On Learning Mixture Models for Permutations

Flavio Chierichetti*
Sapienza University of Rome
Rome, Italy
flavio@di.uniroma1.it

Anirban Dasgupta
IIT
Gandhinagar, India
anirban.dasgupta@gmail.com

Ravi Kumar
Google
Mountain View, CA
ravi.k53@gmail.com

Silvio Lattanzi
Google
New York, NY
silviolat@gmail.com

ABSTRACT

In this paper we consider the problem of learning a mixture of permutations, where each component of the mixture is generated by a stochastic process. Learning permutation mixtures arises in practical settings when a set of items is ranked by different sub-populations and the rankings of users in a sub-population tend to agree with each other. While there is some applied work on learning such mixtures, they have been mostly heuristic in nature.

We study the problem where the permutations in a mixture component are generated by the classical Mallows process in which each component is associated with a center and a scalar parameter. We show that even when the centers are arbitrarily separated, with exponentially many samples one can learn the mixture, provided the parameters are all the same and known; we also show that the latter two assumptions are information-theoretically inevitable. We then focus on polynomial-time learnability and show bounds on the performance of two simple algorithms for the case when the centers are well separated.

Conceptually, our work suggests that while permutations may not enjoy as nice mathematical properties as Gaussians, certain structural aspects can still be exploited towards analyzing the corresponding mixture learning problem.

1. INTRODUCTION

Mixture models have been studied for more than a century [37]. In a mixture model setting, we postulate a probabilistic process for generating samples from a convex combination of a small set of distributions, where the distributions in this set are usually from the same underlying family. The parameters of these distributions are typically unknown. Given a set of independently generated samples from this process, the question is to learn the hidden parameters of the distributions and cluster the samples. Mixture model learning is a standard way to interpret data in many machine learning and data mining applications [31,39]. This elegant notion has enormously influenced widely-used algorithms in computer science including the expectation-maximization and the k-means algorithms.

A large body of literature exists on learning mixtures where the underlying family is the Gaussian distribution. For applications with real-valued data, the Gaussian distribution is a natural candidate to model the vagaries of the data. There are several heuristics (including versions of the expectation-maximization algorithm and versions of the k-means algorithm) for this learning problem; however very little [15] can be said about these heuristics formally. For more details, see the section on related work.

Much less theoretical work has been done on the problem of learning a mixture of permutations. Permutation mixture models arise in many real-world settings where a set of objects is implicitly or explicitly ranked. As a motivating example, consider a user population that (fully) ranks a small set of local restaurants. The population could be composed of several sub-populations where the users in each sub-population might rank the restaurants in a similar manner. For example, a sub-population that places a lot of emphasis on ambiance will choose a ranking where the restaurant ambiance attribute is highlighted. Similarly, a sub-population with an Italian cuisine preference might rank all Italian restaurants above non-Italian restaurants. These characteristics of such sub-populations may or may not be known a priori. Of course, the members within a sub-population need not necessarily agree with each other about the entire ranking; they might have a few disagreements. From an application point of view, say for restaurant recommendation or for targeted advertisement, it is important to identify these sub-populations, the aggregated ranking of a sub-population, and cluster the entire population into such meaningful sub-populations.

A way to model the above scenario is to use a mixture of distributions on permutations. An important question is which family of distributions on permutations is best suited for this purpose, both from theoretical and practical points of view. A compelling candidate is the *Mallows model* [29]: in this model, a center (a permutation) and a parameter (a real number) induce a distribution on the space of all per-

*This author was supported by a Google Focused Award.

ITCS'15, January 11–13, 2015, Rehovot, Israel.
Copyright 2015 ACM 978-1-4503-3333-7/15/01...$15.00.
http://dx.doi.org/10.1145/2688073.2688111.

mutations, where the probability mass on a particular permutation depends on its (Kendall) distance from the center, scaled by the parameter. Mallows model can be thought of as an analog of Gaussians for permutations, where the center plays the role of the mean and the parameter plays the role of the variance. However, unlike multidimensional Gaussian distributions, Mallows model is much less well-behaved with less nice properties and much less is known about the model. Meila and Chen [32] used such a Mallows mixture model for clustering rankings, but their work is restricted to empirical analysis. In particular, they do not have any provable bounds on the performance of their algorithm. There has been other work in the statistics and machine learning community to model rankings as mixtures of other choice models such as the Plackett–Luce and Benter models [9, 20, 21], but it is unclear if these models are easily amenable to algorithmic treatment.

