant reduction from $\Theta(N^2)$ that is guaranteed by the Birhkoff-von Neumann and Caratheodory's theorems. Given that we can expect to find sparse models, we considered the issue of efficient recovery of sparse models. We showed that as long as there is a choice model λ of sparsity K in the signature family that is an ε-fit to the data, we can find a choice model of sparsity

$O(K \log N)$ that is a 2ε-fit to the data in time $O(\exp(\Theta(K \log N)))$ as opposed to the brute-force time complexity of $O(\exp(\Theta(KN \log N)))$. Finally, we justified the existence of a model in signature family that is a good fit to the data by showing that the signature family is appropriately "dense" for a large class of models.

6.1 Future work

Although we proposed a nonparametric approach to choice modeling that can be successfully applied in practice, our work is no panacea for all choice modeling problems. In particular, one merit of a structural/ parametric modeling approach to modeling choice is the ability to extrapolate. That is to say, a nonparametric approach such as ours can start making useful predictions about the interactions of a particular product with other products only once *some* data related to that product is observed. With a structural model, one can hope to say useful things about products never seen before. The decision of whether a structural modeling approach is relevant to the problem at hand or whether the approach we offer is a viable alternative thus merits a careful consideration of the context. Of course, as we have discussed earlier, resorting to a parametric approach will typically require expert input on underlying product features that 'matter', and is thus difficult to automate on a large scale. Thus, reconciling the two approaches would be a natural next step.

More precisely, having learned a choice model that consists of a distribution over a small number of rank lists, there are a number of qualitative insights one might hope to draw. For instance, using fairly standard statistical machinery, one might hope to ask for the product features that most influence choice from among thousands of potential features by understanding which of these features best rationalize the rank lists learned. In a different direction, one may use the distribution learned as a 'prior', and given further interactions with a given customer infer a distribution specialized to that customer via Bayes rule. This is effectively a means to accomplishing 'collaborative filtering'.

There are also interesting directions to pursue from a theoretical perspective:

First, extending our understanding of the limits of identification. In particular, it would be useful to characterize the limits of recoverability for additional families of observable data beyond those discussed in this thesis. Second, we have exhibited an efficient way to find a sparse approximation to the given first-order data. Can we extend the algorithm to other types of partial information like the one captured by transaction data? Finally, the robust approach of predicting revenues presents us with a family of difficult optimization problems for which the present work has presented a generic optimization scheme that is in the spirit of cutting plane approaches. An alternative to this is the development of strong relaxations that yield uniform approximation guarantees (in the spirit of the approximation algorithms literature).

Appendix A

Appendix: Decision Problems

A.1 Proof of Theorem 4

The result of Theorem 4 follows immediately from the following lemma, which we prove below.

Lemma 4. *Let d denote the column rank of matrix A and $\mathcal{Y}$ denote the convex hull of the columns of A. Then, it must be that y belongs to a $d-1$ dimensional subspace, $\|\lambda^{\min}(y)\|_0 \leq d+1$, and*

$$\mathrm{vol}_{d-1}(\mathcal{Y}^{\mathrm{sparse}}) = \mathrm{vol}_{d-1}(\mathcal{Y}),$$

where $\mathcal{Y}^{\mathrm{sparse}} \subset \mathcal{Y}$ denotes the set of all data vectors such that

$$\|\lambda^{\min}(y)\|_0 \leq d+1 \text{ and } \|\lambda^{\mathrm{sparse}}(y)\|_0 \geq d$$

and $\mathrm{vol}_{d-1}(S)$ denotes the $d-1$ dimensional volume of a set S of points.

Proof of Lemma 4 We prove this lemma in two parts: (1) $\mathcal{Y}$ belongs to a $d-1$ dimensional subspace and $\|\lambda^{\min}(y)\|_0 \leq d+1$ for all $y \in \mathcal{Y}$, and (2) $\mathrm{vol}_{d-1}(\mathcal{Y} \backslash \mathcal{Y}^{\mathrm{sparse}}) = 0$.

To prove the first part, note that any data vector $y \in \mathcal{Y}$ belongs to $d-1$ dimensional subspace because A has a d dimensional range space and y belongs to the intersection of the range space of A and the hyperplane $\sum_\sigma \lambda_\sigma = 1$. Let $\tilde{A}$ denote the

augmented matrix, which is obtained by augmenting the last row of matrix A with a row of all 1s. Similarly, let $\tilde{y}$ denote the vector obtained by augmenting vector y with 1. The equality constraints of (3.2) can now be written as $\tilde{y} = \tilde{A}\lambda$, $\lambda \geq 0$. Since A has rank d, the rank of $\tilde{A}$ will be at most $d+1$. Therefore, for any data vector $y \in \mathcal{Y}$, an optimal BFS solution to (3.2) must be such that

$$\|\lambda^{\min}(y)\|_0 \leq d+1, \quad \forall\, y \in \mathcal{Y}. \tag{A.1}$$

Coming to the second part of the proof, for any $r \leq d-1$, let $\mathcal{Y}_r$ denote the set of all data vectors that can be written as a convex combination of at most r columns of matrix A. Let L denote the number of columns of A of size at most r, and let $S_1, S_2, \ldots, S_L$ denote the corresponding subsets of columns of A of size at most r. Then, it is easy to see that $\mathcal{Y}_r$ can be written as the union of disjoint subsets $\mathcal{Y}_r = \cup_{i=1}^{L} \mathcal{Y}_{ri}$, where for each i, $\mathcal{Y}_{ri}$ denotes the set of data vectors that can be written as the convex combination of the columns in subset S_i. For each i, since $\mathcal{Y}_{ri}$ is a polytope residing in $r-1 \leq d-2$ dimensional space, it must follow that $\text{vol}_{d-1}(\mathcal{Y}_{ri}) = 0$. Since L is finite, it follows that $\text{vol}_{d-1}(\mathcal{Y}_r) = 0$. Therefore, we can conclude that

$$\text{vol}_{d-1}\left(y \in \mathcal{Y} \colon \|\lambda^{\text{sparse}}(y)\|_0 \leq d-1\right) = 0 \tag{A.2}$$

The result of the lemma now follows from (A.1) and (A.2).

A.2 The exact approach to solving the robust problem

Here we provide further details on the second approach described for the solution to the (dual of) the robust problem (3.3). In particular, we first consider the case of ranking data, where an efficient representation of the constraints in the dual may be produced. We then illustrate a method that produces a sequence of 'outer-approximations' to (3.5) for general types of data, and thereby allows us to produce a

sequence of improving lower bounding approximations to our robust revenue estimation problem, (3.2). This provides a general procedure to address the task of solving (3.3), or equivalently, (3.2).

A.2.1 A canonical representation for ranking data

Recall the definition of *ranking data* from Section 3.2: This data yields the fraction of customers that rank a given product i as their rth choice. Thus, the partial information vector y is indexed by i,r with $0 \leq i,r \leq N$. For each i,r, y_{ri} denotes the probability that product i is ranked at position r. The matrix A is thus in $\{0,1\}^{N^2 \times N!}$ and for a column of A corresponding to the permutation σ, $A(\sigma)$, we will thus have $A(\sigma)_{ri} = 1$ iff $\sigma(i) = r$. We will now construct an efficient representation of the type (3.5) for this type of data.

Consider partitioning $\mathcal{S}_j(\mathcal{M})$ into $D_j = N$ sets wherein the dth set is given by

$$\mathcal{S}_{jd}(\mathcal{M}) = \{\sigma \in \mathcal{S}_j(\mathcal{M}) \colon \sigma(j) = d\}.$$

and define, as usual, $\mathcal{A}_{jd}(\mathcal{M}) = \{A(\sigma) : \sigma \in \mathcal{S}_{jd}(\mathcal{M})\}$. Thus, $\mathcal{A}_{jd}(\mathcal{M})$ is the set of columns of A whose corresponding permutations rank the jth product as the dth most preferred choice.

It is easily seen that the set $\mathcal{A}_{jd}(\mathcal{M})$ is equal to the set of all vectors x^{jd} in $\{0,1\}^{N^2}$ satisfying:

$$\begin{aligned}
&\sum_{i=0}^{N-1} x_{ri}^{jd} = 1 \quad \text{for } 0 \leq r \leq N-1 \\
&\sum_{r=0}^{N-1} x_{ri}^{jd} = 1 \quad \text{for } 0 \leq i \leq N-1 \\
&x_{ri}^{jd} \in \{0,1\} \quad \text{for } 0 \leq i,r \leq N-1. \\
&x_{dj}^{jd} = 1 \\
&x_{d'i}^{jd} = 0 \qquad \text{for all } i \in \mathcal{M}, i \neq j \text{and } 0 \leq d' < d.
\end{aligned} \tag{A.3}$$

The first three constraints in (A.3) enforce the fact that x^{jd} represents a valid permutation. The penultimate constraint requires that the permutation encoded by x^{jd}, say σ^{jd}, satisfies $\sigma^{jd}(j) = d$. The last constraint simply ensures that $\sigma^{jd} \in \mathcal{S}_j(\mathcal{M})$.

Our goal is, of course, to find a description for $\bar{\mathcal{A}}_{jd}(\mathcal{M})$ of the type (3.5). Now consider replacing the third (integrality) constraint in (A.3)

$$x_{ri}^{jd} \in \{0,1\} \text{ for } 0 \leq i,r \leq N-1$$

with simply the non-negativity constraint

$$x_{ri}^{jd} \geq 0 \text{ for } 0 \leq i,r \leq N-1$$

We claim that the resulting polytope is precisely the convex hull of $\mathcal{A}_{jd}(\mathcal{M}), \bar{\mathcal{A}}_{jd}(\mathcal{M})$. To see this, we note that all feasible points for the resulting polytope satisfy the first, second, fourth and fifth constraint of (A.3). Further, the polytope is integral, being the projection of a matching polytope with some variables forced to be integers (Birkhoff [1946], von Neumann [1953]), so that any feasible solution must also satisfy the third constraint of (A.3). We consequently have an *efficient* canonical representation of the type (3.5), which via (3.7) yields, in turn, an efficient solution to our robust revenue estimation problem (3.2) for ranking data, which we now describe for completeness.

Let us define for convenience the set $\mathcal{V}(\mathcal{M}) = \{(j,d) : j \in \mathcal{M}, 0 \leq d \leq N-1\}$, and for each pair (j,d), the sets $\mathcal{B}(j,d,\mathcal{M}) = \{(i,d') : i \in \mathcal{M}, i \neq j, 0 \leq d' < d\}$. Then, specializing (3.7) to the canonical representation just proposed, we have that the following simple program in the variables α, ν and $\gamma^{jd} \in \mathbb{R}^{2N}$ is, in fact, equivalent to (3.2) for ranking data:

$$\begin{array}{lll} \underset{\alpha,\nu}{\text{maximize}} & \alpha^\top y + \nu & \\ \text{subject to} & \gamma_i^{jd} + \gamma_{N+r}^{jd} \geq \alpha_{ri} & \text{for all } (j,d) \in \mathcal{V}(\mathcal{M}), (i,r) \notin \mathcal{B}(j,d,\mathcal{M}) \\ & \sum_{i \neq j} \gamma_i^{jd} + \sum_{r \neq d} \gamma_{N+r}^{jd} + \nu \leq p_j - \alpha_{dj} & \text{for all } (j,d) \in \mathcal{V}(\mathcal{M}) \end{array} \tag{A.4}$$

A.2.2 Computing a canonical representation: the general case

While it is typically quite easy to 'write down' a description of the sets $\mathcal{A}_{jd}(\mathcal{M})$ as all integer solutions to some set of linear inequalities (as we did for the case of ranking data), relaxing this integrality requirement will typically *not* yield the convex hull of $\mathcal{A}_{jd}(\mathcal{M})$. In this section we describe a procedure that starting with the former (easy to obtain) description, solves a sequence of linear programs that yield improving solutions. More formally, we assume a description of the sets $\mathcal{A}_{jd}(\mathcal{M})$ of the type

$$\mathcal{I}_{jd}(\mathcal{M}) = \{x^{jd} : A_1^{jd}\, x^{jd} \geq b_1^{jd}, \quad A_2^{jd}\, x^{jd} = b_2^{jd}, \quad A_3^{jd}\, x^{jd} \leq b_3^{jd}, \quad x^{jd} \in \{0,1\}^m\} \tag{A.5}$$

This is similar to (3.5), with the important exception that we now allow integrality constraints. Given a set $\mathcal{I}_{jd}(\mathcal{M})$ we let $\bar{\mathcal{I}}^0_{jd}(\mathcal{M})$ denote the polytope obtained by relaxing the requirement $x^{jd} \in \{0,1\}^m$ to simply $x^{jd} \geq 0$. In the case of ranking data, $\bar{\mathcal{I}}^o_{jd}(\mathcal{M}) = \text{conv}(\mathcal{I}_{jd}(\mathcal{M})) = \bar{\mathcal{A}}_{jd}(\mathcal{M})$ and we were done; we begin with an example where this is not the case.

Example 2. *Recall the definition of* comparison data *from Section 3.2. In particular, this data yields the fraction of customers that prefer a given product i to a product j. The partial information vector y is thus indexed by i, j with $0 \leq i, j \leq N; i \neq j$ and for each i, j, $y_{i,j}$ denotes the probability that product i is preferred to product j. The matrix A is thus in $\{0,1\}^{N(N-1)\times N!}$. A column of A, $A(\sigma)$, will thus have $A(\sigma)_{ij} = 1$ if and only if $\sigma(i) < \sigma(j)$.*

Consider $\mathcal{S}_j(\mathcal{M})$, the set of all permutations that would result in a purchase of j assuming $\mathcal{M}$ is the set of offered products. It is not difficult to see that the corresponding set of columns $\mathcal{A}_j(\mathcal{M})$ is equal to the set of vectors in $\{0,1\}^{(N-1)N}$ satisfying the following constraints:

$$
\begin{array}{ll}
x^j_{il} \geq x^j_{ik} + x^j_{kl} - 1 & \text{for all } i,k,l \in \mathcal{N}, i \neq k \neq l \\
x^j_{ik} + x^j_{ki} = 1 & \text{for all } i,k \in \mathcal{N}, i \neq k \\
x^j_{ji} = 1 & \text{for all } i \in \mathcal{M}, i \neq j \\
x^j_{ik} \in \{0,1\} & \text{for all } i,k \in \mathcal{N}, i \neq k
\end{array}
\tag{A.6}
$$

Briefly, the second constraint follows since for any $i,k, i \neq k$, either $\sigma(i) > \sigma(k)$ or else $\sigma(i) < \sigma(k)$. The first constraint enforces transitivity: $\sigma(i) < \sigma(k)$ and $\sigma(k) < \sigma(l)$ together imply $\sigma(i) < \sigma(l)$. The third constraint enforces that all $\sigma \in S_j(\mathcal{M})$ must satisfy $\sigma(j) < \sigma(i)$ for all $i \in \mathcal{M}$. Thus, (A.6) *is a description of the type* (A.5) *with $D_j = 1$ for all j. Now consider the polytope $\bar{\mathcal{I}}^o_j(\mathcal{M})$ obtained by relaxing the fourth (integrality) constraint to simply $x^j_{ik} \geq 0$. Of course, we must have $\bar{\mathcal{I}}^o_j(\mathcal{M}) \supseteq \text{conv}(\mathcal{I}_j(\mathcal{M})) = \text{conv}(\mathcal{A}_j(\mathcal{M}))$. Unlike the case of ranking data, however, $\bar{\mathcal{I}}^o_j(\mathcal{M})$ can in fact be shown to be* non-*integral* [1]*, so that $\bar{\mathcal{I}}^o_j(\mathcal{M}) \neq \text{conv}(\mathcal{A}_j(\mathcal{M}))$ in general.*

We next present a procedure that starting with a description of the form in (A.5), solves a sequence of linear programs each of which yield improving solutions to (3.2) along with bounds on the quality of the approximation:

1. Solve (3.7) using $\bar{\mathcal{I}}^o_{jd}(\mathcal{M})$ in place of $\text{conv}(\mathcal{I}_{jd}(\mathcal{M})) = \bar{\mathcal{A}}_{jd}(\mathcal{M})$. This yields a lower bound on (3.2) since $\bar{\mathcal{I}}^o_{jd}(\mathcal{M}) \supset \bar{\mathcal{A}}_{jd}(\mathcal{M})$. Call the corresponding solution $\alpha_{(1)}, \nu_{(1)}$.

2. Solve the optimization problem $\max \alpha^\top_{(1)} x^{jd}$ subject to $x^{jd} \in \bar{\mathcal{I}}^o_{jd}(\mathcal{M})$ for each pair (j,d). If the optimal solution $\hat{x}^{jd}$ is integral for each (j,d), then stop; the solution computed in the first step is in fact optimal.

3. Otherwise, let $\hat{x}^{jd}$ possess a non-integral component for some (j,d); say $\hat{x}^{jd}_c \in (0,1)$. Partition $\mathcal{A}_{jd}(\mathcal{M})$ on this variable - i.e. define

$$\mathcal{A}_{jd_0}(\mathcal{M}) = \{A(\sigma) : A(\sigma) \in \mathcal{A}_{jd}(\mathcal{M}), A(\sigma)_c = 0\}$$

[1] for $N \geq 5$; the polytope can be shown to be integral for $N \leq 4$

and

$$\mathcal{A}_{jd_1}(\mathcal{M}) = \{A(\sigma) : A(\sigma) \in \mathcal{A}_{jd}(\mathcal{M}), A(\sigma)_c = 1\},$$

and let $\mathcal{I}_{jd_0}(\mathcal{M})$ and $\mathcal{I}_{jd_1}(\mathcal{M})$ represent the corresponding sets of linear inequalities with integer constraints (i.e. the projections of $\mathcal{I}_{jd}(\mathcal{M})$ obtained by restricting x_c^{jd} to be 0 and 1 respectively). Of, course, these sets remain of the form in (A.5). Replace $\mathcal{I}_{jd}(\mathcal{M})$ with $\mathcal{I}_{jd_0}(\mathcal{M})$ and $\mathcal{I}_{jd_1}(\mathcal{M})$ and go to step 1.

