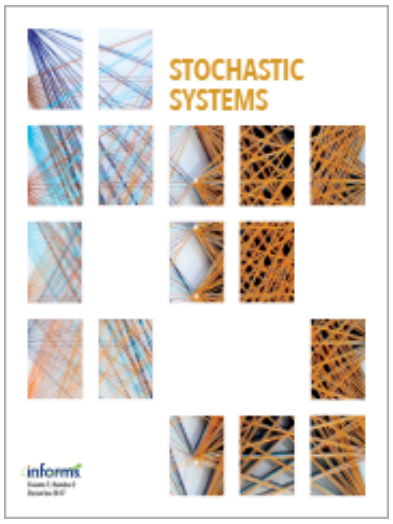
Inferring Sparse Preference Lists from Partial Information

As Published: 10.1287/STSY.2019.0060

Publisher: Institute for Operations Research and the Management Sciences (INFORMS)

Persistent URL: https://hdl.handle.net/1721.1/134414

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context



Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Inferring Sparse Preference Lists from Partial Information

Vivek Farias, Srikanth Jagabathula, Devavrat Shah

To cite this article:
Vivek Farias, Srikanth Jagabathula, Devavrat Shah (2020) Inferring Sparse Preference Lists from Partial Information. Stochastic Systems 10(4):335-360. https://doi.org/10.1287/stsy.2019.0060

http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 10, No. 4, December 2020, pp. 335–360
ISSN 1946-5238 (online)

Inferring Sparse Preference Lists from Partial Information

Vivek Farias,[a] Srikanth Jagabathula,[b] Devavrat Shah[c]

[a] MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; [b] Department of Technology, Operations, and Statistics, New York University Stern School of Business, New York, New York 10012; [c] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Contact: vivekf@mit.edu (VF); sjagabat@stern.nyu.edu, https://orcid.org/0000-0002-4854-3181 (SJ); devavrat@mit.edu (DS)

Received: March 14, 2018
Revised: September 27, 2019
Accepted: April 19, 2020
Published Online in Articles in Advance: October 7, 2020

https://doi.org/10.1287/stsy.2019.0060

Abstract. Probability distributions over rankings are crucial for the modeling and design of a wide range of practical systems. In this work, we pursue a nonparametric approach that seeks to learn a distribution over rankings (aka the ranking model) that is consistent with the observed data and has the sparsest possible support (i.e., the smallest number of rankings with nonzero probability mass). We focus on first-order marginal data, which comprise information on the probability that item i is ranked at position j, for all possible item and position pairs. The observed data may be noisy. Finding the sparsest approximation requires brute force search in the worst case. To address this issue, we restrict our search to, what we dub, the signature family, and show that the sparsest model within the signature family can be found computationally efficiently compared with the brute force approach. We then establish that the signature family provides good approximations to popular ranking model classes, such as the multinomial logit and the exponential family classes, with support size that is small relative to the dimension of the observed data. We test our methods on two data sets: the ranked election data set from the American Psychological Association and the preference ordering data on 10 different sushi varieties.

Funding: Vivek Farias was partially supported by the National Science Foundation [Grant CMMI 1727239]. Srikanth Jagabathula was partially supported by the National Science Foundation [Grant CMMI 1454310]. Devavrat Shah was partially supported by the National Science Foundation [Grant CMMI 1634259, CMMI 1462158, and TRIPODS 1740751].

Keywords: nonparametric models • rankings • sparse • compressive sensing

1. Introduction

1.1. Background

Probability distributions over rankings (or permutations) play a central role in the modeling and the design of a wide range of practical systems, including systems for ranked elections, multiobject tracking, sport and game rankings, prediction markets, and providing personalized recommendations. As an example, consider the system for conducting ranked elections. In a ranked election, the objective is to determine the winning *ranking* of the candidates as opposed to just the winning candidate. In such systems, distributions over rankings can be used to describe the population preferences over the candidates. When there are N candidates contesting the election, there are $N!$ possible rankings of the candidates, and the probability associated with each ranking is interpreted as the fraction of the population for which the particular ranking is the most preferred ranking.

In practice, observed data provide only limited or partial information about the underlying distribution. The challenge then is to estimate the complex distribution, which in the most generality, has $O(N!)$ (or more precisely, $N! - 1$) unknowns, from far fewer data points. To overcome this challenge, one approach is to assume that the distribution obeys a certain *parametric* form with few parameters (see Section 1.2 for details). These parameters can then be estimated from the observed data. Although the parametric approach has found great success in many applications, it suffers from model misspecification issues when the particular assumptions fail to hold in an application at hand.

Another approach is to learn the *simplest* distribution consistent with the observations. This approach is inspired by the ideas of *signal compression* in signal processing, in which large signals, such as image, audio, and video files, are maintained in a compressed form and reconstructed as needed. Such compression, of course, leads to loss in information. However, if the underlying signal has (an approximate) sparse decomposition in an

appropriate basis, then the loss in information is small, and the original signal can be reconstructed from its compressed version with high fidelity; see the discussion on *compressed sensing* in Section 1.3.1.

A definition of simplest distribution or probability mass function (PMF) is that it is well approximated by a PMF with a sparse support. The support of a distribution comprises the rankings with nonzero probability mass, and a PMF with a sparse support puts probability mass on only a few rankings. Assuming that the underlying PMF is indeed sparse, we collect and maintain only partial information about the PMF and reconstruct it as needed. The assumption of sparsity is well motivated in many practical contexts (Donoho 2006, Jagabathula and Shah 2011). For instance, in the context of ranked elections, sparsity assumption implies that the population decomposes into a small number of types, with each type corresponding to a preferred complete ranking of the candidates. Such an observation may be justified by noting that, although there are several candidates, only a few criteria about them determine the preferred rankings of the population.

Motivated by this discussion, we consider the problem of reconstructing the sparsest PMF from first-order partial information. The concrete setup is as follows. Suppose there are N items and a particular problem context is modeled by a PMF λ over the $N!$ permutations or rankings. We collect and maintain partial information about the PMF in the form of the $N \times N$ *first-order* marginal matrix M, where the entry M_{ij} is the probability that item i is ranked at position j. Because λ has about $N!$ degrees of freedom and M has at most N^2 degrees of freedom, multiple PMFs λ are consistent with M. Assuming that the underlying PMF is indeed sparse, we focus on the problem of finding the sparsest PMF λ, the one with the smallest support size, that is consistent with M. Our key objective is to propose a method with strong theoretical guarantees.

The first-order marginals are the most common type of partial information for the ranking PMF (see Jagabathula and Shah 2011 for a discussion on other, more complex, types of partial information). In practice, it is often easy to collect and maintain this information. For instance, in a ranked election setting, where N candidates or options must be rank-ordered by a population of individuals, the population opinion can be elicited by asking each individual from a representative sample if they rank candidate i at position j or not. By contrast, it is cognitively more difficult for an individual to provide a complete ranking of all the N candidates, increasing the chances of mistakes.

Another example comes from the multiobject tracking problem in which goal is track the identities of N objects from noisy measurements of positions and identities. The positions of the objects are tracked by the deployed sensors, but because of measurement noise, when two objects pass by a sensor at the same time, the sensor can measure the positions of the two objects, but may confuse their identities. As a result, one can only maintain a probability distribution over the mappings from the objects to their corresponding positions. Such one-to-one mappings between two sets of the same number of elements are equivalent to rankings, so the resulting probability distribution is an PMF on the $N!$ rankings. Instead of maintaining the entire distribution, one maintains only the first-order partial information and reconstructs the underlying PMF as needed.

Finally, betting markets also serve as sources of the first-order marginals. In horse racing or a ranked election, there are $N!$ possible outcomes. Therefore, a betting mechanism designed to aggregate information about the probabilities of the possible outcomes cannot allow betting on all possible outcomes. A simpler betting mechanism will allow the traders to bet on a collection of outcome pairs such as "candidate i will finish at position j." The resulting bids provide fist-order partial information, from which one wishes to recover the underling PMF over all possible rankings. See Agrawal et al. (2008) for details on such a betting mechanism.

1.2. Related Prior Work

There is a large amount of literature devoted to learning structurally simple ranking PMFs from partial observations. Most prior work has focused on parametric approaches. Given the nature of the topic, the literature is quite diverse, and hence, it is not surprising that the same ranking model appears under different names in different areas. In what follows, we provide a succinct overview of the literature.

1.2.1. Learning Parametric Models. To being with, the monograph by (Diaconis 1988, chapter 9) provides a detailed history of most of the models and references given below. Most of the parametric models proposed in the literature start with the observation that a ranking model is equivalently specified by the order statistics of N random variables $Y_1, \ldots, Y_N$. When items are products and the random variables Y_is are interpreted as utilities assigned to products, the resulting model class is referred to a random utility based model.

In the simplest setting, it is assumed that $Y_i = u_i + X_i$ where the u_i are parameters and the X_i are independent, identically distributed, random variables. Once the distributions of X_is are specified, the model is specified. This class of models was proposed nearly a century ago by Thurstone (1927). A specialization of the

above model when the X_is are assumed to be normal with mean 0 and variance 1 is known as the Thurstone-Mosteller model. This is also known as the *probit* model.

Another specialization of the Thurstone model is realized when X_is are assumed to have Gumbel or logit distributions (one of the extreme value distributions). This model has been credited differently across communities. Holman and Marley established that this model is equivalent (see Yellott (1977) for details) to a generative model where the N objects have positive weights $w_1, \ldots, w_N$ associated with them, and a random permutation of these N objects is generated by recursive selection (without replacement) of objects in the first position, second position and so on with selection probabilities proportional to theirs weights. As per this, the probability of object i being preferred over object j (i.e., object i is ranked higher compared with object j) is $w_i/(w_i + w_j)$. The model in this form is known as the Luce model (Luce, 1959) and the Plackett model (Plackett, 1975).[1] In the context when the partial observations are choice observations (i.e., the observation that an item is chosen from an offered subset of items), this model is called the multinomial logit model (MNL) after McFadden (1973), who called it *conditional logit*; also see Debreu (1960).

The MNL model is of central importance for various reasons. It was introduced by Luce to be consistent with the axiom of *independence from irrelevant alternatives* (IIA). The model was shown to be consistent with the induced preferences assuming a form of random utility maximization framework whose inquiry was started by Marschak (1959) and Marschak and Radner (1972). Very early on, simple statistical tests and simple estimation procedures were developed to fit such a model to observed data (McFadden 1973). The IIA property possessed by the MNL model is not necessarily desirable as evidenced in many empirical scenarios. Despite such structural limitations, the MNL model has been *widely* used across application areas primarily because of the ability to learn the model parameters easily from observed data. For example, see Ben-Akiva and Lerman (1985) and McFadden (1981, 2000) for applications in transportation and Guadagni and Little (1983) and Mahajan and van Ryzin (1999) for applications in operations management and marketing.

With the view to addressing the structural limitations of the MNL model, a number of generalizations to this model have been proposed over the years. Notable among these are the so-called *nested* MNL model and mixtures of MNL models (or MMNL models). These generalizations avoid the IIA property and continue to be consistent with the random utility maximization framework at the expense of increased model complexity; see Ben-Akiva (1973), Ben-Akiva and Lerman (1985), Boyd and Mellman (1980), Cardell and Dunbar (1980), and McFadden and Train (2000) for examples. The interested reader is also referred to an overview article on this line of research by McFadden (2000). Although generalized models of this sort are in principle attractive, their complexity makes them difficult to learn while avoiding the risk of overfitting. More generally, specifying an appropriate parametric model is a difficult task, and the risks associated with mis-specification are costly in practice. For an applied view of these issues, see Bartels et al. (1999), Debreu (1960), and Horowitz (1993). Thus, although these models are potentially valuable in specific well-understood scenarios, the generality of their applicability is questionable.

Most of the development in the utility-based models has occurred in the context when the partial observations are choice observations. Within this context, these models are also called choice models. With a focus on first-order marginals, our work differs from the previous body of work not only in eliminating parametric assumptions but also in the type of the partial observations considered.

As an alternative to the MNL model (and its extensions), one might also consider the parametric family of choice models induced by the exponential family of distributions over permutations. These may be viewed as the models that have maximum entropy among those models that satisfy the constraints imposed by the observed data. The number of parameters in such a model is equal to the number of constraints in the maximum entropy optimization formulation, or equivalently the effective dimension of the underlying data (Koopman 1936, Koopman-Pitman-Darmois theorem). This scaling of the number of parameters with the effective data dimension makes the exponential family obtained via the maximum entropy principle very attractive. Philosophically, this approach imposes on the model only those constraints implied by the observed data. On the flip side, learning the parameters of an exponential family model is a computationally challenging task (Crain 1976, Beran 1979, Wainwright and Jordan 2008) because it requires computing a *partition function* possibly over a complex state space.

1.2.2. Learning Nonparametric Models. Parametric models either impose strong restrictions on the structure of the model and/or are computationally challenging to learn. To overcome these limitations, we consider a nonparametric approach to learning the model from the observed first-order marginals.