1.1 Our contributions

In this paper we study the problem of learning a mixture of distributions on permutations where each distribution in the mixture generated by a Mallows model, with its own center and parameter. As in the Gaussian case, it is unsurprising that the learnability of the problem can heavily depend on how well-separated are the mixture centers. We first address the following question: assuming the centers are arbitrarily placed and assuming the algorithm has access to an unlimited number of samples, what are the conditions under which the mixture can be learned? We show that for learnability, it is information-theoretically necessary that the following two conditions hold: the Mallows parameter is the *same* for all distributions in the mixture and this (single) parameter must be *known* to the learning algorithm. We complement this non-learnability by obtaining an algorithm for learning the centers from (exponentially) many samples, provided the Mallows parameters are all the same and the algorithm knows its value. In fact, we show that this learnability result holds for any exponential distribution of a distance function that is embeddable into ℓ_2^2; this may be of independent interest.

Next, we focus on the cases where the centers are well-separated, where our goal is to use only a polynomial number of samples for learning. We first obtain an algorithm based on single-linkage clustering that can learn the centers, provided the centers are well-separated. The algorithm works by first clustering the samples and then using an existing algorithm to infer the centers for each cluster. We next obtain a different algorithm based on the nearest-neighbor criterion for clustering the samples. This algorithm needs a quadratically-weaker center separation assumption compared to the single-linkage clustering, but uses a stronger assumption of knowing the position of the centers. While both these algorithms are very simple, the difficulty is in proving that their performance can be tied to the separation of the centers.

Our work illustrates that while permutations and the Mallows model are more cumbersome to work with than multidimensional real-valued distributions because of the finite discrete nature of the permutation space (e.g., even counting the number of permutations at a certain distance is not fully resolved), we can use their structural properties (e.g., embeddability, a simple generative process, alternative equivalent representations such as the insertion vector) in order to analyze popular and practically used heuristics and obtain provable bounds. On the other hand, the discrete nature of the space also enables us to establish non-learnability for the most general cases of the mixture model, unlike the Gaussian models, and a larger separation between centers is necessitated due to the lack of independence of the "element-wise perturbations" in the Mallows setting.

1.2 Related work

Given a set of permutations generated according to the Mallows model, the permutation that maximizes the likelihood of this set is in fact the center [41]; finding this permutation is the well-known rank aggregation problem. Braverman and Mossel [8] obtained an algorithm for learning the center given a set of samples from a Mallows distribution; our work can be thought of as a natural extension of this work to a mixture setting. Chierichetti et al. [12] studied the problem of reconstructing the center from samples, where each sample is obtained from a Mallows model with the same center but different parameter. Very recently, Awasthi et al. [5] has considered learning the parameters of Mallows model mixtures and obtained a polynomial time algorithm for the case of two mixtures, by using a clever tensor decomposition. However, [5] considers the case of only two components, and settles the polynomial time learnability question for arbitrary separation of the centers. We, on the other hand, present an existential identifiability result for arbitrary number of components, as well as polynomial time algorithms for well-separated centers.

For the problem of learning a mixture of Gaussians, a tremendous amount of progress has been made on the theoretical front over the last few years, starting with the groundbreaking work of S. Dasgupta [14], followed by a series of impressive results [4, 15, 27, 40], culminating with the recent work on the provably learning Gaussian mixtures via the method of moments [6, 22, 24, 25, 33]. See the recent survey article by Kalai, Moitra, and Valiant [26] for an almost up-to-date overview of this research area.

There have been some work on learning mixtures of other distribution families. With the increasing role of heavy-tailed distributions in contemporary applications, the mixture learning problem has also been studied for such distributions [11, 13]. The results here are somewhat weaker than for the Gaussian case. Another topic that has attracted a lot of attention is learning a mixture of product distributions [19, 35]. The problem of learning mixtures of distributions when the domain is discrete but the distribution itself is allowed to be arbitrary has also been studied: the structured case was investigated by Chan et al. [36] and the unstructured case was investigated by Rabani et al. [38] and Anandkumar et al. [3]. Learning mixture of tree graphical models was considered by Anandkumar et al. [2]. None of the tools/techniques developed in these papers seems to apply to the Mallows mixture problem in particular and to permutations in general. While there has been some empirical work [28, 34], mostly using EM, in learning Mallows mixtures, these do not come with theoretical guarantees.

2. PRELIMINARIES

Let S_n be the symmetric group on $[n] = \{1, \ldots, n\}$. For a permutation $\sigma \in S_n$ and an element i, let $\sigma(i)$ denote the rank of the ith element. For two permutations $\pi, \sigma \in S_n$, let $\mathbb{K}(\pi, \sigma)$ denote the *Kendall tau* distance, which is the number of inversions between π and σ.