The above procedure is akin to a cutting plane method and is clearly finite, but the size of the LP we solve increases (by up to a factor of 2) at each iteration. Nonetheless, each iteration produces a lower bound to (3.2) whose quality is easily measured (for instance, by solving the maximization version of (3.2) using the same procedure, or by sampling constraints in the program (3.3) and solving the resulting program in order to produce an upper bound on (3.2)). Moreover, the quality of our solution improves with each iteration. In our computational experiments with a related type of data, it sufficed to stop after a single iteration of the above procedure.

A.2.3 Explicit LP solved for censored comparison data in Section 3.4

The LP we want to solve is

$$\begin{array}{ll} \underset{\lambda}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\ \text{subject to} & A\lambda = y, \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0. \end{array} \tag{A.7}$$

For the 'censored' comparison data, the partial information vector is indexed by i, j with $0 \leq i, j \leq N-1$, $i \neq j$. For each i, j such that $i \neq 0$, y_{ij} denotes the fraction of customers that prefer product i to both products j and 0; in other words, y_{ij} denotes the fraction of customers that purchase product i when then offer set is $\{i, j, 0\}$. Further, for each $j \neq 0$, y_{0j} denotes the fraction of customers who prefer

the 'no-purchase' option to product j; in fact, y_{0j} is the fraction of customers who don't purchase anything when the set $\{j, 0\}$ is on offer. The matrix A is then in $\{0,1\}^{N(N-1)}$, with the column of A corresponding to permutation σ, $A(\sigma)$, having $A(\sigma)_{ij} = 1$ if $\sigma(i) < \sigma(j)$ and $\sigma(i) < \sigma(0)$ for each $i \neq 0, j$, and $A(\sigma)_{0j} = 1$ if $\sigma(0) < \sigma(j)$ for $j \neq 0$, and $A(\sigma)_{ij} = 0$ otherwise.

For reasons that will become apparent soon, we modify the LP in (A.7) by replacing the constraint $A\lambda = y$ with $A\lambda \geq y$. It is now easy to see the following:

$$\begin{array}{lll} \underset{\lambda}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) & \\ \text{subject to} & A\lambda \geq y, & \\ & \mathbf{1}^\top \lambda = 1, & \\ & \lambda \geq 0. & \end{array} \quad \leq \quad \begin{array}{lll} \underset{\lambda}{\text{minimize}} & \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) & \\ \text{subject to} & A\lambda = y, & \\ & \mathbf{1}^\top \lambda = 1, & \\ & \lambda \geq 0. & \end{array} \tag{A.8}$$

We now take the dual of the modified LP. In order to do that, recall from section 3.3 that $\mathcal{S}_j(\mathcal{M}) = \{\sigma \in S_N \colon \sigma(j) < \sigma(i), \forall i \in \mathcal{M}, i \neq j\}$ denotes the set of all permutations that result in the purchase of the product $j \in \mathcal{M}$ when the offered assortment is $\mathcal{M}$. In addition, $\mathcal{A}_j(\mathcal{M})$ denotes the set $\{A(\sigma) \colon \sigma \in \mathcal{S}_j(\mathcal{M})\}$. Now, the dual of the modified LP is

$$\begin{array}{lll} \underset{\alpha, \nu}{\text{maximize}} & \alpha^\top y + \nu & \\ \text{subject to} & \max\limits_{z^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j, & \text{for each } j \in \mathcal{M} \\ & \alpha \geq 0. & \end{array} \tag{A.9}$$

where α and ν are dual variables corresponding respectively to the data consistency constraints $A\lambda = y$ and the requirement that λ is a probability distribution (i.e. $\mathbf{1}^\top \lambda = 1$) respectively.

Now, consider the following representation of the set $\mathcal{A}_j(\mathcal{M})$, for a fixed j.

$$\begin{aligned}
& z^j_{ik} = \min\left\{x^j_{ik}, x^j_{i0}\right\} && \text{for all } i,k \in \mathcal{N}, i \neq k, i \neq 0 \\
& z^j_{0k} = x^j_{0k} && \text{for all } k \in \mathcal{N}, k \neq 0 \\
& z^j_{ik} \in \{0,1\} && \text{for all } i,k \in \mathcal{N}, i \neq k \\
& x^j_{il} \geq x^j_{ik} + x^j_{kl} - 1 && \text{for all } \ i,k,l \in \mathcal{N}, i \neq k \neq l \\
& x^j_{ik} + x^j_{ki} = 1 && \text{for all } i,k \in \mathcal{N}, i \neq k \\
& x^j_{ji} = 1 && \text{for all } i \in \mathcal{M}, i \neq j \\
& x^j_{ik} \in \{0,1\} && \text{for all } i,k \in \mathcal{N}, i \neq k
\end{aligned} \tag{A.10}$$

The last four constraints are the same as the set of inequalities in (A.6), which correspond to the representation of the set $\mathcal{A}_j(\mathcal{M})$ for comparison data; thus, every point satisfying the set of last four constraints in (A.10) corresponds to a permutation $\sigma \in \mathcal{S}_j(\mathcal{M})$ such that $x^j_{ik} = 1$ if and only if $\sigma(i) < \sigma(k)$. We now claim that the set of points z^j that satisfy the constraints in (A.10) is equal to the set of vectors in $\mathcal{A}_j(\mathcal{M})$. To see that, note that $z^j_{ik} = 1$ if and only if the corresponding $x^j_{ik} = 1$ *and* $x^j_{i0} = 1$, for $i \neq 0$. This implies that $z^j_{ik} = 1$ if and only if i is preferred to k *and* i is preferred to 0. Similarly, $z^j_{0k} = 1$ if and only if $x^j_{0k} = 1$ i.e., 0 is preferred to k.

Let $\bar{\mathcal{I}}_j(\mathcal{M})$ denote the convex hull of the vectors in $\mathcal{A}_j(\mathcal{M})$, equivalently, of the vectors z^j satisfying the set of constraints in (A.10). Let $\bar{\mathcal{I}}^o_j(\mathcal{M})$ be the convex hull of the vectors z^j satisfying the constraints in (A.10) with the constraint $z^j_{ik} = \min\left\{x^j_{ik}, x^j_{i0}\right\}$ replaced by the constraints $z^j_{ik} \leq x^j_{ik}$ and $z^j_{ik} \leq x^j_{i0}$, and the constraint $z^j_{0k} = x^j_{0k}$ replaced by the constraint $z^j_{0k} \leq x^j_{0k}$. Finally, let $\bar{\mathcal{I}}^1_j(\mathcal{M})$ represent the polytope $\bar{\mathcal{I}}^0_j(\mathcal{M})$ with the integrality constraints relaxed to $z^j_{ik} \geq 0$ and $x^j_{ik} \geq 0$. We now have the following relationships:

$$\begin{aligned}
\left\{\alpha \geq 0, \nu : \max_{z^j \in \bar{\mathcal{I}}_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j\right\} &= \left\{\alpha \geq 0, \nu : \max_{z^j \in \bar{\mathcal{I}}^o_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j\right\} \\
&\supseteq \left\{\alpha \geq 0, \nu : \max_{z^j \in \bar{\mathcal{I}}^1_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j\right\}
\end{aligned} \tag{A.11}$$

The first equality follows because $\alpha \geq 0$ and, hence, at the optimal solution, $z_{ik}^j = 1$ if $x_{ik}^j = x_{i0}^j = 1$, and $z_{0k}^j = 1$ if $x_{0k}^j = 1$. It should be now clear that in order to establish this equality we considered the modified LP. The second relationship follows because of the relaxation of constraints. It now follows from (A.8), (A.9) and (A.11) that

$$\begin{aligned}
&\underset{\lambda}{\text{minimize}} \quad \sum_{j \in \mathcal{M}} p_j \lambda_j(\mathcal{M}) \\
&\text{subject to} \quad A\lambda = y, \\
&\qquad\qquad \mathbf{1}^\top \lambda = 1, \\
&\qquad\qquad \lambda \geq 0.
\end{aligned}
\quad \geq \quad
\begin{aligned}
&\underset{\alpha,\nu}{\text{maximize}} \quad \alpha^\top y + \nu \\
&\text{subject to} \quad \max_{z^j \in \mathcal{A}_j(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j, \text{ for each } j \in \mathcal{M} \\
&\qquad\qquad \alpha \geq 0.
\end{aligned}$$

$$\begin{aligned}
&\underset{\alpha,\nu}{\text{maximize}} \quad \alpha^\top y + \nu \\
&\text{subject to} \quad \max_{z^j \in \bar{\mathcal{I}}_j^1(\mathcal{M})} \left(\alpha^\top z^j + \nu\right) \leq p_j, \text{ for each } j \in \mathcal{M} \\
&\qquad\qquad \alpha \geq 0.
\end{aligned} \tag{A.12}$$

Using the procedure described in Section 3.3.3, we solve the last LP in (A.12) by taking the dual of the constraint in the LP. For convenience, we write out the program $\max_{z^j \in \mathcal{I}_j^1(\mathcal{M})} \left(\alpha^\top z^j + \nu\right)$ and the corresponding dual variables we use for each of the constraints.

$$\begin{aligned}
&\underset{z^j}{\text{maximize}} \quad \alpha^\top z^j + \nu \\
&\text{subject to}
\end{aligned}$$

		Dual Variables
$z_{ik}^j - x_{ik}^j \leq 0$	for all $i,k \in \mathcal{N}, i \neq k$	$\Omega 1_{ik}^j$
$z_{ik}^j - x_{i0}^j \leq 0$	for all $i,k \in \mathcal{N}, i \neq k, i \neq 0$	$\Omega 2_{ik}^j$
$x_{ik}^j + x_{kl}^j - x_{il}^j \leq 1$	for all $i,k,l \in \mathcal{N}, i \neq k \neq l$	Γ_{ikl}^j
$x_{ik}^j + x_{ki}^j = 1$	for all $i,k \in \mathcal{N}, i < k$	Δ_{ik}^j
$x_{ji}^j = 1$	for all $i \in \mathcal{M}, i \neq j$	Θ_i^j
$x_{ik}^j, z_{ik}^j \geq 0$	for all $i,k \in \mathcal{N}, i \neq k$	

(A.13)

Let P denote the set $\{(i,k)\colon i \neq k, 0 \leq i,k \leq N-1\}$, and T denote the set

$\{(i,k,l)\colon i \neq k \neq l, 0 \leq i,k,l \leq N-1\}$. Moreover, let $g(a,b,k,j)$ denote $\sum_{k\in\mathcal{N},k\neq a,b}\Gamma^j_{abk}$ $+\sum_{k\in\mathcal{N},k\neq a,b}\Gamma^j_{kab}-\sum_{k\in\mathcal{N},k\neq a,b}\Gamma^j_{akb}$. Then, the LP we solve is

$$\underset{\nu,\alpha}{\text{maximize}} \quad \nu+\textstyle\sum_{(i,k)\in P}\alpha_{ik}y_{ik}$$

subject to

$$\begin{aligned}
&\sum_{(i,k,l)\in T}\Gamma^j_{ikl}+\sum_{(i,k)\in P,i<k}\Delta^j_{ik}+\sum_{i\in\mathcal{M},i\neq j}\Theta^j_i \leq p_j-\nu && \forall\ j\in\mathcal{M}\\
&g(a,b,k,j)+\Delta^j_{ab}-\Omega 1^j_{ab}\geq 0 && \forall\ j\in\mathcal{M},a,b\in\mathcal{N},a<b;\ \text{if } a=j,b\notin\mathcal{M}\\
&g(a,b,k,j)+\Delta^j_{ba}-\Omega 1^j_{ab}\geq 0 && \forall\ j\in\mathcal{M},a,b\in\mathcal{N},a>b,b\neq 0;\ \text{if } a=j,b\notin\mathcal{M}\\
&g(a,b,k,j)+\Delta^j_{ab}+\Theta^j_b-\Omega 1^j_{ab}\geq 0 && \forall\ j\in\mathcal{M},a=j,b\in\mathcal{M},a<b\\
&g(a,b,k,j)+\Delta^j_{ba}+\Theta^j_b-\Omega 1^j_{ab}\geq 0 && \forall\ j\in\mathcal{M},a=j,b\in\mathcal{M},a>b,b\neq 0\\
&g(a,b,k,j)+\Delta^j_{ba}-\textstyle\sum_{k\in\mathcal{N},k\neq a}\Omega 2^j_{ak}\geq 0 && \forall\ j\in\mathcal{M},a\in\mathcal{N},a\neq j,b=0\\
&g(a,b,k,j)+\Delta^j_{ba}+\Theta^j_b-\textstyle\sum_{k\in\mathcal{N},k\neq a}\Omega 2^j_{ak}\geq 0 && \forall\ j\in\mathcal{M},a=j,b=0\\
&\Omega 1^j_{ab}+\Omega 2^j_{ab}\geq\alpha_{ab} && \forall\ j\in\mathcal{M},a,b\in P,a\neq 0,b\neq 0\\
&\Omega 2^j_{ab}\geq\alpha_{ab} && \forall\ j\in\mathcal{M},a\in\mathcal{N}\setminus\{0\},b=0\\
&\Omega 1^j_{ab}\geq\alpha_{ab} && \forall\ j\in\mathcal{M},a=0,b\in\mathcal{N}\setminus\{0\}\\
&\alpha,\Gamma,\Omega 1,\Omega 2\geq 0.
\end{aligned} \tag{A.14}$$

A.3 Case-study: major US automaker

In this section, we provide precise details of how each method was used to produce conversion-rate predictions. We first establish some notation. We let $\mathscr{M}^{\text{training}}$ and $\mathscr{M}^{\text{test}}$ respectively denote the set of assortments used as part of training and test data. Further, for some product $i\in\mathcal{M}$, let $C_{i,\mathcal{M}}$ denote the number of sales observed for product i when assortment $\mathcal{M}$ was on offer at some dealership; $C_{0,\mathcal{M}}$ denotes the number of customers who purchased nothing when $\mathcal{M}$ was on offer. Finally, let

Table A.1: Relevant attributes of DVDs from Amazon.com data and mean utilities of MNL model fit by Rusmevichientong et al. [2010a]

Product ID	Mean utility	Price ($)	Avg. price per disc ($)	# of helpful votes
1	-4.513	115.49	5.7747	462
2	-4.600	92.03	7.6694	20
3	-4.790	91.67	13.0955	496
4	-4.514	79.35	13.2256	8424
5	-4.311	77.94	6.4949	6924
6	-4.839	70.12	14.0242	98
7	-4.887	64.97	16.2423	1116
8	-4.757	49.95	12.4880	763
9	-4.552	48.97	6.9962	652
10	-4.594	46.12	7.6863	227
11	-4.552	45.53	6.5037	122
12	-3.589	45.45	11.3637	32541
13	-4.738	45.41	11.3523	69
14	-4.697	44.92	11.2292	1113
15	-4.706	42.94	10.7349	320

T^{training} (T^{test}) denote the set of tuples $(i, \mathcal{M})$ such that $\mathcal{M} \in \mathscr{M}^{\text{training}}$ ($\mathscr{M}^{\text{test}}$) and the count $C_{i,\mathcal{M}}$ is available; in our case study, we treated count $C_{i,\mathcal{M}}$ as unavailable if either no sales were observed for product i when $\mathcal{M}$ was on offer or $C_{i,\mathcal{M}} \leq 6$, in which case we discarded the sales as too low to be significant[2].

Robust method. Given $\mathscr{M}^{\text{training}}$ and an assortment $\mathcal{M} \in \mathscr{M}^{\text{test}}$, the robust approach predicts the conversion-rate of $\mathcal{M}$ by solving the following LP:

$$\begin{aligned} \underset{\lambda}{\text{minimize}} \quad & \sum_{j \in \mathcal{M}} \mathbb{P}_\lambda(j|\mathcal{M}) \\ \text{subject to} \quad & a_t \leq \mathbb{P}_\lambda(i|\mathcal{M}) \leq b_t, \quad \forall t = (i, \mathcal{M}) \in T^{\text{training}} \\ & \mathbf{1}^\top \lambda = 1, \\ & \lambda \geq 0, \end{aligned} \tag{A.15}$$

where recall that $\mathbb{P}_\lambda(i|\mathcal{M}) = \sum_{\sigma \in \mathcal{S}_i(\mathcal{M})} \lambda(\sigma)$ with $\mathcal{S}_i(\mathcal{M})$ denoting the set

$$\{\sigma \colon \sigma(i) < \sigma(j) \ \ \forall j \in \mathcal{M}, i \neq j\}$$

and $[a_t, b_t]$ denotes the interval to which $\mathbb{P}_\lambda(i|\mathcal{M})$ belongs. The LP in (A.15) is a

[2]In other words, we discard counts that are "too small" in order to avoid noisy sample probability estimates.

slight modification of the LP in (3.2) with the prices p_i all set of 1 and the equalities $y_t = \mathbb{P}_\lambda(\mathcal{M})$ changed to inequalities $a_t \leq \mathbb{P}_\lambda(\mathcal{M}) \leq b_t$. Setting all the prices to 1 has the effect of computing conversion-rate for the assortment, and the inequalities account for the fact that there is uncertainty in the choice probabilities because of finite sample errors. For each tuple $t = (i, \mathcal{M}) \in \mathscr{M}^{\text{training}}$, we computed the left and right end-points as $a_t = \hat{y}_t(1 - z\varepsilon_t)$ and $b_t = \hat{y}_t(1 + z\varepsilon_t)$, where

$$\hat{y}_t = \frac{C_{i\mathcal{M}}}{C_{0\mathcal{M}} + \sum_{j \in \mathcal{M}} C_{j\mathcal{M}}}, \quad \varepsilon_t = \sqrt{\frac{1 - \hat{y}_t}{C_{i\mathcal{M}}}}.$$

Here, $\hat{y}_t$ is the sample average, $\hat{y}_t\varepsilon_t$ is the standard error, and z is a constant multiplier that determines the width of the interval. Different values of z give us approximate confidence intervals for $\mathbb{P}_\lambda(i|\mathcal{M})$. If z is very small, the problem in (A.15) will be infeasible, and on the other hand if z is very large, the estimate produced by (A.15) will be very conservative. For our experiments, we set z to be 3.15, which corresponds to the smallest value of z for which (A.15) was feasible; incidentally, this value of z also corresponds to approximate 99.8% confidence interval for $\mathbb{P}_\lambda(i|\mathcal{M})$.