The given partial data most certainly does not completely identify the underlying model. There are multiple models that are (near) consistent with the given observations, and we need an appropriate model selection criterion. In the parametric approach, one uses the imposed parametric structure as the model selection criterion.

On the other hand, in the nonparametric approach considered in this paper, we use *simplicity*, or more precisely *sparsity* (support size of the distribution), as the criterion for selection: specifically, we select the *sparsest* model or the distribution with the smallest support size from the set of models that are (near) consistent with the observations. This nonparametric approach was first proposed by Jagabathula and Shah (2008, 2011) and developed further by Farias et al. (2009, 2013). Following Farias et al. (2009, 2013) and Jagabathula and Shah (2008, 2011), we restrict ourselves to observations that are in the form of *first-order* marginals.

A major issue with the identification of sparse models from marginal information is the associated computational cost. Recovering a distribution over permutations of N objects, in principle, requires identifying probabilities of $N!$ distinct permutations. The first-order marginal information can be thought of as a linear projection of the ranking model on the $(N-1)^2$ dimensional space. Thus, finding a sparse model consistent with the observations is equivalent to solving a severely under-determined system of linear equations with $N!$ variables, with the aim of finding a sparse solution. As a result, at a first glance, it appears that the computational complexity of *any* procedure should scale with the dimension of the variable space, $N!$.

In Farias et al. (2009, 2013) and Jagabathula and Shah (2008, 2011), the authors identified the so-called *signature condition* on the space of ranking models and showed that whenever a model satisfies the signature condition and *noiseless* marginal data are available, it can be *exactly* recovered in an efficient manner from marginal data with computation cost that scales linearly in the dimension of the marginal data ($(N-1)^2$ for first-order marginals) and exponentially in the sparsity of the model. Indeed, for sparse models, this is excellent. The authors also established that the signature condition is necessary for the recovery of the underlying model, in that if the underlying model does not satisfy the signature condition, then the sparsest model consistent with the observed data are not unique, resulting in nonidentification. Finally, they also established that the *signature family* (the family of ranking models satisfying the signature condition) is *dense* in the space of sparse models: a randomly chosen model with a *small enough* sparsity (support size) satisfies the signature condition with a high probability. The precise sparsity scaling depends on the type of marginal data available; for instance, for the first-order marginals, the authors show that a randomly generated model with sparsity up to $O(N \log N)$ satisfies the signature conditions. In summary, the works of Farias et al. (2009, 2013) and Jagabathula and Shah (2008, 2011) provide a complete characterization of the *exact* recovery of underlying sparse models from *noise-free* partial observations: the true model can be recovered if and only if it belongs to the signature family, and if does belong to the signature family, then it can be recovered efficiently.

1.3. Our Contributions

The main contribution of this paper is to extend prior works (Jagabathula and Shah 2008, 2011; Farias et al. 2009, 2013) to the settings where (a) the available data are not noise free or (b) the underlying model is not sparse. These settings are, of course, closer to practice. Specifically, we consider the problem of finding the sparsest model that is near consistent with—or equivalently, within a distance ε of—the first-order marginal data. We discuss how our methods and results extend to general types of marginal information at the end.

Our objective is to provide strong recovery guarantees and design computationally efficient algorithms. Specifically, we consider the following questions: (a) if the underlying model is K sparse and belongs to the signature family, how can we recover it from noisy first-order information in an efficient manner? (b) If the underlying model does not belong to the signature family and is perhaps even dense, how good is a sparse approximation from the signature family?

The first question pertains to efficient recovery of a sparse model from the signature family when there is noise in the observations. Without any restrictions, a brute-force exhaustive search is the only possible way to recover the sparsest model. Such an exhaustive search has $\binom{N!}{K} \approx \exp(\Theta(KN \log N))$ computational complexity, scaling exponentially in N. Furthermore, prior work (Jagabathula and Shah 2008, 2011; Farias et al. 2009, 2013) has shown that, even in noise-free settings, it is necessary for the ranking model to belong to the signature family for efficient recovery to be possible, and the signature family is dense in the class of sparse models. Motivated by these reasons, we restrict our search to the signature family of distributions. Although existing work has proposed algorithms to recover ranking models from the signature family in noise-free settings, these algorithms do not extend to noisy settings. In fact, it is a priori unclear if even the signature family lends itself to an efficient search. We show that the signature family *can* indeed be searched efficiently.

To elaborate on our result, note that the space of first-order marginal information is equivalent to the space of doubly stochastic matrices (this equivalence is explained in detail in subsequent sections). Given this, finding the sparsest model in the signature family that is near consistent with the given first-order marginal information is equivalent to determining the convex decompositions (in terms of permutations) of all doubly stochastic matrices that are within a ball of small radius around the given observation matrix and choosing

the sparsest convex decomposition. It follows from Birkhoff-von Neumann's celebrated result (Birkhoff 1946, von Neumann 1953) that a doubly stochastic matrix belongs to an $(N-1)^2$ dimensional polytope, also called the Birkhoff polytope, with the permutation matrices as the extreme points and that there is an efficient description of the polytope.

By exploiting the structures of the Birkhoff polytope and the signature family and adapting the multiplicative weights update framework of Plotkin-Shmoys-Tardos (Plotkin et al. 1995), we are able to devise a novel algorithm that reduces the problem of finding a K-sparse model within the signature family that is near consistent with the given first-order marginal information into solving a sequence of linear programs, each of which can be efficiently solved. The algorithm is described in detail in the proof of Theorem 1. Our algorithm finds a model with sparsity at most $O(\varepsilon^{-2}K\log N)$ in the signature family in time $\exp(\Theta(K\log N))$ if there exists a model in signature family with sparsity K that approximates the data well; here, ε is the approximation error.

The complexity $\exp(\Theta(K\log N))$ is exponentially smaller than the brute-force $\exp(\Theta(KN\log N))$ complexity noted previously. This result is similar in spirit to many of the recently developed sparse model learning methods under the framework of *compressed sensing*, which establish that a sparse model can be learned efficiently by solving a convex program.

The second question pertains to the generality of the signature family in the space of all ranking models. In particular, is there a *sparse enough* model in the signature family that is near consistent with the marginal information from a broad class of models? We establish that for a very large class of models, including very dense models such as the models from the MNL or exponential family, there exists a model of sparsity $O(N/\varepsilon^2)$ in the signature family that is within ε (in ℓ_2 norm) of the first-order marginals generated by the underlying model (see Theorem 2).

To put the sparsity bound of N/ε^2 in perspective, our problem involves obtaining the sparsest convex decomposition of a point in the Birkhoff polytope. It follows from Caratheodory's theorem 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 $(N-1)^2+1=\Theta(N^2)$. Of course, the sparsity bound of $O(N/\varepsilon^2)$ is significantly smaller than $\Theta(N^2)$. The reason is that if we allow for an ε error, we can shave off a factor N from the sparsity bound. Specifically, we establish that in as much as first-order marginals are concerned, *any* ranking model can be ε-approximated by a model (*not necessarily in the signature family*) with sparsity or support size $O(N/\varepsilon^2)$. More precisely, we show that any nonnegative valued doubly stochastic matrix can be ε-approximated in the ℓ_2 sense by a convex combination of $O(N/\varepsilon^2)$ permutation matrices (Theorem 3). In such generality, we show that the bound $O(N/\varepsilon^2)$ is tight (in the scaling of N) by exhibiting a doubly stochastic matrix that requires $\Omega(N)$ sparsity to guarantee ε-error.

Combining the results of Theorems 2 and 3, we note that even with the restriction to the signature family, the sparsity scaling is still $O(N/\varepsilon^2)$, implying that we are not losing much as far as the worst-case sparsity scaling is concerned. We also argue in Section 3 after the statement of Theorem 3 that picking an arbitrary solution consistent with the observations most certainly will result in $\Theta(N^2)$ sparsity, which is significantly suboptimal with respect to the optimal sparsity of $O(N/\varepsilon^2)$.

We establish the effectiveness of our approach by applying our sparse model learning procedure to the well-studied ranked election data of the American Psychological Association (APA) (i.e., the data used by in Diaconis 1989). Interestingly enough, through sparse model approximation of the election data, we find structural information in the data similar to that unearthed by Diaconis (1989). The basic premise in Diaconis (1989) was that by looking at linear projections of the ranked election votes, it may be possible to unearth hidden structure in the data. Our sparse approximation captures similar structural information from projected data suggesting the utility of this approach in unearthing nonobvious structural information.

We also test the decision accuracy of the sparse model on the classical *assortment optimization* problem, which deals with the decision of determining the revenue maximizing subset of products from a universe of N products. Demand is modeled in these settings using a distribution λ over the rankings of the N products. Each ranking denotes a preference ordering over the products. When a customer is offered a subset S of products, she is assumed to sample a ranking σ from λ and choose the most preferred offered product. We compare the revenue under the optimal assortment and that under the assortment computed using the sparse distribution. We observe that the loss in revenues from using the sparse distribution is no more than 3.3% on average.

1.3.1. Thematic Relation: Compressed Sensing. This work is thematically related to the recently developed theory of compressed sensing and streaming algorithms and further to classical coding theory and signal processing (Shannon 1949, Nyquist 2002). In the compressive sensing literature (Candes and Tao 2005;

Candes and Romberg 2006; Candes et al. 2006a, b; Donoho 2006), the goal is to estimate a *signal* by means of a minimal number of measurements. Operationally this is equivalent to finding the sparsest signal consistent with the observed (linear) measurements. In the context of coding theory, this corresponds to finding the most likely transmitted code word given the received message (Reed and Solomon 1960, Gallager 1962, Sipser and Spielman 1996, Luby et al. 2001). In the context of streaming algorithms, this task corresponds to maintaining a minimal data structure to implement algorithmic operations (Tropp, 2004, 2006; Cormode and Muthukrishnan 2006; Gilbert et al. 2007; Berinde et al. 2008). Despite the thematic similarity, existing approaches to compressive sensing are ill-suited to the problem at hand; see Jagabathula and Shah (2011), where the authors establish that the generic *restricted null space* condition—a necessary and sufficient condition for the success of the convex optimization in finding sparsest possible solution—is not useful in the setting considered here. In a nutshell, the projections of the signal we observe are a given as opposed to being a design choice. Put another way, the present paper can be viewed as providing a nontrivial extension to the theory of compressive sensing for the problem of efficiently learning distributions over permutations.

1.4. Organization

The rest of the paper is organized as follows. In Section 2, the precise problem statement along with the signature condition is introduced. We also introduce the MNL and exponential family parametric models. The main results of this paper are stated in Section 3. These results are established in Section 4. In Section 5, we study the application of our results to the popular APA and sushi data sets. Using a simple heuristic motivated by the signature condition, we learn sparse approximations of the observed data. We show that the sparse approximations can capture nontrivial structure present in the data and can also lead to accurate decisions. Finally, we conclude in Section 6 with a discussion on how the methods we propose for first-order marginal data extend to general types of marginal data.

2. Setup

Given N objects or items, we are interested in a ranking model or distribution over permutations of these N items. Let S_N denote the space of $N!$ permutations of these N items. A ranking model (equivalently, a distribution over S_N) can then be represented as a vector of $N!$ dimension with nonnegative components, all of them summing up to 1. The observations we consider here are the marginal distributions of the choice model. Specifically, throughout this paper, we primarily restrict ourselves to *first-order* marginal information. We point out how our results extend to general marginal information in the discussion (Section 6).

More precisely, let λ denote a ranking model or a distribution over S_N. Then, the first-order marginal information, $M(\lambda) = [M_{ij}(\lambda)]$, is an $N \times N$ doubly stochastic matrix with nonnegative entries defined as

$$M_{ij}(\lambda) = \sum_{\sigma \in S_N} \lambda(\sigma) \mathbf{1}_{\{\sigma(i)=j\}},$$

where $\sigma \in S_N$ represents a permutation, $\sigma(i)$ denotes the rank of item i under permutation σ, and $\mathbf{1}_{\{x\}}$ is the standard indicator with $\mathbf{1}_{\{\text{true}\}} = \mathbf{1}$ and $\mathbf{1}_{\{\text{false}\}} = \mathbf{0}$.

We assume that there is a *ground-truth* model λ and the observations are a noisy version of M. Specifically, let the observations be $D = M + \eta$ so that $\|\eta\|_2 \le \delta$ for some small enough $\delta > 0$, where $\|\eta\|_2$ denotes the Frobenius norm

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

Without loss of generality, we assume that D is also doubly stochastic (or else, it is possible to transform it into that form). The goal is to learn a distribution $\hat{\lambda}$ over S_N so that it's first-order marginal $M(\hat{\lambda})$ approximates D (and hence M) well and the support of $\hat{\lambda}$, $\|\hat{\lambda}\|_0$ is small. Here

$$\|\hat{\lambda}\|_0 \triangleq \left|\{\sigma \in S_N : \hat{\lambda}(\sigma) > 0\}\right|.$$

Indeed, one way to find such $\hat{\lambda}$ is to solve the following program: for a choice of approximation error $\varepsilon > 0$,

$$\begin{aligned} &\text{minimize} \quad \|\mu\|_0 \quad \text{over} \quad \text{ranking models} \ \mu \\ &\text{such that} \quad \|M(\mu) - D\|_2 \le \varepsilon. \end{aligned} \tag{1}$$

Suppose the sparsest solution has sparsity K. Then, as noted previously, a brute-force exhaustive search is the only possible way to recover the sparsest model when there are no additional restrictions, and such an

approach has $\binom{N!}{K} \approx \exp(\Theta(KN \log N))$ computational complexity, scaling exponentially in N. The question then is whether this could be improved on significantly.