Let $\beta \in (0, \infty)$ be a parameter and let $\sigma \in S_n$ be a fixed permutation. In the *Mallows* model $\mathcal{M}(\sigma, \beta)$ of generating permutations [29], the parameter β and the permutation σ induce a distribution on S_n as follows:

$$\Pr_{\mathcal{M}(\sigma,\beta)}[\pi] = \frac{e^{-\beta\mathbb{K}(\pi,\sigma)}}{Z_\beta},$$

where Z_β is the normalization constant defined as $Z_\beta = \sum_{\pi\in S_n} e^{-\beta\mathbb{K}(\pi,\sigma)}$. We use $\pi \sim \mathcal{M}(\sigma, \beta)$ to denote that π is generated according to $\mathcal{M}(\sigma, \beta)$. When σ is the identity permutation, we simply denote $\mathbb{K}(\sigma, \pi)$ by $\mathbb{K}(\pi)$ and $\mathcal{M}(\sigma, \beta)$ by $\mathcal{M}(\beta)$.

Let $k > 1$ be an integer. Let $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_k\}$ be a set of distinct permutations and let $\beta_1, \ldots, \beta_k$ be their corresponding *parameters*. In the Mallows mixture model setting with k centers, a sample permutation is generated by the following mixture process: pick $\sigma_i \in \boldsymbol{\sigma}$ according to a probability distribution (called *weight* distribution) $\mathcal{W}$ on $\boldsymbol{\sigma}$ and output the permutation generated according to $\mathcal{M}(\sigma_i, \beta_i)$. Here, the $\boldsymbol{\sigma}$ is the set of hidden *centers*. In the mixture model learning problem, we are given a set S of samples generated according to the mixture process and the goal is find the hidden centers. In a variant of the problem, in addition to the samples, we are also given the set $\boldsymbol{\sigma}$ of centers and the goal is to assign each sample to the center it came from. The complexity of the learning task is measured in terms of the size $|S|$ of the samples, the running time of the algorithm, and the quality of inferred centers (i.e., their distance to the true hidden centers). Sometimes, we will deal with the case when $\mathcal{W}$ is the uniform distribution on $\boldsymbol{\sigma}$; we call this the uniform Mallows mixture learning problem.

We need the following two tail bounds from [7], which show that no element deviates too far off from its original position and that the Kendall distance of the sample to the center is also bounded.

THEOREM 1 (BHATNAGAR AND PELED [7]). *For all $\beta > 0$, $i \geq 1$ and $t \geq 1$ if $\pi \sim \mathcal{M}(\beta)$, then*

$$\text{(a)} \quad \Pr_\pi[|\pi(i) - i| \geq t] \leq 2e^{-\beta t},$$

and for all $c > 0$,

$$\text{(b)} \quad \Pr_\pi\left[\mathbb{K}(\pi) > c\frac{n}{\beta}\ln(2nt)\right] < (nt)^{-c}.$$

3. ARBITRARY MIXTURES

In this section we consider the problem of learning the mixtures when the centers can be arbitrarily placed, in particular, when they can be very close to each other. In such settings, we first show that it is necessary to assume that the Mallows parameter β's are the same and are known, for otherwise it is not feasible to identify the components of the mixture. Next, we show that if the β's are the same, then we can learn the mixture for arbitrary separation between the centers provided we have access to sufficiently many samples from the mixture. Note that such a result is not obvious, even for Gaussian mixture models.

3.1 Non-learnability

We first show two easy non-learnability results: the first considers the case of when the parameters can be different and the second considers the case when the parameters, even if all same, are not known to the learning algorithm. Note that in both these results, we will need the centers to sometimes be placed very close (within $\mathbb{K}(\cdot, \cdot) = \Theta(1)$) to each other, i.e., these results are not necessarily true when we assume a super-constant separation between the centers.

First we note that for learnability, the parameters should all be positive for otherwise there is a very simple instance that is not learnable. Indeed, the following two worlds are indistinguishable. In the first world, there are $n!$ centers, all with same β, and $\mathcal{W}$ is uniform. In the second world, there is one center with $\beta = 0$. It is easy to see that both these worlds induce the uniform distribution on S_n.

Therefore, in what follows, we assume that β's are all positive. We now show that if the β's are not all the same, then there are instances of the mixture learning problem that are information-theoretically impossible to learn. We show that this holds in a very strong sense: even if there are only two centers $\{\sigma_1, \sigma_2\}$ and the weight distribution $\mathcal{W}$ is uniform on this set of centers.

LEMMA 2. *The uniform Mallows mixture learning problem with two centers cannot solved in general, without a knowledge of k, β_1 and β_2, regardless of the number of samples.*

PROOF. Suppose that we select the world uniformly at random between the following two possible worlds:

World 1: $k = 1$ with $\sigma = (1, 2), \beta = \ln 2$.

