Computing Choice

Learning Distribution over Permutations

Devavrat Shah

Contents

1 Computing Choice

We discuss the question of learning distribution over permutations of a given set of choices, options or items based on partial observations. This is central to capturing the so called "choice" in a variety of contexts: understanding preferences of consumers over a collection of products based on purchasing and browsing data in the setting of retail and e-commerce, learning public opinion amongst a collection of socio-economic issues based on sparse polling data, deciding a ranking of teams or players based on outcomes of games, electing leaders based on votes, and more generally collaborative decision making based on collective judgement such as accepting paper(s) in a competitive academic conference. The question of learning distribution over permutations arises beyond capturing "choice" as well. For example, tracking a collection of objects using noisy cameras, or aggregating ranking of web-pages using outcomes of multiple search engines. It is only natural that such a topic has been extensively studied in Economics, Political Science and Psychology for more than a century, and more so recently in Computer Science, Electrical Engineering, Statistics and Operations Research.

Here we shall focus on the task of learning distribution over permutations from its marginal distributions of two types: first-order marginals and pair-wise comparisons. There has been a lot of progress made on this topic in the last decade. The ideal goal is to provide a comprehensive overview of the state-of-art on this topic. We shall provide detailed overview of selective aspects, biased by author's perspective of the topic. And provide sufficient pointers to aspects not covered here. We shall emphasize on ability to identify the entire distribution over permutation as well as "best ranking".

1.1 Background

1.1.1 Learning from comparisons

Consider a grocery store around the corner from your home. The owner of the store would like to have the ability to identify exactly what every customer would purchase (or not) given the options available in the store. If such an ability exists, then for example, optimal stocking decisions can be made by the store operator or the net worth of the store can be evaluated. This ability is what one would call as the "choice model" of consumers of the store. Precisely, such a "choice

model" can be viewed as a black-box that spits out the probability of purchase of a particular option when presented with a collection of options.

A canonical fine-grained representation for such a "choice model" is the distribution over permutations of all the possible options (including *no-purchase* option). Then, probability of purchasing a particular option when presented with a collection of options is simply the probability that this particular option has the highest (relative) order or rank amongst all the presented options (including *no-purchase* option).

Therefore, one way to operationalize such a "choice model" is to learn the distribution over permutations of all options that a store owner can stock in the store. Clearly, such a distribution needs to be learnt from the observations or data. The data available to the store owner is the historical transactions as well as what was stocked in the store when each transaction happened. Such data effectively provides a bag of pair-wise comparisons between options: consumer exercises or purchases option A over option B corresponds to a pair-wise comparison A $>$ B or "A is preferred to B".

In summary, to model consumer choice, we wish to learn distribution over permutations of all possible options using observations in terms of a collection of pair-wise comparisons that are consistent with the learnt distribution.

In the context of sports, we wish to go a step further: obtain ranking of sports teams or players based on outcome of games which are simply pair-wise comparisons (between teams or players). Similarly, for the purpose of data-driven policy making, we wish to aggregate people's opinion about socio-economic issues such as modes of transportation based on survey data; for designing online recommendation systems based on historical online activity of individuals, we wish to recommend top few options; or to sort objects based on noisy outcomes of pair-wise comparisons.

1.1.2 Learning from first-order marginals

The task of learning distribution over permutations of options, using different type of partial information comes up in other scenarios. To that end, now suppose the store owner wants to track each consumer's journey in store with the help of cameras. The consumers constantly move within the store as they search through aisles. Naturally, when multiple consumers are in the store, their paths are likely to cross. When paths of two(or more) consumers cross and subsequently follow different trajectories, confusion can arise in the sense that which of the multiple trajectories map to which of the consumers. That is, at each instance of time we need to continue mapping physical locations of consumers observed by cameras with the trajectories of consumers that are being tracked. Equivalently, it's about keeping track of a 'matching' between locations and individuals in a bipartite graph or keeping track of permutations!

In (Huang, Guestrin & Guibas 2009), authors proposed distribution over permutations as the canonical model where a permutation corresponds to "match-

ing" of consumers or trajectories to locations. In such a scenario, due to various constraints and tractability reasons, the information that is available is the likelihoods of each consumer or trajectory to be in a specific location. In the context of distribution over permutation, this corresponds to knowing the "first-order" marginal distribution information that states the probability of a given option being in certain position in the permutation. Therefore, to track consumers in the store, we wish to learn distribution over consumer trajectories that is consistent with this first-order marginal information over time.

In summary, the model to track trajectories of individuals boils down to continually learning distribution over permutations that is consistent with the first-order marginal information and subsequently finding the *most likely* ranking or permutation as per the learnt distribution. It is the very same question that arises in the context of aggregating web-page rankings obtained through result of search from multiple search engines in a computationally efficient manner.

1.1.3 Historical remarks

This fine-grained representation for choice, distribution over permutations, is ancient. Here, we provide a brief historical overview of the use of distribution over permutations as a model for choice and other applications. We also refer to the monograph by (Diaconis 1988, Chapter 9) for a nice historical overview from a Statistician's perspective.

One of the earliest references to model and learn choice using, potentially inconsistent, comparisons is the seminal work by Thurstone (Thurstone 1927). It presents "a law of comparative judgement" or more precisely a simple parametric model to capture the outcomes of a collection of pair-wise comparisons between given options (or *stimuli* in the language of (Thurstone 1927)). This model can be rephrased as an instance of *Random Utility Model* (RUM) as follows (also see (Marschak 1959) and (Marschak & Radner 1972)): given N options, let each option, say i, have inherent utility u_i associated with it; when two options i and j are compared, random variables Y_i, Y_j are sampled and the i is preferred over j iff $Y_i > Y_j$; here $Y_i = u_i + \varepsilon_i$, $Y_j = u_j + \varepsilon_j$ with $\varepsilon_i, \varepsilon_j$ are independent random variables with identical mean.

A specialization of the above model when the ε_is are assumed to be Gaussian with mean 0 and variance 1 for all i is known as the Thurstone-Mosteller model. It is also known as the *probit* model. Another specialization of the Thurstone model is realized when ε_is are assumed to have Gumbel distribution (one of the extreme value distribution). 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 described in detail in Section 1.3.2. It is known as the Luce model (Luce 1959) and the Plackett model (Plackett 1975). 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)

called it *conditional logit*; also see (Debreu 1960). It is worth remarking that when restricted to pair-wise comparisons only, this model matches the Bradley-Terry (Bradley 1953) model, but Bradley-Terry model did not consider the requirement that the pair-wise comparison marginals need to be consistent with an underlying distribution over permutations.

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 & Radner 1972). Very early on, simple statistical tests as well as simple estimation procedures were developed to fit such a model to observed data (McFadden 1973). Now 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* utilized across application areas primarily due to the ability to learn the model parameters easily from observed data. For example, see (McFadden 1981, Ben-Akiva & Lerman 1985, McFadden 2001) for application in transportation and (Guadagni & Little 1983, Mahajan & 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, as well as 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 & Lerman 1985, Boyd & Mellman 1980, Cardell & Dunbar 1980, McFadden & Train 2000) for example. The interested reader is also referred to an overview article on this line of research by (McFadden 2001). While generalized models of this sort are in principle attractive, their complexity makes them difficult to learn while avoiding the risk of over-fitting. 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, Boztug & Muller 1999, Horowitz 1993, Debreu 1960).

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, see (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 (see (Crain 1976), (Beran 1979) and (Wainwright & Jordan 2008)) as it requires computing a "partition function" possibly over a complex state space.

Very recently, (Jagabathula & Shah 2009, Jagabathula & Shah 2011) introduced *non-parametric* sparse model. Here the distribution over permutations is assumed to have sparse (or small) support. While this may not be exactly true, it can be an excellent approximation for the reality and can provide computationally efficient ways to both infer the model (Jagabathula & Shah 2009, Jagabathula & Shah 2011) consistent with observations as well as utilize it for effective decision making (Farias, Jagabathula & Shah 2009, Farias, Jagabathula & Shah 2013).

1.2 Setup

Given N objects or items denoted as $[N] = \{1, \ldots, N\}$, we are interested in distribution over permutations of these N items. A permutation $\sigma : [N] \to [N]$ is one-to-one and onto mapping with $\sigma(i)$ denoting the position or ordering of element $i \in [N]$.

Let S_N denote the space of $N!$ permutations of these N items. The set of distribution over S_N, denoted as $\mathcal{M}(S_N) = \{\lambda : S_N \to [0,1] : \sum_{\sigma \in S_N} \lambda(\sigma) = 1\}$. Given $\lambda \in \mathcal{M}(S_N)$, the first-order marginal information, $M(\lambda) = [M_{ij}(\lambda)]$, is an $N \times N$ doubly stochastic matrix with non-negative entries defined as

$$M_{ij}(\lambda) = \sum_{\sigma \in S_N} \lambda(\sigma) \mathbf{1}_{\{\sigma(i)=j\}}, \tag{1.1}$$

where for $\sigma \in S_N$, $\sigma(i)$ denotes the rank of item i under permutation σ, and $\mathbf{1}_{\{x\}}$ is the standard indicator with $\mathbf{1}_{\{\text{true}\}} = 1$ and $\mathbf{1}_{\{\text{false}\}} = 0$. The comparison marginal information, $C(\lambda) = [C_{ij}(\lambda)]$, is an $N \times N$ matrix with non-negative entries defined as

$$C_{ij}(\lambda) = \sum_{\sigma \in S_N} \lambda(\sigma) \mathbf{1}_{\{\sigma(i)>\sigma(j)\}}. \tag{1.2}$$

By definition, diagonal entries of $C(\lambda)$ are all 0s and $C_{ij}(\lambda) + C_{ji}(\lambda) = 1$ for all $1 \leq i \neq j \leq N$. We shall abuse notation by using $M(\sigma)$ and $C(\sigma)$ to denote the matrices obtained by applying them to distribution with σ having probability 1.

Throughout, we assume that there is a *ground-truth* model λ. We observe marginal information $M(\lambda)$ or $C(\lambda)$, or their noisy versions.

1.2.1 Questions of interest

We are primarily interested in two questions: recovering distribution and producing the ranking, based on the distribution.

Question 1. Recover distribution. The primary goal is to recover λ from observations. Precisely, we observe $D = P(\lambda) + \eta$ where $P(\lambda) \in \{M(\lambda), C(\lambda)\}$ and potentially noisy perturbation η. The noisy perturbation may representation the finite sample error introduced due to forming empirical estimation of $P(\lambda)$ or inability to observe data associated with certain components.

A generic λ has $N!-1$ unknowns while dimension of D is at most N^2. Learning λ from D boils down to finding solution to a set of linear equations where there are at most N^2 linear equations involving $N! - 1$ unknowns. This is highly under-determined system of equations and hence without imposing structural conditions on λ, it is unlikely for us to be able to recover λ faithfully. Therefore, the basic "information" question would be to understand under what structural assumption on λ, is it feasible to recover it from $P(\lambda)$ (i.e. when $\eta = \mathbf{0}$). The next question is to understand the "robustness" of such a recovery condition when we have non-trivial noise, η. And finally, we would like answer the "computational" question associated with it which asks whether such recovery is possible using computation that scales polynomially in N.

Question 2. Produce ranking. An important associated decision question is that of finding the "ranking" or most "relevant" permutation for the underlying λ. To begin with, what is the most "relevant" permutation assuming we know the λ perfectly. This, in a sense, is ill-defined due to the impossibility result of Arrow (Arrow 1950): *there is no ranking algorithm that works for all λ and satisfies certain basic hypothesis expected from any ranking algorithm even when $N = 3$.*

For this reason, like in the context of recovering λ, we will have to impose structure on λ. In particular, the structure that we shall impose seem to suggest a natural answer for ranking or most "relevant" permutation: find σ that has maximal probability, i.e. find $\sigma^*(\lambda)$ where

$$\sigma^*(\lambda) \in \arg\max_{\sigma \in S_N} \lambda(\sigma). \tag{1.3}$$

Again, the goals would include ability to recover $\sigma^*(\lambda)$ (exactly or approximately) using observations when (a) $\eta = \mathbf{0}$, (b) non-trivial η and (c) computationally efficiently.

1.3 Models

We shall consider two types of model here: non-parametric sparse model and parametric random utility model. These models effectively impose constraint on λ that allows for ability to recover them as well as find ranking from the partial information of the form D. As mentioned earlier, there are a large number of models that are studied in literature and not discussed in detail here. We shall provide a brief overview of such result in Section 1.6.

1.3.1 Sparse Model

The support of distribution λ, denoted as $\mathsf{supp}\,(\lambda)$ is defined as

$$\mathsf{supp}\,(\lambda) \triangleq \{\sigma \in S_N : \lambda(\sigma) > 0\}. \tag{1.4}$$

The ℓ_0 norm of λ, denoted as $\|\lambda\|_0$, is defined as

$$\|\lambda\|_0 \triangleq \big|\mathsf{supp}\,(\lambda)\big|. \tag{1.5}$$

We call λ having sparsity K if $K = \|\lambda\|_0$. Naturally, by varying K, all possible $\lambda \in \mathcal{M}(S_N)$ can be captured. In that sense, this is a *non-parametric* model. This model was introduced in (Jagabathula & Shah 2009, Jagabathula & Shah 2011).