Note that depending on the values of $z, \varepsilon_t, \hat{y}_t$, it possible that either $a_t < 0$, or $b_t > 1$, or both $a_t < 0$ and $b_t > 1$. In cases when one of the end-points lies outside the interval $[0, 1]$, effectively only one of the bounds $\mathbb{P}_\lambda(i|\mathcal{M}) \geq a_t$ and $\mathbb{P}_\lambda(i|\mathcal{M}) \leq b_t$ is imposed. Likewise, when both the end-points lie outside the interval $[0, 1]$, the entire constraint $a_t \leq \mathbb{P}_\lambda(i|\mathcal{M}) \leq b_t$ becomes redundant. This implies that depending on the estimate $\hat{y}_t$ and its quality ε_t, the robust approach automatically discards some of "low-quality" data while generating predictions.

MNL model. We assumed the following specific random utility structure for the MNL model: $U_i = V_i + \xi_i$, $i = 0, 1, 2, \ldots, N$, where V_i is the mean utility and ξ_i are i.i.d. Gumbel distributed with location parameter 0 and scale parameter 1, and $N = 14$ is the number of products. With this assumption, we used the package BIOGEME Bierlaire [2003, 2008] to estimate V_i for $i = 1, 2, \ldots, N$ (with V_0 fixed at 0) from training data $\mathscr{M}^{\text{training}}$.

MMNL model. We assumed the following specific random utility structure for the

MMNL model: $U_i = V_i + \beta x_i + \xi_i,\ i = 0, 1, 2, \ldots, 14$, where as before V_i denotes the mean utility and $N = 14$ the number of products, ξ_i are i.i.d. Gumbel with location parameter 0 and scale parameter 1, x_i are dummy features with $x_0 = 0$ and $x_i = 1$ for $i > 0$, and β is Gaussian with mean 0 and variance s^2. Again, fixing V_0 to 0, we used BIOGEME Bierlaire [2003, 2008] to estimate $V_i, i > 0$, and s.

A.4 Proofs of Section 3.6

In this section, we give the proofs of Theorems 2 and 3, and Lemma 1.

A.4.1 Proof of Theorem 2

We want to upper bound $\delta_k(\mathcal{M})$. Since $\delta_k(\mathcal{M}) = \max_{j \in \mathcal{M}} \delta_k(j \mid \mathcal{M})$, we start with upper bounding $\delta_k(j \mid \mathcal{M})$. Using the definitions, we have

$$\delta_k(j \mid \mathcal{M}) = \min_{\mathcal{M}' \subset \mathcal{M}:\ j \in \mathcal{M}';\, |\mathcal{M}'| < k} \frac{B_j(\mathcal{M}'; \mathcal{M})}{\mathbb{P}\left(j \mid \mathcal{M}' \cup \{0\}\right)},$$

where $B_j(\mathcal{M}'; \mathcal{M}) = \sum_{i \in \mathcal{M} \setminus \mathcal{M}'} B_j(\mathcal{M}'; i)$ with

$$B_j(\mathcal{M}'; i) = \mathbb{P}\left(j \mid \mathcal{M}' \cup \{0\}\right) - \mathbb{P}\left(j \mid \mathcal{M}' \cup \{i, 0\}\right).$$

Now, let $\mathcal{M}_j$ denote the set $\{1, 2, \ldots, k-1\} \cup \{j\}$. Clearly, $|\mathcal{M}_j| \leq k$. Therefore, for any $i \in \mathcal{M} \setminus \mathcal{M}_j$, we can write

$$\begin{aligned}
B_j(\mathcal{M}_j; i) &= \frac{w_j}{1 + \sum_{\ell \in \mathcal{M}_j} w_\ell} - \frac{w_j}{1 + w_i + \sum_{\ell \in \mathcal{M}_j} w_\ell} \\
&= \frac{w_j}{1 + \sum_{\ell \in \mathcal{M}_j} w_\ell} \left(1 - \frac{1 + \sum_{\ell \in \mathcal{M}_j} w_\ell}{1 + w_i + \sum_{\ell \in \mathcal{M}_j} w_\ell}\right) \\
&= \mathbb{P}\left(j \mid \mathcal{M}_j \cup \{0\}\right) \frac{w_i}{1 + w_i + \sum_{\ell \in \mathcal{M}_j} w_\ell} \\
&\leq \mathbb{P}\left(j \mid \mathcal{M}_j \cup \{0\}\right) \frac{w_i}{1 + w_1 + w_2 + \cdots w_{k-1}}.
\end{aligned}$$

It now follows that

$$\frac{B_j(\mathcal{M}_j;\mathcal{M})}{\mathbb{P}(j \mid \mathcal{M}_j \cup \{0\})} \leq \frac{\sum_{i\in\mathcal{M}\setminus\mathcal{M}_j} w_i}{1+w_1+\cdots+w_{k-1}} \leq \frac{w_k+w_{k+1}+\cdots+w_C}{1+w_1+\cdots+w_{k-1}}.$$

The result now follows from the definition of $\delta_k(\mathcal{M})$.

A.4.2 Proof of Theorem 3

Fix a product $j \in \mathcal{M}$. Our goal is to upper bound $\delta_k(j \mid \mathcal{M})$. For that, we need the following notation. Given any assortment $\mathcal{M}$, for any β, let $\mathcal{M}_k^\beta$ denote the assortment that minimizes $B_j(\mathcal{M}';\mathcal{M})/L_\beta(j \mid \mathcal{M}' \cup \{0\})$ over all assortments $\mathcal{M}' \subset \mathcal{M}$ such that $j \in \mathcal{M}'$ and $|\mathcal{M}'| \leq k$ for the MNL model with weights $w_i = \exp(\langle \beta, x_i \rangle)$. As discussed above, $\mathcal{M}_k^\beta$ corresponds to the assortment with essentially the largest k weights in $\mathcal{M}$; more precisely we must have $w_\ell \geq w_i$ for all products $\ell \in \mathcal{M}_k^\beta \setminus \{j\}$ and $i \in \mathcal{M} \setminus \mathcal{M}_k^\beta$. In other words, we have:

$$\langle \beta, x_\ell - x_i \rangle \geq 0 \quad \forall\, \ell \in \mathcal{M}_k^\beta \setminus \{j\} \text{ and } i \in \mathcal{M} \setminus \mathcal{M}_k^\beta. \tag{A.16}$$

Clearly, (A.16) is a a polyhedral cone. In fact, if the attribute vectors x_j are all distinct, then it is a pointed cone with 0 as the only extreme point. Now, let $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_V$ denote the different polyhedral cones obtained for different assortments $\mathcal{M}' \subset \mathcal{M}$ such that $|\mathcal{M}'| < k$. One can then write

$$\mathbb{P}(j \mid \mathcal{M} \cup \{0\}) = \sum_{v=1}^{V} \int_{\mathcal{R}_v} L_\beta(j \mid \mathcal{M} \cup \{0\})\, G(d\beta).$$

Further, for each $1 \leq v \leq V$, let $\mathcal{M}_k^v \subset \mathcal{M}$ denote the subset of $k-1$ products with the largest weights according to $w_i^\beta = \exp(\langle \beta, x_i \rangle)$, where $\beta \in \mathcal{R}_v$ (it follows from our definitions that the subset of largest $k-1$ weights remains the same for all $\beta \in \mathcal{R}_v$). Further, let $\mathcal{M}^v$ denote the set $\mathcal{M}_k^v \cup \{j\}$. Also, without loss of generality assume that $\mu \in \mathcal{R}_1$.

Now for any product $i \in \mathcal{M} \setminus \mathcal{M}^1$, we have

$$\begin{aligned}
B_j(\mathcal{M}^1;\mathcal{M}) &= \sum_{i\in\mathcal{M}\setminus\mathcal{M}^1} B_j(\mathcal{M}^1;i) \\
&= \sum_{i\in\mathcal{M}\setminus\mathcal{M}^1} \left(\mathbb{P}\left(j \mid \mathcal{M}^1 \cup \{0\}\right) - \mathbb{P}\left(j \mid \mathcal{M}^1 \cup \{0,i\}\right)\right) \\
&= \sum_{v=1}^{V} \int_{\mathcal{R}_v} \sum_{i\in\mathcal{M}\setminus\mathcal{M}^1} \left(L_\beta\left(j \mid \mathcal{M}^1 \cup \{0\}\right) - L_\beta\left(j \mid \mathcal{M} \cup \{0,i\}\right)\right) G(d\beta) \\
&\leq \sum_{v=1}^{V} \int_{\mathcal{R}_v} \sum_{i\in\mathcal{M}\setminus\mathcal{M}^v} \left(L_\beta\left(j \mid \mathcal{M}^v \cup \{0\}\right) - L_\beta\left(j \mid \mathcal{M}^v \cup \{0,i\}\right)\right) G(d\beta) \\
&\quad + \sum_{v=2}^{V} \int_{\mathcal{R}_v} \sum_{i\in\mathcal{M}\setminus\mathcal{M}^1} L_\beta\left(j \mid \mathcal{M}^1 \cup \{0\}\right) + \sum_{i\in\mathcal{M}\setminus\mathcal{M}^v} L_\beta\left(j \mid \mathcal{M}^v \cup \{0,i\}\right) G(d\beta) \\
&\leq \sum_{v=1}^{V} \int_{\mathcal{R}_v} \delta_k^\beta(j;\mathcal{M}) L_\beta\left(j \mid \mathcal{M}_v \cup \{0\}\right) + 2k(1 - G(\mathcal{R}_1)),
\end{aligned}$$

where the last inequality follows because (a) $\mathcal{M}^v$ is the assortment that maximizes $B_j(\mathcal{M}';\mathcal{M})$ among all assortments such that $j \in \mathcal{M}'$ and $|\mathcal{M}'| \leq k$ for the MNL model with weights $w_i = \exp(\langle \beta, x_i \rangle)$ and $\beta \in \mathcal{R}_v$, and (b) $L_\beta\left(j \mid \mathcal{M}^1 \cup \{0\}\right) \leq 1$, $L_\beta\left(j \mid \mathcal{M}^v \cup \{0,i\}\right) \leq 1$, and $|\mathcal{M}^1|, |\mathcal{M}^v| \leq k$.

In order to finish the proof of the theorem, we only need to establish that

$$1 - G(\mathcal{R}_1) \leq \varepsilon_r.$$

For that, let B^r denote the ball of radius r around μ. Then, $\mathcal{R} \stackrel{\text{def}}{=} \{\rho\beta\colon \rho > 0, \beta \in B^r\}$ denotes the cone formed by taking the convex combination of the origin and B^r. We now claim that $\mathcal{R} \subset \mathcal{R}_1$. Assuming that the claim is true, it immediately follows that $G(\mathcal{R}_1) \geq G(\mathcal{R})$.

In order to prove that $\mathcal{R} \subset \mathcal{R}_1$, consider a vector $v \in \mathcal{R}$. By the definition of $\mathcal{R}$, we can write $v = \rho\beta$ for some $\rho > 0$ and $\beta \in B^r$. Let $\mathcal{M}'$ be the assortment that defines $\mathcal{R}_1$. Then, for any pair of products $\ell \in \mathcal{M}' \setminus \{j\}$ and $i \notin \mathcal{M}'$, we must have

$\langle \mu, x_\ell - x_i \rangle \geq 0$ since $\mu \in \mathcal{R}_i$. It now follows that

$$\begin{aligned}
\frac{1}{\rho}\langle v, x_\ell - x_i \rangle = \langle \beta, x_\ell - x_i \rangle &= \langle \mu, x_\ell - x_i \rangle + \langle \beta - \mu, x_\ell - x_i \rangle \\
&= |\langle \mu, x_\ell - x_i \rangle| + \langle \beta - \mu, x_\ell - x_i \rangle \\
&\overset{(a)}{\geq} |\langle \mu, x_\ell - x_i \rangle| - \|\beta - \mu\| \|x_\ell - x_i\| \\
&\overset{(b)}{\geq} |\langle \mu, x_\ell - x_i \rangle| - r\|x_\ell - x_i\| \\
&= \|x_\ell - x_i\| \left(\frac{|\langle \mu, x_\ell - x_i \rangle|}{\|x_\ell - x_i\|} - r \right) \\
&\overset{(c)}{\geq} 0,
\end{aligned}$$

where (a) follows from Cauchy-Schwarz inequality, (b) follows from the fact that $\beta \in B^r$, and (c) follows from the fact that $r \leq |\langle \mu, x_y - x_z \rangle| / \|x_y - x_z\|$, which follows from the definition of r. It is now straightforward to conclude that $v \in \mathcal{R}_1$ and, hence, $\mathcal{R} \subset \mathcal{R}_1$.

Now consider the case when $\Sigma = \sigma^2 I_D$. Let $(B^r)^c$ denote $\mathbb{R}^D \setminus B^r$. Then,

$$G(\mathcal{R}) \geq G(B^r) = 1 - G((B^r)^c) = 1 - \frac{1}{\sigma^D (2\pi)^{D/2}} \int_{(B^r)^c} \exp\left(\frac{\|\beta - \mu\|^2}{2\sigma^2} \right) d\beta.$$

Using the transformation $\beta = \mu + \rho\theta$, where θ is a point on the unit sphere S^{D-1}, we can write

$$\begin{aligned}
G((B^r)^c) &= \frac{1}{\sigma^D (2\pi)^{D/2}} \int_{\rho > r, \theta \in S^{D-1}} \exp\left(-\frac{\rho^2}{2\sigma^2} \right) d\rho d\theta \\
&= \frac{1}{\sigma^D (2\pi)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2} \right) \int_{\theta \in S^{D-1}} d\theta \int_{\rho > r} \exp\left(-\frac{\rho^2 - r^2}{2\sigma^2} \right) d\rho \\
&= \frac{\text{area}(S^{D-1})}{\sigma^D (2\pi)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2} \right) \int_{\rho > r} \exp\left(-\frac{(\rho - r)(\rho + r)}{2\sigma^2} \right) d\rho \\
&\leq \frac{\text{area}(S^{D-1})}{\sigma^D (2\pi)^{D/2}} \exp\left(-\frac{r^2}{2\sigma^2} \right) \int_{\rho > r} \exp\left(-\frac{(\rho - r)2r}{2\sigma^2} \right) d\rho \\
&= C(D) \frac{\exp\left(-\frac{r^2}{2\sigma^2} \right)}{r}.
\end{aligned}$$

where $C(D) = \sigma^2 \text{area}(S^{D-1}) / (\sigma^D (2\pi)^{D/2})$. Here the inequality follows from the fact

that $(\rho - r)(\rho + r) \geq (\rho - r)(2r)$ for $\rho \geq r$, and the last equality follows from the fact that $\int_{\rho > r} \exp\left(-(\rho - r)r/\sigma^2\right) = \sigma^2/r$. The surface area of S^{D-1} is $2\pi^{D/2}/\Gamma(D/2)$, where $\Gamma(\cdot)$ is Euler's Gamma function. Therefore, $C(D) = 1/(2^{(D/2-1)}\sigma^{D-2}\Gamma(D/2))$. The result of theorem now follows.

A.4.3 Proof of Lemma 1

Suppose the underlying choice model is described by the distribution λ over permutations. Now, given two assortments $\mathcal{M}' \subset \mathcal{M}$ such that $j \in \mathcal{M}'$, let $\mathcal{E}_j$ denote the set of permutations σ such that $\sigma(j) < \sigma(\ell)$ for all products $\ell \in \mathcal{M}'$. Further, for any $i \in \mathcal{M} \setminus \mathcal{M}'$, let $\mathcal{E}_i$ denote the set of permutations σ such that $\sigma(i) < \sigma(j)$ and $\sigma(j) < \sigma(\ell)$ for all products $\ell \in \mathcal{M}' \cup \{0\}$. We can now write

$$\begin{aligned}
\mathbb{P}\left(j \mid \mathcal{M}' \cup \{0\}\right) - \mathbb{P}\left(j \mid \mathcal{M} \cup \{0\}\right) &= \lambda\left(\bigcup_{i \in \mathcal{M} \setminus \mathcal{M}'} \mathcal{E}_i\right) \\
&\leq \sum_{i \in \mathcal{M} \setminus \mathcal{M}'} \lambda(\mathcal{E}_i) \\
&= \sum_{i \in \mathcal{M} \setminus \mathcal{M}'} \mathbb{P}\left(j \mid \mathcal{M}' \cup \{0\}\right) - \mathbb{P}\left(j \mid \mathcal{M}' \cup \{0, i\}\right) \\
&= \sum_{i \in \mathcal{M} \setminus \mathcal{M}'} B_j(\mathcal{M}'; i),
\end{aligned}$$

where the inequality follows from the union bound. The result of the lemma now follows.

Appendix B

Appendix: Learning choice models

B.1 Proof of Theorem 7

The proof of Theorem 7 requires us to establish two claims: whenever λ satisfies Condition 1, (i) the sparsest-fit algorithm recovers λ from $y = A\lambda$ and (ii) λ is the unique solution to the program in (5.1). We establish these two claims in that order.