World 2: $k = 2$ with $\sigma_1 = (1, 2), \beta_1 = \ln(14)$ and $\sigma_2 = (2, 1), \beta_2 = \ln(3/2)$; $\mathcal{W}$ is uniform on $\{\sigma_1, \sigma_2\}$.

The mixture distribution $P_1(\cdot)$ in World 1 is

$$P_1(1, 2) = \frac{1}{1 + e^{-\beta}} \quad \text{and} \quad P_1(2, 1) = 1 - P_1(1, 2).$$

In World 2, since $\mathcal{W}$ is uniform, the mixture distribution $P_2(\cdot)$ is

$$P_2(1, 2) = \frac{1}{2}\cdot\left(\frac{1}{1 + e^{-\beta_1}} + \frac{e^{-\beta_2}}{1 + e^{-\beta_2}}\right) = \frac{1}{1 + e^{-\beta}} = P_1(1, 2).$$

Since the two distributions are identical, no algorithm can distinguish between the two worlds. Note that this argument can work for any $\beta_1 > 0$. Then, for any $0 < \beta < \ln\left(3 - \frac{8}{e^{\beta_1}+3}\right)$, set $\beta_2 = \ln\frac{1-e^\beta+2e^{\beta_1}}{e^{\beta_1}\cdot\left(e^\beta-1\right)+2e^\beta}$. □

We next show that even if all the parameters are the same, if the algorithm does not know the value of the parameter, the Mallows mixture problem cannot be solved in general.

LEMMA 3. *Let $n \geq 2$ and let all the centers have the same $\beta > 0$. The Mallows mixture learning problem cannot be solved in general if β is unknown, regardless of the number of samples.*

PROOF. Fix any $\beta > 0$. Let $p_e(\sigma, \beta)$ be the probability of obtaining a permutation from $\mathcal{M}(\sigma, \beta)$ that has an even Kendall τ distance to σ. Then,

$$p_e(\sigma, \beta) = \sum_{\pi|\mathbb{K}(\pi,\sigma)\in 2\mathbb{Z}} \frac{e^{-\beta\mathbb{K}(\pi,\sigma)}}{Z_\beta} = p_e(\beta),$$

since the sum has the same terms and hence the same value for any choice of σ. Note that $\frac{1}{2} < p_e(\beta) < 1$. It is also easy to see that $p_e(\beta)$ is continuous and increasing in β, $p_e(0) = \frac{1}{2}$, and $\lim_{\beta\to\infty} p_e(\beta) = 1$.

Let $0 < \beta_1 < \beta_2$ be chosen arbitrarily. Let

$$\alpha = \frac{p_e(\beta_1) + p_e(\beta_2) - 1}{2p_e(\beta_2) - 1}; \quad \frac{1}{2} < \alpha < 1.$$

Let $A_n \subseteq S_n$ be the set of the *even* permutations on $[n]$, i.e., permutations π such that $\mathbb{K}(\pi, \mathbf{1})$ is even where $\mathbf{1}$ is the identity permutation. It follows $|A_n| = |S_n|/2$.

We select the world uniformly at random between the following two possible worlds:

World 1: create a center for each $\sigma \in A_n$ and each center has the parameter β_1; $\mathcal{W}$ is uniform on A_n.

World 2: create a center for each $\sigma \in S_n$ and each center has the parameter β_2. $\mathcal{W}$ is defined as:

$$\mathcal{W}(\pi) = \begin{cases} \alpha/|A_n| & \pi \in A_n, \\ (1-\alpha)/|A_n| & \pi \in S \setminus A_n. \end{cases}$$

We now calculate the probability of any given even permutation in both worlds. The probability of any given even permutation in World 1 is $p_e(\beta_1)/|A_n|$ and in World 2 is

$$\begin{aligned} &\frac{\alpha}{|A_n|} \cdot p_e(\beta_2) + \frac{(1-\alpha)}{|A_n|} \cdot (1 - p_e(\beta_2)) \\ &= \frac{1}{|A_n|} \cdot (\alpha(2p_e(\beta_2) - 1) + 1 - p_e(\beta_2)) \\ &= \frac{p_e(\beta_1)}{|A_n|}, \end{aligned}$$

by our choice of α. Analogously, it can be seen that the probability of any given odd permutation in both worlds is $\frac{1-p_e(\beta_1)}{|A_n|}$. Since both these worlds induce the same distribution on S_n, no algorithm can distinguish between the two worlds. □

Note that even though $k = n!/2$ in the above proof, since it holds for any $n \geq 2$, k need not be particularly large for the instance. Also, even though we defined $\mathcal{W}$ to be non-uniform in World 2, this was for simplicity: for infinitely many β's, by replicating the centers appropriately, we can create an equivalent instance for World 2 where $\mathcal{W}$ is uniform.