The goal would be to learn a sparsest possible λ that is consistent with observations. Formally, this corresponds to solving

$$\begin{array}{lll} \mathsf{minimize} & \|\mu\|_0 \quad \mathsf{over} \quad \mu \in \mathcal{M}(S_N) \\ \mathsf{such\ that} & P(\mu) \approx D, \end{array} \tag{1.6}$$

where $P(\mu) \in \{D(\mu), C(\mu)\}$ depending upon the type of information considered. We discuss known results about recovering sparse model in Section 1.4.

1.3.2 Random Utility Model (RUM)

We consider the random utility model (RUM) that in effect was considered in the "law of comparative judgement" by Thurstone (Thurstone 1927). Formally, each option $i \in [N]$ has a deterministic utility u_i associated with it. The random utility, Y_i associated with option $i \in [N]$ obeys the form

$$Y_i = u_i + \varepsilon_i, \tag{1.7}$$

where ε_i are independent random variables across all $i \in [N]$ – they represent "random perturbation" of the "inherent utility" u_i. We assume that all ε_i have identical mean across all $i \in [N]$, but can have varying distribution. The specific form of distribution gives rise to different types of models. We shall describe few popular examples of this in what shall follow. Before we do that, we explain how this setup gives rise to distribution over permutation by describing generative form of the distribution. Specifically, to generate a random permutation over the N options, we first sample random variable $Y_i,\ i \in [N]$ independently. Then, we sort $Y_1, \ldots, Y_N$ in decreasing order[1] and this sorted order of indices $[N]$ provides the permutation. Now we describe two popular examples of this model.

Probit Model. Let ε_i have Gaussian distribution with mean 0 and variance σ_i^2 for $i \in [N]$. Then, the resulting model is known as the Probit Model. In the homogenous setting, we shall assume that $\sigma_i^2 = \sigma^2$ for all $i \in [N]$.

[1] We shall assume that the distribution of ε_i, $i \in [N]$ have *densities* and hence ties never happen between Y_i, Y_j for any $i \neq j \in [N]$.

Multinomial Logit (MNL) Model. Let ε_i have Gumbel distribution with mode μ_i and scaling parameter $\beta_i > 0$, i.e. the PDF of ε_i is given by

$$f(x) = \frac{1}{\beta_i} \exp(-(z + \exp(-z))), \text{ where } z = \frac{x - \mu_i}{\beta_i}, \text{ for } x \in \mathbb{R}. \tag{1.8}$$

In the homogenous setting, $\mu_i = \mu$ and $\beta_i = \beta$ for all $i \in [N]$. In this scenario, the resulting distribution over permutation turns out to be equivalent to the following generative model.

Let $w_i > 0$ be parameter associated with $i \in [N]$. Then the probability of permutation $\sigma \in S_N$ is given by (for example, see (Marden 1995))

$$\mathbb{P}(\sigma) = \prod_{j=1}^{N} \frac{w_{\sigma^{-1}(j)}}{w_{\sigma^{-1}(j)} + w_{\sigma^{-1}(j+1)} + \cdots + w_{\sigma^{-1}(N)}}. \tag{1.9}$$

Above, $\sigma^{-1}(j) = i$ iff $\sigma(i) = j$. Specifically, for $i \neq j \in [N]$,

$$\mathbb{P}(\sigma(i) > \sigma(j)) = \frac{w_i}{w_i + w_j}. \tag{1.10}$$

We provide a simple explanation of the above, seemingly mysterious, relationship between two very different descriptions of the MNL model.

LEMMA 1.1 *Let $\varepsilon_i, \varepsilon_j$ be independent random variables with Gumbel distribution with mode μ_i, μ_j respectively with scaling parameters $\beta_i = \beta_j = \beta > 0$. Then, $\Delta_{ij} = \varepsilon_i - \varepsilon_j$ has Logistic distribution with parameters $\mu_i - \mu_j$ (location) and β (scale).*

The proof of Lemma 1.1 follows by, for example, using the characteristic function associated with Gumbel distribution along with property of Gamma function ($\Gamma(1+z)\Gamma(1-z) = z\pi/sin(\pi z)$) and then identifying characteristic function of Logistic distribution.

Back to our model, when we compare the random utilities associated with options i and j, Y_i and Y_j respectively, we assume the corresponding random perturbation to be homogenous, i.e. $\mu_i = \mu_j = \mu$ and $\beta_i = \beta_j = \beta > 0$. Therefore, Lemma 1.1 suggests that

$$\begin{aligned}
\mathbb{P}\big(Y_i > Y_j\big) &= \mathbb{P}\big(\varepsilon_i - \varepsilon_j > u_j - u_i\big) \\
&= \mathbb{P}\big(\mathsf{Logistic}(0, \beta) > u_j - u_i\big) \\
&= 1 - \mathbb{P}\big(\mathsf{Logistic}(0, \beta) < u_j - u_i\big) \\
&= 1 - \frac{1}{1 + \exp\big(-\frac{u_j - u_i}{\beta}\big)} \\
&= \frac{\exp\big(\frac{u_i}{\beta}\big)}{\exp\big(\frac{u_i}{\beta}\big) + \exp\big(\frac{u_j}{\beta}\big)} \\
&= \frac{w_i}{w_i + w_j},
\end{aligned} \tag{1.11}$$

where $w_i = \exp\big(\frac{u_i}{\beta}\big),\ w_j = \exp\big(\frac{u_j}{\beta}\big)$.

Learning the model and ranking. For random utility model, the question of learning model from data effectively boils down to learning the model parameters from observations. In the context of homogenous model, i.e. ε_i in (1.7) have identical distribution across all $i \in [N]$, the primary interest is in learning the inherent utility parameters u_i, for $i \in [N]$. The question of recovering ranking, on the other hand, is about recovering $\sigma \in S_N$ which is the sorted (decreasing) order of the inherent utilities $u_i, i \in [N]$: for example, if $u_1 \geq u_2 \geq \cdots \geq u_N$, then the ranking is the identity permutation.

1.4 Sparse Model

In this section, we describe the conditions under which we can learn the underlying sparse distribution using first-order marginal and comparison marginal information. We divide the presentation into two parts: first, we consider access to exact or noise-less marginals for exact recovery; and then, we discuss its robustness.

We can recover ranking in terms of the most-likely permutation once we have recovered the sparse model by simply sorting the likelihoods of the permutations in the support of the distribution, which require time $O(K \log K)$ where K is the sparsity of the model. Therefore, the key question in the context of sparse model is recovery of distribution, which we shall focus in the remainder of this section.

1.4.1 Exact marginals: infinite samples

We are interested in understanding when is it feasible to recover underlying distribution λ given access to its marginal information $M(\lambda)$ or $C(\lambda)$. As mentioned earlier, one way to recover such a distribution using exactly marginal information is to solve (1.6) with equality constraint of $P(\lambda) = D$ where $P(\lambda) \in \{M(\lambda), C(\lambda)\}$ depending upon the type of marginal information.

We can view the unknown λ as a high-dimensional vector in $\mathbb{R}^{N!}$ which is sparse. That is, $\|\lambda\|_0 \ll N!$. The observations are marginals of λ, either first-order marginals $M(\lambda)$ or comparison marginals $C(\lambda)$. They can be viewed as linear projections of the λ vector of dimension N^2 or $N(N-1)$. Therefore, recovering sparse model from marginal information boils down to recovering a sparse vector in high-dimensional space (here $N!$ dimensional) based on a small number of linear measurements of the sparse vector. That is, we wish to recover $x \in \mathbb{R}^n$ from observation $y = Ax$ where $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m \times n}$ with $m \ll n$. In the best case, one can hope to recover x uniquely as long as $m \sim \|x\|_0$.

Such a question has been well studied in the context of sparse model learning from linear measurements of the signal in the signal processing and it has been popularized under the umbrella term of *compressed sensing*, see for example (Candes & Tao 2005, Candes, Romberg & Tao 2006*a*, Candes & Romberg 2006,

Candes, Romberg & Tao 2006*b*, Donoho 2006). It has been argued that such recovery is possible as long as A satisfies certain conditions, for example *Restricted Isoperimetric Property* (RIP) (see (Candes et al. 2006*a*, Berinde, Gilbert, Indyk, Karloff & Strauss 2008)) with $m \sim K \log n/K$, then the ℓ_0 optimization problem,

$$\min_z \|z\|_0 \quad \text{over} \quad y = Az,$$

recovers the true signal x as long as the sparsity of x, $\|x\|_0$ is at most K. The remarkable fact about an RIP-like condition is that it not only recovers the sparse signal using the ℓ_0 optimization, but it can be done using computationally efficient procedure, a linear program.

Impossibility of recovering distribution even for $N = 4$. Back to our setting, $n = N!$, $m = N^2$ and we wish to understand up to what level of sparsity of λ can we recover it. The key difference here is in the fact that A is *not designed* but *given*. Therefore, the question is whether A has *nice* property such as RIP that can allow for sparse recovery. To that end, consider the following simple counter example that shows that it is impossible to recover a sparse model uniquely even with support size 3 using the ℓ_0 optimization (Jagabathula & Shah 2009, Jagabathula & Shah 2011).

Example 1.4.1.1 (Impossibility) For $N = 4$, consider the four permutations $\sigma_1 = [1 \to 2, 2 \to 1, 3 \to 3, 4 \to 4]$, $\sigma_2 = [1 \to 1, 2 \to 2, 3 \to 4, 4 \to 3]$, $\sigma_3 = [1 \to 2, 2 \to 1, 3 \to 4, 4 \to 3]$ and $\sigma_4 = \mathsf{id} = [1 \to 1, 2 \to 2, 3 \to 3, 4 \to 4]$, i.e. the identity permutation. It is easy to check that

$$M(\sigma_1) + M(\sigma_2) = M(\sigma_3) + M(\sigma_4).$$

Now suppose that $\lambda(\sigma_i) = p_i$, where $p_i \in [0, 1]$ for $1 \leq i \leq 3$, and $\lambda(\sigma) = 0$ for all other $\sigma \in S_N$. Without loss of generality, let $p_1 \leq p_2$. Then,

$$p_1 M(\sigma_1) + p_2 M(\sigma_2) + p_3 M(\sigma_3) = (p_2 - p_1) M(\sigma_2) + (p_3 + p_1) M(\sigma_3) + p_1 M(\sigma_4).$$

Here, note that $\{M(\sigma_1), M(\sigma_2), M(\sigma_3)\}$ are linearly independent, yet the sparsest solution is not unique. Therefore, it is not feasible to recover sparse model uniquely.

Note that the above example, can be extended for any $N \geq 4$ by simply having identity permutation for all elements larger than 4 in the above example. Therefore, for any N with support size 3, we can not always recover them uniquely.

Signature condition for recovery. The example 1.4.1.1 suggests that it is not feasible to expect RIP-like condition for the 'projection matrix' corresponding to the first-order marginals or comparison marginals so that *any* sparse probability distribution can be recovered. The next best thing we can hope for is the ability to recover *almost all* of the sparse probability distributions. This leads us to the *signature condition* of the matrix for a given sparse vector which, as we shall see, allows for recovery of the particular sparse vector (Jagabathula & Shah 2009, Jagabathula & Shah 2011).

CONDITION 1.2 (Signature Conditions) Given matrix $A \in \mathbb{R}^{m\times n}$ is said to satisfy *signature codition* with respect to index set $S \subset \{1,\dots,n\}$ if for each $i \in S$, there exists $j(i) \in [m]$ such that $A_{j(i)i} \neq 0$ and $A_{j(i)i'} = 0$ for all $i' \neq i, i' \in S$.

The signature condition allows for recovery of sparse vector using a simple 'peeling' algorithm. We summarize the recovery result followed by algorithm that will imply the result.

THEOREM 1.3 *Let $A \in \{0,1\}^{m\times n}$ with all of its columns being distinct. Let $x \in \mathbb{R}^n_{\geq 0}$ be such that A satisfies signature condition with respect to the set $\mathsf{supp}(x)$. Let the non-zero components of x, i.e. $\{x_i : i \in \mathsf{supp}(x)\}$ be such that for any two distinct $S_1, S_2 \subset \mathsf{supp}(x)$, $\sum_{i\in S_1} x_i \neq \sum_{i'\in S_2} x_{i'}$. Then, x can be recovered from y where $y = Ax$.*

Proof. To establish Theorem 1.3, we shall describe the algorithm that recovers x under conditions of the theorem and simultaneously argue its correctness. To that end, the algorithm starts by sorting components of y. Since $A \in \{0,1\}^{m\times n}$, for each $j \in [m]$, $y_j = \sum_{i\in S_j} x_i$, with $S_j \subseteq \mathsf{supp}(x)$. Due to signature condition, for each $i \in \mathsf{supp}(x)$, there exists $j(i) \in [m]$ such that $y_{j(i)} = x_i$. If we can identify $j(i)$ for each i, we recover the values x_i, but not necessarily the position i. To identify position, we identify the ith column of matrix A, and since columns of matrix A are all distinct, this will help us identify the position. This will require use of property that for any $S_1 \neq S_2 \subset \mathsf{supp}(x)$, $\sum_{i\in S_1} x_i \neq \sum_{i'\in S_2} x_{i'}$. This implies, to begin with, all non-zero elements of x are distinct. Without loss of generality, let the non-zero elements of x be $x_1,\dots,x_K$ with $K = |\mathsf{supp}(x)|$ such that $0 < x_1 < \cdots < x_K$.