The sparsest-fit algorithm works. Let $\sigma_1, \sigma_2, \ldots, \sigma_K$ be the permutations in the support and $\lambda_1, \lambda_2, \ldots, \lambda_K$ be their corresponding probabilities. Since we assumed that λ satisfies the "signature" condition, for each $1 \leq i \leq K$, there exists a $d(i)$ such that $y_{d(i)} = \lambda_i$. In addition, the linear independence condition guarantees that the condition in the **if** statement of the algorithm is not satisfied whenever $d = d(i)$. To see why, suppose the condition in the **if** statement is true; then, we will have $\lambda_{d(i)} - \sum_{i \in T} \lambda_i = 0$. Since $d(i) \notin T$, this clearly violates the linear independence condition. Therefore, the algorithm correctly assigns values to each of the λ_is. We now prove that the $A(\sigma)$s that are returned by the algorithm do indeed correspond to the σ_is. For that, note that the condition in the **if** statement being true implies that y_d is a linear combination of a subset T of the set $\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$. Again, the linear independence condition guarantees that such a subset T, if exists, is unique. Thus, when the condition in the **if** statement is true, only the permutations with $A(\sigma)_d = 1$ are in the set T. Similarly, when the condition in the **if** statement is false, then it

follows from the signature and linear independence conditions that only for σ_i, we have $A(\sigma)_{d(i)} = 1$. From this, we conclude that the algorithm correctly finds the true underlying distribution.

Unique solution to ℓ_0 optimization. Suppose, to arrive at a contradiction, assume that there exists a distribution μ over the permutations such that $y = A\mu$ and $\|\mu\|_0 \leq \|\lambda\|_0$. Let $v_1, v_2, \ldots, v_K$ and $u_1, u_2, \ldots, u_L$ denote the values that λ and μ take on their respective supports. It follows from our assumption that $L \leq K$. In addition, since λ satisfies the signature condition, there exist $1 \leq d(i) \leq m$ such that $y_{d(i)} = v_i$, for all $1 \leq i \leq K$. Thus, since $y = A\mu$, for each $1 \leq i \leq K$, we can write $v_i = \sum_{j \in T(i)} u_j$, for some $T(i) \subseteq \{1, 2, \ldots, L\}$. Equivalently, we can write $v = Bu$, where B is a $0-1$ matrix of dimensions $K \times L$. Consequently, we can also write $\sum_{i=1}^{k} v_i = \sum_{j=1}^{L} \zeta_j u_j$, where ζ_j are integers. This now implies that $\sum_{j=1}^{L} u_j = \sum_{j=1}^{L} \zeta_j u_j$ since $\sum_{i=1}^{K} v_i = \sum_{j=1}^{L} u_j = 1$.

Now, there are two possibilities: either all the ζ_js are > 0 or some of them are equal to zero. In the first case, we prove that μ and λ are identical, and in the second case we arrive at a contradiction. In the case when $\zeta_j > 0$ for all $1 \leq j \leq L$, since $\sum_j u_j = \sum_j \zeta_j u_j$, it should follow that $\zeta_j = 1$ for all $1 \leq j \leq L$. Thus, since $L \leq K$, it should be that $L = K$ and $(u_1, u_2, \ldots, u_L)$ is some permutation of $(v_1, v_2, \ldots, v_K)$. By relabeling the u_js, if required, without loss of generality, we can say that $v_i = u_i$, for $1 \leq i \leq K$. We have now proved that the values of λ and μ are identical. In order to prove that they have identical supports, note that since $v_i = u_i$ and $y = A\lambda = A\mu$, μ must satisfy the signature and the linear independence conditions. Thus, the algorithm we proposed accurately recovers μ and λ from y. Since the input to the algorithm is only y, it follows that $\lambda = \mu$.

Now, suppose that $\zeta_j = 0$ for some j. Then, it follows that some of the columns in the B matrix are zeros. Removing those columns of B, we can write $v = \tilde{B}\tilde{u}$ where $\tilde{B}$ is B with the zero columns removed and $\tilde{u}$ is u with u_js such that $\zeta_j = 0$ removed. Let $\tilde{L}$ be the size of $\tilde{u}$. Since at least one column was removed $\tilde{L} < L \leq K$. The condition $\tilde{L} < K$ implies that the elements of vector v are not linearly independent i.e., we can find integers c_i such that $\sum_{i=1}^{K} c_i v_i = 0$. This is a contradiction, since this

condition violates our linear independence assumption. The result of the theorem now follows.

B.2 Proof of Theorem 8

We prove this theorem by showing that when two permutations, say σ_1, σ_2, are chosen uniformly at random, then with a high probability, the sum of columns of the A matrix corresponding to the permutations $A^\rho(\sigma_1) + A^\rho(\sigma_2)$ can be decomposed in at least two ways. For that, note that a permutation can be represented using cycle notation, e.g. for $N = 4$, the permutation $1 \mapsto 2, 2 \mapsto 1, 3 \mapsto 4, 4 \mapsto 3$ can be represented as a composition of two cycles $(12)(34)$. We call two cycles *distinct* if they have no elements in common, e.g. the cycles (12) and (34) are distinct. Given two permutations σ_1 and σ_2, let $\sigma_{1,2} = \sigma_1\sigma_2$ be their composition.

Now consider two permutations σ_1 and σ_2 such that they have distinct cycles. For example, $\sigma_1 = (1, 2)$ and $\sigma_2 = (3, 4)$ are permutations with distinct cycles. Then $\sigma_{1,2} = \sigma_1\sigma_2 = (12)(34)$. We first prove the theorem for $\rho = (N-1, 1)$ and then extend it to a general ρ; thus, we fix the partition $\rho = (N-1, 1)$. Then, we have:

$$A^\rho(\sigma_1) + A^\rho(\sigma_2) = A^\rho(\sigma_{1,2}) + A^\rho(\mathrm{id}) \tag{B.1}$$

where σ_1 and σ_2 have distinct cycles and id is the identity permutation. Now, assuming that $p_1 \leq p_2$, consider the following:

$$\begin{aligned} & p_1 A^\rho(\sigma_1) + p_2 A^\rho(\sigma_2) \\ = \; & p_1 A^\rho(\sigma_{1,2}) + p_1 A^\rho(\mathrm{id}) + (p_2 - p_1) A^\rho(\sigma_2). \end{aligned}$$

Thus, given $M^\rho(\lambda) = p_1 A^\rho(\sigma_1) + p_2 A^\rho(\sigma_2)$, it can be decomposed in two distinct ways with both having the same ℓ_1 norm. Of course, the same analysis can be carried out when λ has a sparsity K. Thus, we conclude that whenever λ has two permutations with distinct cycles in its support, the ℓ_1 minimization solution is not unique. Therefore, to establish claim of Theorem 8, it is sufficient to prove that when

we choose two permutations uniformly at random, they have distinct cycles with a high probability.

To this end, let $\mathcal{E}$ denote the event that two permutations chosen uniformly at random have distinct cycles. Since permutations are chosen uniformly at random, $\mathbb{P}(\mathcal{E})$ can be computed by fixing one of the permutations to be id. Then, $\mathbb{P}(\mathcal{E})$ is the probability that a permutation chosen at random has more than one cycle.

Let us evaluate $\mathbb{P}(\mathcal{E}^c)$. For that, consider a permutation having exactly one cycle with the cycle containing L elements. The number of such permutations will be $\binom{N}{L}(L-1)!$. This is because we can choose the l elements that form the cycle in $\binom{N}{L}$ ways and the L numbers can be arranged in the cycle in $(L-1)!$ ways. Therefore,

$$\mathbb{P}(\mathcal{E}^c) = \frac{1}{N!}\sum_{L=1}^{N}\binom{N}{L}(L-1)! = \sum_{L=1}^{N}\frac{1}{L(N-L)!} \tag{B.2}$$

Now, without loss of generality let's assume that N is even. Then,

$$\sum_{L=1}^{N/2}\frac{1}{L(N-L)!} \leq \sum_{L=1}^{N/2}\frac{1}{\left(\frac{N}{2}\right)!} = \frac{1}{\left(\frac{N}{2}-1\right)!} \tag{B.3}$$

The other half of the sum becomes

$$\sum_{L=N/2}^{N}\frac{1}{L(N-L)!} \leq \sum_{k=0}^{N/2}\frac{1}{\frac{N}{2}k!} \leq \frac{2}{N}\sum_{k=0}^{\infty}\frac{1}{k!} \leq \frac{O(1)}{N} \tag{B.4}$$

Putting everything together, we have

$$\mathbb{P}(\mathcal{E}) \geq 1 - \mathbb{P}(\mathcal{E}^c) \geq 1 - \left(\frac{1}{\left(\frac{N}{2}-1\right)!} + \frac{O(1)}{N}\right)$$
$$\rightarrow 1 \text{ as } N \rightarrow \infty.$$

Thus, Theorem 8 is true for $\rho = (N-1, 1)$.

In order to extend the proof to a general ρ, we observe that the standard cycle notation for a permutation we discussed above can be extended to a general ρ partition. Specifically, for any given ρ, observe that a permutation can be imagined

as a perfect matching in a $D_\rho \times D_\rho$ bipartite graph, which we call the ρ-bipartite graph and denote it by $G^\rho = (V_1^\rho \times V_2^\rho, E^\rho)$; here V_1^ρ and V_2^ρ respectively denote the left and right vertex sets with $|V_1^\rho| = |V_2^\rho| = D_\rho$ with a node for every ρ partition of N. Let $t_1, t_2, \ldots, t_{D_\rho}$ denote the D_ρ ρ-partitions of N; then, the nodes in V_1^ρ and V_2^ρ can be labeled by $t_1, t_2, \ldots, t_{D_\rho}$. Since every perfect matching in a bipartite graph can be decomposed into its corresponding distinct cycles (the cycles can be obtained by superposing the bipartite graph corresponding to identity permutation with the ρ-bipartite graph of the permutation), every permutation can be written as a combination of distinct cycles in its ρ-bipartite graph. The special case of this for $\rho = (N-1, 1)$ is the standard cycle notation we discussed above; for brevity, we call the ρ-bipartite graph for $\rho = (N-1, 1)$ the standard bipartite graph.

In order to prove the theorem for a general ρ, using an argument similar to above, it can be shown that it is sufficient to prove that a randomly chosen permutation contains at least two distinct cycles in its ρ-bipartite graph with a high probability. For that, it is sufficient to prove that a permutation with at least two distinct cycles in its standard bipartite graph has at least two distinct cycles in its ρ-bipartite graph for any general ρ. The theorem then follows from the result we established above that a randomly chosen permutation has at least two distinct cycles in its standard bipartite graph with a high probability.

To that end, consider a permutation, σ, with at least two distinct cycles in the standard bipartite graph. Let $A := (a_1, a_2, \ldots, a_{\ell_1})$ and $B := (b_1, b_2, \ldots, b_{\ell_2})$ denote the first two cycles in the standard bipartite graph; clearly, $\ell_1, \ell_2 \geq 2$ and at least one of ℓ_1, ℓ_2 is $\leq N/2$. Without loss of generality we assume that $\ell_2 \leq N/2$. Let $\rho = (\rho_1, \rho_2, \ldots, \rho_s)$. Since $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_s$, we have $\rho_s \leq N/2$. First, we consider the case when $\rho_s < N/2$. Now consider the ρ-partition, t_1, of N constructed as follows: a_1 placed in the sth partition, a_2 in the first partition, all the elements of the second cycle $b_1, b_2, \ldots, b_{\ell_2}$ arbitrarily in the first $s-1$ partitions and the rest placed arbitrarily. Note that such a construction is possible by the assumption on ρ_s. Let t_1' denote $\sigma(t_1)$; then, $t_1' \neq t_1$ because t_1 does not contain a_2 in the sth partition while t_1' contains $\sigma(a_1) = a_2$ in the sth partition. Thus, the partition t_1 belongs to a cycle

that has a length of at least 2 partitions. Thus, we have found one cycle, which we denote by C_1. Now consider a second partition t_2 constructed as follows: b_1 placed in the sth partition, b_2 in the first and the rest placed arbitrarily. Again, note that $\sigma(t_2) \neq t_2$. Thus, t_2 belongs to a cycle of length at least 2, which we denote by C_2. Now we have found two cycles C_1, C_2, and we are left with proving that they are distinct. In order to establish the cycles are distinct, note that none of the partitions in cycle C_1 can be t_2. This is true because, by construction, t_2 contains b_1 in the sth partition while none of the partitions in C_1 can contain any elements from the cycle B in the sth partition. This finishes the proof for all ρ such that $\rho_s < N/2$.

We now consider the case when $\rho_s = N/2$. Since $\rho_1 \geq \rho_s$, it follows that $s = 2$ and $\rho = (N/2, N/2)$. For $\ell_2 < N/2$, it is still feasible to construct t_1 and t_2, and the theorem follows from the arguments above. Now we consider the case when $\ell_1 = \ell_2 = N/2$; let $\ell := \ell_1 = \ell_2$. Note that now it is infeasible to construct t_1 as described above. Therefore, we consider $t_1 = \{a_1, b_2, \ldots, b_\ell\}\{b_1, a_2, \ldots, a_\ell\}$ and $t_2 = \{b_1, a_2, \ldots, a_\ell\}\{a_1, b_2, \ldots, b_\ell\}$. Clearly, $t_1 \neq t_2$, $\sigma(t_1) \neq t_1$ and $\sigma(t_2) \neq t_2$. Thus, t_1 and t_2 belong to two cycles, C_1 and C_2, each with length at least 2. It is easy to see that these cycles are also distinct because every $\rho-$partition in the cycle C_1 will have only one element from cycle A in the first partition and, hence, C_1 cannot contain the $\rho-$partition t_2. This completes the proof of the theorem.

B.3 Proof of Theorem 9

First, we note that, irrespective of the form of observed data, the choice model generated from random model $R(K, \mathscr{C})$ satisfies the linear independence condition with probability 1. The reason is as follows: the values $\lambda(\sigma_i)$ obtained from the random model are i.i.d uniformly distributed over the interval $[a, b]$. Therefore, the vector $(\lambda(\sigma_1), \lambda(\sigma_2), \ldots, \lambda(\sigma_K))$ corresponds to a point drawn uniformly at random from the hypercube $[a, b]^K$. In addition, the set of points that satisfy $\sum_{i=1}^{K} c_i \lambda(\sigma_i) = 0$ lie in a lower-dimensional space. Since c_is are bounded, there are only finitely many such sets of points. Thus, it follows that with probability 1, the choice model generated

satisfies the linear independence condition.

The conditions under which the choice model satisfies the signature condition depends on the form of observed data. We consider each form separately.

Comparison information. For each permutation σ, we truncate its corresponding column vector $A(\sigma)$ to a vector of length $N/2$ by restricting it to only the disjoint unordered pairs: $\{0,1\},\{2,3\},\ldots,\{N-2,N-1\}$. Denote the truncated binary vector by $A'(\sigma)$. Let $\tilde{A}$ denote the matrix A with each column $A(\sigma)$ truncated to $A'(\sigma)$. Clearly, since $\tilde{A}$ is just a truncated form of A, it is sufficient to prove that $\tilde{A}$ satisfies the "signature" condition.

For brevity, let L denote $N/2$, and, given K permutations, let B denote the $L \times K$ matrix formed by restricting the matrix $\tilde{A}$ to the K permutations in the support. Then, it is easy to see that a set of K permutations satisfies the "signature" condition iff there exist K rows in B such that the $K \times K$ matrix formed by the K rows is a permutation matrix.

Let $R_1, R_2, \ldots, R_J$ denote all the subsets of $\{1,2,\ldots,m\}$ with cardinality K; clearly, $J = \binom{L}{K}$. In addition, let B^j denote the $K \times K$ matrix formed by the rows of B that are indexed by the elements of R_j. Now, for each $1 \leq j \leq J$, when we generate the matrix B by choosing K permutations uniformly at random, let $\mathcal{E}_j$ denote the event that the $K \times K$ matrix B^j is a permutation matrix and let $\mathcal{E}$ denote the event $\cup_j \mathcal{E}_j$. We want to prove that $\mathbb{P}(\mathcal{E}) \to 1$ as $N \to \infty$ as long as $K = o(\log N)$. Let X_j denote the indicator variable of the event $\mathcal{E}_j$, and X denote $\sum_j X_j$. Then, it is easy to see that $\mathbb{P}(X = 0) = \mathbb{P}((\mathcal{E})^c)$. Thus, we need to prove that $\mathbb{P}(X = 0) \to 0$ as $N \to \infty$ whenever $K = o(\log n)$. Now, note the following:

$$\mathrm{Var}(X) \geq (0 - \mathbb{E}[X])^2 \mathbb{P}(X = 0)$$

It thus follows that $\mathbb{P}(X = 0) \leq \mathrm{Var}(X)/(\mathbb{E}[X])^2$. We now evaluate $\mathbb{E}[X]$. Since X_js are indicator variables, $\mathbb{E}[X_j] = \mathbb{P}(X_j = 1) = \mathbb{P}(\mathcal{E}_j)$. In order to evaluate $\mathbb{P}(\mathcal{E}_j)$, we restrict our attention to the $K \times K$ matrix B^j. When we generate the entries of matrix B by choosing K permutations uniformly at random, all the elements of B

will be i.i.d Be(1/2) i.e., uniform Bernoulli random variables. Therefore, there are 2^{K^2} possible configurations of B^j and each of them occurs with a probability $1/2^{K^2}$. Moreover, there are $K!$ possible $K \times K$ permutation matrices. Thus, $\mathbb{P}(\mathcal{E}_j) = K!/2^{K^2}$. Thus, we have:

$$\mathbb{E}[X] = \sum_{j=1}^{J} \mathbb{E}[X_j] = \sum_{j=1}^{J} \mathbb{P}(\mathcal{E}_j) = \frac{JK!}{2^{K^2}}. \tag{B.5}$$

Since $J = \binom{L}{K}$, it follows from Stirling's approximation that $J \geq L^K/(eK)^K$. Similarly, we can write $K! \geq K^K/e^K$. It now follows from (B.5) that

$$\mathbb{E}[X] \geq \frac{L^K}{e^K K^K} \frac{K^K}{e^K} \frac{1}{2^{K^2}} = \frac{L^K}{e^{2K} 2^{K^2}}. \tag{B.6}$$

We now evaluate $\mathrm{Var}(X)$. Let ρ denote $K!/2^{K^2}$. Then, $\mathbb{E}[X_j] = \rho$ for all $1 \leq j \leq J$. We can write,

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \sum_{i=1}^{J} \sum_{j=1}^{J} \mathbb{P}(X_i = 1, X_j = 1) - J^2\rho^2.$$

Suppose $|R_i \cap R_j| = r$. Then, the number of possible configurations of B^i and B^j is $2^{(2K-r)K}$ because, since there is an overlap of r rows, there are $2K - r$ distinct rows and, of course, K columns. Since all configurations occur with the same probability, it follows that each configuration occurs with a probability $1/2^{(2K-r)K}$, which can also be written as $2^{rK}\rho^2/(K!)^2$. Moreover, the number of configurations in which both B^i and B^j are permutation matrices is equal to $K!(K-r)!$, since, fixing the configuration of B^i will leave only $K - r$ rows of B^j to be fixed.