3.2 Uniform parameters

In this section we obtain algorithms that can learn Mallows mixtures for arbitrary separation between centers as long as all the parameters are the *same* and are *known*. In conjunction with the non-learnability results in Lemma 2 and Lemma 3, these assumptions about the parameters are inevitable. The main idea in the algorithm is to use the invertibility of a certain Gram matrix.

Let $|X| = n$, and let $d : X \times X \to [0, \infty)$ be a function. We say d is isometrically embeddable into ℓ_2^2 if and only if there exists n vectors $x_1, \ldots, x_n$ such that $d(i,j) = ||x_i - x_j||^2$. The Kendall distance can be isometrically embedded into ℓ_1, and therefore into ℓ_2^2. Indeed, let $I(\sigma)$ be the $\binom{n}{2}$ length vector that in the $\{i,j\}$th position is 1 if and only if $i < j$ and $\sigma(i) > \sigma(j)$; it follows $\mathbb{K}(\pi, \sigma) = |I(\pi) - I(\sigma)|$.

A real matrix M is *positive definite* (resp., positive semidefinite) if, for every non-zero real vector x, it holds $xMx^T > 0$ (resp., $xMx^T \geq 0$). We denote $M \succ 0$ (resp., $M \succeq 0$) to denote M is positive definite (resp., positive semidefinite). The *Gram matrix* of the real vectors $x_1, \ldots, x_n$ is defined as $G_{i,j} = \langle x_i, x_j \rangle$ and the *Hadamard exponential* $\widehat{A}$ of a matrix A is defined as $\hat{A}_{i,j} = e^{A_{i,j}}$. We will use the following folklore results.

FACT 4. *If G is a Gram matrix, then $G \succeq 0$.*

LEMMA 5 (THEOREM 7.5.9 [23]). *If A is a positive semidefinite matrix and if A does not contain two equal rows, then its Hadamard exponential satisfies $\widehat{A} \succ 0$ and hence it is non-singular.*

Using these, we now prove the following.

THEOREM 6. *Let $x_1, \ldots, x_k$ be pairwise different vectors and let M be the $k \times k$ matrix such that:*

$$M_{i,j} = \frac{e^{-\beta\,||x_i - x_j||^2}}{\sum_{\sum_{\ell=1}^{k}} e^{-\beta\,||x_i - x_\ell||^2}}.$$

Then, M is non-singular.

Therefore, the $n! \times n!$ matrix $N_{i,j} = \frac{e^{-\beta\,\mathbb{K}(\sigma_i, \sigma_j)}}{\sum_{\ell=1}^{n!} e^{-\beta\,\mathbb{K}(\sigma_i, \sigma_\ell)}}$ is non-singular.

PROOF. The latter claim follows from the former since the Kendall tau distance is isometrically embeddable into ℓ_2^2, and since $\mathbb{K}(\sigma_i, \sigma_j) = 0$ if and only if $i = j$. We now prove the former claim for an arbitrary matrix M.

Let M'' be the Gram matrix corresponding to the vectors $x_1, \ldots, x_k$, i.e., $M''_{i,j} = \langle x_i, x_j \rangle$. We use two properties of M''. First, from Fact 4, we know that $M'' \succeq 0$ and hence for any $\beta > 0$, we have

$$2\beta M'' \succeq 0. \tag{1}$$

Second, we argue that M'' does not contain two equal rows. Indeed, assume the contrary and let the ith and the jth row ($j \neq i$) of M'' be equal, i.e., $\langle x_i, x_\ell \rangle = \langle x_j, x_\ell \rangle$ for all ℓ. By choosing $\ell = i$, we have that $||x_i||^2 = \langle x_i, x_i \rangle = \langle x_j, x_i \rangle$ and by choosing $\ell = j$, we have $\langle x_i, x_j \rangle = \langle x_j, x_j \rangle = ||x_j||^2$. Thus, $||x_i||^2 = \langle x_i, x_i \rangle = \langle x_j, x_j \rangle = ||x_j||^2$. By the Cauchy–Schwarz inequality, we know that for $\langle x_i, x_j \rangle^2$ to equal $||x_i||^2 \cdot ||x_j||^2$, we need x_i and x_j to be linearly dependent. Since $||x_i||^2 = ||x_j||^2$, this implies $x_i = x_j$, which is a contradiction, since we assumed that the x_ℓ's are pairwise different.