2.1. Question 1

Is it possible to solve (1) with a running time complexity that is far better than $O(\exp(KN \log N))$, at least for a reasonable large class of observations D?

We obtain a faster algorithm by restricting our search to models that belong to the signature family that was introduced in earlier work (Jagabathula and Shah 2008, 2011; Farias et al. 2009, 2013), defined as follows:

2.1.1. Signature Family. A distribution 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 j.

The above definition states that each element in the support of λ has its signature in the data. In other words, we solve the following program instead:

$$\begin{array}{lll} \text{minimize} & \|\mu\|_0 & \text{over} \quad \text{ranking models in signature family } \mu \\ \text{such that} & \|M(\mu) - D\|_2 \leq \varepsilon. & \end{array} \tag{2}$$

The structure of the family allows for efficient search. In particular, we show that the program (2) can be solved approximately with computational complexity scaling polynomially in N but exponentially in K, where K is the optimal value of (2).

Given that we can efficiently search the signature family for sparse solutions, the next natural question is how general the signature family is.

2.2. Question 2

Is there a sparse enough model in the signature family that is near consistent with the marginal information from a broad class of ranking models?

We answer this question by showing that when the underlying ranking model is the MNL model or the exponential family of models, then there exists a ranking model of sparsity $O(N/\varepsilon^2)$ in the signature family that is within ε (in ℓ_2 norm) of the first-order marginals generated by the underlying model. We now describe the two parametric models, MNL and exponential family, before we describe our answers to the previous questions.

2.2.1. MNL Model. Here we describe the version of the model as introduced by Luce (1959) and Plackett (1975). This is a parametric model with N positive valued parameters, one each associated with each of the N items. Let $w_i > 0$ be parameter associated with item i. Then the probability of permutation $\sigma \in S_N$ is given by (Marden 1995)

$$\mathbb{P}_w(\sigma) = \prod_{j=1}^{N} \frac{w_{\sigma^{-1}(j)}}{w_{\sigma^{-1}(j)} + w_{\sigma^{-1}(j+1)} + \cdots + w_{\sigma^{-1}(N)}}. \tag{3}$$

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

2.2.2 Exponential Family Model. Now we describe an exponential family of distributions over permutations. The exponential family is parameterized by N^2 parameters θ_{ij} for $1 \leq i, j \leq N$. Given such a vector of parameters θ, the probability of a permutation σ is given by

$$\begin{aligned} \mathbb{P}_\theta(\sigma) &\propto \exp\left(\sum_{1 \leq i,j \leq N} \theta_{ij}\sigma_{ij}\right), \\ &= \frac{1}{Z(\theta)} \exp\left(\sum_{1 \leq i,j \leq N} \theta_{ij}\sigma_{ij}\right), \end{aligned} \tag{4}$$

where $Z(\theta) = \sum_{\sigma \in S_N} \exp(\sum_{1 \leq i,j \leq N} \theta_{ij}\sigma_{ij})$; $\sigma_{ij} = 1$ if and only if (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 previously described 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, Wainwright and Jordan (2008) for details on this correspondence between parameters of exponential family and its marginals.

3. Main Results

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

3.1. On Question 1 (Efficient Algorithm to Solve (1))

We show that by restricting ourselves to the signature family, we can improve the running time complexity from $O(\exp(\Theta(KN\log N))$ (when using the brute-force algorithm) to $O(\exp(\Theta(K\log N)))$—effectively shaving off a factor of N from the exponent—to obtain an approximate solution. More precisely, we can establish the following result.

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

The proof of Theorem 1 is constructive in the sense that it proposes an algorithm to find a sparse model with the stated guarantees. The algorithm uses structural properties of the Birkhoff polytope and the models in the signature family along with an adaptation of the multiplicative weights update framework of Plotkin-Shmoys-Tardos (Plotkin et al. 1995) to reduce (2) into a sequence of linear programs, each of which can be solved efficiently. The result of Theorem 1 establishes that as long as there is a model of sparsity K in the signature family that is an ε-fit to the observations D, we can shave off a factor of N in the exponent from the running time complexity at the cost of introducing a factor of $\log N$ in the sparsity.

The theorem applies to any combination of K and ε that satisfy the conditions of the theorem. Therefore, the computational effort expended can be controlled by running the algorithm with a smaller value of K, but at the expense of a worse approximation guarantee ε.

3.2. On Question 2 (Generality of the Signature Family)

The guarantee of Theorem 1 is contingent on the existence of a sparse 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 D, there are two possibilities. First, 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. Second, 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 previously 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 an exponential family model.

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

1. *either λ is an MNL model with parameters $w_1, \ldots, w_N$ (and without loss of generality $w_1 \le w_2 \le \ldots \le w_N$) such thatfor $L = N^\delta$ for some $\delta \in (0, 1)$;*

$$\frac{w_N}{\sum_{k=1}^{N-L} w_k} \le \frac{\sqrt{\log N}}{N}, \tag{5}$$

2. *or λ is an exponential family model 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 $1 \le i_k, j_k \le N$ for $1 \le k \le 4$)*

$$\frac{\exp(\theta_{i_1 j_1} + \theta_{i_2 j_2})}{\exp(\theta_{i_3 j_3} + \theta_{i_4 j_4})} \le \sqrt{\log N}. \tag{6}$$

Then, there exists a $\hat{\lambda}$ in the signature family such that: $\|D - M(\hat{\lambda})\|_2 \le 2\varepsilon$ and $\|\hat{\lambda}\|_0 = O(N/\varepsilon^2)$.

Remark. The conditions (5) and (6) 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 2 that (under appropriate conditions) the rich families of MNL and exponential models can be approximated by sparse models in signature families as far as first-order marginals are concerned. Both families induce distributions that have 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.

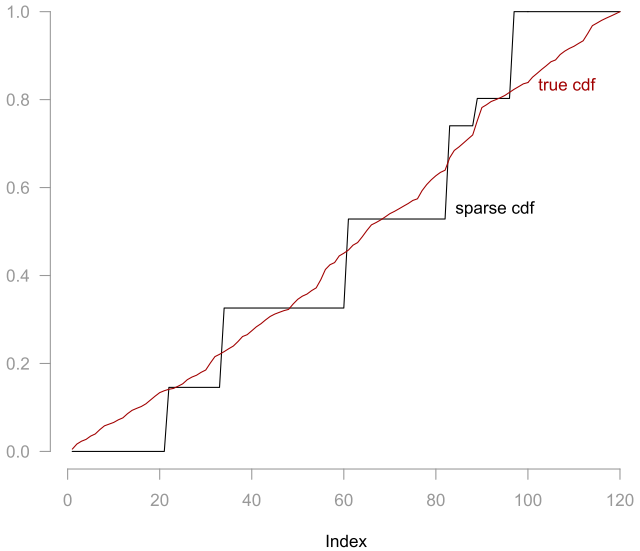
The conditions (5) and (6) imply that for dense distributions under the MNL and the exponential family models to be approximated by a sparse distribution in the signature family, the dense distribution cannot be too *spiky* and must be *close* to the uniform distribution. Intuitively, we need such a condition because of the following reason. The permutations in the support of a distribution belonging to the signature family must each have a signature edge, for which the permutations should all be well separated or spread out. Therefore, a member of the signature family will be able to approximate the underlying dense distribution only if the dense distribution is also spread out or, equivalently, if the dense distribution is close to a uniform distribution.

To understand these conditions better, let us focus on condition (5). It can be seen that for this condition to be satisfied, it is sufficient for $w_N/w_1 \le \sqrt{\log N}$. That is, the largest and smallest attraction values are no more than a factor of $\sqrt{\log N}$ apart. This condition is satisfied in settings where there is sufficient heterogeneity in population preferences such that each item is preferred by a nontrivial fraction of the population. Such a condition will be satisfied, for instance, in a market with horizontally differentiated products (say, different flavors of the same product) or when we restrict our attention to a subset of popular items. As an example, in Section 5, we discuss the APA ranked election setting in which the population votes for five candidates. It can be seen from Figure 1 that there is no clear consensus among the population on any particular rank ordering of candidates and the distribution is very spread out.

To put the sparsity bound $O(N/\varepsilon^2)$ in context, we establish that given *any* doubly stochastic matrix D and $\varepsilon > 0$, there exists a model λ with sparsity $O(N/\varepsilon^2)$ such that $\|M(\lambda) - D\|_2 \le \varepsilon$. 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. More precisely, we have the following theorem.

Theorem 3. *For any doubly stochastic matrix D and $\varepsilon \in (0,1)$, there exists a model λ such that $\|\lambda\|_0 \le N/\varepsilon^2$ and $\|M(\lambda) - D\|_2 \le \varepsilon$.*

Figure 1. Comparison of the CDFs of the True Distribution and the Sparse Approximation We Obtain for the APA Data Set



Note. 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.

We emphasize here that this result holds for *any* doubly stochastic matrix D. 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 D with all of its entries equal to $1/N$. Then, any model λ with $o(N)$ support can have at most $N * o(N) = o(N^2)$ nonzero entries, which in turn means that the ℓ_2 error $\|M(\lambda) - D\|_2$ is at least $\sqrt{(N^2 - o(N^2))/N^2} \approx 1$ for large N.

In the light of Theorem 3, the sparsity scaling of $O(N/\varepsilon^2)$ that we established in Theorem 1 implies that we are not losing much in terms of worst-case sparsity scaling by the restriction to the signature family.

The result of Theorem 3 also justifies why picking an arbitrary model (by solving the convex relaxation, obtained by replacing the ℓ_0-norm by the ℓ_1-norm, of the program in (1)) can be significantly suboptimal for our problem. Specifically, suppose we are given a doubly stochastic matrix D and a tolerance parameter $\varepsilon > 0$. Then, all the consistent models λ, which satisfy $\|M(\lambda) - D\|_2 \le \varepsilon$, 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 Farias et al. (2013, theorem 1), we can show that the set of doubly stochastic matrices $\tilde{D}$ such that $\|\tilde{D} - D\|_2 \le \varepsilon$ and can be written as $M(\lambda) = \tilde{D}$ 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).

We prove Theorem 3 in Section 4 using the probabilistic method. An alternate way to obtain a sparsity bound is to use the Frank-Wolfe (FW) theory (see Jaggi 2013 and references therein). The FW algorithm is an iterative algorithm designed for constrained convex optimization. In each iteration, the algorithm linearizes the convex objective at the current iterate and then moves toward the minimizer of the linear approximation over the constraint space. To apply the FW theory, we can obtain an approximate convex decomposition of a doubly stochastic matrix D by solving the following convex program

$$\begin{aligned} &\text{minimize} \quad \|M - D\|_2^2 \quad \text{over } M \\ &\text{such that} \quad M \in \text{ Birkhoff polytope.} \end{aligned} \tag{7}$$

Because the Birkhoff polytope has an efficient description, this convex program can be solved efficiently using the FW algorithm. In each iteration, the FW algorithm adds at most one permutation to the solution, because, after K iterations, the algorithm finds a solution with sparsity at most K. (Jaggi 2013, theorem 1) shows that the FW algorithm obtains an ε approximation in $\|\cdot\|_2$ norm in C/ε^2 iterations, where C is the curvature constant. When the objective function is $\|\cdot\|_2^2$, the curvature constant can be shown to be equal to the twice of the diameter of the Birkhoff polytope:

$$\begin{aligned} C = \text{maximize} \quad & 2\|X - Y\|_2^2 \quad \text{over } X, Y \\ \text{such that} \quad & X, Y \text{ are doubly stochastic.} \end{aligned}$$

Because $|X_{ij} - Y_{ij}| \le 1$ for all $1 \le i, j \le N$, a trivial upper bound for C is $2N^2$, which gives a $O(N^2/\varepsilon^2)$ sparsity bound. Clearly, this is a factor N off from the bound of $O(N/\varepsilon^2)$ we derived in Theorem 3. A more careful argument provides us with a tighter bound. Note that

$$\|X - Y\|_2^2 = \sum_{1 \le i,j \le N} (X_{ij} - Y_{ij})^2 \le \sum_{1 \le i,j \le N} |X_{ij} - Y_{ij}| \le \sum_{1 \le i,j \le N} (X_{ij} + Y_{ij}) = 2N,$$

where the first inequality follows because $0 \le |X_{ij} - Y_{ij}| \le 1$ for all $1 \le i, j \le N$, the second inequality follows from the triangle inequality for $|\cdot|$, and the last equality from $\sum_{i,j} X_{ij} = \sum_{ij} Y_{ij} = N$. We thus obtain the upper bound $C \le 4N$, resulting in the sparsity guarantee $4N/\varepsilon^2$, which is a factor 4 off the guarantee we obtain in Theorem 3.