Now consider the smallest, non-zero element of y. Let it be y_{j_1}. From property of x, it follows that y_{j_1} must be the smallest non-zero element of x, x_1. The second smallest component of y that is distinct from x_1, let it be y_{j_2}, must be x_2. The third distinct smallest component, y_{j_3} however could be $x_1 + x_2$ or x_3. Since we know x_1, x_2, and due to property of x, we can identify whether y_{j_3} is x_3 or not. Iteratively, we consider the kth distinct smallest value of y, say y_{j_k}. Then, it either equals sum of subset of already identified components of x or the next smallest unidentified component of x due to signature property and non-negativity of x. In summary, by the time we are done going through all non-zero components of y in the increasing order as described above, we will recover all the non-zero elements of x in the increasing order as well as the corresponding columns of A. This is because iteratively we identify for each y_j the set $S_j \subset \mathsf{supp}(x)$ such that $y_j = \sum_{i\in S_j} x_i$. That is, $A_{ji} = 1$ for all $i \in S_j$ and 0 otherwise. This completes the proof. □

Now we remark on the computation complexity of 'peeling' algorithm described above. It runs in at most m iterations. In each iteration, it tries to effectively solve a subset sum problem whose computation cost is at most $O(2^K)$

where $K = \|x\|_0$ is the sparsity of x. In addition, the additional step for sorting components of y costs $O(m \log m)$. In summary, the computation cost is $O(2^K + m \log m)$. Notice that, somewhat surprisingly this does not depend on n at all. In contrast, the linear programing based approach used for sparse signal recovery in the context of compressed sensing literature, the computational complexity scales, at least linearly in n, the ambient dimension of the signal. For example, if this were applicable to our setting, it would scale as $N!$ which is simply prohibitive.

Recovering distribution using first-order marginals via signature condition. We shall utilize the signature condition 1.2 in the context of recovering distribution over permutations from its first-order marginals. Again, given the counter example 1.4.1.1, it is not feasible to recover *all* sparse models even with sparsity 3 from first-order marginals uniquely. However, with aid of signature condition, we will argue that it is feasible to recover *most* sparse models with reasonably large sparsity.

To that end, let $A^f \in \{0,1\}^{N^2 \times N!}$ denote the first-order marginal matrix that maps the $N!$ dimensional vector corresponding to distribution over permutations to N^2 dimensional vector corresponding to the first-order marginals of the distribution. We state the signature property of A^f next.

LEMMA 1.4 *Let S be a randomly chosen subset of $\{1, \ldots, N!\}$ of size K. Then the first-order marginal matrix A^f satisfies signature condition with respect to S with probability $1 - o(1)$ as long as $K \leq (1-\epsilon)N \log N$ for any $\epsilon > 0$.*

The proof of the above Lemma can be found in (Jagabathula & Shah 2009, Jagabathula & Shah 2011). The Lemma 1.4 and Theorem 1.3 immediately imply the following result.

THEOREM 1.5 *Let $S \subset S_N$ be a randomly chosen subset of S_N of size K, denoted as $S = \{\sigma_1, \ldots, \sigma_K\}$. Let $p_1, \ldots, p_K$ be chosen from a joint distribution with a continuous density over subspace of $[0,1]^K$ corresponding to $p_1 + \cdots + p_K = 1$. Let λ be distribution over S_N such that*

$$\lambda(\sigma) = \begin{cases} p_k & \text{if } \sigma = \sigma_k,\ k \in [K] \\ 0 & \text{otherwise.} \end{cases} \tag{1.12}$$

Then, λ can be recovered from its first-order marginal distribution with probability $1 - o(1)$ as long as $K \leq (1-\epsilon)N \log N$ for a fixed $\epsilon > 0$.

The proof of Theorem 1.5 can be found in (Jagabathula & Shah 2009, Jagabathula & Shah 2011). In a nutshell, it states that most sparse distribution over permutations with sparsity up to $N \log N$ can be recovered from its first-order marginals. This is in sharp contrast with counter example 1.4.1.1 which states that for any N, distribution with sparsity 3 can not be recovered uniquely.

Recovering distribution using comparison marginals via signature condition. Next, we utilize the signature condition 1.2 in the context of recovering distribution

over permutations from its comparison marginals. Let $A^c \in \{0,1\}^{N(N-1)\times N!}$ denote the comparison marginal matrix that maps the $N!$ dimensional vector corresponding to distribution over permutations to N^2 dimensional vector corresponding to the comparison marginals of the distribution. We state the signature property of A^c next.

LEMMA 1.6 *Let S be a randomly chosen subset of $\{1,\ldots,N!\}$ of size K. Then the comparison marginal matrix A^c satisfies signature condition with respect to S with probability $1-o(1)$ as long as $K = o(\log N)$.*

The proof of the above Lemma can be found in (Farias et al. 2009, Farias et al. 2013). The Lemma 1.6 and Theorem 1.3 immediately imply the following result.

THEOREM 1.7 *Let $S \subset S_N$ be a randomly chosen subset of S_N of size K, denoted as $S = \{\sigma_1,\ldots,\sigma_K\}$. Let $p_1,\ldots,p_K$ be chosen from a joint distribution with a continuous density over subspace of $[0,1]^K$ corresponding to $p_1+\cdots+p_K = 1$. Let λ be distribution over S_N such that*

$$\lambda(\sigma) = \begin{cases} p_k & \text{if } \sigma = \sigma_k,\ k \in [K] \\ 0 & \text{otherwise.} \end{cases} \tag{1.13}$$

Then, λ can be recovered from its first-order marginal distribution with probability $1-o(1)$ as long as $K = o(\log N)$.

The proof of Theorem 1.7 can be found in (Farias et al. 2009, Farias et al. 2013). It suggests that it is feasible to recover sparse model with growing support size with N as long as it is $o(\log N)$. However, it is exponentially smaller than recoverable support size compared to first-order marginal. This seem to be related to the fact that first-order marginal relatively information rich compared to the comparison marginal.

1.4.2 Noisy marginals: finite samples

Thus far, we have considered setup where we had access to exact marginal distribution information. Instead, suppose we have access to marginal distributions formed based on empirical distribution of finite samples from the underlying distribution. This can be viewed as access to "noisy" marginal distribution. Specifically, given distribution λ, we observe $D = P(\lambda)+\eta$ where $P(\lambda) \in \{M(\lambda), C(\lambda)\}$ depending upon the type of marginal information and η being noise such that some norm of η, e.g. $\|\eta\|_2$ or $\|\eta\|_\infty$ is bounded above by δ, with $\delta > 0$ being small if we have access to enough samples. The δ represents the error observed due to access to finitely many samples and is assumed to known.

For example, if we have access to n independent samples for each marginal entry (e.g. i ranked in position j for first-order marginal or i compared better than j for comparison marginal) as per λ, and we create empirical estimation of each entry in $M(\lambda)$ or $C(\lambda)$, then using Chernoff Bound for Binomial distribution

and Union bound over a collection of event, it can be argued that $\|\eta\|_\infty \leq \delta$ with probability $1-\delta$ as long as $n \sim \frac{1}{\delta^2}\log\frac{4N}{\delta}$ for each entry. Using more sophisticated method from Matrix Estimation literature, it is feasible to obtain better estimation of $M(\lambda)$ or $C(\lambda)$ from fewer samples of entries and even when some of the entries are entirely unobserved as long as $M(\lambda)$ or $C(\lambda)$ has *structure*. This is beyond the scope of this exposition, however we refer an interested reader to see (Chatterjee et al. 2015, Song, Lee, Li & Shah 2016, Borgs, Chayes, Lee & Shah 2017, Shah, Balakrishnan, Guntuboyina & Wainwright 2016) as well as discussion in Section 1.6.4.

Given this, the goal is to recover *sparse* distribution whose marginals are close to the observations. Precisely, we wish to find distribution $\hat{\lambda}$ such that $\|P(\hat{\lambda}) - D\|_2 \leq f(\delta)$ and $\|\hat{\lambda}\|_0$ is small. Here, ideally we would like $f(\delta) = \delta$ but we may settle for any f such that $f(\delta) \to 0$ as $\delta \to 0$.

Following the line of reasoning in Section 1.4.1, we shall assume that there is a sparse model λ^{s} with respect to which the marginal matrix satisfies *signature condition* 1.2 and $\|P(\lambda^{\mathsf{s}}) - D\|_2 \leq \delta$. The goal would be produce estimate $\hat{\lambda}$ so that $\|\hat{\lambda}\|_0 = \|\lambda\|_0$ and $\|P(\hat{\lambda}) - D\|_2 \leq f(\delta)$.

This is the exact analog of the robust recovery of sparse signal in the context of compressed sensing where the RIP-like condition allowed recovery of sparse approximation to the original signal from linear projections through linear optimization. The computational complexity of such an algorithm scales, at least linearly in n, the ambient dimension of the signal. As discussed earlier, in our context this would lead to computation cost scaling as $N!$, which is prohibitive. The exact recovery algorithm discussed in Section 1.4.1 has computation cost $O(2^K + N^2 \log N)$ in the context of recovering sparse model satisfying signature condition. The brute force search for sparse model will lead to cost at least $\binom{N!}{K} \approx (N!)^K$ or $\exp(\Theta(KN\log N))$ for $K \ll N!$. The question is, if it is possible to get rid of dependence on $N!$, and ideally scaling of $O(2^K + N^2 \log N)$ as in the case of exact model recovery.

In what follows, we describe conditions on noise under which the algorithm described in Section 1.4.1 is robust. This requires assumption that underlying ground truth distribution is sparse and satisfies signature condition. This recovery result requires noise to be *small*. Such a recovery in a higher noise regime remains broadly remains open; initial progress towards it is made in (Farias, Jagabathula & Shah 2012).

Robust recovery under signature condition: low noise regime. Recall that the 'peeling' algorithm recovers the sparse model when *signature condition* is satisfied using exact marginals. Here, we discuss robustness of the 'peeling' algorithm under noise. Specifically, we argue that 'peeling' algorithm as described is robust as long as noise is 'low'. We formalize this in the statement below.

THEOREM 1.8 *Let $A \in \{0,1\}^{m\times n}$ with all of its columns being distinct. Let $x \in \mathbb{R}^n_{\geq 0}$ be such that A satisfies signature condition with respect to the set $\mathsf{supp}(x)$. Let the non-zero components of x, i.e. $\{x_i : i \in \mathsf{supp}(x)\}$ be such that*

for any $S_1 \neq S_2 \subset \mathsf{supp}\,(x)$,

$$\big|\sum_{i\in S_1} x_i - \sum_{i'\in S_2} x_{i'}\big| > 2\delta K, \tag{1.14}$$

for some $\delta > 0$. *Then, given* $y = Ax + \eta$ *with* $\|\eta\|_\infty < \delta$, *it is feasible to find* $\hat{x}$ *so that* $\|\hat{x} - x\|_\infty \leq \delta$.

Proof To establish Theorem 1.8, we shall utilize effectively the same algorithm as that utilized for establishing Theorem 1.3 in the proof of Theorem 1.3. However, we will have to deal with the 'error' in measurement y delicately.

To begin with, following arguments in the proof of Theorem 1.3, it follows that all non-zero elements of x are distinct. Without loss of generality, let the non-zero elements of x be $x_1, \ldots, x_K$ with $K = |\mathsf{supp}\,(x)| \geq 2$ such that $0 < x_1 < \cdots < x_K$; $x_i = 0$ for $K+1 \leq i \leq n$. From (1.14), it follows for that $x_{i+1} \geq x_i + 4\delta$ for $1 \leq i < K$ and $x_1 \geq 2\delta$. Therefore,

$$x_k \geq (k-1)4\delta + 2\delta, \tag{1.15}$$

for $1 \leq k \leq K$. Next, we shall argue that, inductively, it is feasible to find $\hat{x}_i$, $1 \leq i \leq n$ so that $|\hat{x}_i - x_i| \leq \delta$ for $1 \leq i \leq K$ and $\hat{x}_i = 0$ for $K+1 \leq i \leq n$.

Now since $A \in \{0,1\}^{m\times n}$, for each $j \in [m]$, $y_j = \sum_{i\in S_j} x_i + \eta_j$, with $S_j \subseteq \mathsf{supp}\,(x)$ and $|\eta_j| \leq \delta$. From (1.14), it follows for $x_1 > 2\delta$. Therefore, if $S_j \neq \emptyset$ then $y_j > \delta$. That is, we will start by restricting to indices $J^1 = \{j \in [m] : y_j > \delta\}$.