Let $M' = \widehat{2\beta M''}$, i.e., the Hadamard exponential of $2\beta M''$ given by $M'_{i,j} = e^{2\beta \langle x_i, x_j \rangle}$. By (1) and since M'' (and hence $2\beta M''$) has no two identical rows, using Lemma 5, we get

$$M' \succ 0 \Longrightarrow \det M' > 0. \tag{2}$$

Finally, using these, we will show M is non-singular by showing its determinant is non-zero. The determinant of M can be written as:

$$\begin{aligned} \det M &= \sum_{\pi \in S_n} \operatorname{sgn}(\pi) \cdot \frac{e^{-2\beta \sum_{i=1}^{n} ||x_i||^2 + 2\beta \sum_{i=1}^{n} \langle x_i, x_{\pi(i)} \rangle}}{\prod_{i=1}^{n} \sum_{\sigma_\ell \in \boldsymbol{\sigma}} e^{-\beta\,\mathbb{K}(\sigma_i, \sigma_\ell)}} \\ &= \frac{e^{-2\beta \sum_{i=1}^{n} ||x_i||^2}}{\prod_{i=1}^{n} \sum_{\sigma_\ell \in \boldsymbol{\sigma}} e^{-\beta\,\mathbb{K}(\sigma_i, \sigma_\ell)}} \\ &\qquad \cdot \sum_{\pi \in S_n} \operatorname{sgn}(\pi) \cdot e^{2\beta \sum_{i=1}^{n} \langle x_i, x_{\pi(i)} \rangle}. \end{aligned}$$

Note that the first term in this product is positive and the second term is precisely $\det M'$, which is also positive using (2). □

THEOREM 7. *If $\beta > 0$ is known and is the same for all the centers and if sufficiently many samples are given, then we can learn the Mallows mixture model with probability $1 - o(1)$.*

PROOF. Consider the matrix N of Theorem 6. Since it is invertible, let

$$u = \left|\left|N^{-1}\right|\right|_\infty .$$

For any i, j, since $\mathbb{K}(\sigma_i, \sigma_j) \le \binom{n}{2}$, we have

$$\begin{aligned} N_{i,j} &= \frac{e^{-\beta\,\mathbb{K}(\sigma_i,\sigma_j)}}{\sum_{\sigma_\ell \in \boldsymbol{\sigma}} e^{-\beta\,\mathbb{K}(\sigma_i,\sigma_\ell)}} \\ &\ge \frac{e^{-\beta\binom{n}{2}}}{\sum_{\sigma_\ell \in \boldsymbol{\sigma}} e^{-\beta\,\mathbb{K}(\sigma_i,\sigma_\ell)}} \\ &\ge \frac{1}{n!} \cdot e^{-\beta\binom{n}{2}} \ge e^{-(1+\beta)\,n^2} \\ &= v. \end{aligned}$$

Finally, let η be the minimum non-zero probability in $\mathcal{W}$:

$$\eta = \min_{\sigma \in \boldsymbol{\sigma}, \mathcal{W}(\sigma) > 0} \mathcal{W}(\sigma).$$

Now, let C_t be the column vector of length $n!$ indexed by $\pi \in S_n$ such that the πth entry contains the fraction of times that the permutation π was produced by the mixture model, if it were sampled t times. Let $\mathcal{M}$ be the vector of length $n!$ such that the πth entry is the expected number of times the mixture model outputs π.

By the Chernoff bound, for each $\epsilon > 0$, there exists $\gamma = \gamma(\epsilon)$ such that if $t \ge v^{-\gamma} = e^{\gamma(1+\beta)n^2}$, then with probability $1 - o(1)$, we will have

$$(1 - \epsilon)\mathcal{M}(\pi) \le C_t(\pi) \le (1 + \epsilon)\mathcal{M}(\pi),$$

uniformly for each $\pi \in S_n$. By choosing t, the number of samples, to be $O\left(\frac{(3u \cdot n!)^2 \log n}{v\eta^3}\right)$, we can assume $\epsilon = \frac{\eta}{3u \cdot n!}$.

Let $\overline{\mathcal{W}} = N^{-1}C_t$. Observe that, since $N \cdot \mathcal{W} = \mathcal{M}$, we have

$$\overline{\mathcal{W}(\pi)} = N^{-1}\mathcal{M} + \xi,$$

where $|\xi| \le \epsilon u \cdot n! = \frac{\eta}{3}$. Therefore, the set of π's such that $\overline{\mathcal{W}(\pi)} \ge \frac{2}{3}\eta$, is the correct set of unknown centers with probability $1 - o(1)$. □

In fact, the above proof naturally extends to the following more general setting. Given a set X of elements, let $d : X \times X \to [0, \infty)$ be a semi-metric, i.e., let $d(x, y) = d(y, x)$ for each $x, y \in X$, and $d(x, y) = 0$ iff $x = y$. Given X, d, $\beta > 0$, and some $x \in X$, we define the probability distribution $P_x = P_{x,X,d,\beta}$ as follows:

$$P_x(y) = \frac{e^{-\beta\, d(x,y)}}{\sum_{z \in X} e^{-\beta\, d(x,z)}} \qquad \forall y \in X.$$