For a fixed R_i, we now count the number of subsets R_j such that $|R_i \cap R_j| = r$. We construct an R_j by first choosing r rows from R_i and then choosing the rest from $\{1, 2, \ldots, l\} \setminus R_i$. We can choose r rows from the subset R_i of K rows in $\binom{K}{r}$ ways,

and the remaining $K-r$ rows in $\binom{L-K}{K-r}$ ways. Therefore, we can now write:

$$\begin{aligned}
\sum_{j=1}^{J}\mathbb{P}(X_i=1,X_j=1) &= \sum_{r=0}^{K}\binom{K}{r}\binom{L-K}{K-r}K!(K-r)!\frac{2^{rK}\rho^2}{(K!)^2} \\
&\leq \rho^2\sum_{r=0}^{K}\binom{L}{K-r}\frac{2^{rK}}{r!}, \quad \text{Using } \binom{L-K}{K-r}\leq\binom{L}{K-r} \\
&= \binom{L}{K}\rho^2+\rho^2\sum_{r=1}^{K}\binom{L}{K-r}\frac{2^{rK}}{r!} \\
&\leq J\rho^2+\rho^2L^K\sum_{r=1}^{K}\left(\frac{e2^K}{L}\right)^r\frac{1}{r^r(K-r)^{K-r}}
\end{aligned}$$

The last inequality follows from Stirling's approximation: $\binom{L}{K-r}\leq(L/(K-r))^{K-r}$ and $r!\geq(r/e)^r$; in addition, we have used $J=\binom{L}{K}$. Now consider

$$\begin{aligned}
r^r(K-r)^{K-r} &= \exp\{r\log r+(K-r)\log(K-r)\} \\
&= \exp\{K\log K-KH(r/K)\} \\
&\geq \frac{K^K}{2^K}
\end{aligned}$$

where $H(x)$ is the Shannon entropy of the random variable distributed as $\text{Be}(x)$, defined as $H(x)=-x\log x-(1-x)\log(1-x)$ for $0<x<1$. The last inequality follows from the fact that $H(x)\leq\log 2$ for all $0<x<1$. Putting everything together, we get

$$\begin{aligned}
\text{Var}(X) &= \sum_{i=1}^{J}\left[\sum_{j=1}^{J}\mathbb{P}(X_i=1,X_j=1)\right]-\mathbb{E}[X]^2 \\
&\leq J\left[J\rho^2+\rho^2L^K\frac{2^K}{K^K}\sum_{r=1}^{K}\left(\frac{e2^K}{L}\right)^r\right]-J^2\rho^2 \\
&= \frac{J\rho^2 2^K L^K}{K^K}\sum_{r=1}^{K}\left(\frac{e2^K}{L}\right)^r
\end{aligned}$$

We can now write,

$$\begin{aligned}
\mathbb{P}(X=0) &\leq \frac{\mathrm{Var}(X)}{(\mathbb{E}[X])^2} \\
&\leq \frac{1}{J^2\rho^2}\frac{J\rho^2 2^K L^K}{K^K}\sum_{r=1}^{K}\left(\frac{e2^K}{L}\right)^r \\
&= \frac{1}{J}\,\frac{2^K L^K}{K^K}\,\frac{e2^K}{L}\sum_{r=0}^{K-1}\left(\frac{e2^K}{L}\right)^r \\
&\leq \frac{e^K K^K}{L^K}\,\frac{2^K L^K}{K^K}\,\frac{e2^K}{L}\sum_{r=0}^{K-1}\left(\frac{e2^K}{L}\right)^r, \quad \text{Using } J=\binom{L}{K}\leq\left(\frac{L}{eK}\right)^K \\
&= e\frac{(4e)^K}{L}\sum_{r=0}^{K-1}\left(\frac{e2^K}{L}\right)^r
\end{aligned}$$

It now follows that for $K = o(\log L/\log(4e))$, $\mathbb{P}(X=0)\to 0$ as $N\to\infty$. Since, by definition, $L = N/2$, this completes the proof for the case of comparison information.

Top-set information. For this type of data, let A^1 denote the $N \times N!$ submatrix of the matrix A that corresponds to the fraction of customers who have product i as their top-choice for each i. Note that it is sufficient to prove that $A^{(1)}$ satisfies the signature property with a high probability; therefore, we ignore the comparison data and focus only on the data corresponding to the fraction of customers that have product i as their top choice, for every product i. For brevity, we abuse the notation and denote $A^{(1)}$ by A and the corresponding part of the data vector $y^{(1)}$ by y. Clearly, y is of length N and so is each column vector $A(\sigma)$. Every permutation σ ranks only one product in the first position. Hence, for every permutation σ, exactly one element of the column vector $A(\sigma)$ is 1 and the rest are zeros.

In order to obtain a bound on the support size, we reduce this problem to a balls-and-bins setup. For that, imagine K balls being thrown uniformly at random into N bins. In our setup, the K balls correspond to the K permutations in the support and the N bins correspond to the N products. A ball is thrown into bin i provided the permutation corresponding to the ball ranks product i to position 1. Our random model chooses permutations independently; hence, the balls are thrown independently. In addition, a permutation chosen uniformly at random ranks a given product i to position 1 with probability $1/N$. Therefore, each ball is thrown uniformly

at random.

In the balls-and-bins setup, the signature condition translates into all K balls falling into different bins. By "Birthday Paradox" McKinney [1966], the K balls falls into different bins with a high probability provided $K = o(\sqrt{N})$. This finishes the proof for the case of top-set information.

First-order information. Our interest is in recovering a random distribution λ from partial information $M^{\rho}(\lambda)$ for $\rho = (N-1, 1)$. To this end, let

$$K = \|\lambda\|_0, \quad \text{supp}\,(\lambda) = \{\sigma_k \in S_N : 1 \leq k \leq K\},$$
$$\text{and} \quad \lambda(\sigma_k) = p_k, 1 \leq k \leq K.$$

Here σ_k and p_k are randomly chosen as per the random model $R(K, \mathscr{C})$ described in Section 3.2. For $\rho = (N-1, 1)$, $D_\rho = N$; thus, $M^{\rho}(\lambda)$ is an $N \times N$ matrix with its (i,j)th entry being

$$M^{\rho}(\lambda)_{ij} = \sum_{k:\sigma_k(j)=i} p_k, \qquad \text{for} \qquad 1 \leq i, j \leq N.$$

In order to establish the result for the first-order information, we prove that as long as $K \leq C_1 N \log N$ with $C_1 = 1 - \varepsilon$, λ satisfies the signature condition with probability $1 - o(1)$ for any fixed $\varepsilon > 0$. To this end, let $4\delta = \varepsilon$ so that $C_1 \leq 1 - 4\delta$. Let $\mathcal{E}_k$ be the event that σ_k satisfies the signature condition, for $1 \leq k \leq K$. Under the random model $R(K, \mathscr{C})$, since K permutations are chosen from S_N independently and uniformly at random, it follows that $\mathbb{P}(\mathcal{E}_k)$ is the same for all k. Therefore, by union bound, it is sufficient to establish that $K\mathbb{P}(\mathcal{E}_1^c) = o(1)$. Since we are interested in $K = O(N \log N)$, it is sufficient to establish $\mathbb{P}(\mathcal{E}_1^c) = O(1/N^2)$. Finally, once again due the symmetry, it is sufficient to evaluate $\mathbb{P}(\mathcal{E}_1)$ assuming $\sigma_1 = \text{id}$, i.e. $\sigma_1(i) = i$ for $1 \leq i \leq N$. Define

$$\mathscr{F}_j = \{\sigma_k(j) \neq j, \quad \text{for} \quad 2 \leq k \leq K\}, \quad \text{for} \quad 1 \leq j \leq N.$$

It then follows that

$$\mathbb{P}(\mathcal{E}_1) = \mathbb{P}\left(\cup_{j=1}^{n} \mathscr{F}_j\right).$$

Therefore, for any $L \leq N$, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}\left(\cap_{j=1}^{N} \mathscr{F}_j^c\right) \\
&\leq \mathbb{P}\left(\cap_{j=1}^{L} \mathscr{F}_j^c\right) \\
&= \mathbb{P}(\mathscr{F}_1^c)\left[\prod_{j=2}^{L} \mathbb{P}\left(\mathscr{F}_j^c \,\middle|\, \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\right)\right]. \qquad \text{(B.7)}
\end{aligned}$$

We next show that for the selection $L = N^{1-\delta}$, the RHS of (B.7) is bounded above by $\exp(-N^\delta) = O(1/N^2)$. That will complete the proof for this case.

For that, we start by bounding $\mathbb{P}(\mathscr{F}_1^c)$:

$$\begin{aligned}
\mathbb{P}(\mathscr{F}_1^c) &= 1 - \mathbb{P}(\mathscr{F}_1) \\
&= 1 - \left(1 - \frac{1}{N}\right)^{K-1}. \qquad \text{(B.8)}
\end{aligned}$$

The last equality follows because all permutations are chosen uniformly at random. For $j \geq 2$, we now evaluate $\mathbb{P}\left(\mathscr{F}_j^c \,\middle|\, \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\right)$. Given $\cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c$, for any $k, 2 \leq k \leq K$, $\sigma_k(j)$ will take a value from $N - j + 1$ values, possibly including j, uniformly at random. Thus, we obtain the following bound:

$$\mathbb{P}\left(\mathscr{F}_j^c \,\middle|\, \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\right) \leq 1 - \left(1 - \frac{1}{N-j+1}\right)^{K-1}. \qquad \text{(B.9)}$$

From (B.7)-(B.9), we obtain that

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &\leq \prod_{j=1}^{L} \left(1 - \left(1 - \frac{1}{N-j+1}\right)^{K-1}\right) \\
&\leq \left[1 - \left(1 - \frac{1}{N-L}\right)^{K}\right]^{L} \\
&\leq \left[1 - \left(1 - \frac{1}{N-L}\right)^{C_1 N \log N}\right]^{L}, \qquad \text{(B.10)}
\end{aligned}$$

where we have used $K \leq C_1 N \log N$ in the last inequality. Since $L = N^{1-\delta}$, $N - L = N(1 - o(1))$. Using the standard fact $1 - x = e^{-x}(1 + O(x^2))$ for small $x \in [0, 1)$, we have

$$\left(1 - \frac{1}{N-L}\right) = \exp\left(-\frac{1}{N-L}\right)\left(1 + O\left(\frac{1}{N^2}\right)\right). \tag{B.11}$$

Finally, observe that

$$\left(1 + O\left(\frac{1}{N^2}\right)\right)^{C_1 N \log N} = \Theta(1).$$

Therefore, from (B.10) and (B.11), it follows that

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &\leq \left[1 - \Theta\left(\exp\left(-\frac{C_1 \log N}{1 - N^{-\delta}}\right)\right)\right]^L \\
&\leq \left[1 - \Theta\left(\exp\left(-(C_1 + \delta)\log N\right)\right)\right]^L \\
&= \left[1 - \Theta\left(\frac{1}{N^{C_1+\delta}}\right)\right]^L \\
&\leq \exp\left(-\Theta\left(\frac{L}{N^{C_1+\delta}}\right)\right) \\
&= \exp\left(-\Omega(N^{2\delta})\right),
\end{aligned} \tag{B.12}$$

where we have used the fact that $1 - x \leq e^{-x}$ for $x \in [0, 1]$ and $L = N^{1-\delta}$, $C_1 \leq 1 - 4\delta$. From (B.12), it follows that $\mathbb{P}(\mathcal{E}_1) = O(1/N^2)$. This completes the proof for this case.

The result of the theorem now follows.

B.4 Proof of Theorem 10

Our interest is in recovering the random distribution λ from partial information $M^\rho(\lambda)$ for $\rho = (N - v, v)$. As in proof of Theorem 9, we use the notation

$$K = \|\lambda\|_0, \quad \text{supp}\,(\lambda) = \{\sigma_k \in S_N : 1 \leq k \leq K\},$$
$$\text{and} \quad \lambda(\sigma_k) = p_k, 1 \leq k \leq K.$$

Here σ_k and p_k are randomly chosen as per the random model $R(K, \mathscr{C})$ described in Section 3.2. For $\rho = (N - v, v)$, $D_\rho = \frac{N!}{(N-v)!v!} \sim N^v$ and $M^\rho(\lambda)$ is an $D_\rho \times D_\rho$ matrix.

To establish the result of Theorem 10, we shall prove that as long as $K \leq C_1 N^v \log N$ with $0 < C_1 < \frac{1}{v!}$ a constant, λ satisfies the signature and linear independence conditions with probability $1 - o(1)$. As noted in the proof of Theorem 7, the linear independence is satisfied with probability 1 under $R(K, \mathscr{C})$. Therefore, we are left with establishing the signature property.

To this end, for the purpose of bounding, without loss of generality, let us assume that $K = \frac{(1-2\delta)}{v!} N^v \log N$ for some $\delta > 0$. Set $L = N^{1-\delta}$. Following arguments similar to those in the proof of Theorem 9, it will be sufficient to establish that $\mathbb{P}(\mathcal{E}_1^c = O(1/N^{2v})$; where $\mathcal{E}_1$ is the event that permutation $\sigma_1 = \text{id}$ satisfies the signature condition.

To this end, recall that $M^\rho(\lambda)$ is a $D_\rho \times D_\rho$ matrix. Each row (and column) of this matrix corresponds to a distinct ρ partition of N : $t_i, 1 \leq i \leq D_\rho$. Without loss of generality, let us order the D_ρ ρ partitions of N so that the ith partition, t_i, is defined as follows: $t_1 = \{1, \ldots, N - v\}\{N - v + 1, \ldots, N\}$, and for $2 \leq i \leq L$,

$$t_i = \{1, \ldots, N - iv, N - (i-1)v + 1, \ldots, N\} \\ \{N - iv + 1, \ldots, N - (i-1)v\}.$$

Note that since $\sigma_1 = \text{id}$, we have $\sigma_1(t_i) = t_i$ for all $1 \leq i \leq D_\rho$. Define

$$\mathscr{F}_j = \{\sigma_k(t_j) \neq t_j, \quad \text{for} \quad 2 \leq k \leq K\}, \quad \text{for} \quad 1 \leq j \leq D_\rho.$$

Then it follows that

$$\mathbb{P}(\mathcal{E}_1) = \mathbb{P}\left(\cup_{j=1}^{D_\rho} \mathscr{F}_j\right).$$

Therefore,

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}\left(\cap_{j=1}^{D_\rho}\mathscr{F}_j^c\right)\\ &\leq \mathbb{P}\left(\cap_{j=1}^{L}\mathscr{F}_j^c\right)\\ &= \mathbb{P}(\mathscr{F}_1^c)\left[\prod_{j=2}^{L}\mathbb{P}\left(\mathscr{F}_j^c\,\middle|\,\cap_{\ell=1}^{j-1}\mathscr{F}_\ell^c\right)\right]. \end{aligned} \tag{B.13}$$

First, we bound $\mathbb{P}(\mathscr{F}_1^c)$. Each permutation $\sigma_k, k \neq 1$, maps $t_1 = \{1,\ldots,N-v\}\{N-v+1,\ldots,N\}$ to $\{\sigma_k(1),\ldots,\sigma_k(N-v)\}\{\sigma_k(N-v+1),\ldots,\sigma_k(N)\}$. Therefore, $\sigma_k(t_1) = t_1$ iff σ_k maps set of elements $\{N-v+1,\ldots,N\}$ to the same set of elements. Therefore,

$$\begin{aligned}\mathbb{P}(\sigma_k(t_1) = t_1) &= \frac{1}{\binom{N}{v}}\\ &= \frac{v!}{\prod_{\ell=0}^{v-1}(N-\ell)}.\\ &\leq \frac{v!}{(N-Lv)^v}. \end{aligned} \tag{B.14}$$

Therefore, it follows that

$$\begin{aligned}\mathbb{P}(\mathscr{F}_1^c) &= 1-\mathbb{P}(\mathscr{F}_1)\\ &= 1-\mathbb{P}(\sigma_k(t_1)\neq t_1,\ 2\leq k\leq K)\\ &= 1-\prod_{k=2}^{K}(1-\mathbb{P}(\sigma_k(t_1)=t_1))\\ &\leq 1-\left(1-\frac{v!}{(N-Lv)^v}\right)^K. \end{aligned} \tag{B.15}$$