Let j_1 be index in J such that $y_{j_1} \in \arg\min_{j\in J^1} y_j$. We set $\hat{x}_1 = y_{j_1}$ and $A_{j_1 1} = 1$. To justify this, we next argue that $y_{j_1} = x_1 + \eta_{j_1}$ and hence $|\hat{x}_1 - x_1| < \delta$. By signature condition, for each $i \in \mathsf{supp}\,(x)$, there exists $j(i) \in [m]$ such that $|y_{j(i)} - x_i| \leq \delta$ and hence $j(i) \in J^1$ since $y_{j(i)} \geq x_i - \delta > \delta$. Let $J(1) = \{j \in J^1 : y_j = x_1 + \eta_j\}$. Clearly, $J(1) \neq \emptyset$. Effectively, we want to argue that $j_1 \in J(1)$. To that end, suppose not. Then, there exists $S \subset \mathsf{supp}\,(x)$, such that $S \neq \emptyset$, $S \neq \{1\}$ and $y_{j_1} = \sum_{i\in S} x_i + \eta_{j_1}$. Then

$$\begin{aligned} y_{j_1} &> \sum_{i\in S} x_i - \delta, \quad \text{since } |\eta_{j_1}| < \delta, \\ &> x_1 + 2K\delta - \delta, \quad \text{by (1.14)}, \\ &\geq x_1 + \delta, \quad \text{since } K \geq 1, \qquad (1.16) \\ &\geq y_j, \qquad (1.17) \end{aligned}$$

for any $j \in J(1) \subset J^1$. But this is a contradiction since $y_{j_1} \leq y_j$ for all $j \in J$. That is, $S = \{1\}$ or $y_{j_1} = x_1 + \eta_{j_1}$. Thus, we have found $\hat{x}_1 = y_{j_1}$ such that $|\hat{x}_1 - x_1| < \delta$.

Now for any $j \in J^1$ with $y_j = \sum_{i\in S} x_i + \eta_j$ with $S \cap \{2, \ldots, K\} \neq \emptyset$, we have

with notation $x(S) = \sum_{i \in S} x_i$

$$\begin{aligned} |\hat{x}_1 - y_j| &= |x_1 - y_j + \hat{x}_1 - x_1| \\ &= |x_1 - x(S) - \eta_j + \hat{x}_1 - x_1| \\ &\geq |x_1 - x(S)| - |\eta_j| - |\hat{x}_1 - x_1| \\ &> 2K\delta - \delta - \delta \\ &\geq 2\delta. \end{aligned} \tag{1.18}$$

And if $S = \{1\}$, then $|\hat{x}_1 - y_j| < 2\delta$. Therefore, we set

$$A_{j1} = 1 \quad \text{if} \quad |\hat{x}_1 - y_j| < 2\delta, \text{ for } j \in J^1,$$

and

$$J^2 \leftarrow J^1 \backslash \{j \in J^1 : |y_j - \hat{x}_1| < 2\delta\}.$$

Clearly,

$$j \in J^2 \;\Leftrightarrow\; j \in [m],\; y_j = x(S) + \eta_j, \text{ such that } S \cap \{2, \ldots, K\} \neq \emptyset.$$

Now suppose, inductively we have found $\hat{x}_1, \ldots, \hat{x}_k$, $1 \leq k < K$ so that $|\hat{x}_i - x_i| < \delta$ for $1 \leq i \leq k$ and

$$j \in J^{k+1} \;\Leftrightarrow j \in [m],\; y_j = x(S) + \eta_j, \text{ such that } S \cap \{k+1, \ldots, K\} \neq \emptyset.$$

To establish inductive step, we suggest to set $j_{k+1} \in \arg\min_{j \in J^{k+1}} \{y_j\}$, $\hat{x}_{k+1} = y_{j_{k+1}}$ and

$$J^{k+2} \leftarrow J^{k+1} \backslash \{j \in J^{k+1} : |y_j - \hat{x}_1| < (k+1)\delta\}.$$

We shall argue that $|\hat{x}_{k+1} - x_{k+1}| < \delta$ by showing that $y_{j_{k+1}} = x_{k+1} + \eta_{j_{k+1}}$ and establishing

$$j \in J^{k+2} \;\Leftrightarrow\; j \in [m],\; y_j = x(S) + \eta_j, \text{ such that } S \cap \{k+2, \ldots, K\} \neq \emptyset.$$

To that end, let $y_{j_{k+1}} = x(S) + \eta_{j_{k+1}}$. By inductive hypothesis, $S \subset \mathsf{supp}\,(x)$ and $S \cap \{k+1, \ldots, K\} \neq \emptyset$. Suppose $S \neq \{k+1\}$. Then,

$$\begin{aligned} y_{j_{k+1}} &> x(S) - \delta \\ &= (x(S) - x_{k+1}) + x_{k+1} - \delta \\ &\geq 2\delta + x_{k+1} - \delta \\ &= \delta + x_{k+1}, \\ &> y_j, \end{aligned}$$

for any $j \in J(k+1) \equiv \{j \in J^{k+1} : y_j = x_{k+1} + \eta_j\}$. In above, we have used the fact that since $S \cap \{k+1, \ldots, K\} \neq \emptyset$ and $S \neq \{k+1\}$, it must be that $x(S) \geq \min\{x_1 + x_{k+1}, x_{k+2}\}$. In either case, using (1.15), it follows that $x(S) - x_{k+1} \geq 2\delta$. We note that due to signature condition and inductive hypothesis about J^{k+1}, it follows that $J(k+1) \neq \emptyset$. But $y_{j_{k+1}}$ is the minimal value of y_j for $j \in J^{k+1}$. This is a contradiction. Therefore, $S = \{k+1\}$. That is, $\hat{x}_{k+1} = y_{j_{k+1}}$ satisfies $|\hat{x}_{k+1} - x_{k+1}| < \delta$.

Now, consider any set $S \subset \{1, \ldots, k+1\}$ and any $j \in J^{k+2}$ such that $y_j = \sum_{i \in S'} x_i + \eta_j$ with $S' \cap \{k+2, \ldots, K\} \neq \emptyset$. Using notation $\hat{x}(S) = \sum_{i \in S} \hat{x}_i$, we have

$$\begin{aligned} |\hat{x}(S) - y_j| &= |x(S) - y_j + \hat{x}(S) - x(S)| \\ &= |x(S) - x(S') - \eta_j + \hat{x}(S) - x(S)| \\ &\geq |x(S) - x(S')| - |\eta_j| - |\hat{x}(S) - x(S)| \\ &> 2K\delta - (1 + |S|)\delta \\ &\geq (|S| + 1)\delta, \end{aligned} \tag{1.19}$$

where we used the fact that $|S| + 1 \leq K$. Therefore, if we set

$$J^{k+2} \leftarrow J^{k+1} \backslash \{j \in J^{k+1} : |y_j - \hat{x}(S)| \leq (|S|+1)\delta, \text{ for some } S \subset \{1, \ldots, k+1\}\},$$

it follows that

$$j \in J^{k+2} \;\Leftrightarrow\; j \in [m],\; y_j = x(S) + \eta_j, \text{ such that } S \cap \{k+2, \ldots, K\} \neq \emptyset.$$

This completes the induction step. And establishes the desired result that we can recover $\hat{x}$ so that $\|\hat{x} - x\|_\infty \leq \delta$. □

Naturally, as before, Theorem 1.8 implies *robust* versions of Theorems 1.5 and 1.7. In particular, if we are forming empirical estimation of $M(\lambda)$ or $C(\lambda)$ based on independently drawn samples, then by simple application of Chernoff bound along with a union bound will imply that it may be sufficient to have samples that scale as $\delta^{-2} \log N$ to have $\hat{M}(\lambda)$ or $\hat{C}(\lambda)$ so that $\|\hat{M}(\lambda) - M(\lambda)\|_\infty < \delta$ or $\|\hat{C}(\lambda) - C(\lambda)\|_\infty < \delta$ with high probability (i.e. $1 - o_N(1)$). Then, as long as λ satisfies condition (1.14) in addition to the signature condition, then Theorem 1.8 guarantees approximate recovery as discussed above.

Robust recovery under signature condition: high noise regime. Theorem 1.8 provides conditions under which 'peeling' algorithm manages to recover the distribution as long as the element in the support are far enough. Put it other way, for a given x, the error tolerance needs to be *small enough* compared to the *gap* that is implicitly defined by (1.14) for recovery to be feasible.

Here, we make an attempt to go beyond such restrictions. In particular, assuming that the noisy observations come from a signature family, we will be satisfied if we recover *any* signature family that is consistent with observations. For this, we shall assume the knowledge of sparsity K.

Now, we need to learn $\mathsf{supp}\,(x)$, i.e. positions of x that are non-zero and the non-zero values in those positions. The determination of $\mathsf{supp}\,(x)$ corresponds to selecting the columns of A. Now if A satisfies signature condition with respect to $\mathsf{supp}\,(x)$, then we can simply choose the entries in the positions of y corresponding to the signature component. If the choice of $\mathsf{supp}\,(x)$ is correct then this will provide estimate $\hat{x}$ so that $\|\hat{x} - x\|_2 \leq \delta$. In general, if we assume that there exists x such that A satisfies signature condition with respect to $\mathsf{supp}\,(x)$ with

$K = \|x\|_0$ and $\|y - Ax\|_2 \leq \delta$, then an approach is to find $\hat{x}$ such that $\|\hat{x}\|_0 = K$, A satisfies signature condition with respect to $\mathsf{supp}\,(\hat{x})$ and it minimizes $\|y - A\hat{x}\|_2$.

In summary, we are solving a combinatorial optimization problem over the space of columns of A that collectively satisfy the signature condition. Formally, the space of subsets of columns of A of size K can be encoded through a binary valued matrix $Z \in \{0,1\}^{m \times m}$ as follows: all but K columns of Z are zero, and the non-zero columns of Z are distinct columns of A collectively satisfying the signature condition. Precisely, for any $1 \leq i_1 < i_2 < \cdots < i_K \leq m$, representing as the signature columns, the variable Z should satisfy

$$Z_{i_j i_j} = 1, \text{ for } 1 \leq j \leq K \tag{1.20}$$

$$Z_{i_j i_k} = 0, \text{ for } 1 \leq j \neq k \leq K \tag{1.21}$$

$$[Z_{ai_j}]_{a \in [m]} \in \mathsf{col}(A), \text{ for } 1 \leq j \leq K \tag{1.22}$$

$$Z_{ab} = 0, \text{ for } a \in [m],\ b \notin \{i_1, \ldots, i_K\}. \tag{1.23}$$

In above, $\mathsf{col}(A) = \{[A_{ij}]_{i \in [m]} : 1 \leq j \leq n\}$ represents the set of columns of matrix A. Then, the optimization problem of interest is

$$\begin{aligned} &\mathsf{minimize}\ \|y - Zy\|_2 \quad \mathsf{over} \quad Z \in \{0,1\}^{m \times m} \\ &\mathsf{such\ that}\ Z\ \mathsf{satisfies\ constraints}\ (1.20) - (1.23). \end{aligned} \tag{1.24}$$

The constraint set (1.20)-(1.23) can be viewed as disjoint union of $\binom{m}{K}$ sets each one corresponding to choice of $1 \leq i_1 < \cdots < i_K \leq m$. For each such choice, we can solve the optimization (1.24) and choose the best solution across all of them. That is, the computation cost is $O(m^K)$ times the cost of solving the optimization problem (1.24). The complexity of solving (1.24) fundamentally depends on the constraint (1.22) – it captures the structural complexity of describing the column set of matrix A.

A natural convex relaxation of the optimization problem (1.24) involves replacing (1.22) and $Z \in \{0,1\}^{m \times m}$ by

$$[Z_{ai_j}]_{a \in [m]} \in \mathsf{convex\text{-}hull}(\mathsf{col}(A)), \text{ for } 1 \leq j \leq K; \quad Z \in [0,1]^{m \times m}. \tag{1.25}$$

In above, for any set S,

$$\mathsf{convex\text{-}hull}(S) \equiv \Big\{ \sum_{\ell=1}^{Q} a_\ell x_\ell : a_\ell \geq 0, x_\ell \in S \text{ for } \ell \in [Q],\ \sum_{\ell} a_\ell = 1, \text{ for } Q \geq 2. \Big\}.$$

In the best case, it may be feasible to solve the optimization with the convex relaxation efficiently. However, the relaxation may not yield solution that is achieved at the extreme points of the $\mathsf{convex\text{-}hull}(\mathsf{col}(A))$ which is what we desire. This is due to the fact that the objective, ℓ_2 norm of error, we are considering is strictly convex. To overcome this challenge, we can replace ℓ_2 by ℓ_1. And the

constraints of interest are, for a given $\varepsilon > 0$ of

$$Z_{i_j i_j} = 1, \text{ for } 1 \le j \le K \tag{1.26}$$

$$Z_{i_j i_k} = 0, \text{ for } 1 \le j \ne k \le K \tag{1.27}$$

$$y_i - (Zy)_i \le \varepsilon, \text{ for } 1 \le i \le m \tag{1.28}$$

$$y_i - (Zy)_i \ge -\varepsilon, \text{ for } 1 \le i \le m \tag{1.29}$$

$$[Z_{ai_j}]_{a\in[m]} \in \mathsf{convex\text{-}hull}(\mathsf{col}(A)), \text{ for } 1 \le j \le K \tag{1.30}$$

$$Z_{ab} = 0, \text{ for } a \in [m],\ b \notin \{i_1, \ldots, i_K\}. \tag{1.31}$$

This results in the Linear Program

$$\begin{aligned} &\mathsf{minimize} \sum_{i,j=1}^{m} \zeta_{ij} Z_{ij} \quad \mathsf{over} \quad Z \in [0,1]^{m\times m} \\ &\mathsf{such\ that} \quad Z \text{ satisfies constraints } (1.26) - (1.31), \end{aligned} \tag{1.32}$$

In above, $\zeta = [\zeta_{ij}] \in [0,1]^{m\times m}$ is a random vector with each of its component chosen by drawing a number from $[0,1]$ uniformly at random. The purpose of choosing ζ is to obtain unique solution, if it is feasible. Note that, when feasible, the solution is achieved at the extreme point which happens to be the valid solution of interest. We can solve (1.32) iteratively for choice of $\varepsilon = 2^{-q}$ for $q \ge 0$ till we fail to find a feasible solution. The value of ε before that will be the smallest (within factor 2) error tolerance that can is feasible within signature family. Therefore, the cost of finding such solution is within $O(\log 1/\varepsilon)$ times the cost of solving the Linear Program (1.32). The cost of Linear Program (1.32) depends on the complexity of the convex relaxation of the set $\mathsf{col}(A)$. If it is indeed simple enough, then we can solve the (1.32) efficiently.