Given a set $C \subseteq X$, and a probability distribution $\mathcal{W}$ over C, we define the mixture E of the $\{P_x\}_{x \in C}$ as:

$$E(y) = \sum_{x \in C} \mathcal{W}(x) \cdot P_x(y) \qquad \forall y \in X.$$

Observe that E is a probability distribution over X. The proof of Theorem 7 can be extended to show that if d can be isometrically embedded into ℓ_2^2, if all the centers' parameters are equal to $\beta > 0$, and if β is known, then given sufficiently many samples, it is possible to guess exactly the set C of centers that make up the mixture E with probability $1 - o(1)$.

4. WELL-SEPARATED MIXTURES

Theorem 7 in Section 3 states that with enough samples, one can learn the Mallows mixture model for arbitrary centers if all the β's are the same and known. Unfortunately, the resulting algorithm requires a superpolynomial number of samples and hence has a superpolynomial running time.

In this section we focus on developing efficient algorithms that are provably correct if the centers are well-separated, i.e., there is a minimal Kendall tau distance between every pair of centers. We focus on two cases: in the first, the algorithm does not know the position of the centers. In the second case, the algorithm knows the centers and the goal is to cluster the given samples with respect to the given centers.

4.1 Unknown centers

In this section we assume that the algorithm does not know the location of the centers (or the parameters). Nevertheless, we show that we can reconstruct clusters corresponding to different centers and then use the samples in the clusters to estimate the respective centers.

Let $\beta_1 \le \cdots \le \beta_k$ be the Mallows parameters, and let $\sigma_1, \ldots, \sigma_k \in S_n$ be the respective centers.

THEOREM 8. *Let t be the number of samples. If for each $1 \le i < j \le k$ we have $\mathbb{K}(\sigma_i, \sigma_j) \ge \Omega\left(\frac{n \log(nt)}{\beta_1}\right)$, then we can learn the Mallows mixture model with probability $1 - \frac{1}{(nt)^{\Theta(1)}}$.*

PROOF. The algorithm we use is the so-called single-linkage clustering, which is a popular practical heuristic. Each of the t samples starts as a singleton. Repeat the following until we obtain k clusters: select clusters C_1, C_2 such that $\min_{\pi_1 \in C_1} \min_{\pi_2 \in C_2} \mathbb{K}(\pi_1, \pi_2)$ is the minimum and merge C_1 and C_2 into a single cluster.

To prove the correctness of this algorithm, we appeal to Theorem 1(b). From this, if the minimum distance between the centers is at least $c' \frac{n}{\beta_1} \ln(2nt)$, for $c' > 2c$, then we are guaranteed that no two samples coming from different centers will end up in the same cluster. Thus, it is possible to guess correctly for each pair of samples if they were produced by the same center (by the mixture process) with probability at least $1 - 1/(nt)^{\Theta(1)}$.

After obtaining this clustering, for the final step of computing the centers themselves, we use the polynomial algorithm of Braverman and Mossel [8, Theorem 7] to use the samples in each cluster in order to estimate the centers. □

4.2 Known centers

In this section we still focus on the case of well-separated centers, when the centers are known in advance to the algorithm; the parameters $\beta.$ need not be known to the algorithm. Interestingly we can show that it is possible to obtain an algorithm that for some parameters outperforms the algorithm in Theorem 8.

The core intuition behind the algorithm is that even if the average distance between a center and a sample is large, a sample should be closer to the center that generated it than to any other center. In particular, we analyze the following natural algorithm: assign each sample to its nearest center. In the rest of this section we show that, under a separation assumption, this algorithm recovers the correct clustering.

Before describing our results, we recall some properties of the Mallows model. It is well known that a permutation can be sampled from the Mallows distribution using a simple process. For completeness, we describe the process; a full proof of why it corresponds to the $\mathcal{M}(\beta)$ distribution is available in [17].

Insertion process P. Define $q = e^{-\beta}$. We consider the elements $1, \dots, n$ in this order. For each i, π_i will denote a permutation over the elements 1 to i. Define $\pi_1(1) = 1$. We then define π_i in terms of π_{i-1} as follows. First the entry at the ith position of π_i, i.e., $\pi_i(i)$ is chosen using the following random process:

$$\Pr[\pi_i(i) = j] = \frac{(1/q)^{j-1}}{1 + 1/q + \cdots + (1/q)^{i-1}}, \text{ for } j \in \{1, \dots, i\}. \tag{3}$$

Then, for s such that $\pi_{i-1}(s) < \pi_i(i)$, set $\pi_i(s) = \pi_{i-1}(s)$ and else set $\pi_i(s) = \pi_{i-1}(s) + 1$. Finally, $\pi = \pi_n$.