Next we evaluate $\mathbb{P}\left(\mathscr{F}_j^c\,\middle|\,\cap_{\ell=1}^{j-1}\mathscr{F}_\ell^c\right)$ for $2\leq j\leq L$. Given $\cap_{\ell=1}^{j-1}\mathscr{F}_\ell^c$, we have (at least partial) information about the action of $\sigma_k, 2\leq k\leq K$ over elements $\{N-(j-1)v+1,\ldots,N\}$. Conditional on this, we are interested in the action of σ_k on t_j, i.e. $\{N-jv+1,\ldots,N-jv+v\}$. Specifically, we want to (upper) bound the probability that these elements are mapped to themselves. Given $\cap_{\ell=1}^{j-1}\mathscr{F}_\ell^c$, each σ_k will map $\{N-jv+1,\ldots,N-jv+v\}$ to one of the $\binom{N-(j-1)v}{v}$ possibilities with equal probability. Further, $\{N-jv+1,\ldots,N-jv+v\}$ is not a possibility. Therefore, for

the purpose of upper bound, we obtain that

$$\begin{aligned}\mathbb{P}\left(\mathscr{F}_j^c \,\middle|\, \cap_{\ell=1}^{j-1}\mathscr{F}_\ell^c\right) &\leq 1-\left(1-\frac{1}{\binom{N-(j-1)v}{v}}\right)^{K-1} \\ &\leq 1-\left(1-\frac{v!}{(N-Lv)^v}\right)^{K}. \end{aligned} \tag{B.16}$$

From (B.13)-(B.16), we obtain that

$$\mathbb{P}(\mathcal{E}_1^c) \leq \left[1-\left(1-\frac{v!}{(N-Lv)^v}\right)^{K}\right]^{L}. \tag{B.17}$$

Now $Lv = o(N)$ and hence $N - Lv = N(1-o(1))$. Using $1-x = e^{-x}(1+O(x^2))$ for small $x \in [0,1)$, we have

$$\begin{aligned}&\left(1-\frac{v!}{(N-Lv)^v}\right) \\ =\ &\exp\left(-\frac{v!}{(N-Lv)^v}\right)\left(1+O\left(\frac{1}{N^{2v}}\right)\right). \end{aligned} \tag{B.18}$$

Finally, observe that since $K = O(N^v \log N)$,

$$\left(1+O\left(\frac{1}{N^{2v}}\right)\right)^{K} = \Theta(1).$$

Thus, from (B.17) and (B.18), it follows that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1^c) &\leq \left[1-\Theta\left(\exp\left(-\frac{Kv!}{N^v(1-Lv/N)^v}\right)\right)\right]^{L} \\ &\leq \left[1-\Theta\left(\exp\left(-\frac{(1-2\delta)\log N}{(1-N^{-\delta}v)^v}\right)\right)\right]^{L} \\ &\leq \left[1-\Theta\left(\exp\left(-(1-3\delta/2)\log N\right)\right)\right]^{L} \\ &\leq \left[1-\Theta\left(\frac{1}{N^{1-3\delta/2}}\right)\right]^{L} \\ &\leq \exp\left(-\Omega(LN^{-1+3\delta/2})\right) \\ &\leq \exp\left(-\Omega(N^{\delta/2})\right) \\ &= O\left(\frac{1}{N^{2v}}\right). \end{aligned} \tag{B.19}$$

In above, we have used the fact that $1-x \leq e^{-x}$ for $x \in [0,1]$ and choice of $L = N^{1-\delta}$. This completes the proof of Theorem 10.

B.5 Proof of Theorem 11

So far we have obtained the sharp result that λ satisfies the signature and linear independence conditions up to sparsity essentially $\frac{1}{v!}N^v \log N$ for ρ with $\rho_1 = N - v$ where $v = O(1)$. Now we investigate this further when v scales with N, i.e. $v = \omega(1)$. Let $\rho_1 = N - \eta$ with $\eta \leq N^{\frac{2}{9}-\delta}$ for some $\delta > 0$. For such $\rho = (\rho_1, \ldots, \rho_s)$,

$$\begin{aligned} D_\rho &= \frac{N!}{\prod_{i=1}^s \rho_i!} \\ &\leq \frac{N!}{\rho_1!} \\ &\leq N^{N-\rho_1} = N^\eta. \end{aligned} \tag{B.20}$$

Our interest is in the case when $K \leq (1-\varepsilon) D_\rho \log\log D_\rho$ for any $\varepsilon > 0$. For this, the structure of arguments will be similar to those used in Theorems 9 and 10. Specifically, it will be sufficient to establish that $\mathbb{P}(\mathcal{E}_1^c) = O(1/D_\rho^2)$, where $\mathcal{E}_1$ is the event that permutation $\sigma_1 = \text{id}$ satisfies the signature condition.

To this end, we order the rows (and corresponding columns) of the $D_\rho \times D_\rho$ matrix $M^\rho(\lambda)$ in a specific manner. Specifically, we are interested in the $L = 3N^{\frac{4}{9}-2\delta}\log^3 N$ rows that we call $t_\ell, 1 \leq \ell \leq L$ and they are as follows: the first row, t_1 corresponds to a partition where elements $\{1, \ldots, \rho_1\}$ belong to the first partition and $\{\rho_1+1, \ldots, N\}$ are partitioned into remaining $s-1$ parts of size $\rho_2, \ldots, \rho_s$ in that order. The partition t_2 corresponds to the one in which the first part contains the ρ_1 elements $\{1, \ldots, N-2\eta, N-\eta+1, \ldots, N\}$, while the other $s-1$ parts contain $\{N-2\eta+1, \ldots, N-\eta\}$ in that order. More generally, for $3 \leq \ell \leq L$, t_ℓ contains $\{1, \ldots, N-\ell\eta, N-(\ell-1)\eta+1, \ldots, N\}$ in the first partition and remaining elements $\{N-\ell\eta+1, \ldots, N-(\ell-1)\eta\}$ in the rest of the $s-1$ parts in that order. By our choice of L, $L\eta = o(N)$ and, hence, the above is well defined. Next, we bound $\mathbb{P}(\mathcal{E}_1^c)$ using these L rows.

Now $\sigma_1 = \text{id}$ and hence $\sigma_1(t_i) = t_i$ for all $1 \leq i \leq D_\rho$. Define

$$\mathscr{F}_j = \{\sigma_k(t_j) \neq t_j, \quad \text{for} \quad 2 \leq k \leq K\}, \quad \text{for} \quad 1 \leq j \leq D_\rho.$$

Then it follows that

$$\mathbb{P}(\mathcal{E}_1) = \mathbb{P}\left(\cup_{j=1}^{D_\rho} \mathscr{F}_j\right).$$

Therefore,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}\left(\cap_{j=1}^{D_\rho} \mathscr{F}_j^c\right) \\
&\leq \mathbb{P}\left(\cap_{j=1}^{L} \mathscr{F}_j^c\right) \\
&= \mathbb{P}(\mathscr{F}_1^c)\left[\prod_{j=2}^{L} \mathbb{P}\left(\mathscr{F}_j^c \,\middle|\, \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\right)\right]. \qquad \text{(B.21)}
\end{aligned}$$

First, we bound $\mathbb{P}(\mathscr{F}_1^c)$. Each permutation $\sigma_k, 1 \leq k \leq K$ maps t_1 to one of the D_ρ possible other ρ partitions with equal probability. Therefore, it follows that

$$\mathbb{P}(\sigma_k(t_1) = t_1) = \frac{1}{D_\rho}. \qquad \text{(B.22)}$$

Thus,

$$\begin{aligned}
\mathbb{P}(\mathscr{F}_1^c) &= 1 - \mathbb{P}(\mathscr{F}_1) \\
&= 1 - \mathbb{P}(\sigma_k(t_1) \neq t_1, \ 2 \leq k \leq K) \\
&= 1 - \prod_{k=2}^{K} (1 - \mathbb{P}(\sigma_k(t_1) = t_1)) \\
&= 1 - \left(1 - \frac{1}{D_\rho}\right)^K. \qquad \text{(B.23)}
\end{aligned}$$

Next we evaluate $\mathbb{P}\left(\mathscr{F}_j^c \,\middle|\, \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\right)$ for $2 \leq j \leq L$. Given $\cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c$, we have (at least partial) information about the action of $\sigma_k, 2 \leq k \leq K$ over elements $\{N - (j-1)\eta + 1, \ldots, N\}$. Conditional on this, we are interested in the action of σ_k on t_j. Given the partial information, each of the σ_k will map t_j to one of at least $D_{\rho(j)}$ different options with equal probability for $\rho(j) = (\rho_1 - (j-1)\eta, \rho_2, \ldots, \rho_r)$ – this is because

the elements $1, \ldots, \rho_1 - (j-1)\eta$ in the first part and all elements in the remaining $r-1$ parts are mapped completely randomly conditional on $\cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c$. Therefore, it follows that

$$\mathbb{P}\left(\mathscr{F}_j^c \middle| \cap_{\ell=1}^{j-1} \mathscr{F}_\ell^c\right) \leq 1 - \left(1 - \frac{1}{D_{\rho(j)}}\right)^K. \tag{B.24}$$

From (B.21)-(B.24), we obtain that

$$\begin{aligned} \mathbb{P}(\mathcal{E}_1^c) &\leq \prod_{j=1}^{L} \left[1 - \left(1 - \frac{1}{D_{\rho(j)}}\right)^K\right] \\ &\leq \left[1 - \left(1 - \frac{1}{D_{\rho(L)}}\right)^K\right]^L. \end{aligned} \tag{B.25}$$

In above we have used the fact that

$$D_\rho = D_{\rho(1)} \geq \ldots \geq D_{\rho(L)}.$$

Consider

$$\begin{aligned} \frac{D_{\rho(j)}}{D_{\rho(j+1)}} &= \frac{(N-(j-1)\eta)!\ (\rho_1 - j\eta)!}{(N-j\eta)!\ (\rho_1 - (j-1)\eta)!} \\ &= \prod_{\ell=0}^{\eta-1} \frac{(N-(j-1)\eta-\ell)}{(\rho_1-(j-1)\eta-\ell)} \\ &= \left(\frac{N}{\rho_1}\right)^\eta \prod_{\ell=0}^{\eta-1} \frac{1 - \frac{(j-1)\eta-\ell}{N}}{1 - \frac{(j-1)\eta-\ell}{\rho_1}} \end{aligned} \tag{B.26}$$

Therefore, it follows that

$$\frac{D_{\rho(1)}}{D_{\rho(L)}} = \left(\frac{N}{\rho_1}\right)^{(L-1)\eta} \prod_{\ell=0}^{(L-1)\eta} \frac{1 - \frac{\ell}{N}}{1 - \frac{\ell}{\rho_1}}. \tag{B.27}$$

Using $1 + x \leq e^x$ for any $x \in (-1,1)$, $1 - x \geq e^{-2x}$ for $x \in (0, 1/2)$ and $L\eta = o(N)$,

we have that for any $\ell, 0 \leq \ell \leq (L-1)\eta$

$$\begin{aligned}
\frac{1-\frac{\ell}{N}}{1-\frac{\ell}{\rho_1}} &= \frac{1-\frac{\ell}{N}+\frac{\ell}{\rho_1}-\frac{\ell^2}{N\rho_1}}{1-\frac{\ell^2}{\rho_1^2}} \\
&\leq \exp\left(-\frac{\ell^2-\ell\eta}{N\rho_1}+\frac{2\ell^2}{\rho_1^2}\right) \\
&\leq \exp\left(\frac{\ell\eta}{N\rho_1}+\frac{2\ell^2}{\rho_1^2}\right). \qquad \text{(B.28)}
\end{aligned}$$

Therefore, we obtain

$$\frac{D_{\rho(1)}}{D_{\rho(L)}} \leq \left(\frac{N}{\rho_1}\right)^{L\eta} \exp\left(\Theta\left(\frac{L^2\eta^3}{N\rho_1}+\frac{2L^3\eta^3}{\rho_1^2}\right)\right). \qquad \text{(B.29)}$$

Now

$$\begin{aligned}
\left(\frac{N}{\rho_1}\right)^{L\eta} &= \left(1+\frac{\eta}{\rho_1}\right)^{L\eta} \\
&\leq \exp\left(\frac{L\eta^2}{\rho_1}\right). \qquad \text{(B.30)}
\end{aligned}$$

It can be checked that for given choice of L, η, we have $L\eta^2 = o(\rho_1)$, $L^3\eta^3 = o(\rho_1^2)$ and $L^2\eta^3 = o(N\rho_1)$. Therefore, in summary we have that

$$\frac{D_{\rho(1)}}{D_{\rho(L)}} = 1 + o(1). \qquad \text{(B.31)}$$

Using similar approximations to evaluate the bound on RHS of (B.25) along with

(B.20) yields,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_1^c) &\leq \exp\left(-L\exp\left(-\frac{K}{D_{\rho(L)}}\right)\right) \\
&= \exp\left(-L\exp\left(-(1-\varepsilon)\log\log D_\rho(1+o(1))\right)\right) \\
&\leq \exp\left(-L\exp\left(-\log\log D_\rho\right)\right) \\
&= \exp\left(-\frac{L}{\log D_\rho}\right) \\
&= \exp\left(-\frac{3N^{\frac{4}{9}-2\delta}\log^3 N}{\log D_\rho}\right) \\
&\leq \exp\left(-2\log D_\rho\right) \\
&= \frac{1}{D_\rho^2}. \qquad \text{(B.32)}
\end{aligned}$$

This completes the proof of Theorem 11.

B.6 Proof of Theorem 12

For a general ρ, we establish a bound on the sparsity of λ for which λ satisfies the signature and linear independence conditions with a high probability. Let $\rho = (\rho_1, \ldots, \rho_s), s \geq 2$ with $\rho_1 \geq \cdots \geq \rho_s \geq 1$. As before, let

$$\begin{aligned}
K = \|\lambda\|_0, \quad \text{supp}\,(\lambda) = \{\sigma_k \in S_N : 1 \leq k \leq K\}, \\
\text{and} \quad \lambda(\sigma_k) = p_k, 1 \leq k \leq K.
\end{aligned}$$

Here σ_k and p_k are randomly chosen as per the random model $R(K, \mathscr{C})$ described in Section 3.2. And, we are given partial information $M^\rho(\lambda)$ which is a $D_\rho \times D_\rho$ matrix with

$$D_\rho = \frac{N!}{\prod_{i=1}^s \rho_i!}.$$

Finally, recall the definitions $\alpha = (\alpha_i)_{1 \leq i \leq s}$ with $\alpha_i = \rho_i/n, 1 \leq i \leq s$, and

$$H(\alpha) = -\sum_{i=1}^s \alpha_i \log \alpha_i, \quad \text{and} \quad H'(\alpha) = -\sum_{i=2}^s \alpha_i \log \alpha_i.$$

As usual, it is sufficient to establish that λ satisfies the signature condition because the linear independence condition with a high probability since values p_ks are drawn at random from a finite interval.

For the ease of exposition, we introduce the notion of a ρ-bipartite graph: it is a complete bipartite graph $G^\rho = (V_1^\rho \times V_2^\rho, E^\rho)$ with vertices V_1^ρ, V_2^ρ having a node each for a distinct ρ partition of N and thus $|V_1^\rho| = |V_2^\rho| = D_\rho$. Action of a permutation $\sigma \in S_N$, represented by a 0/1 valued $D_\rho \times D_\rho$ matrix, is equivalent to a perfect matching in G^ρ. In this notation, a permutation σ satisfies the signature condition with respect to a collection of permutations, if and only if there is an edge in the matching corresponding to σ that is not present in any other permutation's matching.

Let $\mathcal{E}_L$ denote the event that $L \geq 2$ permutations chosen uniformly at random satisfy the signature condition. In order to establish the result of Theorem 12, we need to show that $\mathbb{P}(\mathcal{E}_K^c) = o(1)$ as long as $K \leq K_1^*(\rho)$ where $K_1^*(\rho)$ is as defined in (5.3). For that, we bound $\mathbb{P}(\mathcal{E}_{L+1}^c | \mathcal{E}_L)$ for $L \geq 1$. Now consider the bipartite graph, G_L^ρ, which is subgraph of G^ρ, formed by the superimposition of the perfect matchings corresponding to the L random permutations, $\sigma_i, 1 \leq i \leq L$. Now, the probability of $\mathcal{E}_{L+1}^c$ given that $\mathcal{E}_L$ has happened is equal to the probability that a new permutation, generated uniformly at random, has its perfect matching so that all its edges end up overlapping with those of G_L^ρ. Therefore, in order to evaluate this probability we count the number such permutations.

In order to simplify the exposition, we first count the number of such permutations for the cases when $\rho = (N-1, 1)$ and $\rho = (N-2, 2)$. Later, we extend the analysis to a general ρ. As mentioned before, for $\rho = (N-1, 1)$, the corresponding G^ρ is a complete bipartite graph with N nodes on left and right. With a bit of abuse of notation, the left and right vertices be labeled $1, 2, \ldots, N$. Now each permutation, say $\sigma \in S_N$, corresponds to a perfect matching in G^ρ with an edge from left i to right j if and only if $\sigma(i) = j$. Now, consider G_L^ρ, the superimposition of all the perfect matching of the given L permutations. We want to count (or obtain an upper bound on) the number of permutations such that all the edges in their corresponding perfect

matchings overlap with the edges of G_L^ρ. Now, each permutation maps a vertex on left to a vertex on right. In the graph G_L^ρ, each vertex i on the left has degree of at most L. Therefore, if we wish to choose a permutation so that all the edges of its corresponding perfect matching overlap with those of G_ρ^L, it has at most L choices for each vertex on left. There are N vertices in total on left. Therefore, the total number of choices is bounded above by L^N. From this, we conclude that for $\rho = (N-1, 1)$,

$$\mathbb{P}(\mathcal{E}_{L+1}^c | \mathcal{E}_L) \leq \frac{L^N}{N!}.$$

In a similar manner, when $\rho = (N-2, 2)$, the complete bipartite graph G^ρ has $D_\rho = \binom{N}{2}$ nodes on the left and right; each permutation corresponds to a perfect matching in this graph. We label each vertex, on left and right, in G^ρ by unordered pairs $\{i, j\}$, for $1 \leq i < j \leq N$. Again, we wish to bound given $\mathbb{P}(\mathcal{E}_{L+1}^c | \mathcal{E}_L)$. For this, let G_L^ρ, a subgraph of G^ρ, be obtained by the union of edges that belong to the perfect matchings of given L permutations. We would like to count the number possible permutations such that all the edges in their corresponding perfect matchings overlap with the edges of G_L^ρ. For this, we consider the $\lfloor n/2 \rfloor$ pairs $\{1,2\}, \{3,4\}, \ldots, \{2\lfloor n/2 \rfloor - 1, 2\lfloor N/2 \rfloor\}$. Now if N is even then they end up covering all N elements. If not, we consider the last, Nth element, $\{N\}$ as an additional set.