As it turns out, for the case of *first-order* marginals, the convex hull of $\mathsf{col}(A)$ is succinct due to the classical result by Birkhoff and Von Neumann which characterizes the convex relaxation of permutation matrices though linear number of equalities in the size of permutation, here N. Each of the element in $\mathsf{col}(A)$ corresponds to (flattened) permutation matrix. Therefore, its convex hull is simply that of (flattened) permutation matrices and thus leading to succinct description. This results in polynomial time algorithm for solving (1.32). In summary, we conclude the following (see (Farias et al. 2012) for details).

THEOREM 1.9 *For given observation vector $y \in [0,1]^m$, if there exists a distribution μ in* signature familty *of support size K such that the corresponding projection is within $\varepsilon \in (0,1]$ of y in terms of ℓ_∞ norm, then it can be found through an algorithm with computation cost $O(N^{\Theta(K)} \log 1/\varepsilon)$.*

Open Question. The efficient computation (or inability to do so) for finding approximate distribution in the signature family for the pair-wise comparison marginals, equivalent to Theorem 1.9, is not known.

On Universality of Signature Family. Thus far, we have focused on developing

algorithms for learning sparse model with signature condition. The sparse model is a natural approximation for generic distribution over permutation. In Theorems 1.5 and 1.7, we effectively argued that a model with randomly chosen sparse support satisfies signature condition as long as sparsity is not too large. However, it is not clear if sparse model with signature condition is a good approximation beyond such setting. For example, is there a sparse model with signature condition that has approximates the marginal information of the simple parametric model such as the Multinomial Logit Model (MNL) well?

To that end, recently (Farias et al. 2012) have established the following representation result which we state without proof here.

THEOREM 1.10 *Let λ be a MNL model with parameters $w_1, \ldots, w_N$ (and without loss of generality, let $0 < w_1 < \cdots < w_N$) such that*

$$\frac{w_N}{\sum_{k=1}^{N-L} w_k} \leq \frac{\sqrt{\log N}}{N}, \tag{1.33}$$

for some $L = N^\delta$ for some $\delta \in (0,1)$. Then there exists $\hat{\lambda}$ such that $|\mathsf{supp}\left(\hat{\lambda}\right)| = O(N/\varepsilon^2)$, $\hat{\lambda}$ satisfies signature condition with respect to the first-order marginals and $\|M(\lambda) - M(\hat{\lambda})\|_2 \leq \varepsilon$.

1.5 Random Utility Model (RUM)

We discuss recovery of exact model for MNL model and recovery of ranking for generic random utility model with homogenous random perturbation.

1.5.1 Exact marginals: infinite samples

Given the exact marginal information $M(\lambda)$ or $C(\lambda)$ for λ, we wish to recover the parameters of the model when λ is MNL, and we wish to recover ranking when λ is generic random utility model. We first discuss recovery of MNL for both types of marginal information and then discuss recovery of ranking for generic model.

Recovering MNL: first-order marginals. Without loss of generality, let us assume that the parameters $w_1, \ldots, w_N$ are normalized so that $\sum_i w_i = 1$. Then, under the MNL model per (1.9),

$$\mathbb{P}\big(\sigma(i) = 1\big) = w_i. \tag{1.34}$$

That is, the first row of the first-order marginal matrix $M(\lambda) = [M_{ij}(\lambda)]$ precisely provides the parameters of the MNL model!

Recovering MNL: comparison marginals. Under the MNL model, as per (1.11), for any $i \neq j \in [N]$,

$$\mathbb{P}\big(\sigma(i) > \sigma(j)\big) = \frac{w_i}{w_i + w_j}. \tag{1.35}$$

The comparison marginals, $C(\lambda)$ provides access to $\mathbb{P}\big(\sigma(i) > \sigma(j)\big)$ for all $i \neq j \in [N]$. Using these, we wish to recover parameters $w_1, \ldots, w_N$.

Next, we describe a reversible Markov chain over N states whose stationary distribution is precisely the parameters of our interest and its transition Kernel utilizes the $C(\lambda)$. This alternative representation provides an intuitive algorithm for recovering the MNL parameters, and more generally what is known as the *Rank Centrality* (Negahban, Oh & Shah 2012, Negahban, Oh & Shah 2016).

To that end, the Markov chain of interest has N states. The transition Kernel or transition probability matrix $Q = [Q_{ij}] \in [0,1]^{N \times N}$ of the Markov chain is defined using comparison marginals $C = C(\lambda)$ as follows:

$$Q_{ij} = \begin{cases} \frac{C_{ji}}{2N}, \text{ if } i \neq j \\ 1 - \sum_{j \neq i} \frac{C_{ji}}{2N}, \text{ if } i = j. \end{cases} \tag{1.36}$$

The Markov chain has unique stationary distribution because (a) Q is irreducible, since $C_{ij}, C_{ji} > 0$ for all $i \neq j$ as long as $w_i > 0$ for all $i \in [N]$, and (b) $Q_{ii} > 0$ by definition for all $i \in [N]$ and hence it is aperiodic. Further $\mathbf{w} = [w_i]_{i \in [N]} \in [0,1]^N$ is a stationary distribution since it satisfies the detailed balanced condition, i.e. for any $i \neq j \in [N]$

$$\begin{aligned} w_i Q_{ij} = w_i \frac{C_{ji}}{2N} &= w_i \frac{w_j}{2N(w_i + w_j)} \\ &= w_j \frac{w_i}{2N(w_i + w_j)} = w_j \frac{C_{ij}}{2N} \\ &= w_j Q_{ji}. \end{aligned} \tag{1.37}$$

Thus, by finding the stationary distribution of Markov chain as defined above, we can find parameters of the MNL. And this boils down to finding the largest eigenvector of Q which can be done using various efficient algorithms including the standard power-iteration method.

We note that the algorithm to finding parameters of MNL does not need to have access to *all* entries of C. Suppose, $E \subset \{(i,j) : i \neq j \in [N]\}$ be a subset of all possible $\binom{N}{2}$ pairs for which we have access to C. Let us define Markov chain with Q such that for $i \neq j \in [N]$, Q_{ij} is defined as per (1.36) if $(i,j) \in E$ (we assume $(i,j) \in E$ then $(j,i) \in E$ because $C_{ji} = 1 - C_{ij}$ by definition), else $Q_{ij} = 0$; and $Q_{ii} = 1 - \sum_{j \neq i} Q_{ij}$. The resulting Markov chain is aperiodic since by definition $Q_{ii} > 0$. Therefore, as long as the resulting Markov chain is irreducible, then it has unique stationary distribution. Now the Markov chain is irreducible if effectively all N states are reachable from each other via transitions $\{(i,j),(j,i) : (i,j) \in E\}$. That is, there is data that compares any two $i \neq j \in [N]$ through potentially chain of comparisons. Which, in a sense, is a minimal requirement to have consistent ranking across all $i \in [N]$. Once we have this, again it follows that the stationary distribution is given by $\mathbf{w} = [w_i]_{i \in [N]} \in [0,1]^N$ since the detailed balanced equation (1.37) holds for all $i \neq j \in [N]$ with $(i,j) \in E$.

Recovering ranking for homogenous RUM. As mentioned in Section 1.3.2, we wish to recover ranking or ordering of inherent utilities for homogenous random utility model. That is, if $u_1 \geq \cdots \geq u_N$, then the ranking of interest is identity, i.e. $\sigma \in S_N$ such that $\sigma(i) = i$ for all $i \in [N]$. Recall that, in the homogenous RUM random perturbation ε_i in (1.7) have identical distribution for all $i \in [N]$. We shall assume that the distribution of the random perturbation is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. Operationally, for any $t_1 < t_2 \in \mathbb{R}$,

$$\mathbb{P}\big(\varepsilon_1 \in (t_1, t_2)\big) > 0. \tag{1.38}$$

The following is the key characterization of homogenous RUM with (1.38) that will enable recovery of ranking from marginal data (both comparison and first-order); also see (Ammar & Shah 2011, Ammar & Shah 2012).

LEMMA 1.11 *Consider homogenous RUM with property* (1.38)*. Then, for* $i \neq j \in [N]$*,*

$$u_i > u_j \Leftrightarrow \mathbb{P}\big(Y_i > Y_j\big) > \frac{1}{2}. \tag{1.39}$$

Further, for any $k \neq i, j \in [N]$*,*

$$u_i > u_j \Leftrightarrow \mathbb{P}\big(Y_i > Y_k\big) > \mathbb{P}\big(Y_j > Y_k\big). \tag{1.40}$$

Proof. By definition,

$$\mathbb{P}\big(Y_i > Y_j\big) = \mathbb{P}\big(\varepsilon_i - \varepsilon_j > u_j - u_i\big). \tag{1.41}$$

Since $\varepsilon_i, \varepsilon_j$ are independent and identically distribution with property (1.38), their difference random variable $\varepsilon_i - \varepsilon_j$ has 0 mean, symmetric and with property (1.38). That is, 0 is its unique median as well. That is, for any $t > 0$,

$$\mathbb{P}\big(\varepsilon_i - \varepsilon_j > t\big) = \mathbb{P}\big(\varepsilon_i - \varepsilon_j < -t\big) < \frac{1}{2}. \tag{1.42}$$

This concludes that

$$u_i > u_j \Leftrightarrow \mathbb{P}\big(Y_i > Y_j\big) > \frac{1}{2}.$$

Similarly,

$$\mathbb{P}\big(Y_i > Y_k\big) = \mathbb{P}\big(\varepsilon_i - \varepsilon_k > u_k - u_i\big), \tag{1.43}$$

$$\mathbb{P}\big(Y_j > Y_k\big) = \mathbb{P}\big(\varepsilon_j - \varepsilon_k > u_k - u_j\big). \tag{1.44}$$

Now $\varepsilon_i - \varepsilon_k$ and $\varepsilon_j - \varepsilon_k$ are identically distributed with property (1.38). That is, it has strictly monotonically increasing cumulative distribution function (CDF). Therefore, (1.40) follows immediately. □

Recovering ranking: comparison marginals. From (1.39) of Lemma 1.11, using comparison marginals $C(\lambda)$, we can recover ranking of $[N]$ that corresponds to

the ranking of their inherent utility for generic homogenous RUM as follows. For each $i \in [N]$, assign rank as

$$\mathsf{rank}(i) = N - \Big|\Big\{j \in [N] : j \neq i,\ C_{ij} > \frac{1}{2}\Big\}\Big|. \tag{1.45}$$

From Lemma 1.11, it immediately follows that the rank provides the ranking of $[N]$ as desired.

We also take a note that (1.40) of Lemma 1.11 suggests an alternative way (which will turn out to be robust and more useful) to find the same ranking. To that end, for each $i \in [N]$, define score as

$$\mathsf{score}(i) = \frac{1}{N-1}\sum_{k \neq i} C_{ik}. \tag{1.46}$$

From (1.40) of Lemma 1.11, it follows that for any $i \neq j \in [N]$,

$$\mathsf{score}(i) > \mathsf{score}(j) \Leftrightarrow u_i > u_j. \tag{1.47}$$

That is, by ordering $[N]$ in decreasing order of score values, we obtain the desired ranking.

Recovering ranking: first-order marginals. We are given first-order marginal data matrix, $M = M(\lambda) \in [0,1]^{N \times N}$ where the M_{ij} represents $\mathbb{P}(\sigma(i) = j)$ under λ for $i, j \in [N]$. To recover ranking under generic homogenous RUM using M, we shall introduce the notion of *Borda count*, cf. see (Emerson 2013). Precisely, for any $i \in [N]$

$$\mathsf{borda}(i) = \mathbb{E}[\sigma(i)] = \sum_{j \in [N]} \mathbb{P}(\sigma(i) = j)j = \sum_{j \in [N]} jM_{ij}. \tag{1.48}$$

That is, $\mathsf{borda}(i)$ can be computed using M for any $i \in [N]$. Recall that, we argued earlier that the $\mathsf{score}(\cdot)$ (in decreasing order) provides desired ordering or ranking of $[N]$. However, computing score required access to comparison marginals C. And it's not feasible to recover C from M.