The FW theory does not provide any tightness guarantee for the sparsity bound of $4N/\varepsilon^2$. However, as we argued above, $O(N/\varepsilon^2)$ sparsity bound is indeed tight. Moreover, solving the program in (7) can provide a significantly suboptimal solution to (1), especially when the optimal sparsity is $\ll N$.

Finally, the guarantee of Theorem 1 cannot be obtained by applying the FW algorithm to solve the problem in (7) by replacing the objective function with $\|M - D\|_\infty$ because the curvature constant for the $\|\cdot\|_\infty$ norm is unbounded.

In the next sections, we present the proofs of Theorems 1–3 before we present the results of our empirical study.

4. Proofs

4.1. Proof of Theorem 1

We are given a doubly stochastic observation matrix D. Suppose there exists a model μ in the signature family such that $\|\mu\|_0 = K$ and $\|M(\mu) - D\|_2 \le \varepsilon$. Then, the algorithm we describe finds a model $\hat{\lambda}$ such that $\|\hat{\lambda}\|_0 = O(\varepsilon^{-2}K \log N)$ and $\|M(\hat{\lambda}) - D\|_\infty \le 2\varepsilon$ in time $\exp(\Theta(K \log N))$. The algorithm proposes a way to search over the signature family efficiently. Before we describe the algorithm, we introduce a representation of the models in the signature family, which allows us to reduce the problem into solving a collection of linear programs (LPs).

4.1.1. Representation of the Signature Family. We start by developing a representation of models from the signature family that is based on their first order marginal information. All the relevant variables are represented as vectors in N^2 dimension. For example, the data matrix $D = [D_{ij}]$ is represented as an N^2 dimensional vector with components indexed by tuples for the ease of exposition: D_{ij} will be denoted as $D_{(i,j)}$ and the dimensions will be ordered as per the lexicographic ordering of the tuples, that is, $(i,j) < (i',j')$ iff $i < i'$ or $i = i'$ and $j < j'$. Therefore, D in a column vector form is

$$D = \left[D_{(1,1)}\ D_{(1,2)} \dots D_{(1,N)}\ D_{(2,1)} \dots D_{(N,N)}\right]^T.$$

In a similar manner, we represent a permutation $\sigma \in S_N$ as a 0–1 valued N^2 dimensional vector as $\sigma = [\sigma_{(i,j)}]$ with $\sigma_{(i,j)} = 1$ if $\sigma(i) = j$ and 0 otherwise.

Now consider a model in the signature family with support K. Suppose it has the support $\sigma^1, \dots, \sigma^K$ with their respective probabilities $p_1, \dots, p_K$. Because the model belongs to the signature family, the K permutations have distinct *signature* components. For each 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 $\sigma^{k'}(i_k) \ne j_k$ (i.e., $\sigma^{k'}_{(i_k,j_k)} = 0$) for all $k' \ne k$ and $1 \le k' \le K$. Now let $M = [M_{(i,j)}]$ be the first-order marginals of this model. Then, it is clear from our notation that $M_{(i_k,j_k)} = p_k$ for $1 \le k \le K$ and $M_{(i,j)}$ is a summation of a subset of the K values $p_1, \dots, p_K$, for any other (i,j), $1 \le i,j \le N$.

The previous discussion leads to the following representation of a model from the signature family. Each model is represented by an $N^2 \times N^2$ matrix with 0–1 entries, say $Z = [Z_{(i,j)(i',j')}]$ for $1 \le i,j,i',j' \le N$: in $Z_{(i,j)(i',j')}$, the index (i,j) represents the row index and the index (i',j') represents the column index. The ranking model with support K is identified with its K signature components (i_k, j_k), $1 \le k \le K$. The corresponding Z has all $N^2 - K$ columns corresponding to indices other than these K tuples equal to 0. The columns corresponding to (i_k, j_k), for all $1 \le k \le K$, indices are nonzero with each column representing a permutation consistent with signature condition: for each (i_k, j_k), $1 \le k \le K$,

$$Z_{(i,j)(i_k,j_k)} \in \{0,1\}, \quad \text{for all } 1 \le i,j \le N, \tag{8}$$

$$Z_{(i_k,j_k)(i_k,j_k)} = 1, \tag{9}$$

$$Z_{(i,j)(i_k,j_k)} = 0, \quad \text{if } (i,j) \in \{(i_{k'}, j_{k'}) : 1 \le k' \le K,\ k' \ne k\}, \tag{10}$$

$$\sum_{\ell=1}^{N} Z_{(i,\ell)(i_k,j_k)} = 1, \quad \sum_{\ell=1}^{N} Z_{(\ell,j)(i_k,j_k)} = 1, \quad \text{for all } 1 \le i,j \le N. \tag{11}$$

Observe that (9) and (10) enforce the signature condition, whereas (11) enforces the permutation structure. In summary, given a set of K distinct pairs of indices, (i_k, j_k) for $1 \le k \le K$ with $1 \le i_k, j_k \le N$, (8)–(11) represent the collection of all possible sets of permutations in the signature family with these indices as their signature components.

Given the previous representation, the problem of finding a model of support K within the signature family that is within an ε-ball of the observed first-order marginal data, D, may be summarized as finding a Z satisfying (8)–(11) and, in addition, satisfying $\|D - ZD\|_2 \le \varepsilon$. The remainder of this section will be devoted to solving this problem tractably.

4.1.2. Efficient Representation of the Signature Family. Indeed, a signature family 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 for models with support K is to first pick a set of K tuples (there are $\binom{N^2}{K}$ such sets) and then for that particular set of K tuples, search among all Zs satisfying (8)–(11). It will be the complexity of this procedure that essentially drives the complexity of our approach. To this end we begin with the following observation: the problem of optimizing a linear functional of Z subject to the constraints (8)–(11) is equivalent to optimizing the functional over the constraints:

$$Z_{(i,j)(i_k,j_k)} \in [0,1], \quad \text{for all } 1 \le i,j \le N, \tag{12}$$

$$Z_{(i_k,j_k)(i_k,j_k)} = 1, \tag{13}$$

$$Z_{(i,j)(i_k,j_k)} = 0, \quad \text{if } (i,j) \in \{(i_{k'},j_{k'}) : 1 \le k' \le K,\ k' \ne k\}, \tag{14}$$

$$\sum_{\ell=1}^{N} Z_{(i,\ell)(i_k,j_k)} = 1, \quad \sum_{\ell=1}^{N} Z_{(\ell,j)(i_k,j_k)} = 1, \quad \text{for all } 1 \le i,j \le N. \tag{15}$$

It is easy to see that the points described by the set of Equations (8)–(11) are contained in the polytope described by Equations (12)–(15). Thus, in order to justify our observation, it suffices to show that the polytope is the convex hull of points satisfying (8)–(11). However, this immediately follows from the Birkhoff-Von Neumann theorem.

4.1.3. Searching the Signature Family. We now describe the main algorithm that will establish the result of Theorem 1. The algorithm succeeds in finding a model $\hat{\lambda}$ with sparsity $\|\hat{\lambda}\|_0 = O(\varepsilon^{-2}K\log N)$ and error $\|M(\hat{\lambda}) - D\|_\infty \le 2\varepsilon$ if there exists a model μ in signature family with sparsity K that is near consistent with D in the sense that $\|M(\mu) - D\|_2 \le \varepsilon$ (note that $\|\cdot\|_\infty \le \|\cdot\|_2$). The computation cost scales as $\exp(\Theta(K\log N))$. Our algorithm uses the so-called *multiplicative weights* algorithm used within the framework developed by Plotkin et al. (1995) for fractional packing (also see Arora et al. 2012). The algorithm starts by going over all possible $\binom{N^2}{K}$ subsets of possible signature components in any order until the desired choice model $\hat{\lambda}$ is found or all the possible subsets are exhausted. In the latter case, we declare the infeasibility of finding a K sparse model in the signature family that is near consistent. Now consider any such set of K signature components, (i_k,j_k) with $1 \le k \le K$. By the definition of the signature family, the values $D_{(i_k,j_k)}$ for $1 \le k \le K$ are probabilities of the K permutations in the support. Therefore, we check if $1 - \varepsilon \le \sum_{k=1}^{K} D_{(i_k,j_k)} \le 1 + \varepsilon$. If not, we reject this set of K tuples as signature components and move to the next set. If yes, we continue toward finding a model with these K as signature components and the corresponding probabilities.

The model of interest to us, and represented by a Z satisfying (8)–(11), should be such that $D \approx ZD$. Put another way, we are interested in finding a Z such that

$$D_{(i,j)} - \varepsilon \ \le \sum_{k=1}^{K} Z_{(i,j)(i_k,j_k)} D_{(i_k,j_k)} \ \le \ D_{(i,j)} + \varepsilon, \quad \text{for all } \ 1 \le i,j \le N^2, \tag{16}$$

$$Z \text{ satisfies } \ (8) - (11). \tag{17}$$

This is precisely the setting considered by Plotkin-Shmoys-Tardos (Plotkin et al. 1995): Z is required to satisfy a certain collection of difficult linear inequalities (16) and a certain other collection of easy convex constraints (17) (easy, because these constraints can be replaced by (12)–(15), which provide a relaxation with no integrality gap as discussed earlier). If there is a feasible solution satisfying (16) and (17), then the procedure in Plotkin et al. (1995) finds a Z that satisfies (16) approximately and (17) exactly. Otherwise, the procedure provides a certificate of the infeasibility of the program; that is, a certificate showing that no signature choice model approximately consistent with the data and with the K signature components in question exists. We describe the 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 (16) as $2N^2$ inequalities denoted by $Az \ge b$ with A being a $2N^2 \times KN^2$ matrix and b being a $2N^2$ dimensional vector. Finally, the set of z satisfying (17) is denoted $\mathcal{P}$. Thus, we are interested in finding $z \in \mathcal{P}$ such that $Az \ge b$.

The framework in Plotkin et al. (1995) essentially tries to solve the *Lagrangian* relaxation of $Az \ge b$ 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 $a_\ell^T z \ge b_\ell$ for $1 \le \ell \le 2N^2$ (where a_ℓ is the ℓth row of A). We update the weights iteratively:

let $t \in \{0, 1, \ldots\}$ represent the index of the iteration. Initially, $t = 0$ and $p_\ell(0) = 1$ for all ℓ. Given $p(t) = [p_\ell(t)]$, we find z^t by solving the linear program

$$\begin{aligned} &\text{maximize} \ \sum_\ell p_\ell(t)\left(a_\ell^T z - b_\ell\right) \\ &\text{over} \ \ z \in \text{co}(\mathcal{P}). \end{aligned} \tag{18}$$

By our earlier discussion, $\text{co}(\mathcal{P})$ is the polyhedron defined by the linear inequalities (12)–(15), so that optimal *basic* solutions to the LP above are optimal solutions to the optimization problem obtained if one replaced $\text{co}(\mathcal{P})$ with simply $\mathcal{P}$. Now in the event that the LP is infeasible, or if its optimal value is negative, we declare immediately that there does not exist a K-sparse choice model with the K signature components in question that is approximately consistent with the observed data; this is because the program is a relaxation to (16) and (17) in that (16) has been relaxed via the *lagrange* multiplier $p(t)$. Furthermore, if the original program was feasible, then our LP should have a solution of nonnegative value because the weights $p(t)$ are nonnegative. Assuming we do *not* declare infeasibility, the solution z_t obtained is a K-sparse choice model whose signature components correspond to the K components we began the procedure with.

Assuming that the linear program is feasible, and given an optimal basic feasible solution z_t, the weights $p(t+1)$ are obtained as follows: for $\delta = \min(\varepsilon/8, 1/2)$, we set:

$$p_\ell(t+1) = p_\ell\left(1 - \delta\left(a_\ell^T z^t - b_\ell\right)\right). \tag{19}$$

The update (19) suggests that if the ℓth inequality is not satisfied, we should increase the penalty imposed by $p_\ell(t)$ (in proportion to the degree of violation) or else, if it is satisfied, we decrease the penalty imposed by $p_l(t)$ in proportion to the slack in the constraint. Now, $a_\ell^T z^t - b_\ell \in [-2, 2]$. To see this note that, first, $b_\ell \in [0, 1]$, because it corresponds to an entry in a nonnegative doubly stochastic matrix D. Furthermore, $a_\ell^T z^t \in [0, 1 + \varepsilon]$ because it corresponds to the summation of a subset of K nonnegative entries $D_{(i_k, j_k)}$, $1 \le k \le K$ and by choice we have made sure that the sum of these K entries is at most $1 + \varepsilon$. Hence, the multiplicative update to each of the $p_\ell(\cdot)$ is by a factor of at most $(1 \pm 2\delta)$ in a single iteration. Such a bound on the relative change of these weights is necessary for the success of the algorithm.