Now, using a similar argument as before, we conclude that there are at most $L^{\lfloor N/2 \rfloor}$ ways of mapping each of these $\lfloor N/2 \rfloor$ pairs such that all of these edges overlap with the edges of G_L^ρ. Note that this mapping fixes what each of these $\lceil N/2 \rceil$ unordered pairs get mapped to. Given this mapping, there are 2! ways of fixing the order in each unordered pair. For example, if an unordered pair $\{i, j\}$ maps to unordered pair $\{k, l\}$ there there are $2! = 2$ options: $i \mapsto k$, $j \mapsto l$ or $i \mapsto l$, $j \mapsto k$. Thus, once we fix the mapping of each of the $\lceil N/2 \rceil$ disjoint unordered pairs, there can be at most $(2!)^{\lceil N/2 \rceil}$ permutations with the given mapping of unordered pairs. Finally, note that once the mapping of these $\lfloor N/2 \rfloor$ pairs is decided, if N is even that there is no element that is left to be mapped. For N odd, since mapping of the $N-1$ elements is decided, so is that of $\{N\}$. Therefore, in summary in both even N or odd N case, there are at

most $L^{\lfloor N/2 \rfloor}(2!)^{\lceil N/2 \rceil}$ permutations that have all of the edge of corresponding perfect matching in G^{ρ} overlapping with the edges of G_L^{ρ}. Therefore,

$$\mathbb{P}(\mathcal{E}_{L+1}^c|\mathcal{E}_L) \leq \frac{L^{\lfloor N/2 \rfloor}(2!)^{\lfloor N/2 \rfloor}}{N!}.$$

Now consider the case of general $\rho = (\rho_1, \rho_2, \ldots, \rho_s)$. Let $T = \lfloor N/(N-\rho_1) \rfloor$ and $J = N - T(N-\rho_1)$. Clearly, $0 \leq J < N - \rho_1$. Now we partition the set $\{1, 2, \ldots, N\}$ into $T+1$ partitions covering all the elements: $\{1, \ldots, N-\rho_1\}, \ldots, \{(N-\rho_1)(T-1)+1, \ldots, (N-\rho_1)T\}$ and $\{(N-\rho_1)T+1, \ldots, N\}$. As before, for the purpose of upper bounding the number of permutations that have corresponding perfect matchings in G^{ρ} overlapping with edges of G_L^{ρ}, each of the first T partitions can be mapped in L different ways; in total at most L^T ways. For each of these mappings, we have options at the most

$$(\rho_2!\rho_3! \ldots \rho_s!)^T.$$

Given the mapping of the first T partitions, the mapping of the J elements of the $T+1$st partition is determined (without ordering). Therefore, the additional choice is at most $J!$. In summary, the total number of permutations can be at most

$$L^T \left(\prod_{i=2}^{s} \rho_i! \right)^T J!.$$

Using this bound, we obtain

$$\mathbb{P}\left(\mathcal{E}_{L+1}^c|\mathcal{E}_L\right) \leq \frac{1}{N!} L^T \left(\prod_{i=2}^{s} \rho_i! \right)^T J!. \tag{B.33}$$

Let,

$$x_L \triangleq \frac{1}{N!} L^T \left(\prod_{i=2}^{s} \rho_i! \right)^T J!.$$

Note that $\mathcal{E}_{k+1} \subset \mathcal{E}_k$ for $k \geq 1$. Therefore, it follows that

$$\begin{aligned} \mathbb{P}(\mathcal{E}_K) &= \mathbb{P}(\mathcal{E}_K \cap \mathcal{E}_{K-1}) \\ &= \mathbb{P}(\mathcal{E}_K|\mathcal{E}_{K-1})\,\mathbb{P}(\mathcal{E}_{K-1}). \end{aligned} \tag{B.34}$$

Recursive application of argument behind (B.34) and fact that $\mathbb{P}(\mathcal{E}_1) = 1$, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_K) &= \mathbb{P}(\mathcal{E}_1) \prod_{L=1}^{K-1} \mathbb{P}(\mathcal{E}_{L+1}|\mathcal{E}_L) \\
&= \prod_{L=1}^{K-1} \left(1 - \mathbb{P}\left(\mathcal{E}_{L+1}^c|\mathcal{E}_L\right)\right) \\
&= \prod_{L=1}^{K-1} (1 - x_L) \\
&\geq 1 - \left(\sum_{L=1}^{K-1} x_L\right). \qquad \text{(B.35)}
\end{aligned}$$

Using (B.33), it follows that $x_{k+1} \geq x_k$ for $k \geq 1$. Therefore,

$$\begin{aligned}
\sum_{L=2}^{K} x_L &\leq K x_K \\
&\leq \frac{1}{N!} K^{T+1} \left(\prod_{i=2}^{s} \rho_i!\right)^T J! \\
&= \frac{1}{N!} K^{T+1} \left(\frac{N!}{\rho_1! D_\rho}\right)^T J! \\
&= \frac{K^{T+1}}{D_\rho^T} \left(\frac{N!}{\rho_1!}\right)^T \frac{J!}{N!} \\
&= \frac{K^{T+1}}{D_\rho^T} \left(\frac{N!}{\rho_1!(N-\rho_1)!}\right)^T \frac{J!((N-\rho_1)!)^T}{N!}. \qquad \text{(B.36)}
\end{aligned}$$

Since $N = J + T(N - \rho_1)$, we have a binomial and a multinomial coefficient in RHS of (B.36). We simplify this expression by obtaining an approximation for a multinomial coefficient through Stirling's approximation. For that, first consider a general multinomial coefficient $v!/(k_1!k_2!\ldots k_\ell!)$ with $v = \sum_i k_i$. Then, using the

Stirling's approximation $\log n! = n \log n - n + 0.5 \log n + O(1)$, for any n, we obtain

$$\begin{aligned} & \log \left(\frac{v!}{k_1! k_2! \ldots k_\ell!} \right) \\ = \; & v \log v - v + 0.5 \log v + O(1) - \\ & \sum_{i=1}^{\ell} (k_i \log k_i - k_i + 0.5 \log k_i + O(1)) \\ = \; & m \sum_{i=1}^{\ell} \frac{k_i}{v} \log \frac{v}{k_i} + 0.5 \log \frac{v}{k_1 k_2 \ldots k_\ell} - O(\ell) \end{aligned}$$

Thus, we can write

$$\begin{aligned} & T \log \frac{N!}{\rho_1!(N-\rho_1)!} \\ = \; & TN\alpha_1 \log \frac{1}{\alpha_1} + TN(1-\alpha_1) \log \frac{1}{1-\alpha_1} \\ & + 0.5 \log \frac{1}{N^T \alpha_1^T (1-\alpha_1)^T} - O(T) \end{aligned} \tag{B.37}$$

where $\alpha_1 = \rho_1/N$. Similarly, we can write

$$\begin{aligned} & \log \frac{N!}{J!((N-\rho_1)!)^T} \\ = \; & N\delta \log \frac{1}{\delta} + TN(1-\alpha_1) \log \frac{1}{1-\alpha_1} \\ & + 0.5 \log \frac{1}{N^T \delta (1-\alpha_1)^T} - O(T) \end{aligned} \tag{B.38}$$

where $\delta = J/N$. It now follows from (B.37) and (B.38) that

$$\begin{aligned} & T \log \frac{N!}{\rho_1!(N-\rho_1)!} - \log \frac{N!}{J!((N-\rho_1)!)^T} \\ = \; & -TN\alpha_1 \log \alpha_1 + \delta N \log \delta \\ & + 0.5 \log \frac{\delta}{\alpha_1^T} + O(T) \end{aligned} \tag{B.39}$$

Since $\delta < 1$, $\delta N \log \delta \leq 0$ and $\log(\delta/\alpha_1^T) \leq -T \log \alpha_1$. Thus, we can write

$$\begin{aligned} & T \log \frac{N!}{\rho_1!(N-\rho_1)!} - \log \frac{N!}{J!((N-\rho_1)!)^T} \\ \leq\ & TN\alpha_1 \log(1/\alpha_1) + O(T \log(1/\alpha_1)) \\ =\ & O(TN\alpha_1 \log(1/\alpha_1)) \end{aligned} \tag{B.40}$$

It now follows from (B.36), (B.39) and (B.40) that

$$\begin{aligned} & \log \left(\sum_{L=2}^{K} x_L \right) \\ \leq & (T+1) \log K - T \log D_\rho + O(TN\alpha_1 \log(1/\alpha_1)) \end{aligned} \tag{B.41}$$

Therefore, for $\mathbb{P}(\mathcal{E}_K) = 1 - o(1)$, a sufficient condition is

$$\begin{aligned} & \log K + \frac{c \log N}{T+1} \\ \leq & \frac{T}{T+1} \log D_\rho - \frac{T}{T+1} O(N\alpha_1 \log(1/\alpha_1)) \end{aligned} \tag{B.42}$$

for some $c > 0$. We now claim that $\log N = O(TN\alpha_1 \log(1/\alpha_1))$. The claim is clearly true for $\alpha_1 \to \theta$ for some $0 < \theta < 1$. Now suppose $\alpha_1 \to 1$. Then, $T \geq 1/(1-\alpha_1) - 1 = \alpha_1/(1-\alpha_1) = x$, say. This implies that $T\alpha_1 \log(1/\alpha_1) \geq \alpha_1 x \log(1 + 1/x) \to 1$ as $\alpha_1 \to 1$. Thus, $TN\alpha_1 \log(1/\alpha_1) = N(1 + o(1))$ for $\alpha_1 \to 1$ as $N \to \infty$. Hence, the claim is true for $\alpha_1 \to 1$ as $N \to \infty$. Finally, consider $\alpha_1 \to 0$ as $N \to \infty$. Note that the function $h(x) = x \log(1/x)$ is increasing on $(0, \epsilon)$ for some $0 < \epsilon < 1$. Thus, for N large enough, $N\alpha_1 \log(1/\alpha_1) \geq \log N$ since $\alpha_1 \geq 1/N$. Since $T \geq 1$, it now follows that $TN\alpha_1 \log(1/\alpha_1) \geq \log N$ for N large enough and $\alpha_1 \to 0$. This establishes the claim.

Since $\log N = O(TN\alpha_1 \log(1/\alpha_1))$, it now follows that (B.42) is implied by

$$\begin{aligned} \log K &\leq \frac{T}{T+1} \log D_\rho - \frac{T}{T+1} O(N\alpha_1 \log(1/\alpha_1)) \\ &= \frac{T}{T+1} \log D_\rho \left[1 - \frac{O(N\alpha_1 \log(1/\alpha_1))}{\log D_\rho} \right] \end{aligned} \tag{B.43}$$

Now consider $D_\rho = N!/(\rho_1!\rho_2!\ldots\rho_s!)$. Then, we claim that for large N

$$\log D_\rho \geq 0.5NH(\alpha). \tag{B.44}$$

In order to see why the claim is true, note that Stirling's approximation suggests,

$$\begin{aligned} \log N! &= N\log N - N + 0.5\log N + O(1), \\ \log \rho_i! &= \rho_i \log \rho_i - \rho_i + 0.5\log \rho_i + O(1). \end{aligned}$$

Therefore,

$$\log D_\rho \geq NH(\alpha) + 0.5\log(N/\rho_1) - \sum_{i=2}^{r} 0.5(O(1) + \log \rho_i).$$

Now consider,

$$\begin{aligned} &\rho_i \log(N/\rho_i) - \log \rho_i - O(1) \\ &= \left(\rho_i - \frac{\log \rho_i}{\log(N/\rho_i)}\right)\log(N/\rho_i) - O(1) \end{aligned} \tag{B.45}$$

Since $\rho_i \leq N/2$ for $i \geq 2$, $\log(N/\rho_i) \geq \log 2$. Thus, the first term in the RHS of (B.45) is non-negative for any $\rho_i \geq 1$. In addition, for every ρ_i, either $\rho_i - \log \rho_i \to \infty$ or $\log(N/\rho_i) \to \infty$ as $N \to \infty$. Therefore, the term on the RHS of (B.45) is asymptotically non-negative. Hence,

$$\log D_\rho \geq 0.5NH(\alpha). \tag{B.46}$$

Thus, it now follows from (B.44) that (B.43) is implied by

$$\log K \leq \frac{T}{T+1}\log D_\rho \left[1 - \frac{O(\alpha_1 \log(1/\alpha_1))}{H(\alpha)}\right].$$

That is, we have the signature condition satisfied as long as

$$K = O\left(D_\rho^{\gamma(\alpha)}\right), \tag{B.47}$$

where

$$\gamma(\alpha) \;=\; \frac{T}{T+1}\left[1 - C'\frac{H(\alpha) - H'(\alpha)}{H(\alpha)}\right], \tag{B.48}$$

and C' is some constant. This completes the proof of theorem.

B.7 Proof of Theorem 13: Limitation on Recovery

In order to make a statement about the inability of *any* algorithm to recover λ from $M^{\rho}(\lambda)$, we rely on the formalism of classical information theory. In particular, we establish a bound on the sparsity of λ beyond which recovery is not asymptotically reliable (precise definition of asymptotic reliability is provided below).

B.7.1 Information theory preliminaries

Here we recall some necessary Information Theory preliminaries. Further details can be found in the book by Cover and Thomas Cover and Thomas [2006].

Consider a discrete random variable X that is uniformly distributed over a finite set $\mathscr{X}$. Let X be *transmitted* over a *noisy* channel to a receiver; suppose the receiver receives a random variable Y, which takes values in a finite set $\mathscr{Y}$. Essentially, such "transmission over noisy channel" setup describes any two random variables X, Y defined through a joint probability distribution over a common probability space.

Now let $\hat{X} = g(Y)$ be an estimation of the transmitted information that the receiver produces based on the observation Y using some function $g : \mathscr{Y} \to \mathscr{X}$. Define probability of error as $p_{\text{err}} = \mathbb{P}(X \neq \hat{X})$. Since X is uniformly distributed over $\mathscr{X}$, it follows that

$$p_{\text{err}} = \frac{1}{|\mathscr{X}|}\sum_{x\in\mathscr{X}} \mathbb{P}(g(Y) \neq x | x). \tag{B.49}$$

Recovery of X is called asymptotically reliable if $p_{\text{err}} \to 0$ as $|\mathscr{X}| \to \infty$. Therefore, in order to show that recovery is not asymptotically reliable, it is sufficient to prove that p_{err} is bounded away from 0 as $|\mathscr{X}| \to \infty$. In order to obtain a lower bound on

p_{err}, we use Fano's inequality:

$$H(X|\hat{X}) \leq 1 + p_{\text{err}} \log|\mathscr{X}|. \tag{B.50}$$

Using (B.50), we can write

$$\begin{aligned} H(X) &= I(X;\hat{X}) + H(X|\hat{X}) \\ &\leq I(X;\hat{X}) + p_{\text{err}} \log|\mathscr{X}| + 1 \\ &\overset{(a)}{\leq} I(X;Y) + p_{\text{err}} \log|\mathscr{X}| + 1 \\ &= H(Y) - H(Y|X) + p_{\text{err}} \log|\mathscr{X}| + 1 \\ &\leq H(Y) + p_{\text{err}} \log|\mathscr{X}| + 1, \end{aligned} \tag{B.51}$$

where we used $H(Y|X) \geq 0$ for a discrete[1] valued random variable. The inequality (a) follows from the data processing inequality: if we have Markov chain $X \to Y \to \hat{X}$, then $I(X;\hat{X}) \leq I(X;Y)$. Since $H(X) = \log|\mathscr{X}|$, from (B.51) we obtain

$$p_{\text{err}} \geq 1 - \frac{H(Y)+1}{\log|\mathscr{X}|}. \tag{B.52}$$

Therefore, to establish that probability of error is bounded away from zero, it is sufficient to show that

$$\frac{H(Y)+1}{\log|\mathscr{X}|} \leq 1 - \delta, \tag{B.53}$$

for any fixed constant $\delta > 0$.

B.7.2 Proof of theorem 13.

Our goal is to show that when K is large enough (in particular, as claimed in the statement of Theorem 13), the probability of error of *any* recovery algorithm is uniformly bounded away from 0. For that, we first fix a recovery algorithm, and then

[1]The counterpart of this inequality for a continuous valued random variable is not true. This led us to study the limitation of recovery algorithm over model $R(K,L)$ rather than $R(K,\mathscr{C})$.

utilize the above setup to show that recovery is not asymptotically reliable when K is large. Specifically, we use (B.53), for which we need to identify random variables X and Y.