On the other hand, intuitively it seems that borda (in increasing order) provides ordering over $[N]$ that might be what we want. Next, we state a simple invariant that ties $\mathsf{score}(i)$ and $\mathsf{borda}(i)$, which will lead to the conclusion that we can recover desired ranking by sorting $[N]$ in increasing order of borda count (Ammar & Shah 2011, Ammar & Shah 2012).

LEMMA 1.12 *For any $i \in [N]$ and any distribution over permutations,*

$$\mathsf{borda}(i) + (N-1)\mathsf{score}(i) = N. \tag{1.49}$$

Proof Consider any permutation $\sigma \in S_N$. For any $i \in [N]$, $\sigma(i)$ denotes the position in $[N]$ that i is ranked to. That is, $N - \sigma(i)$ is precisely the number of elements in $[N]$ (and not equal to i) that are ranked below i. Formally,

$$N - \sigma(i) = \sum_{j \neq i} \mathbf{1}(\sigma(i) > \sigma(j)). \tag{1.50}$$

Taking expectation on both sides with respect to the underlying distribution over permutations and re-arranging terms, we obtain

$$N = \mathbb{E}[\sigma(i)] + \sum_{j \neq i} \mathbb{P}(\sigma(i) > \sigma(j)). \tag{1.51}$$

Using definitions from (1.46) and (1.48), we have

$$N = \mathsf{borda}(i) + (N-1)\mathsf{score}(i). \tag{1.52}$$

This completes the proof. □

1.5.2 Noisy marginals: finite samples

Now we consider setup where we have access to marginal distributions formed based on empirical distribution of finite samples from the underlying distribution. This can be viewed as access to "noisy" marginal distribution. Specifically, given distribution λ, we observe $D = P(\lambda) + \eta$ where $P(\lambda) \in \{M(\lambda), C(\lambda)\}$ depending upon the type of marginal information and η being noise such that $\|\eta\|_2 \leq \delta$, with $\delta > 0$ being small if we have access to enough samples. The δ represents the error observed due to access to finitely many samples and is assumed to be known.

As before, we wish to recover the parameters of the model when λ is MNL, and we wish to recover ranking when λ is generic random utility model. We first discuss recovery of MNL for both types of marginal information and then discuss recovery of ranking for generic model.

Recovering MNL: first-order marginals. As discussed in Section 1.5.1, we can recover parameters of MNL, $\mathbf{w} = [w_i]_{i \in [N]}$, using the first row of first-order marginal matrix, $M = M(\lambda)$ by simply setting $w_i = M_{1i}$. Since we have access to $M_{i1} + \eta_{i1}$, a simple estimator is to set $\hat{w}_i = M_{1i} + \eta_{1i}$. Then, it follows that

$$\|\hat{\mathbf{w}} - \mathbf{w}\|_2 = \|\eta_{1\cdot}\|_2 \;\leq\; \|\eta\|_2 \;\leq\; \delta. \tag{1.53}$$

That is, using the same algorithm for estimating parameter as in the case of access to the exact marginals, we obtain an estimator which seem to have reasonably good property.

Recovering MNL: comparison marginals. We shall utilize the noisy comparison data to create a Markov chain as in Section 1.5.1. The stationary distribution of this noisy or perturbed Markov chain, will be a good approximation of the original Markov chain, i.e. the true MNL parameters. This will lead to a good estimator of MNL model using noisy comparison data.

To that end, we have access to noisy comparison marginals $\widehat{C} = C + \eta$. To keep things generic, we shall assume that we have access to comparison for subset of pairs. Let $E \subset \{(i,j) : i \neq j \in [N]\}$ denote the subset of all possible $\binom{N}{2}$ pairs for which we have access to noisy comparison marginals and we shall assume

that $(i,j) \in E$ iff $(j,i) \in E$. Define $d_i = |\{j \in [N] : j \neq i, (i,j) \in E\}$ and $d_{\max} = \max_i d_i$. Define noisy Markov chain with transition matrix $\widehat{Q}$ as

$$\widehat{Q}_{ij} = \begin{cases} \frac{\widehat{C}_{ji}}{2d_{\max}} & \text{if } i \neq j,\ (i,j) \in E, \\ 1 - \frac{1}{2d_{\max}} \sum_{j:(i,j)\in E} \widehat{C}_{ji} & \text{if } i = j, \\ 0, & \text{if } (i,j) \notin E. \end{cases} \tag{1.54}$$

We shall assume that E is such that the resulting Markov chain with transition matrix $\widehat{Q}$ is irreducible; it is aperiodic since $\widehat{Q}_{ii} > 0$ for all $i \in [N]$ by definition (1.54). As before, it can be verified that this noisy Markov chain is reversible and has unique stationary distribution that satisfies the detailed balanced condition. Let $\widehat{\pi}$ denote this stationary distribution. The corresponding ideal Markov chain has transition matrix Q defined as

$$Q_{ij} = \begin{cases} \frac{C_{ji}}{2d_{\max}} & \text{if } i \neq j,\ (i,j) \in E, \\ 1 - \frac{1}{2d_{\max}} \sum_{j:(i,j)\in E} C_{ji} & \text{if } i = j, \\ 0, & \text{if } (i,j) \notin E. \end{cases} \tag{1.55}$$

It is reversible and has unique stationary distribution $\pi = \mathbf{w}$. We want to bound $\|\widehat{\pi} - \pi\|$.

By definition of $\widehat{\pi}$ being stationary distribution of $\widehat{Q}$, we have that

$$\widehat{\pi}^T \widehat{Q} = \widehat{\pi}^T. \tag{1.56}$$

We can find $\widehat{\pi}$ using power-iteration algorithm. Precisely, let $\nu_0 \in [0,1]^N$ be a probability distribution as our initial guess. Iteratively, for iteration $t \geq 0$

$$\nu_{t+1}^T = \nu_t^T \widehat{Q}. \tag{1.57}$$

We make the following claim, cf. (Negahban et al. 2012, Negahban et al. 2016).

LEMMA 1.13 *For any $t \geq 1$,*

$$\frac{\|\nu_t - \pi\|}{\|\pi\|} \leq \left(\rho^t \frac{\|\nu_0 - \pi\|}{\|\pi\|} + \frac{1}{1-\rho}\|\Delta\|_2\right)\sqrt{\frac{\pi_{\max}}{\pi_{\min}}}. \tag{1.58}$$

Here $\Delta = \widehat{Q} - Q$, $\pi_{\max} = \max_i \pi_i$, $\pi_{\min} = \min_i \pi_i$; $\lambda_{\max}(Q)$ be the second largest (in norm) eigenvalue of Q; $\rho = \lambda_{\max}(Q) + \|\Delta\|_2\sqrt{\pi_{\max}/\pi_{\min}}$ and it is assumed that $\rho < 1$.

Before we provide the proof of this Lemma, let us consider its implications. It is quantifying the robustness of the our approach for identifying the parameters of MNL using comparison data. Specifically, since $\lim_{t\to\infty} \nu_t \to \widehat{\pi}$, (1.58) implies

$$\frac{\|\widehat{\pi} - \pi\|}{\|\pi\|} \leq \frac{1}{1-\rho}\|\Delta\|_2\sqrt{\frac{\pi_{\max}}{\pi_{\min}}}. \tag{1.59}$$

Thus, the operator or spectral norm of perturbation matrix $\Delta = \widehat{Q} - Q$ determines the error in our ability to learn parameters using the above mentioned,

Rank Centrality, algorithm. By definition

$$|\widehat{Q}_{ij} - Q_{ij}| \leq \begin{cases} \frac{\left|\widehat{C}_{ij} - C_{ji}\right|}{2d_{\max}} & \text{if } i \neq j,\ (i,j) \in E, \\ \frac{1}{2d_{\max}} \left| \sum_{j:(i,j)\in E} (\widehat{C}_{ij} - C_{ji}) \right| & \text{if } i = j, \\ 0 & \text{if } (i,j) \notin E. \end{cases} \tag{1.60}$$

Therefore, it follows that

$$\begin{aligned} \|\Delta\|_F^2 = \sum_{i,j} \Delta_{ij}^2 \ &= \ \sum_{i,j} (\widehat{Q}_{ij} - Q_{ij})^2 \\ &= \frac{1}{4d_{\max}^2} \sum_{i \neq j} (\widehat{C}_{ij} - C_{ji})^2 + \frac{1}{4d_{\max}^2} \sum_i \Big| \sum_{j:(i,j)\in E} (\widehat{C}_{ij} - C_{ji}) \Big|^2 \\ &\leq \frac{1}{4d_{\max}^2} \sum_{i,j} (\widehat{C}_{ij} - C_{ji})^2 + \frac{1}{4d_{\max}^2} \sum_i d_{\max} \sum_{j:(i,j)\in E} (\widehat{C}_{ij} - C_{ji})^2 \\ &\leq \frac{d_{\max} + 1}{4d_{\max}^2} \sum_{i,j} (\widehat{C}_{ij} - C_{ji})^2 \\ &= \frac{d_{\max} + 1}{4d_{\max}^2} \|\eta\|_F^2 \ \leq \ \frac{1}{2d_{\max}} \|\eta\|_F^2, \end{aligned} \tag{1.61}$$

where η is the error in comparison marginals, i.e. $\eta = \widehat{C} - C$. Thus, if $\|\eta\|_F^2 \leq \delta^2$, then since $\|\Delta\|_2 \leq \|\Delta\|_F$, we have that

$$\frac{\|\widehat{\pi} - \pi\|}{\|\pi\|} \leq \frac{1}{1-\rho} \frac{\delta}{2d_{\max}} \sqrt{\frac{\pi_{\max}}{\pi_{\min}}}, \tag{1.62}$$

with

$$|\rho - \lambda_{\max}(Q)| \leq \frac{\delta}{2d_{\max}} \sqrt{\pi_{\max}/\pi_{\min}}. \tag{1.63}$$

Therefore, if Q has a good spectral gap, i.e. $1 - \lambda_{\max}(Q)$ is large enough, and δ is small enough, then the estimate $\widehat{\pi}$ is a good proxy of π, the true parameters. The precisely role of "graph structure" induced by observed entries, E, comes in play in determining $\lambda_{\max}(Q)$. This, along with implications on sample complexity for random sampling model is discussed in details in (Negahban et al. 2012, Negahban et al. 2016).

Proof of Lemma 1.13. Define inner product space induced by π. For any $u, v \in \mathbb{R}^N$, define inner product $\langle \cdot, \cdot \rangle_\pi$ as

$$\langle u, v \rangle_\pi = \sum_i u_i v_i \pi_i. \tag{1.64}$$

This defines norm $\|u\|_\pi = \sqrt{\langle u, u \rangle_\pi}$ for all $u \in \mathbb{R}$. Let $L^2(\pi)$ denote the space of all vectors with finite $\|\cdot\|_\pi$ norm endowed with inner product $\langle \cdot, \cdot \rangle_\pi$. Then, for

any $u, v \in L^2(\pi)$

$$\begin{aligned}\langle u, Qv\rangle_\pi &= \sum_i u_i \big(\sum_j Q_{ij} v_j\big)\pi_i \\ &= \sum_{i,j} u_i v_j \pi_i Q_{ij} \;=\; \sum_{i,j} u_i v_j \pi_j Q_{ji} \\ &= \sum_j \pi_j v_j \big(\sum_i Q_{ji} u_i\big) \;=\; \langle Qu, v\rangle_\pi. \end{aligned} \tag{1.65}$$

That is, Q is *self-adjoint* over $L^2(\pi)$. For a self-adjoint matrix Q over $L^2(\pi)$, define norm

$$\|Q\|_{2,\pi} = \max_u \frac{\|Qu\|_\pi}{\|u\|_\pi}. \tag{1.66}$$

It can be verified that for any $u \in \mathbb{R}^N$ and Q,

$$\sqrt{\pi_{\min}}\|u\|_2 \le \|u\|_\pi \le \sqrt{\pi_{\max}}\|u\|_2 \tag{1.67}$$

$$\sqrt{\frac{\pi_{\min}}{\pi_{\max}}}\|Q\|_2 \le \|Q\|_\pi \le \sqrt{\frac{\pi_{\max}}{\pi_{\min}}}\|Q\|_2. \tag{1.68}$$

Consider a symmetrized version of Q as $S = \Pi^{\frac{1}{2}} Q \Pi^{-\frac{1}{2}}$, where $\Pi^{\pm\frac{1}{2}}$ is the $N \times N$ diagonal matrix with ith entry on the diagonal being $\pi_i^{\pm\frac{1}{2}}$. The symmetry of S follows due to detailed balance property of Q, i.e. $\pi_i Q_{ij} = \pi_j Q_{ji}$ for all i, j. Since Q is probability transition matrix, by Perron-Frobenius theorem we have that its eigenvalues are in $[-1, 1]$ with top eigenvalue being 1 and unique. Let they be $1 = \lambda_1 > \lambda_2 \ge \ldots \lambda_N > -1$. Let the corresponding (left) eigenvectors of Q be $v_1, \ldots, v_N$. By definition $v_1 = \pi$. Therefore, $u_i = \Pi^{-\frac{1}{2}} v_i$ are (left) eigenvectors of S with eigenvalue λ_i for $1 \le i \le N$, since

$$\begin{aligned} u_i^T S = (\Pi^{-\frac{1}{2}} v_i)^T \Pi^{\frac{1}{2}} Q \Pi^{-\frac{1}{2}} \;&=\; v_i^T Q \Pi^{-\frac{1}{2}} \\ = \lambda_i v_i^T \Pi^{-\frac{1}{2}} \;&=\; \lambda_i (\Pi^{-\frac{1}{2}} v_i)^T \;=\; \lambda_i u_i. \end{aligned} \tag{1.69}$$