Now, assume we have not declared infeasibility for all $t \le T$ and consider the sequence of solutions, z^t. Furthermore, set $T = 64\varepsilon^{-2}\ln(2N^2) = O(\varepsilon^{-2}\log N)$, and define $\hat{z} = \frac{1}{T}\sum_{t=0}^{T-1} z^t$. Then, we have via corollary 4 in Arora et al. (2012) (see also their section 3.2), that

$$a_\ell^T \hat{z} \ge b_\ell - \varepsilon, \quad \text{for all } 1 \le \ell \le 2N^2. \tag{20}$$

Now $\hat{z}$ corresponds to a choice model (call it $\hat{\lambda}$) with support over at most $O(KT) = O(\frac{K}{\varepsilon^2}\log N)$ permutations because each z^t is a choice model with support over K permutations in the signature family. Furthermore, (20) implies that $\|M(\hat{\lambda}) - D\|_\infty \le 2\varepsilon$.

Finally, the computational complexity of the previously described algorithm, for a given subset of K signature components, is polynomial in N. Therefore, the overall computational cost of the previously described algorithm is dominated by term $\binom{N^2}{K}$, which is at most N^{2K}. That is, for any $K \ge 1$, the overall computation cost of the algorithm is bounded above by $\exp(\Theta(K\log N))$. This completes the proof of Theorem 1.

4.1.4. Using the Algorithm. It is not a priori clear if for given observation D, there exists a model in the signature family of sparsity K within some small error $\varepsilon > 0$ with $\varepsilon \le \varepsilon_0$ where ε_0 is most error we wish to tolerate. The natural way to adapt the previous 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.

4.2. Proof of Theorem 3

We prove this theorem using the probabilistic method. Given the doubly stochastic matrix D, there exists a model (by Birkhoff-von Neumann's result) λ such that $M(\lambda) = D$. 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 $\|M(\hat{\lambda}) - D\|_2 \le \varepsilon$. Therefore, there must exist a 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 ranking model) that puts $1/T$ probability mass over each of the sampled permutations.

Now consider a pair of indices $1 \le i, j \le N$. Let X_{ij}^t denote the indicator variable of the event that $\sigma_t(i) = j$. Because 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 $1 \le t \le T$. Furthermore,

$$\mathbb{P}\left(X_{ij}^t = 1\right) = \mathbb{E}\left[X_{ij}^t\right] = D_{ij}.$$

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 D_{ij}, denoted by $B(T, D_{ij})$. Therefore, with respect to the randomness of sampling,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T} X_{ij}^t - D_{ij}\right)^2\right] &= \frac{1}{T^2}\mathrm{Var}\left(B(T, D_{ij})\right) \\ &= \frac{1}{T^2} T D_{ij}(1 - D_{ij}) \\ &\le \frac{D_{ij}}{T}, \end{aligned} \tag{21}$$

where we used the fact that $D_{ij} \in [0, 1]$ for all $1 \le i, j \le N$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\left\|M(\hat{\lambda}) - D\right\|_2^2\right] &= \mathbb{E}\left[\sum_{ij}\left(\frac{1}{T}\sum_{t=1}^{T} X_{ij}^t - D_{ij}\right)^2\right] \\ &\le \sum_{ij}\frac{D_{ij}}{T} \\ &= \frac{N}{T}, \end{aligned} \tag{22}$$

where the last equality follows from the fact that D is a doubly stochastic matrix and hence its entries sum up to N. From (22), it follows that by selecting $T = N/\varepsilon^2$, the error in approximating the first-order marginals, $\|M(\hat{\lambda}) - D\|_2$, is within ε on average. Therefore, the existence of such a model follows by the probabilistic method. This completes the proof of Theorem 3.

4.3. Proof of Theorem 2

We prove Theorem 2 using the probabilistic method as well. Suppose that we observe D, which is a noisy version of the first-order marginal $M(\lambda)$ of the underlying model λ. As per the hypothesis of Theorem 2, we shall assume that λ satisfies one of the two conditions: either it is from MNL model or from exponential family with regularity condition on its parameters. For such λ, we establish the existence of a sparse model $\hat{\lambda}$ that satisfies the signature conditions and approximates $M(\lambda)$ (and hence approximates D) well.

As in the proof of Theorem 3, consider T permutations drawn independently and in an identical manner from distribution λ. Let $\hat{\lambda}$ be the empirical distribution of these T samples as considered previously. Following the arguments in the proof of Theorem 3, we obtain (similar to (22)) that

$$\mathbb{E}\left[\left\|M(\hat{\lambda}) - M(\lambda)\right\|_2^2\right] \le \frac{N}{T}. \tag{23}$$

For the choice of $T = 4N/\varepsilon^2$, using Markov's inequality, we can write

$$\mathbb{P}\left(\left\|M(\hat{\lambda}) - M(\lambda)\right\|_2^2 \ge \varepsilon^2\right) \le \frac{1}{4}. \tag{24}$$

Because $\|M(\lambda) - D\|_2 \le \varepsilon$, it follows that $\|M(\hat{\lambda}) - D\|_2 \le 2\varepsilon$ with probability at least 3/4.

Next, we show that the $\hat{\lambda}$ thus generated satisfies the signature condition 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 2 with probability at least 1/4.

To that end, let E_t be the event that σ_t satisfies the signature condition with respect to set $(\sigma_1, \ldots, \sigma_T)$. Because all the permutations $\sigma_1, \ldots, \sigma_T$ are chosen in an i.i.d. manner, the probabilities of all the events are identical.

We wish to show that $\mathbb{P}(\cup_{1\le t\le T} E_t^c) \le 1/2$. This will follow from establishing $T\mathbb{P}(E_1^c) \le 1/2$. To establish this, it is sufficient to show that $\mathbb{P}(E_1^c) \le 1/N^2$ because $T = 4N/\varepsilon^2$.

To that end, suppose σ_1 is such that $\sigma_1(1) = i_1, \ldots, \sigma_1(N) = i_N$. Let $F_j = \{\sigma_t(j) \ne i_j,\ 2 \le t \le T\}$. Then by the definition of the signature condition, it follows that

$$E_1 = \bigcup_{j=1}^{N} F_j.$$

Therefore,

$$\begin{aligned}\mathbb{P}(E_1^c) &= \mathbb{P}\left(\bigcap_{j=1}^{N} F_j^c\right)\\ &\le \mathbb{P}\left(\bigcap_{j=1}^{L} F_j^c\right)\\ &= \mathbb{P}(F_1^c)\prod_{j=2}^{L}\mathbb{P}\left(F_j^c\,\middle|\,\bigcap_{\ell=1}^{j-1} F_\ell^c\right).\end{aligned} \tag{25}$$

We will establish that the right side of (25) is bounded above by $O(1/N^2)$ and hence $T\mathbb{P}(E_1^c) = O(\varepsilon^{-2}/N) \ll 1/2$ for N large enough as desired. To establish this bound of $O(1/N^2)$ under two different conditions stated in Theorem 2, we consider in turn the two cases: (i) λ belongs to the MNL family with condition (5) satisfied, and (ii) λ belongs to the max-ent exponential family model with condition (6) satisfied.

4.3.1. Bounding (25) Under MNL Model with (5). Let $L = N^\delta$ for some $\delta > 0$ as in the hypothesis of Theorem 2 under which (5) holds. Now

$$\begin{aligned}\mathbb{P}(F_1^c) &= 1 - \mathbb{P}(F_1)\\ &= 1 - \mathbb{P}(\sigma_t(1) \ne i_1;\ 2 \le t \le T)\\ &= 1 - \mathbb{P}(\sigma_2(1) \ne i_1)^{T-1}\\ &= 1 - \left(1 - \frac{w_{i_1}}{\sum_{k=1}^{N} w_k}\right)^{T-1}.\end{aligned} \tag{26}$$

For $j \ge 2$, in order to evaluate $\mathbb{P}(F_j^c | \cap_{\ell=1}^{j-1} F_\ell^c)$, we shall evaluate $1 - \mathbb{P}(F_j | \cap_{\ell=1}^{j-1} F_\ell^c)$. To evaluate $\mathbb{P}(F_j \,|\, \cap_{\ell=1}^{j-1} F_\ell^c)$, note that the conditioning event $\cap_{\ell=1}^{j-1} F_\ell^c$ suggests that for each σ_t, $2 \le t \le T$, 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 $\cap_{\ell=1}^{j-1} F_\ell^c$, 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. Because by assumption (without loss of generality), $w_1 < \cdots < w_N$, it follows that $\sum_{k\in X} w_k \ge \sum_{k=1}^{N-j+1} w_k$. Therefore,

$$\mathbb{P}\left(F_j\,\middle|\,\bigcap_{\ell=1}^{j-1} F_\ell^c\right) \ge \left(1 - \frac{w_{i_j}}{\sum_{k=1}^{N-j+1} w_k}\right)^{T-1}. \tag{27}$$

Therefore, it follows that

$$\begin{aligned}\mathbb{P}(E_1^c) &\le \prod_{j=1}^{L}\left[1 - \left(1 - \frac{w_{i_j}}{\sum_{k=1}^{N-j+1} w_k}\right)^{T-1}\right]\\ &\le \prod_{j=1}^{L}\left[1 - \left(1 - \frac{w_N}{\sum_{k=1}^{N-j+1} w_k}\right)^{T-1}\right]\\ &\le \prod_{j=1}^{L}\left[1 - \left(1 - \frac{w_N}{\sum_{k=1}^{N-L+1} w_k}\right)^{T-1}\right]\\ &= \left[1 - \left(1 - \frac{w_N}{\sum_{k=1}^{N-L+1} w_k}\right)^{T-1}\right]^{L}.\end{aligned} \tag{28}$$

Let $W(L,N) = w_N/(\sum_{k=1}^{N-L+1} w_k)$. By hypothesis of Theorem 2, it follows that $W(L,N) \le \sqrt{\log N}/N$ and $L = N^\delta$. Therefore, from above it follows that

$$\begin{aligned}\mathbb{P}(E_1^c) &\le \left[1 - \left(1 - \frac{\sqrt{\log N}}{N}\right)^{T-1}\right]^L \\ &\le \left[1 - \Theta\left(\exp\left(-\frac{T\sqrt{\log N}}{N}\right)\right)\right]^L, \end{aligned} \tag{29}$$

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 because $T = N/\varepsilon$, $(1 + O(\log N/N^2))^T = 1 + o(1) = \Theta(1)$. Now

$$\exp\left(-\frac{T\sqrt{\log N}}{N}\right) = \exp\left(-4\sqrt{\log N}/\varepsilon^2\right) \ll 1. \tag{30}$$

Therefore, using the inequality $1 - x \le \exp(-x)$ for $x \in [0,1]$, we have

$$\mathbb{P}(E_1^c) \le \exp\left(-L\exp\left(-4\sqrt{\log N}/\varepsilon^2\right)\right). \tag{31}$$

Because $L = N^\delta$ for some $\delta > 0$ and $\exp(-4\sqrt{\log N}/\varepsilon^2) = o(N^{\delta/2})$ for any $\delta > 0$, it follows that

$$\begin{aligned}\mathbb{P}(E_1^c) &\le \exp\left(-\Theta\left(N^{\delta/2}\right)\right) \\ &\le O\left(1/N^2\right). \end{aligned} \tag{32}$$

Therefore, it follows that all the T samples satisfy the signature condition with respect to each other with probability at least $O(1/N) \le 1/4$ for N large enough. Therefore, we have established the existence of the desired sparse model in signature family. This completes the proof of Theorem 2 under MNL model with condition (5).

4.3.2. Bounding (25) Under Exponential Family Model with (6). As before, let $L = N^\delta$ for some $\delta > 0$ (the choice of $\delta > 0$ here is arbitrary; for simplicity, we shall think of this δ as being the same as that used previously). Now

$$\begin{aligned}\mathbb{P}(F_1^c) &= 1 - \mathbb{P}(F_1) \\ &= 1 - \mathbb{P}(\sigma_t(1) \ne i_1;\ 2 \le t \le T) \\ &= 1 - \mathbb{P}(\sigma_2(1) \ne i_1)^{T-1}. \end{aligned} \tag{33}$$

To bound the right side of (33), we need to carefully understand the implication of (6) on the exponential family distribution. To start with, suppose parameters θ_{ij} are equal for all $1 \le i,j \le N$. In that case, it is easy to see that all permutations have equal $(1/N!)$ probability assigned and hence the probability $\mathbb{P}(\sigma_2(1) \ne i_1)$ equals $1 - 1/N$. However, such an evaluation (or bounding) is not straightforward because the form of exponential family involves the *partition* function. To that end, consider $1 \le i \ne i' \le N$. Now by the definition of exponential family (and σ_2 is chosen as per it),

$$\begin{aligned}\mathbb{P}(\sigma_2(1) = i) &= \frac{1}{Z(\theta)}\left[\sum_{\sigma \in S_N(1 \to i)} \exp\left(\sum_{kl} \theta_{kl}\sigma_{kl}\right)\right] \\ &= \frac{\exp(\theta_{1i})}{Z(\theta)}\left[\sum_{\sigma \in S_N(1 \to i)} \exp\left(\sum_{k \ne 1, l} \theta_{kl}\sigma_{kl}\right)\right]. \end{aligned} \tag{34}$$

Here, $S_N(1 \to i)$ denotes the set of all permutations in S_N that map 1 to i:

$$S_N(1 \to i) = \{\sigma \in S_N : \sigma(1) = i\}.$$

Given this, it follows that

$$\frac{\mathbb{P}(\sigma_2(1)=i)}{\mathbb{P}(\sigma_2(1)=i')}=\frac{\exp(\theta_{1i})\left[\sum_{\sigma\in S_N(1\to i)}\exp\left(\sum_{k\neq 1,l}\theta_{kl}\sigma_{kl}\right)\right]}{\exp(\theta_{1i'})\left[\sum_{\rho\in S_N(1\to i')}\exp\left(\sum_{k\neq 1,l}\theta_{kl}\rho_{kl}\right)\right]}. \tag{35}$$

Next, we will 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). Under this mapping, suppose $\sigma\in S_N(1\to i)$ is mapped to $\rho\in S_N(1\to i')$. Then we shall have that

$$\exp\left(\sum_{kl}\sigma_{kl}\theta_{kl}\right)\le\sqrt{\log N}\exp\left(\sum_{kl}\rho_{kl}\theta_{kl}\right). \tag{36}$$

This, along with (35) will imply that

$$\frac{\mathbb{P}(\sigma_2(1)=i)}{\mathbb{P}(\sigma_2(1)=i')}\le\sqrt{\log N}. \tag{37}$$

This in turn implies that for any i, $\mathbb{P}(\sigma_2(1)=i)\le\sqrt{\log N}/N$, which we shall use in bounding (33).