To this end, for a given K and L, let λ be generated as per the random model $R(K, L)$. Let random variable X represent the support of function λ i.e., X takes values in $\mathscr{X} = S_N^K$. Given ρ, let $M^\rho(\lambda)$ be the partial information that the recovery algorithm uses to recover λ. Let random variable Y represent $M^\rho(\lambda)$, the $D_\rho \times D_\rho$ matrix. Let $h = h(Y)$ denote the estimate of λ, and $g = g(Y) = \text{supp}\,(h)$ denote the estimate of the support of λ produced by the given recovery algorithm. Then,

$$\begin{aligned} \mathbb{P}\,(h \neq \lambda) &\geq \mathbb{P}\,(\text{supp}\,(h) \neq \text{supp}\,(\lambda)) \\ &= \mathbb{P}\,(g(Y) \neq X)\,. \end{aligned} \tag{B.54}$$

Therefore, in order to uniformly lower bound the probability of error of the recovery algorithm, it is sufficient to lower bound its probability of making an error in recovering the support of λ. Therefore, we focus on

$$p_{\text{err}} = \mathbb{P}\,(g(Y) \neq X)\,.$$

It follows from the discussion in Section B.7.1 that in order to show that p_{err} is uniformly bounded away from 0, it is sufficient to show that for some constant $\delta > 0$

$$\frac{H(Y) + 1}{\log|\mathscr{X}|} \leq 1 - \delta. \tag{B.55}$$

Observe that $|\mathscr{X}| = (N!)^K$. Therefore, using Stirling's approximation, it follows that

$$\log|\mathscr{X}| = (1 + o(1))KN \log N. \tag{B.56}$$

Now $Y = M^\rho(\lambda)$ is a $D_\rho \times D_\rho$ matrix. Let $Y = [Y_{ij}]$ with $Y_{ij}, 1 \leq i, j \leq D_\rho$, taking values in $\{1, \ldots, KL\}$; it is easy to see that $H(Y_{ij}) \leq \log(KL)$. Therefore, it follows

that

$$\begin{aligned} H(Y) &\leq \sum_{i,j=1}^{D_\rho} H(Y_{ij}) \\ &\leq D_\rho^2 \log(KL) = D_\rho^2 (\log K + \log L). \end{aligned} \tag{B.57}$$

For small enough constant $\delta > 0$, it is easy to see that the condition of (B.55) will follow if K satisfies the following two inequalities:

$$\frac{D_\rho^2 \log K}{KN \log N} \leq \frac{1}{3}(1+\delta) \quad \Leftarrow \quad \frac{K}{\log K} \geq \frac{3(1-\delta/2)D_\rho^2}{N \log N}, \tag{B.58}$$

$$\frac{D_\rho^2 \log L}{KN \log N} \leq \frac{1}{3}(1+\delta) \quad \Leftarrow \quad K \geq \frac{3(1-\delta/2)D_\rho^2 \log L}{N \log N}. \tag{B.59}$$

In order to obtain a bound on K from (B.58), consider the following: for large numbers x, y, let $y = (c+\varepsilon)x \log x$, for some constants $c, \varepsilon > 0$. Then, $\log y = \log x + \log\log x + \log(c+\varepsilon)$ which is $(1+o(1))\log x$. Therefore,

$$\frac{y}{\log y} = \frac{c+\varepsilon}{1+o(1)} x \geq cx, \tag{B.60}$$

for $x \to \infty$ and constants $c, \varepsilon > 0$. Also, observe that $y/\log y$ is a non-decreasing function; hence, it follows that for $y \geq (c+\varepsilon)x \log x$, $y/\log y \geq cx$ for large x. Now take $x = \frac{D_\rho^2}{N \log N}$, $c = 3$, $\varepsilon = 1$ and $y = K$. Note that $D_\rho \geq N$ for all ρ of interest; therefore, $x \to \infty$ as $N \to \infty$. Hence, (B.58) is satisfied for the choice of

$$K \geq \frac{4D_\rho^2}{N \log N} \left(\log \frac{D_\rho^2}{N \log N} \right). \tag{B.61}$$

From (B.55), (B.58), (B.59), and (B.61) it follows that the probability of error of any algorithm is at least $\delta > 0$ for N large enough and any ρ if

$$K \geq \frac{4D_\rho^2}{N \log N} \left[\log \left(\frac{D_\rho^2}{N \log N} \vee L \right) \right]. \tag{B.62}$$

This completes the proof of the theorem.

B.8 Proof of Lemma 3

Here we present the proof of Lemma 3. For that, first consider the limit $\alpha_1 \uparrow 1$. Specifically, let $\alpha_1 = 1-\varepsilon$, for a very small positive ε. Then, $\sum_{i=2}^{s} \alpha_i = 1-\alpha_1 = \varepsilon$. By definition, we have $H'(\alpha)/H(\alpha) \leq 1$; therefore, in order to prove that $H'(\alpha)/H(\alpha) \rightarrow 1$ as $\alpha_1 \uparrow 1$, it is sufficient to prove that $H'(\alpha)/H(\alpha) \geq 1 - o(1)$ as $\alpha_1 \uparrow 1$. For that, consider

$$\begin{aligned} \frac{H'(\alpha)}{H(\alpha)} &= \frac{H'(\alpha)}{\alpha_1 \log(1/\alpha_1) + H'(\alpha)} \\ &= 1 - \frac{\alpha_1 \log(1/\alpha_1)}{\alpha_1 \log(1/\alpha_1) + H'(\alpha)}. \end{aligned} \tag{B.63}$$

In order to obtain a lower bound, we minimize $H'(\alpha)/H(\alpha)$ over $\alpha \geq 0$. It follows from (B.63) that, for a given $\alpha_1 = 1-\varepsilon$, $H'(\alpha)/H(\alpha)$ is minimized for the choice of $\alpha_i, i \geq 2$ that minimizes $H'(\alpha)$. Thus, we maximize $\sum_{i=2}^{r} \alpha_i \log \alpha_i$ subject to $\alpha_i \geq 0$ and $\sum_{i=2}^{s} \alpha_i = 1 - \alpha_1 = \varepsilon$. Here we are maximizing a convex function over a convex set. Therefore, maximization is achieved on the boundary of the convex set. That is, the maximum is $\varepsilon \log \varepsilon$; consequently, the minimum value of $H'(\alpha) = \varepsilon \log(1/\varepsilon)$. Therefore, it follows that for $\alpha_1 = 1-\varepsilon$,

$$\begin{aligned} 1 \geq \frac{H'(\alpha)}{H(\alpha)} &\geq 1 - \frac{-(1-\varepsilon)\log(1-\varepsilon)}{\varepsilon \log(1/\varepsilon) - (1-\varepsilon)\log(1-\varepsilon)} \\ &\approx 1 - \frac{\varepsilon}{\varepsilon \log(1/\varepsilon) + \varepsilon} \\ &\approx 1 - \frac{1}{1 + \log(1/\varepsilon)} \\ &\stackrel{\varepsilon \rightarrow 0}{\rightarrow} 1. \end{aligned} \tag{B.64}$$

To prove a similar claim for $\alpha_1 \downarrow 0$, let $\alpha_1 = \varepsilon$ for a small, positive ε. Then, it follows that $r = \Omega(1/\varepsilon)$ since $\sum_{i=1}^{s} \alpha_i = 1$ and $\alpha_1 \geq \alpha_i$ for all $i, 2 \leq i \leq s$. Using a convex maximization based argument similar to the one we used above, it can be checked that $H'(\alpha) = \Omega(\log(1/\varepsilon))$. Therefore, it follows that $\alpha_1 \log(1/\alpha_1)/H'(\alpha) \rightarrow 0$ as $\alpha_1 \downarrow 0$. That is, $H'(\alpha)/H(\alpha) \rightarrow 1$ as $\alpha_1 \downarrow 0$. This completes the proof of Lemma 3.

Bibliography

Who had the 'worst year in washington?'. http://voices.washingtonpost.com/thefix/worst-week-in-washington/worst-year-in-washington.html.

The absolute top 101 songs. http://new.wfnx.com/supplements/2011/topsongs/battle/.

P. Anand. The philosophy of intransitive preference. *The Economic Journal*, pages 337–346, 1993.

S. P. Anderson, A. De Palma, and J. F. Thisse. *Discrete choice theory of product differentiation.* MIT press, Cambridge, MA, 1992.

S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta algorithm and applications. *Manuscript*, 2005.

P. P. Belobaba and C. Hopperstad. Boeing/MIT simulation study: PODS results update. In *1999 AGIFORS Reservations and Yield Management Study Group Symposium, April*, pages 27–30, 1999.

M. E. Ben-Akiva. *Structure of passenger travel demand models.* PhD thesis, Department of Civil Engineering, MIT, 1973.

M. E. Ben-Akiva and S. R. Lerman. *Discrete choice analysis: theory and application to travel demand.* CMIT press, Cambridge, MA, 1985.

R. Beran. Exponential models for directional data. *The Annals of Statistics*, pages 1162–1178, 1979.

R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. pages 798 –805, sep. 2008.

M. Bierlaire. BIOGEME: a free package for the estimation of discrete choice models. In *Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland*, 2003.

M. Bierlaire. An introduction to BIOGEME Version 1.7. 2008.

G. Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman Rev. Ser. A*, 5:147–151, 1946.

H.D. Block and J. Marschak. *Random orderings and stochastic theories of responses.* Contributions to Probability and Statistics, Stanford University Press, Stanford, California., 1960.

J. H. Boyd and R. E. Mellman. The effect of fuel economy standards on the u.s. automotive market: An hedonic demand analysis. *Transportation Research Part A: General*, 14(5-6):367 – 378, 1980.

G. Calafiore and M. C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.

E. J. Candes and J. Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Computational Mathematics*, 6(2): 227–254, 2006.

E. J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006a.

E.J. Candes, J.K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59 (8), 2006b.

N. S. Cardell and F. C. Dunbar. Measuring the societal impacts of automobile downsizing. *Transportation Research Part A: General*, 14(5-6):423 – 434, 1980.

S.R. Chandukala, J. Kim, and T. Otter. *Choice Models in Marketing.* Now Publishers Inc, 2008.

G. Cormode and S. Muthukrishnan. Combinatorial algorithms for compressed sensing. *Lecture Notes in Computer Science*, 4056:280, 2006.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition.* Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, July 2006. ISBN 0471241954. URL `http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20\&path=ASIN/0471241954`.

B.R. Crain. Exponential models, maximum likelihood estimation, and the haar condition. *Journal of the American Statistical Association*, pages 737–740, 1976.

D.E. Critchlow, M.A. Fligner, and J.S. Verducci. Probability models on rankings. *Journal of Mathematical Psychology*, 35(3):294–318, 1991.

G. Darmois. Sur les lois de probabilités a estimation exhaustive, cr acad. *Sc. Paris*, 200:1265–1266, 1935.

G. Debreu. Review of r. d. luce, 'individual choice behavior: A theoretical analysis'. *American Economic Review*, 50:186–188, 1960.

P. Diaconis. *Group representations in probability and statistics.* Institute of Mathematical Statistics Hayward, CA, 1988.

P. Diaconis. A generalization of spectral analysis with application to ranked data. *The Annals of Statistics*, 17(3):949–979, 1989. ISSN 0090-5364.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52 (4):1289–1306, 2006.

J. Edmonds and R.M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM)*, 19(2):248–264, 1972.

V. Farias, S. Jagabathula, and D. Shah. A Data-Driven Approach to Modeling Choice. In *Proceedings of Neural Information Processing Systems (NIPS), Vancouver, Canada*, 2009.

V. Farias, S. Jagabathula, and D. Shah. A nonparametric approach to modeling choice with limited data. *Submitted to Management Science*, 2010.

V. F. Farias, S. Jagabathula, and D. Shah. Inferring Sparse Preference Lists From Partial Information. *ArXiv e-prints*, November 2010.

P.C. Fishburn and S.J. Brams. Paradoxes of preferential voting. *Mathematics Magazine*, 56(4):207–214, 1983. ISSN 0025-570X.

R.A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309, 1922.

R.A. Fisher and L.H.C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge Univ Press, 1928.

R. Gallager. Low-density parity-check codes. *Information Theory, IRE Transactions on*, 8(1):21–28, 1962.

G. Gallego, G. Iyengar, R. Phillips, and A. Dubey. Managing flexible products on a network. Working Paper, 2006.

A. C. Gilbert, M. J. Strauss, J. A. Tropp, and R. Vershynin. One sketch for all: fast algorithms for compressed sensing. In *STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 237–246, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-631-8.

V. Goyal, R. Levi, and D. Segev. Near-optimal algorithms for the assortment planning problem under dynamic substitution and stochastic demand. Submitted, June 2009.

P. M. Guadagni and J. D. C. Little. A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3):203–238, 1983.

E.J. Gumbel. *Statistics of extremes*. Columbia University Press, 1966.

D. A. Hensher and W. H. Greene. The mixed logit model: the state of practice. *Transportation*, 30(2):133–176, 2003.

J. Huang, C. Guestrin, and L. Guibas. Efficient inference for distributions on permutations. *Advances in Neural Information Processing Systems*, 20:697–704, 2008.

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

S. Jagabathula and D. Shah. Inferring rankings under constrained sensing. In *Proceedings of Neural Information Processing Systems (NIPS), Vancouver, Canada*, 2008.

S. Jagabathula and D. Shah. Inferring rankings under constrained sensing. *IEEE Transactions on Information Theory*, October 2011.

A. G. Kök, M. L. Fisher, and R. Vaidyanathan. Assortment planning: Review of literature and industry practice. *Retail Supply Chain Management*, pages 1–55, 2008.

R. Kondor, A. Howard, and T. Jebara. Multi-object tracking with representations of the symmetric group. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

BO Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39(399):10, 1936.

M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi, and D. A. Spielman. Improved low-density parity-check codes using irregular graphs. *IEEE Transactions on Information Theory*, 47(2):585–598, 2001.

R.D. Luce. *Individual choice behavior: A theoretical analysis*. Wiley, New York, 1959.

R.D. Luce. The choice axiom after twenty years. *Journal of Mathematical Psychology*, 15(3):215–233, 1977.

R.D. Luce and P. Suppes. Preference, utility, and subjective probability. *Handbook of mathematical psychology*, 3:249–410, 1965.

S. Mahajan and G. J. van Ryzin. On the relationship between inventory costs and variety benefits in retail assortments. *Management Science*, 45(11):1496–1509, 1999.

CL Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

J.I. Marden. *Analyzing and modeling rank data.* Chapman & Hall/CRC, 1995. ISBN 0412995212.

J. Marschak. Binary choice constraints on random utility indicators. *Cowles Foundation Discussion Papers*, 1959.

V. Marzano and A. Papola. On the covariance structure of the cross-nested logit model. *Transportation Research Part B: Methodological*, 42(2):83 – 98, 2008.

A. Mas-Colell, M. D. Whinston, and J. R. Green. *Microeconomic Theory.* Oxford University Press, 1995.

D. McFadden. Conditional logit analysis of qualitative choice behavior. 1973.

D. McFadden and K. Train. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470, September 2000.

Daniel McFadden. Disaggregate behavioral travel demand's rum side – a 30-year retrospective. In *IN TRAVEL BEHAVIOR RESEARCH: THE LEADING*, pages 17–64. Elsevier, 2000.

E. H. McKinney. Generalized birthday problem. *American Mathematical Monthly*, pages 385–387, 1966.

N. Megiddo. Combinatorial optimization with rational objective functions. *Mathematics of Operations Research*, pages 414–424, 1979.

S. Micali and V. Vazirani. An $o\left(\sqrt{|V||e|}\right)$ algorithm for finding maximum matching in general graphs. In *IEEE FOCS*, 1980.

F. Mosteller and J. Tukey. Data analysis, including statistics. *The Collected Works of John W. Tukey: Philosophy and principles of data analysis, 1965-1986*, page 601, 1987.

S. Muthukrishnan. *Data Streams: Algorithms and Applications.* Foundations and Trends in Theoretical Computer Science. Now Publishers, 2005.

H. Nyquist. Certain topics in telegraph transmission theory. *Proceedings of the IEEE*, 90(2):280–305, 2002.

EJG Pitman. Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 32, pages 567–579. Cambridge Univ Press, 1936.

RL Plackett. The analysis of permutations. *Applied Statistics*, 24(2):193–202, 1975. ISSN 0035-9254.

S.A. Plotkin, D.B. Shmoys, and É. Tardos. Fast approximation algorithms for fractional packing and covering problems. In *IEEE FOCS*, 1991.

R. M. Ratliff, V. Rao, C. P. Narayan, and K. Yellepeddi. A multi-flight recapture heuristic for estimating unconstrained demand from airline bookings. *Journal of Revenue and Pricing Management*, 7(2):153–171, 2008.

I. S. Reed and G. Solomon. Polynomial codes over certain finite fields. *Journal of the Society for Industrial and Applied Mathematics*, pages 300–304, 1960.

P. Rusmevichientong and H. Topaloglu. Technical Note: Robust Logit Assortments. 2009.

P. Rusmevichientong, B. Van Roy, and P. Glynn. A nonparametric approach to multiproduct pricing. *Operations Research*, 54(1), 2006.

P. Rusmevichientong, Z.J. Max Shen, and D.B. Shmoys. A ptas for capacitated sum-of-ratios optimization. *Operations Research Letters*, 37(4):230–238, 2009.

P. Rusmevichientong, Z.J.M. Shen, and D.B. Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research*, 58(6):1666–1680, 2010a.

P. Rusmevichientong, D. Shmoys, and H. Topaloglu. Assortment optimization with mixtures of logits. Technical report, Tech. rep., School of IEOR, Cornell University, 2010b.

D. Saure and A. Zeevi. Optimal dynamic assortment planning. Columbia GSB Working Paper, 2009.

C. E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37 (1):10–21, 1949.

M. Sipser and D. A. Spielman. Expander codes. *IEEE Transactions on Information Theory*, 42:1710–1722, 1996.

K. A. Small. A discrete choice model for ordered alternatives. *Econometrica: Journal of the Econometric Society*, 55(2):409–424, 1987.

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

K. T. Talluri and G. J. van Ryzin. *The Theory and Practice of Revenue Management.* Springer Science+Business Media, 2004b.

L. Thurstone. A law of comparative judgement. *Psychological Reviews*, 34:237–286, 1927.

J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051, 2006.

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

G. J. van Ryzin and G. Vulcano. Computing virtual nesting controls for network revenue management under customer choice behavior. *Manufacturing & Service Operations Management*, 10(3):448–467, 2008.

J. von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. *In Contributions to the theory of games*, 2, 1953.

P. Vovsha. Cross-nested logit model: an application to mode choice in the Tel-Aviv metropolitan area. *Transportation Research Record*, 1607:6–15, 1997.

G. Vulcano, G. van Ryzin, and W. Chaar. Om practice—choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing & Service Operations Management*, 12(3):371–392, 2010.

M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

B. Wierenga. *Handbook of marketing decision models.* Springer Verlag, 2008.

J. I. Yellott. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109 – 144, 1977.

W.T. Ziemba and D.B. Hausch. *Beat the racetrack.* Harcourt Brace Jovanovich, 1984.