That is, $u_1 = \pi^{\frac{1}{2}}$ or $\Pi^{-\frac{1}{2}} u_1 = \mathbf{1}$. By singular value decomposition, we can write $S = S_1 + S_{\setminus 1}$ where $S_1 = \lambda_1 u_1 u_1^T$ and $S_{\setminus 1} = \sum_{i=2}^N \lambda_i u_i u_i^T$. That is,

$$Q = \Pi^{-\frac{1}{2}} S \Pi^{\frac{1}{2}} \;=\; \Pi^{-\frac{1}{2}} S_1 \Pi^{\frac{1}{2}} + \Pi^{-\frac{1}{2}} S_{\setminus 1} \Pi^{\frac{1}{2}} = \mathbf{1}\pi^T + \Pi^{-\frac{1}{2}} S_{\setminus 1} \Pi^{\frac{1}{2}}. \tag{1.70}$$

Recalling notation $\Delta = \widehat{Q} - Q$, we can write (1.57) as

$$\begin{aligned} \nu_{t+1}^T - \pi^T = \nu_t^T \widehat{Q} - \pi^T Q \;&=\; (\nu_t - \pi)^T (Q + \Delta) + \pi^T \Delta \\ &= (\nu_t - \pi)^T (\mathbf{1}\pi^T + \Pi^{-\frac{1}{2}} S_{\setminus 1} \Pi^{\frac{1}{2}}) + (\nu_t - \pi)^T \Delta + \pi^T \Delta \\ &= (\nu_t - \pi)^T \Pi^{-\frac{1}{2}} S_{\setminus 1} \Pi^{\frac{1}{2}} + (\nu_t - \pi)^T \Delta + \pi^T \Delta, \end{aligned} \tag{1.71}$$

where we used the fact that $(\nu_t - \pi)^T \mathbf{1} = 0$ since both ν_t and π are probability vector. Now, for any matrix M, $\|\Pi^{-\frac{1}{2}} M \Pi^{\frac{1}{2}}\|_{\pi,2} = \|M\|_2$. Therefore,

$$\|\nu_{t+1}^T - \pi^T\|_{\pi,2} \le \|\nu_t - \pi\|_{\pi,2} \Big(\|S_{\setminus 1}\|_2 + \|\Delta\|_{\pi,2} \Big) + \|\pi^T \Delta\|_{\pi,2}. \tag{1.72}$$

By definition $\|S_{\setminus 1}\|_2 = \max\Big(\lambda_2, |\lambda_N|\Big) = \lambda_{\max}(Q)$. Let $\gamma = \Big(\lambda_{\max}(Q) + \|\Delta\|_{\pi,2}\Big)$. Then

$$\|\nu_t^T - \pi^T\|_{\pi,2} \leq \gamma^t \|\nu_0 - \pi\|_{\pi,2} + \Big(\sum_{s=0}^{t-1} \gamma^s\Big) \|\pi^T \Delta\|_{\pi,2}. \tag{1.73}$$

Using bounds in (1.67)-(1.68), we have

$$\gamma \leq \lambda_{\max}(Q) + \|\Delta\|_2 \sqrt{\frac{\pi_{\max}}{\pi_{\min}}} \;\equiv\; \rho \tag{1.74}$$

$$\|\nu_t^T - \pi^T\|_2 \leq \frac{1}{\sqrt{\pi_{\min}}} \|\nu_t^T - \pi^T\|_{\pi,2} \tag{1.75}$$

$$\|\nu_0 - \pi\|_{\pi,2} \leq \sqrt{\pi_{\max}} \|\nu_0 - \pi\|_2 \tag{1.76}$$

$$\|\pi^T \Delta\|_{\pi,2} \leq \|\pi\|_2 \|\Delta\|_2 \sqrt{\pi_{\max}}. \tag{1.77}$$

Therefore, we conclude

$$\frac{\|\nu_t^T - \pi^T\|}{\|\pi\|} \leq \Big[\rho^t \frac{\|\nu_0 - \pi\|}{\|\pi\|} + \Big(\sum_{s=0}^{t-1} \rho^s\Big) \|\Delta\|_2\Big] \sqrt{\frac{\pi_{\max}}{\pi_{\min}}}. \tag{1.78}$$

This completes the proof by bounding $\sum_{s=0}^{t-1} \rho^s = \frac{1-\rho^t}{1-\rho} \leq \frac{1}{1-\rho}$ for $\rho < 1$. □

Recovering ranking: comparison marginals. We will consider recovering ranking from noisy comparison marginals, $\widehat{C} = C + \eta$ using the scores as in (1.46) defined using noisy marginals. That is, for $i \in [N]$ define

$$\widehat{\mathsf{score}}(i) = \frac{1}{N-1} \sum_{k \neq i} \widehat{C}_{ik}. \tag{1.79}$$

Then error in score for i is

$$\begin{aligned} \mathsf{error}(i) = |\widehat{\mathsf{score}}(i) - \mathsf{score}(i)| \;&=\; \frac{1}{N-1} \Big| \sum_{k \neq i} \widehat{C}_{ik} - C_{ik} \Big| \\ &\leq \frac{1}{N-1} \sum_{k \neq i} |\widehat{C}_{ik} - C_{ik}| \;\leq\; \frac{1}{N-1} \|\eta_{i\cdot}\|_1, \end{aligned} \tag{1.80}$$

where $\eta_{i\cdot} = [\eta_{ik}]_{k \in [N]}$. Therefore, the relative order of any pair of $i, j \in [N]$ is preserved under noisy score as long as

$$\mathsf{error}(i) + \mathsf{error}(j) < |\mathsf{score}(i) - \mathsf{score}(j)|. \tag{1.81}$$

That is,

$$\|\eta_{i\cdot}\|_1 + \|\eta_{j\cdot}\|_1 < (N-1) |\mathsf{score}(i) - \mathsf{score}(j)|. \tag{1.82}$$

In summary, (1.82) provides robustness property of ranking algorithm based on noisy comparison to be able to recover true relative order for each pair of i, j; and subsequently entire ranking.

Recovering ranking: first-order marginals. We consider using Borda count for

finding ranking using noisy first-order marginals. Precisely, given noisy first-order marginals $\widehat{M} = M + \eta$, we define noisy Borda count for $i \in [N]$ as

$$\widehat{\mathsf{borda}}(i) = \sum_{k \in [N]} k \widehat{M}_{ik}. \tag{1.83}$$

Then, error in Borda count for i is

$$\begin{aligned} \mathsf{error}(i) = |\widehat{\mathsf{borda}}(i) - \mathsf{borda}(i)| \ &\leq \ \sum_{k \in [N]} k |\widehat{M}_{ik} - M_{ik}| \\ &= \sum_{k \in [N]} k |\eta_{ik}| \ = \ \mathsf{borda}^{\eta^+}(i). \end{aligned} \tag{1.84}$$

That is, $\mathsf{error}(i)$ is like computing Borda count for i using $\eta^+ \equiv [|\eta_{ik}|]$, which we define as $\mathsf{borda}^{\eta^+}(i)$. Then, the relative order of any pair $i, j \in [N]$ per noisy Borda count is preserved if

$$\mathsf{borda}^{\eta^+}(i) + \mathsf{borda}^{\eta^+}(j) < |\mathsf{borda}(i) - \mathsf{borda}(j)|. \tag{1.85}$$

In summary, (1.85) provides robustness property of ranking algorithm based on noisy first-order marginal to be able to recover the relative order for a given pair i, j; and subsequently for the entire ranking.

1.6 Discussion

We discussed learning distribution over permutations from marginal information. In particular, we focussed on marginal information of two types: first-order marginals, and pair-wise marginals. We discussed the conditions for recovering distribution as well as ranking associated with the distribution under two model classes: *sparse* models and *random utility models* (RUM). For all of these settings, we discussed the settings where we had access to *exact* marginal information as well as *noisy* marginals. There has been a lot of progress made, especially in the past decade on this topic in various directions. Here, we point out some of the prominent directions and provide associated references.

1.6.1 Beyond first-order and pair-wise marginals

To start with, learning distribution for different types of marginal information has been discussed in (Jagabathula & Shah 2011) where authors discuss relationship between the level of sparsity and the type of marginal information available. Specifically, through connecting marginal information with the spectral representation of the permutation group, authors find that as the higher order marginal information is made available, the distribution with larger support size can be recovered with tantalizingly similar relationship between dimensionality of the higher order information and the recoverability of support size just like in the first-order marginal information. A reader is referred to (Jagabathula &

Shah 2011) for more details and some of the open questions. We note that this collection of results utilize the *signature* condition discussed in Section 1.4.1.

1.6.2 Learning MNL beyond Rank Centrality

The work in (Negahban et al. 2012, Negahban et al. 2016) for recovering parameters of MNL from noisy observations has led to exciting subsequent works in the recent years. In particular, in (Hajek, Oh & Xu 2014) authors argue that the Maximum Likelihood Estimation has similar performance as the *Rank Centrality* that we discussed in Section 1.5.2. There has been work to find refined estimation of parameters, for example, restricting to top few parameters, see (Chen & Suh 2015, Chen, Fan, Ma & Wang 2017, Jang, Kim, Suh & Oh 2016). We also note interesting algorithmic generalization of Rank Centrality that has been discussed in (Maystre & Grossglauser 2015) through connection to continuous time representation of the reversible Markov chain considered in Rank Centrality.

1.6.3 Mixture of MNL

The RUM discussed in detail here has weakness that all options are parameterized by one parameter. This does not allow for heterogeneity in options in terms of multiple "modes" of preferences or rankings. Putting it other way, RUM captures a sliver of the space of all possible distributions over permutations. A natural way to generalize such a model is to consider mixture of RUM models. Specific instance is mixture of MNL models, which is known as Mixture MNL or Mixed MNL model. It can be argued that such a mixed MNL can approximate any distribution over permutations with enough number of mixture components. This is because we can approximate a distribution with unit mass on a permutation by an MNL model by choosing parameters appropriately. Therefore, it makes sense to understand when can we learn Mixed MNL model from pair-wise ranking or more generally partial rankings. In (Ammar, Oh, Shah & Voloch 2014), authors considered this question and effectively identified impossibility result that suggests that pair-wise information is not sufficient to learn mixtures in general. They provided lower bounds that related the number of mixture components, number of choices (here N) and the length of partial rankings observed. For *separable* mixture MNL model, they provide natural clustering based solution for recovery. Such a recovery approach has been further refined in the context of collaborative ranking (Oh, Thekumparampil & Xu 2015, Lu & Negahban 2015) through use of convex optimization based methods and imposing "low-rank" structure on the model parameter matrix to enable recovery. In another line of work, using higher moment information for *separable* mixture model, (Oh & Shah 2014) provided a tensor decomposition based approach for recovering the mixture MNL model.

1.6.4 Matrix Estimation for De-noising Noisy Marginals

In a very different view, the first-order marginal and pair-wise marginal information considered here can be viewed as matrix of observations with an underlying structure. Or more generally, we have access to noisy observation of an underlying ground matrix with structure. The structure is implied by the underlying distribution over permutation generating it. Therefore recovering *exact* marginal information from *noisy* marginal information can be viewed as recovering a matrix, with structure, based on its noisy and potentially partial view. In (Chatterjee et al. 2015), this view was considered for de-noising pair-wise comparison marginal data. In (Chatterjee et al. 2015), it was argued that when the ground-truth comparison marginal matrix satisfies certain *stochastic transitivity* condition, for example implied by MNL model, the true pair-wise marginal matrix can be recovered from noisy, partial observations. This work has been further studied in a sequence of recent works including (Shah, Balakrishnan, Guntuboyina & Wainwright 2016, Shah, Balakrishnan & Wainwright 2016).

1.6.5 Active learning and noisy sorting

Active learning view to ranking using pair-wise comparisons with noisy observations has been well studied for a long time. For example, (Adler, Gemmell, Harchol-Balter, Karp & Kenyon 1994) considered design of "tournament" in presence noisy pair-wise outcomes in the adaptive settings assuming the noisy comparisons satisfied the MNL model. When there is "geometry" imposed on the space of preferences, very efficient adaptive algorithms can be designed for searching using comparisons, for example (Jamieson & Nowak 2011, Karbasi, Ioannidis & Massoulié 2015). Another associated line of works is that of noisy sorting. For example, see (Braverman & Mossel 2008).

It is worth noting that the variation of online learning in the context of "bandit" setting is known as "dueling bandits" wherein comparisons between pair of arms is provided and one is to use this to find the top arm. This, again can be viewed as finding top element from pair-wise comparisons in an online setting with the goal to minimize "regret". For example, see (Yue & Joachims 2009, Jamieson, Katariya, Deshpande & Nowak 2015, Dudk, Hofmann, Schapire, Slivkins & Zoghi 2015).