To that end, we consider the following mapping from $S_N(1\to i)$ to $S_N(1\to i')$. Consider a $\sigma\in S_N(1\to i)$. By definition $\sigma(1)=i$. 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,

$$\begin{aligned}\frac{\exp(\sum_{kl}\sigma_{kl}\theta_{kl})}{\exp(\sum_{kl}\rho_{kl}\theta_{kl})}&=\frac{\exp(\theta_{1i}+\theta_{qi'})}{\exp(\theta_{1i'}+\theta_{qi})}\\&\le\sqrt{\log N},\end{aligned} \tag{38}$$

where the last inequality follows from condition (6) in the statement of Theorem 2. From the previous discussion, we conclude that

$$\begin{aligned}\mathbb{P}(F_1^c)&=1-\mathbb{P}(\sigma_2(1)\neq i_1)^{T-1}\\&\le 1-\left(1-\frac{\sqrt{\log N}}{N}\right)^{T-1}.\end{aligned} \tag{39}$$

For $j\ge 2$, in order to evaluate $\mathbb{P}(F_j^c|\cap_{\ell=1}^{j-1}F_\ell^c)$, we evaluate $1-\mathbb{P}(F_j|\cap_{\ell=1}^{j-1}F_\ell^c)$. To evaluate $\mathbb{P}(F_j|\cap_{\ell=1}^{j-1}F_\ell^c)$, note that the conditioning event $\cap_{\ell=1}^{j-1}F_\ell^c$ suggests that for each σ_t, $2\le t\le T$, some assignments (ranks) for first $j-1$ items are given and we need to find the probability that the jth item of each of the $\sigma_2,\dots,\sigma_T$ are not mapped to i_j. Therefore, given $\cap_{\ell=1}^{j-1}F_\ell^c$, we wish to evaluate (an upper bound on) probability of $\sigma_2(j)$ mapping i_j given that we know assignments of $\sigma_2(1),\dots,\sigma_2(j-1)$. By the form of the exponential family, conditioning on the assignments $\sigma_2(1),\dots,\sigma_2(j-1)$, effectively we have an exponential family on the space of permutations of remaining $N-j+1$ element. 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 previously to show that $\mathbb{P}(\sigma_2(1)=i)\le\sqrt{\log N}/N$, it follows that

$$\begin{aligned}\mathbb{P}\left(\sigma_2(j)=i_j\middle|\bigcap F_j^c\right)&\le\frac{\sqrt{\log N}}{N-j+1}\\&\le\frac{2\sqrt{\log N}}{N},\end{aligned} \tag{40}$$

where we have used the fact that $j\le L=N^\delta\le N/2$ (for N large enough). Therefore, it follows that

$$\mathbb{P}(E_1^c)\le\left[1-\left(1-\frac{2\sqrt{\log N}}{N}\right)^{T-1}\right]^L. \tag{41}$$

From here on, using arguments identical to those used above (under MNL model), we conclude that

$$\begin{aligned}\mathbb{P}(E_1^c)&\le\exp\left(-\Theta\left(N^{\delta/2}\right)\right)\\&\le O(1/N^2).\end{aligned} \tag{42}$$

This completes the proof for max-ent exponential family with condition (6) and hence that of Theorem 2.

5. Empirical Study

This section is devoted to answering the following, inherently empirical, question:

Can sparse models fit to limited information about the underlying true model be used to effectively uncover information one would otherwise uncover with ostensibly richer data?

In this section, we describe two empirical studies we conducted that support an affirmative answer to this question. The objective of the first study is to determine whether the recovered sparse distribution retains useful structural information present in the underlying dense distribution. The second study determines whether the decisions made using the sparse distribution are accurate. For the purposes of the first study, we used the well-known APA data set, first used by Diaconis (1989), and for the second study, we used the sushi preference data set, described in detail in Kamishima et al. (2005). We describe our findings and conclusions from each of these studies next.

5.1. APA Data Set: Capturing the Structure of the Underlying Distribution

The APA data set 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 that, 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, that is, the fraction of people who rank candidate i to position r. As discussed by Diaconis (1989), the first-order marginals retain useful underlying structure such as the following: (1) candidate 3 has a lot of love (28% of the first-position vote) and hate (23% of the last-position vote) votes; (2) candidate 1 is strong in second position (26% of the vote) and has a low hate vote (15% of last-position vote); and (3) voters seem indifferent about candidate 5.

Having collected only first-order information, our goal will be answer natural questions such as: who should win the election? or what is the socially preferred ranking of the candidates? Of course, there is not a definitive manner in which these questions might be answered. However, having a complete distribution over permutations affords us the flexibility of using any of the several rank aggregation systems available. In order to retain this flexibility, we will fit a sparse distribution to the partial information and then use this sparse distribution as input to the rank aggregation system of choice 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 an aim to understand the type of structure sparse models can capture, we first considered the first-order marginal information of the data set (or distribution). We let λ denote the underlying *true* distribution corresponding to the 5,738 complete rankings of the five candidates. The 5×5 first-order marginal matrix D is given in Table 1.

For this D, we ran a *heuristic* version of the algorithm described in Section 4.1. Roughly speaking, the heuristic tries to find in a *greedy* manner a sparse model in the signature family that approximates the observed data. It runs very fast (polynomial in N) and seems to perform essentially as good as guaranteed by

Table 1. First-Order Marginal Matrix Where the Entry Corresponding to Candidate i and Rank j Is the Percentage of Voters Who Ranked Candidate i at Position j

	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

the algorithm of Section 4.1. However, we are unable to prove any guarantees for it. To keep exposition simple and avoid distraction, we skip description of the heuristic from here (an interested reader can find the heuristic in Jagabathula 2011). Using the heuristic, we obtained 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 this 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, that is, 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. The support size of $\hat{\lambda}$ is only six, 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

$$\frac{1}{25}\sum_{1\le i,j\le 5}\frac{\left|M(\hat{\lambda})_{ij}-D_{ij}\right|}{D_{ij}}.$$

This measure of error, being relative, is more stringent that measuring additive error. The main conclusion we can draw from the small relative error we obtained is that the heuristic we used can successfully to find sparse models that are a good fit to the data in interesting practical cases.

5.1.1. Structural Conclusions. Now that we have 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. More importantly, we will attempt to compare these conclusions with conclusions drawn from what is ostensibly richer data.

5.1.1.1. Comparing Cumulative Distribution Functions. We begin with comparing the staircase curves of the cumulative distribution functions (CDFs) of the actual distribution λ and the sparse approximation $\hat{\lambda}$ in Figure 1. 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 how well the sparse model approximates the true CDF.

Now, one is frequently interested in a functional of the underlying ranking model such as determining a winner, or perhaps, determining a socially preferred ranking. We next compare the conclusions drawn from applying certain functionals to the sparse model we have learned with the conclusions drawn from applying the same functional to what is ostensibly richer data:

5.1.1.2. Winner Determination. Consider a functional of the distribution over rankings meant to capture the most socially preferred ranking. There are many such functionals, and the Hare system provides one such example. When applied to the sparse model we have learned, the Hare systems yields the permutation 13245 as the socially preferred ranking. Now, one may use the Hare system to determine a winner with all of the voting data. These data are substantially richer than the first order marginal information used by our approach. In particular, they contain 5,738 votes, each a total ordering of the candidates (from which our first order marginal information was derived), and approximately 10,000 additional votes for partial rankings of the same candidates. Applying the Hare system here also yields 1 as the winning candidate as reported by Diaconis (1989).

Rank Aggregation. In addition to determining a winner, the Hare system applied to a model also yields an aggregate permutation that, one may argue, represents the aggregate opinions of the population in a fair way. Now, as reported previously, the Hare system applied to our sparse model yields the permutation 13245. As it turns out, this permutation is in remarkable agreement with conclusions drawn by Diaconis using *higher*-order

partial information derived from the same set of 5,738 votes used here. In particular, using second-order marginal data, that is, information on the fraction of voters that ranked candidates $\{i,j\}$ to positions $\{k,l\}$ (without accounting for order in the latter set) for all distinct i,j,k,l yields the following conclusion, paraphrased from Diaconis (1989): There is a strong effect for candidates $\{1,3\}$ to be ranked first and second and for candidates $\{4,5\}$ to be ranked fourth and fifth, with candidate 2 in the middle. Diaconis goes on to provide some color to this conclusion by explaining that voting is typically along partisan lines (academicians versus clinicians), and as such, these groups tend to fall behind the candidate groups $\{1,3\}$ and $\{4,5\}$. Simultaneously, these candidate groups also receive hate votes wherein they are voted as the least preferred by the voters in the opposing camp. Candidate 2 is apparently something of a compromise candidate. Remarkably, we have arrived at the very same permutation using the first order data.

Sparse Support Size. It is somewhat tantalizing to relate the support size (six) of the sparse model learned with the structure observed in the data set by Diaconis (1989) discussed in our last point: there are effectively three types (groups) of candidates, viz. $\{2\}$, $\{1,3\}$, and $\{4,5\}$, in the eyes of the partisan voters. Therefore, all votes effectively exhibit an ordering/preference over these three groups primarily and therefore effectively the votes are representing $3! = 6$ distinct preferences. This is precisely the size of the support of our sparse approximation; of course, this explanation is not perfect since the permutations in the model learned split up these groups.

5.2. Sushi Preference Data Set: Accuracy of Assortment Decisions

We now use the sushi preference data set to show that the sparse distribution recovered from first-order marginal data can result in accurate operational decisions, such as the assortment decision. We start with a brief primer on the assortment optimization problem. Then, we describe the sushi preference data set that we use for our study, after which we present our results and conclusions.

5.2.1. Assortment Optimization. Retailers routinely face the task of deciding which subset or *assortment* of products to offer to each customer. They choose this subset from a much larger universe of N products that are available on the market, typically with the objective to maximize their revenue or profit from each visiting customer. This decision problem is referred to in the literature as the *assortment optimization* problem, and it is one of the most commonly studied operational decision problems.

In its most general form, the assortment problem can be formulated as follows. We let r_j denote the revenue or profit earned by the retailer for a unit sale of product j and $\mathbb{P}(j,S)$ the probability that a customer chooses product j from offer set $S \subseteq [N]$ (where $[N]$ denotes the short-form notation for $\{1,2,\ldots,N\}$) of products. Suppose that the objective of the retailer is to maximize the expected revenue or profit from each customer visit. Then, to make the decision, the retailer must solve the following set optimization problem: $\max\{\sum_{j\in S} r_j\mathbb{P}(j,S) : S \subseteq \mathcal{S}\}$, where $\mathcal{S} \subseteq 2^{[N]}$ is the collection of feasible assortments. The constraints on the assortments are driven by practical business considerations. The simplest, yet nontrivial, constraint is a constraint on the size of the offered assortment. For instance, in ecommerce settings, the retailer often offers a fixed number of products, as determined by the available screen real estate. If C is the number of products the retailer must offer, then $\mathcal{S} = \{S \subseteq [N] : |S| = C\}$. In other settings, the retailer can offer either C or less than C products, in which case the feasible set becomes $\mathcal{S} = \{S \subseteq [N] : |S| \le C\}$.

In order to determine the optimal assortment, the retailer must predict the demand $\mathbb{P}(j,S)$ she would observe for each product j in response to each subset S of products she offers. To model the demand, a discrete choice model is commonly used. The most well-studied choice model class is the so-called random utility maximization (RUM) class. As shown in Farias et al. (2013), the RUM model is equivalently described by a distribution λ over the rankings of the N products. Each ranking σ represents the preference ordering of the customer over the N products, with lower ranked products assumed to be preferred over the higher ranked products. When offered a subset S of products, each arriving customer samples a rank ordering σ according to distribution λ and chooses the most preferred product according to σ from among the offered set of products; that is, the customer chooses product j such that $\sigma(j) < \sigma(i)$ for all $i \in S \setminus \{j\}$. Formally, the demand $\mathbb{P}(j,S)$ can be expressed as

$$\mathbb{P}(j,S) = \sum_{\sigma\in S_N} \lambda(\sigma)\cdot \mathbf{1}_{\{\sigma(j)<\sigma(i)\,\forall\, i\in S, i\neq j\}}.$$

Given the choice model λ, the retailer then computes the expected revenue or profit $R^{\lambda}(S)$ for each offer set S as

$$R^{\lambda}(S) = \sum_{j \in S} r_j \cdot \left(\sum_{\sigma \in S_N} \lambda(\sigma) \cdot \mathbf{1}_{\{\sigma(j) < \sigma(i) \ \forall \ i \in S, i \neq j\}} \right),$$

and then determines the optimal assortment by searching over the set of all feasible subsets. The computational complexity of the search depends on the structure of the choice model and that of the constraints. See Jagabathula (2014) for an overview of the known results on carrying out this search either exactly or approximately in an efficient fashion.

The underlying distribution λ may be complex. Therefore, instead, the retailer could use a sparse distribution that is fit to partial information to make the assortment decision. This raises the question of the accuracy of the resulting decision. To answer this question, we use the popular sushi preference data and assess the accuracy on the assortment decision when the retailer is constrained to offer exactly C products, for some $1 \leq C \leq N$. We describe the data set next.

5.2.2. Sushi Preference Data Set. This is a publicly available data set[2] consisting of preference orderings of different people collected over different sushi types. The complete details of the data set are described in Kamishima et al. (2005). We use a part of the data set that consists of 5,000 complete rankings over $N = 10$ varieties of sushi. These data were collected through a survey in which each respondent provided a complete ranking of the 10 varieties of sushi in the order of their preference. This data set provides a distribution over permutations, where the weight of each permutation is equal to the fraction of survey respondents who provided that particular ranking. As done earlier, we let λ denote this ground truth distribution that is based on the 5,000 rankings. Table 2 provides the list of the 10 sushi varieties and the corresponding 10×10 first-order marginal matrix D.

We used the greedy heuristic described in Jagabathula (2011) to fit a sparse distribution to this first-order marginal data. Using the heuristic, we obtained the sparse distribution over rankings $\hat{\lambda}$ presented in Table 3. Each row in the table corresponds to a ranking of the 10 sushi varieties ordered in the decreasing order of their preference from left to right. The number at the end of the row is the probability weight $\hat{\lambda}(\sigma)$ for the particular ranking σ. The sparse distribution has a support size of 18. With only 18 rankings, compared with the 5,000 rankings in the ground truth distribution, we are able to obtain an average relative error of 0.102 in approximating the first-order marginal data, where the average relative error is defined as

$$\frac{1}{100} \sum_{1 \leq i,j \leq 10} \frac{\left| M(\hat{\lambda})_{ij} - D_{ij} \right|}{D_{ij}}.$$

5.2.3. Accuracy of Assortment Decisions. It is clear that the sparse distribution provides a good approximation of the first-order marginal data. However, what is not clear is if it can provide a good assortment decision. For that, we compute the optimality gap from using the assortment obtained using the sparse distribution.

Table 2. First-Order Marginal Matrix Where the Entry Corresponding to Sushi i and Rank j is the Percentage of Survey Respondents Who Ranked Sushi i at Position j

	Rank									
Sushi	1	2	3	4	5	6	7	8	9	10
ebi	9	11	12	14	14	12	10	8	6	4
anago	11	11	11	11	10	10	10	10	8	8
maguro	8	15	16	15	14	12	8	6	4	2
ika	5	6	9	11	13	14	13	12	10	7
uni	15	13	10	7	6	6	6	6	11	20
ikura	11	13	13	12	9	8	8	7	10	9
tamago	4	4	5	7	9	10	13	15	17	16
toro	34	20	13	9	7	5	4	4	2	2
tekka-maki	2	5	9	12	14	15	17	14	9	3
kappa-maki	1	1	2	3	5	8	11	17	22	30

Table 3. Distribution over Rankings Obtained by Fitting a Sparse Distribution to the First-Order Marginal Data in Table 2

Permutations										Weights
ikura	kappa	tamago	ika	uni	ebi	anago	tekka	maguro	toro	0.0134
tekka	anago	kappa	tamago	ebi	ikura	uni	ika	toro	maguro	0.0225
ebi	tamago	toro	kappa	anago	uni	ika	ikura	maguro	tekka	0.0268
tamago	maguro	tekka	ikura	ika	anago	uni	toro	kappa	ebi	0.0381
ika	maguro	tamago	ebi	kappa	uni	toro	anago	tekka	ikura	0.0428
anago	tekka	maguro	tamago	uni	toro	ika	ikura	ebi	kappa	0.0487
uni	toro	ebi	tekka	ika	ikura	anago	maguro	kappa	tamago	0.0645
ebi	ika	maguro	ikura	toro	anago	tekka	uni	tamago	kappa	0.0647
toro	ikura	ebi	uni	tekka	tamago	maguro	kappa	anago	ika	0.0754
maguro	uni	ika	toro	tamago	kappa	ikura	ebi	tekka	anago	0.0791
toro	uni	tekka	maguro	ika	ebi	tamago	anago	kappa	ikura	0.0444
toro	uni	ika	tekka	tamago	maguro	ebi	anago	ikura	kappa	0.0105
toro	anago	uni	ika	ikura	maguro	ebi	tekka	tamago	kappa	0.0937
uni	maguro	toro	ebi	tekka	tamago	anago	ika	ikura	kappa	0.0213
uni	toro	maguro	ebi	anago	tekka	tamago	ika	ikura	kappa	0.0750
ikura	toro	anago	maguro	ebi	tekka	tamago	kappa	ika	uni	0.1063
toro	ebi	ikura	anago	maguro	ika	kappa	tekka	uni	tamago	0.1099
anago	ikura	toro	tekka	maguro	ebi	ika	tamago	kappa	uni	0.0629

Notes. Each row is a permutation in the support of the distribution. The 10 sushi varieties are listed in the decreasing order of their preference from left to right; the names of sushi varieties tekka-maki and kappa-maki have been shortened to tekka and kappa, respectively. The number at the end of the row is the probability weight of the corresponding permutation. There is a total of 18 rows; that is, the support size of the sparse distribution is 18.

We consider the following setup. We assume that the ground truth distribution λ collected from the 5,000 respondents is representative of the population preferences over these 10 varieties of the sushi. Assuming that each arriving customer must be offered exactly C varieties of sushi, we consider the problem of determining the subset of sushi varieties that maximizes the expected revenue. We assume that when offered a subset of sushi varieties, an arriving customer samples a ranking σ from distribution λ and then chooses the most preferred offered sushi variety according to σ. In our setup, we assume that the customer does not have an outside option; that is, the customer will always purchase something. This assumption easily holds when all of the 10 sushi varieties are acceptable to all the customers at the current prices (note that the sushi prices are fixed and we can only change the offer sets), which is reasonable because these are the most popular sushi varieties. Our analysis can be readily replicated when one has access to information on the outside option.

For each value of offer set size C, we find the optimal assortments:

$$S^* \in \arg\max_{S\subseteq[N]:|S|=C} R^{\lambda}(S) \text{ and } \hat{S} \in \arg\max_{S\subseteq[N]:|S|=C} R^{\hat{\lambda}}(S),$$

Table 4. Comparison of the Revenues from Optimal Assortments and the Approximations Obtained Using the Sparse Distribution at All Possible Offer Set Sizes

Size	Revenue		Rel. error
	Optimal	Sparse approx.	
1	4.49	4.49	0.00%
2	4.14	4.08	1.58%
3	3.88	3.64	6.30%
4	3.70	3.62	2.17%
5	3.55	3.37	5.21%
6	3.41	3.21	5.98%
7	3.30	3.16	4.40%
8	3.21	3.03	5.38%
9	3.12	3.05	2.28%
10	3.02	3.02	0.00%

Notes. Revenues are reported in normalized price units, as reported in Kamishima et al. (2005). We observe that the average relative error across the sizes is about 3.3%, indicating that sparse distributions have good decision accuracy.

where we use the *normalized* price available in the data set for each sushi variety j as its revenue r_j. Table 4 presents our results. We consider all possible offer set sizes, $C = 1, 2, \ldots, 10$. For each offer set size, we report the corresponding optimal revenue $R^\lambda(S^*)$, the true revenue $R^\lambda(\hat{S})$ from offering the estimate $\hat{S}$, and the relative error, defined as $|R^\lambda(S^*) - R^\lambda(\hat{S})|/R^\lambda(S^*)$, from making the suboptimal decision. We note from our results that even with just 18 permutations, the decision obtained by the sparse model is always within 6.5% of the optimal revenue. The average error across the 10 offer set sizes is only 3.3%, indicating that sparse models have good decision accuracy.

Figure 2 graphically represents the exact and the approximate optimal assortments. Each row corresponds to an offer set size and each column corresponds to a sushi variety. For each row, all the black squares represent the sushi varieties in both S^* and $\hat{S}$, that is, the set $S^* \cap \hat{S}$. The dark gray squares represent the sushi varieties in the optimal assortment but not in the estimate, that is, the set $S^* \setminus \hat{S}$. Finally, the light gray squares represent $\hat{S} \setminus S^*$. The white squares represent sushi varieties that were not in either of the sets S^* or $\hat{S}$. As is evident from the figure, most of the squares are black, visually indicating that the assortments obtained using the sparse distribution result in decisions that are close approximations of the true decisions.

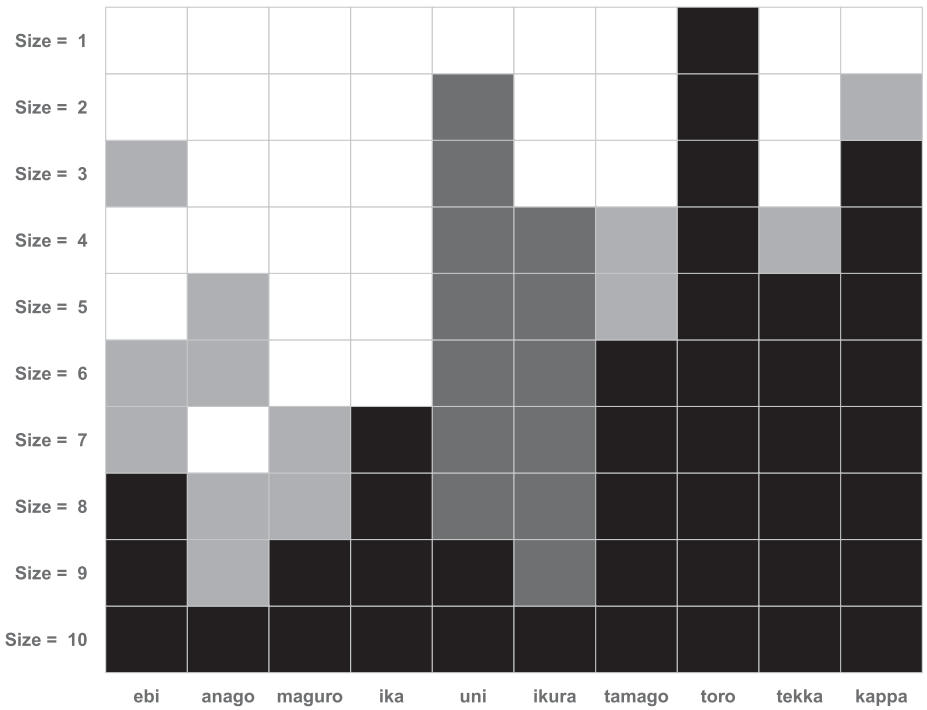
6. Discussion

6.1. Summary

Distributions over rankings form an integral part of various important decision-making tasks. Sparse approximations to the underlying true ranking models based on (information-) limited observed data are particularly attractive as they are simple and hence easy to integrate in complex decision-making tasks. In addition, learning sparse models from marginal data provides a nonparametric approach to maintaining and learning complex distributions over rankings. This paper has taken important steps toward establishing a sparse model approximation as a viable option.