In our view, dueling bandit modeled using distribution over permutations, i.e. the outcome of pair-wise comparisons of arms are consistent with underlying distribution over permutations, provides an exciting direction for making progress towards online matrix estimation.

1.6.6 And it continues...

There is a lot more that is tangentially related to this topic. For example, the question related to ranking or selecting winner in an election is fundamental

to so many disciplines and each (sub-)discipline brings different perspective to the question that makes this topic rich and exciting. The statistical challenges related to learning the distribution over permutations is very recent as clear from exposition provided here. The scale and complexity of the distribution over permutations ($N!$ for N options) makes it challenging from computational view point and thus providing a fertile ground for emerging interaction between statistics and computation. The rich group structure embedded within permutation group makes it an exciting arena for development of algebraic statistics, cf. (Kondor, Howard & Jebara 2007).

The statistical philosophy of max-entropy model learning leads to learning parametric distribution from an exponential family. This brings in rich connections to the recent advances in learning and inference on graphical models. For example, fitting such a model using first-order marginals boils down to computing partition function over the space of matchings or permutations; which can be computationally efficiently solved due to somewhat recent progress in computing permanent (Jerrum, Sinclair & Vigoda 2004). In contrast, learning such a model efficiently in the context of pair-wise marginal is not easy due to connection to feedback arc set problem.

We make note of an interesting connection: a mode computation heuristic based on maximum weight matching in bipartite graph using first-order marginal turns out to be a "first-order" approximation of the mode of such a distribution, see (Ammar & Shah 2012). On the other hand, using pair-wise comparison marginals, there is a large number of heuristics to compute ranking including the Rank Centrality algorithm discussed in detail or more generally variety of spectral methods considered for ranking including (Saaty & Hu 1998, Dwork, Kumar, Naor & Sivakumar 2001), and more recently (Rajkumar & Agarwal 2014, Azari, Parks & Xia 2012).

And this exciting list of work continues to grow even as author completes these final keystrokes and as you get inspired to immerse yourselves in this fascinating topic of Computing Choice.

References

Adler, M., Gemmell, P., Harchol-Balter, M., Karp, R. M. & Kenyon, C. (1994), Selection in the presence of noise: The design of playoff systems, *in* 'SODA', pp. 564–572.

Ammar, A., Oh, S., Shah, D. & Voloch, L. F. (2014), What's your choice?: learning the mixed multi-nomial, *in* 'ACM SIGMETRICS Performance Evaluation Review', Vol. 42, ACM, pp. 565–566.

Ammar, A. & Shah, D. (2011), Ranking: Compare, don't score, *in* 'Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on', IEEE, pp. 776–783.

Ammar, A. & Shah, D. (2012), Efficient rank aggregation using partial data, *in* 'ACM SIGMETRICS Performance Evaluation Review', Vol. 40, ACM, pp. 355–366.

Arrow, K. J. (1950), 'A difficulty in the concept of social welfare', *Journal of political economy* **58**(4), 328–346.

Azari, H., Parks, D. & Xia, L. (2012), Random utility theory for social choice, *in* 'Advances in Neural Information Processing Systems', pp. 126–134.

Bartels, K., Boztug, Y. & Muller, M. M. (1999), Testing the multinomial logit model. Working Paper.

Ben-Akiva, M. E. (1973), Structure of passenger travel demand models, PhD thesis, Department of Civil Engineering, MIT.

Ben-Akiva, M. E. & Lerman, S. R. (1985), *Discrete choice analysis: theory and application to travel demand*, CMIT press, Cambridge, MA.

Beran, R. (1979), 'Exponential models for directional data', *The Annals of Statistics* **7**(6), 1162–1178.

Berinde, R., Gilbert, A. C., Indyk, P., Karloff, H. & Strauss, M. J. (2008), Combining geometry and combinatorics: A unified approach to sparse signal recovery, pp. 798 –805.

Borgs, C., Chayes, J., Lee, C. E. & Shah, D. (2017), Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation, *in* 'Advances in Neural Information Processing Systems', pp. 4715–4726.

Boyd, J. H. & Mellman, R. E. (1980), 'The effect of fuel economy standards on the u.s. automotive market: An hedonic demand analysis', *Transportation Research Part A: General* **14**(5-6), 367 – 378.

Bradley, R. A. (1953), 'Some statistical methods in taste testing and quality evaluation', *Biometrics* **9**, 22–38.

Braverman, M. & Mossel, E. (2008), Noisy sorting without resampling, *in* 'Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms', Society for Industrial and Applied Mathematics, pp. 268–276.

Candes, E. J. & Romberg, J. (2006), 'Quantitative robust uncertainty principles and optimally sparse decompositions', *Foundations of Computational Mathematics* **6**(2), 227–254.

Candes, E. J., Romberg, J. K. & Tao, T. (2006*a*), 'Stable signal recovery from incomplete and inaccurate measurements', *Communications on Pure and Applied Mathematics* **59**(8).

Candes, E. J., Romberg, J. & Tao, T. (2006*b*), 'Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information', *IEEE Transactions on Information Theory* **52**(2), 489–509.

Candes, E. J. & Tao, T. (2005), 'Decoding by linear programming', *Information Theory, IEEE Transactions on* **51**(12), 4203–4215.

Cardell, N. S. & Dunbar, F. C. (1980), 'Measuring the societal impacts of automobile downsizing', *Transportation Research Part A: General* **14**(5-6), 423 – 434.

Chatterjee, S. et al. (2015), 'Matrix estimation by universal singular value thresholding', *The Annals of Statistics* **43**(1), 177–214.

Chen, Y., Fan, J., Ma, C. & Wang, K. (2017), 'Spectral method and regularized mle are both optimal for top-k ranking', *arXiv preprint arXiv:1707.09971* .

Chen, Y. & Suh, C. (2015), Spectral mle: Top-k rank aggregation from pairwise comparisons, *in* 'International Conference on Machine Learning', pp. 371–380.

Crain, B. (1976), 'Exponential models, maximum likelihood estimation, and the haar condition', *Journal of the American Statistical Association* **71**, 737–745.

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

Diaconis, P. (1988), *Group representations in probability and statistics*, Institute of Mathematical Statistics Hayward, CA.

Donoho, D. L. (2006), 'Compressed sensing', *IEEE Transactions on Information Theory* **52**(4), 1289–1306.

Dudk, M., Hofmann, K., Schapire, R. E., Slivkins, A. & Zoghi, M. (2015), Contextual dueling bandits, *in* P. Grnwald, E. Hazan & S. Kale, eds, 'Proceedings of The 28th Conference on Learning Theory', Vol. 40 of *Proceedings of Machine Learning Research*, PMLR, Paris, France, pp. 563–587.

Dwork, C., Kumar, R., Naor, M. & Sivakumar, D. (2001), Rank aggregation methods for the web, *in* 'Proceedings of the 10th international conference on World Wide Web', ACM, pp. 613–622.

Emerson, P. (2013), 'The original borda count and partial voting', *Social Choice and Welfare* **40**(2), 353–358.

Farias, V. F., Jagabathula, S. & Shah, D. (2012), Sparse choice models, *in* 'Information Sciences and Systems (CISS), 2012 46th Annual Conference on', IEEE, pp. 1–28.

Farias, V., Jagabathula, S. & Shah, D. (2009), A data-driven approach to modeling choice, *in* 'Neural Information Processing Systems'.

Farias, V., Jagabathula, S. & Shah, D. (2013), 'A non-parametric approach to modeling choice with limited data', *Management Science* .

Guadagni, P. M. & Little, J. D. C. (1983), 'A logit model of brand choice calibrated on scanner data', *Marketing science* **2**(3), 203–238.

Hajek, B., Oh, S. & Xu, J. (2014), Minimax-optimal inference from partial rankings, *in* 'Advances in Neural Information Processing Systems', pp. 1475–1483.

Horowitz, J. L. (1993), 'Semiparametric estimation of a work-trip mode choice model', *Journal of Econometrics* **58**, 49–70.

Huang, J., Guestrin, C. & Guibas, L. (2009), 'Fourier theoretic probabilistic inference over permutations', *Journal of machine learning research* **10**(May), 997–1070.

Jagabathula, S. & Shah, D. (2009), Inferring rankings under constrained sensing, *in* 'Advances in Neural Information Processing Systems', pp. 753–760.

Jagabathula, S. & Shah, D. (2011), 'Inferring rankings under constrained sensing', *IEEE Transactions on Information Theory* **57**(11), 7288–7306.

Jamieson, K. G., Katariya, S., Deshpande, A. & Nowak, R. D. (2015), Sparse dueling bandits., *in* 'AISTATS'.

Jamieson, K. G. & Nowak, R. (2011), Active ranking using pairwise comparisons, *in* 'Advances in Neural Information Processing Systems', pp. 2240–2248.

Jang, M., Kim, S., Suh, C. & Oh, S. (2016), 'Top-k ranking from pairwise comparisons: When spectral ranking is optimal', *arXiv preprint arXiv:1603.04153* .

Jerrum, M., Sinclair, A. & Vigoda, E. (2004), 'A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries', *Journal of the ACM (JACM)* **51**(4), 671–697.

Karbasi, A., Ioannidis, S. & Massoulié, L. (2015), 'From small-world networks to comparison-based search', *IEEE Transactions on Information Theory* **61**(6), 3056–3074.

Kondor, R., Howard, A. & Jebara, T. (2007), Multi-object tracking with representations of the symmetric group, *in* 'Artificial Intelligence and Statistics', pp. 211–218.

Koopman, B. (1936), 'On distributions admitting a sufficient statistic', *Transactions of the American Mathematical Society* **39**(3), 399–409.

Lu, Y. & Negahban, S. N. (2015), Individualized rank aggregation using nuclear norm regularization, *in* 'Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on', IEEE, pp. 1473–1479.

Luce, R. (1959), *Individual choice behavior: A theoretical analysis*, Wiley, New York.

Mahajan, S. & van Ryzin, G. J. (1999), 'On the relationship between inventory costs and variety benefits in retail assortments', *Management Science* **45**(11), 1496–1509.

Marden, J. (1995), *Analyzing and modeling rank data*, Chapman & Hall/CRC.

Marschak, J. (1959), 'Binary choice constraints on random utility indicators', *Cowles Foundation Discussion Papers* .

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

Maystre, L. & Grossglauser, M. (2015), Fast and accurate inference of plackett–luce models, *in* 'Advances in neural information processing systems', pp. 172–180.

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

McFadden, D. (1981), 'Econometric models of probabilistic choice, in "Structural Analysis of Discrete Data with Econometric Applications,"(CF Manski and D. McFadden, Eds.)'.

McFadden, D. (2001), 'Disaggregate behavioral travel demands rum side', *Travel Behaviour Research* pp. 17–63.

McFadden, D. & Train, K. (2000), 'Mixed MNL models for discrete response', *Journal of Applied Econometrics* **15**(5), 447–470.

Negahban, S., Oh, S. & Shah, D. (2012), Iterative ranking from pair-wise comparisons, *in* 'Advances in neural information processing systems', pp. 2474–2482.

Negahban, S., Oh, S. & Shah, D. (2016), 'Rank centrality: Ranking from pairwise comparisons', *Operations Research* **65**(1), 266–287.

Oh, S. & Shah, D. (2014), Learning mixed multinomial logit model from ordinal data, *in* 'Advances in Neural Information Processing Systems', pp. 595–603.

Oh, S., Thekumparampil, K. K. & Xu, J. (2015), Collaboratively learning preferences from ordinal data, *in* 'Advances in Neural Information Processing Systems', pp. 1909–1917.

Plackett, R. (1975), 'The analysis of permutations', *Applied Statistics* **24**(2), 193–202.

Rajkumar, A. & Agarwal, S. (2014), A statistical convergence perspective of algorithms for rank aggregation from pairwise data, *in* 'International Conference on Machine Learning', pp. 118–126.

Saaty, T. L. & Hu, G. (1998), 'Ranking by eigenvector versus other methods in the analytic hierarchy process', *Applied Mathematics Letters* **11**(4), 121–125.

Shah, N. B., Balakrishnan, S. & Wainwright, M. J. (2016), Feeling the bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons, *in* 'Information Theory (ISIT), 2016 IEEE International Symposium on', IEEE, pp. 1153–1157.

Shah, N., Balakrishnan, S., Guntuboyina, A. & Wainwright, M. (2016), Stochastically transitive models for pairwise comparisons: Statistical and computational issues, *in* 'International Conference on Machine Learning', pp. 11–20.

Song, D., Lee, C. E., Li, Y. & Shah, D. (2016), Blind regression: Nonparametric regression for latent variable models via collaborative filtering, *in* 'Advances in Neural Information Processing Systems', pp. 2155–2163.

Thurstone, L. (1927), 'A law of comparative judgement', *Psychological Reviews* **34**, 237–286.

Wainwright, M. & Jordan, M. (2008), 'Graphical models, exponential families, and variational inference', *Foundations and Trends® in Machine Learning* **1**(1-2), 1–305.

Yellott, J. I. (1977), 'The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution', *Journal of Mathematical Psychology* **15**(2), 109 – 144.

Yue, Y. & Joachims, T. (2009), Interactively optimizing information retrieval systems as a dueling bandits problem, *in* 'Proceedings of the 26th Annual International Conference on Machine Learning', ACM, pp. 1201–1208.