As the first main result, we considered the problem of efficiently recovering a sparse model that is near consistent with the observed first-order marginals. We showed that the signature family of models lends itself

Figure 2. Graphical Representation of the Optimal Assortments Under All Possible Offer Set Sizes



Notes. For each offer set size, the row represents the optimal and approximate offer sets, where the approximation is obtained using the sparse distribution. The names of sushi varieties tekka-maki and kappa-maki have been shortened to tekka and kappa, respectively. A black square represents a product that is part of both the optimal and approximate offer sets. A dark gray square represents a product that is part of the optimal offer set and not the approximate offer set. A light gray square represents a product that is part of the approximate offer set but not the optimal offer set. All white squares represent products that are not part of either offer set. The presence of mostly black squares visually indicates that the sparse distribution obtains a close approximation of the true decision.

to efficient search. In particular, if there is a model λ of sparsity K in the signature family that is an ε-fit to the data, then we can find a model of $\hat{\lambda}$ of sparsity at most $O(\varepsilon^{-2}K\log N)$ that is an ε-fit to the data in time $\exp(\Theta(K\log N))$, as opposed to the brute-force time complexity of $\exp(\Theta(KN\log N))$. The computational efficiency is obtained by exploiting the particular structures of the Birkhoff polytope and the signature family. In prior work, the signature condition was shown to be necessary for efficient and *exact* recovery of the underlying model from noise-free data. This work establishes the robustness of the signature condition.

As the second main result, we showed that the signature family is general and restricting our search to the signature family results in no loss of worst-case sparsity. Specifically, when the underlying ranking model belongs to the MNL or the exponential family, then under appropriate regularity conditions, there exists a model in the signature family with sparsity $O(N/\varepsilon^2)$ that is an ε-fit to the first-order marginal data. We also showed that the first-order marginals from *any* model can be ε-approximated with a model (not necessarily from the signature family) with sparsity $O(N/\varepsilon^2)$; this bound is tight (in the scaling of N) because there exists a doubly stochastic matrix that requires $\Omega(N)$ sparsity to guarantee ε error. As a result, even with the restriction to the signature family, there is no loss in terms of worst-case sparsity scaling.

In the recently popular compressive sensing literature, the restricted null space condition has been shown to be necessary and sufficient for efficient learning of sparse models via linear programs. It was shown in the past that this restricted null space condition (or effectively a linear programming relaxation of our problem) is ineffective in learning sparse ranking models. In that sense, this work shows that signature conditions are another set of sufficient conditions that help learn sparse ranking models computationally efficiently.

Our result currently does not establish the optimality of the scaling of the sparsity level with respect to the error term ε. Based on our proof technique, we conjecture that the scaling is indeed tight. Establishing this rigorously is a promising avenue for future work.

6.2. Beyond First-Order Marginals

Theorems 3 and 2, which establish the generality of the signature family, can be extended in a reasonably straightforward way to other types of marginal information (Jagabathula and Shah 2011 is the basis of this assertion). The key result that strongly exploits the structure of the first-order marginals, however, is the algorithm in Theorem 1. The computational efficiency of the algorithm relies on the Birkhoff-Von Neumann result and will not readily extend to other types of marginal information. The algorithm extends to higher-order marginals with possibly a computationally complex oracle to check the feasibility of a signature model with respect to the higher-order marginal. Indeed, it would be an important direction for future research to overcome this computational threshold by possibly developing better computational approximations. The heuristic used in Section 5 is quite efficient (polynomial in N) for first-order marginals. It is primarily inspired by the exact recovery algorithm based on the signature condition used in our earlier work. We strongly believe that such a heuristic is likely to provide a computationally efficient procedure for higher-order marginal data.

6.3. Signature Condition and Computational Efficiency

We established that the signature condition enables efficient search for a sparse model, effectively knocking off a factor N from the computational complexity of the, otherwise best, brute-force algorithm. However, it is worth asking the question whether alternative algorithms that do not rely on the signature condition can provide a speedup relative to brute-force search. As it turns out, the following can be shown: Assume there exists a sparse model, with sparsity K, that approximates the observed first-order marginals (i.e., the doubly stochastic matrix) within accuracy ε. Then, we can recover a sparse model (not necessarily in the signature family) with sparsity at most K that approximates the observed first-order marginals to within ε in time $(\frac{1}{\varepsilon})^K \times \exp(K\log K)$. If K is close to N, then this computational cost is worse by a factor of $(\frac{1}{\varepsilon})^K$ compared with the approach that relies on the signature condition in Section 4.1. Furthermore, this alternate heuristic has no natural generalization to observed data outside of the first-order marginal information. In contrast, the approach we have followed, by relying on the structure afforded by the signature family, suggests a fast (polynomial in N) heuristic that can potentially be applied for many distinct types of marginal data; this is the very heuristic used in Section 5. Establishing theoretical guarantees for this heuristic remain an important direction for future research.

Now we provide a brief description of the algorithm hinted at previously. The algorithm is similar to that described in Section 4.1. Specifically, it tries to find K permutations and their associated probabilities, so that the resulting distribution has first-order marginals that are near consistent with the observations. Now the K unknown permutations are represented through their linear relaxations implied by the Birkhoff-Von

Neumann result, that is, (8) and (11). Under the signature condition, the associated probabilities were discovered implicitly by means of (9) and (10). However, without the signature condition, the only option we have is to search through all possible values for these probabilities. Because we are interested in approximation accuracy of ε, it suffices to check $(K/\varepsilon)^K$ such probability vectors. For a given such probability vector, we are left with the problem of searching for K permutations with these probabilities that have their corresponding first-order marginals well approximated by the observed data. Using similar ideas described in Section 4.1, we can find a sparse model with sparsity K efficiently (within time polynomial in N) if there existed a sparse model with sparsity K and the particular quantized probability vector that approximated the observations sufficiently well. Because the algorithm will start search over increasing values of K and for a given K, over all $(K/\varepsilon)^K$ distinct probability vectors, the effective computation cost will be dominated by the largest K value encountered by the algorithm. This effectively completes the explanation of the algorithm and its computational cost.

Endnotes

[1] It is worth noting a very similar Bradley-Terry (Bradley 1953) model, where each object i has weight $w_i > 0$ associated with it; this model is, however, different from the model proposed by Plackett and Luce in the probabilities it assigns to each of the permutations.

[2] See http://www.kamishima.net/sushi/.

References

Agrawal S, Wang Z, Ye Y (2008) Parimutuel betting on permutations. Papadimitriou C, Zhang S, eds. *WINE 2008: Internet and Network Economics.* (Springer, Berlin), 126–137.

Arora S, Hazan E, Kale S (2012) The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing.* 8(6):121–164.

Bartels K, Boztuæg Y, Müller MM (1999) Testing the multinomial logit model. Working paper, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin.

Ben-Akiva ME (1973) Structure of passenger travel demand models. PhD thesis, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Ben-Akiva ME, Lerman SR (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand* (CMIT Press, Cambridge, MA).

Beran R (1979) Exponential models for directional data. *Ann. Statist.* 7(6):1162–1178.

Berinde R, Gilbert AC, Indyk P, Karloff H, Strauss MJ (2008) Combining geometry and combinatorics: A unified approach to sparse signal recovery. *Proc. 46th Annual Allerton Conf. on Communication, Control, and Computing* (IEEE, New York), 798–805.

Birkhoff G (1946) Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman Rev. Ser. A* 5:147–151.

Boyd JH, Mellman RE (1980) The effect of fuel economy standards on the u.s. automotive market: An hedonic demand analysis. *Transporation Res. Part A General* 14(5–6):367–378.

Bradley RA (1953) Some statistical methods in taste testing and quality evaluation. *Biometrics* 9:22–38.

Candes EJ, Romberg J (2006) Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations Comput. Math.* 6(2):227–254.

Candes EJ, Tao T (2005) Decoding by linear programming. *IEEE Trans. Inform. Theory* 51(12):4203–4215.

Candes EJ, Romberg J, Tao T (2006a) Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory* 52(2):489–509.

Candes EJ, Romberg JK, Tao T (2006b) Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* 59(8):1207–1223.

Cardell NS, Dunbar FC (1980) Measuring the societal impacts of automobile downsizing. *Transporation Res. Part A General* 14(5–6):423–434.

Cormode G, Muthukrishnan S (2006) Combinatorial algorithms for compressed sensing. *Lecture Notes Comput. Sci.* 4056:280.

Crain B (1976) Exponential models, maximum likelihood estimation, and the haar condition. *J. Amer. Statist. Assoc.* 71:737–745.

Debreu G (1960) Review of r. d. luce, 'individual choice behavior: A theoretical analysis'. *Amer. Econom. Rev.* 50:186–188.

Diaconis P (1988) *Group Representations in Probability and Statistics* (Institute of Mathematical Statistics, Hayward, CA).

Diaconis P (1989) A generalization of spectral analysis with application to ranked data. *Ann. Statist.* 17(3):949–979.

Donoho DL (2006) Compressed sensing. *IEEE Trans. Inform. Theory* 52(4):1289–1306.

Farias V, Jagabathula S, Shah D (2009) A data-driven approach to modeling choice. Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY), 504–512.

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

Fishburn PC, Brams SJ (1983) Paradoxes of preferential voting. *Math. Magazine* 56(4):207–214.

Gallager R (1962) Low-density parity-check codes.*IEEE Trans. Inform. Theory* 8(1):21–28.

Gilbert AC, Strauss MJ, Tropp JA, Vershynin R (2007) One sketch for all: Fast algorithms for compressed sensing. *Proc. 39th Annu. ACM Sympos. on Theory of Computing* (ACM, New York), 237–246.

Guadagni PM, Little JDC (1983) A logit model of brand choice calibrated on scanner data. *Marketing Sci.* 2(3):203–238.

Horowitz JL (1993) Semiparametric estimation of a work-trip mode choice model. *J. Econometrics* 58:49–70.

Jagabathula S (2011) Nonparametric choice modeling: Applications to operations management. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Jagabathula S (2014). Assortment optimization under general choice. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2512831.

Jagabathula S, Shah D (2008) Inferring rankings under constrained sensing. Koller D, Schuurmans D, Bengio Y, Bottou L, eds. *Advances in Neural Information Processing Systems 21* (Curran Associates, Inc., Red Hook, NY), 753–760.

Jagabathula S, Shah D (2011) Inferring rankings under constrained sensing. *IEEE Trans. Inform. Theory* 57(11):7288–7306.

Jaggi M (2013) Revisiting Frank-Wolfe: Projection-free sparse convex optimization. Dasgupta S, McAllester D, eds. *Proc. of the 30th Internat. Conf. on Machine Learn*. (Proceedings of Machine Learning Research, Atlanta, Georgia), 427–435

Kamishima T, Kazawa H, Akaho S (2005) Supervised ordering-an empirical survey. *Proc. 5th IEEE Internat. Conf. on Data Mining (ICDM'05)* (IEEE, New York), 673–676.

Koopman BO (1936) On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.* 39(3):399–409.

Luby MG, Mitzenmacher M, Shokrollahi MA, Spielman DA (2001) Improved low-density parity-check codes using irregular graphs. *IEEE Trans. Inform. Theory* 47(2):585–598.

Luce RD (1959) *Individual Choice Behavior: A Theoretical Analysis* (Wiley, New York).

Mahajan S, van Ryzin GJ (1999) On the relationship between inventory costs and variety benefits in retail assortments. *Management Sci.* 45(11):1496–1509.

Marden JI (1995) *Analyzing and Modeling Rank Data* (Chapman & Hall/CRC, New York).

Marschak J (1959) *Binary Choice Constraints on Random Utility Indicators* (Cowles Foundation Discussion Papers, New Haven, CT).

Marschak J, Radner R (1972) *Economic Theory of Teams* (Yale University Press, New Haven, CT).

McFadden D (1973) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed. *Frontiers in Econometrics* (Academic Press, New York), 105–142.

McFadden D (1981) Econometric models of probabilistic choice. Manski CF, McFadden D, eds. *Structural Analysis of Discrete Data with Econometric Applications* (MIT Press, Cambridge, MA), 198–272.

McFadden D (2000) Disaggregate Behavioral Travel Demands Rum Side. *9th Internat. Conf. on Travel Behaviour Res.* (Queensland, Australia), 38.

McFadden D, Train K (2000) Mixed MNL models for discrete response. *J. Appl. Econometrics* 15(5):447–470.

Nyquist H (2002) Certain topics in telegraph transmission theory. *Proc. IEEE* 90(2):280–305.

Plackett RL (1975) The analysis of permutations. *Appl. Statist*. 24(2):193–202.

Plotkin SA, Shmoys DB, Tardos É (1995) Fast approximation algorithms for fractional packing and covering problems. *Math. of Oper. Res*. 20(2):257–301.

Reed IS, Solomon G (1960) Polynomial codes over certain finite fields. *J. Soc. Ind. Appl. Math.* 8(2):300–304.

Shannon CE (1949) Communication in the presence of noise. *Proc. IRE* 37(1):10–21.

Sipser M, Spielman DA (1996) Expander codes. *IEEE Trans. Inform. Theory* 42:1710–1722.

Thurstone L (1927) A law of comparative judgement. *Psychol. Rev.* 34:237–286.

Tropp JA (2004) Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* 50(10):2231–2242.

Tropp JA (2006) Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory* 52(3):1030–1051.

von Neumann J (1953) A certain zero-sum two-person game equivalent to the optimal assignment problem. *Contributions to the Theory of Games*. 2:5–12.

Wainwright MJ, Jordan MI (2008) Graphical models, exponential families, and variational inference. *Foundations Trends Machine Learning* 1(1–2):1–305.

Yellott JI (1977) The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *J. Math. Psychol.* 15(2):109–144.