We first state the following result from [12]. Let $s_{\beta,k,i}$ be the probability that $\pi(i) > \pi(i+k)$ and let $s'_{\beta,k,i} = \frac{1}{2} - s_{\beta,k,i}$.

Theorem 9 (Lemma 4 [12]). *$s_{\beta,k,i}$ is independent of i and for all k, $s_{\beta,k} \le s_{\beta,1} < \frac{\beta + e^{-\beta} - 1}{e^{\beta} + e^{-\beta} - 2}$. Furthermore, for $\beta > 0$, $s'_{\beta,k} \ge s'_{\beta,1} \ge \Theta(\min(\beta, 1))$. If $\beta = \Omega(1)$, $s'_{\beta,k} > 1 - \Theta(\beta e^{-\beta})$.*

Finally, we will also need the following form of the method of bounded differences [18]; the original result is due to McDiarmid [30].

Theorem 10. *Let f be a function of n random variables $X_1, \dots, X_n$, each X_i taking values in a set A_i, such that $E[f]$ is bounded. Assume that*

$$m \le f(X_1, \dots, X_n) \le M.$$

Let $\mathcal{B}$ be any event and let c_i be maximum effect of f assuming $\mathcal{B}$, i.e.,

$$|E[f \mid \mathbf{X}_{i-1}, X_i = a_i, \mathcal{B}] - E[f \mid \mathbf{X}_{i-1}, X_i = a'_i, \mathcal{B}]| \le c_i.$$

Then

$$\Pr[f > E[f] + t] \le \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right) + \Pr[\mathcal{B}^c],$$

and

$$\Pr[f < E[f] - t] \le \exp\left(-\frac{t^2}{\sum_i c_i^2}\right) + \Pr[\mathcal{B}^c].$$

Now we are ready to prove our main result: when the centers are known, then with only a $\tilde{O}(\sqrt{n}/\beta s'_{\beta,1})$ separation between the permutations, it is possible to label most points accurately. We first show a claim for two centers, one of them being the identity permutation. Then we extend the result to an arbitrary set of permutations.

Lemma 11. *Let $k = 2$ with σ_1 being the identity permutation and $\sigma_2 = \sigma$. For any permutation $\pi \sim \mathcal{M}(\beta)$, consider the following random variable Δ:*

$$\Delta = \mathbb{K}(\pi, \sigma) - \mathbb{K}(\pi).$$

If

$$\mathbb{K}(\sigma) \ge \frac{\log(1/\delta)}{s'_{\beta,1}} \min\left(n^{3/2}, \frac{c\sqrt{n}\log n}{\beta}\right),$$

then with probability $1 - \delta - n^{-c}$, $\Delta > 0$.

Proof. We analyze the expectation of Δ and then use Theorem 10 to obtain the high probability result. Let $\pi \sim \mathcal{M}(\beta)$ and consider the random variable $\Delta = \mathbb{K}(\pi, \sigma) - \mathbb{K}(\pi)$. Let the indicator variables x_{ij} and y_{ij} be defined as follows: for $j < i$, $x_{ij} = \mathbb{1}[\sigma(j) > \sigma(i)]$ and $y_{ij} = \mathbb{1}[\pi(j) > \pi(i)]$. Also, abusing notation, let $x_i = \sum_{j<i} x_{ij}$ be the ith coordinate of the inversion vector of σ. Now, consider that the Mallows permutation π has been generated according the Mallows process P. Thus, each random variable x_i depends solely on the position of the ith element in the corresponding step. Hence the random variable x_i's are mutually independent. Thus,

$$\begin{aligned}\Delta = \sum_i \sum_{j<i} &\mathbb{1}[\sigma(j) < \sigma(i) \text{ and } \pi(j) > \pi(i))] \\ &+ \mathbb{1}[\sigma(j) > \sigma(i) \text{ and } \pi(j) < \pi(i))] - \mathbb{1}[\pi(j) > \pi(i))] \\ = &\sum_i \sum_{j<i} (1 - x_{ij}) y_{ij} + x_{ij}(1 - y_{ij}) - y_{ij} \\ = &\sum_i \sum_{j<i} x_{ij} - 2x_{ij} y_{ij} \\ = &\mathbb{K}(\sigma) - 2\sum_i \sum_{j<i} x_{ij} y_{ij}.\end{aligned}$$

Consider the random variable $S = \sum_i \sum_{j<i} x_{ij} y_{ij}$. Using Theorem 9, for $j < i$, $E[y_{ij}] = s_{\beta,i-j} \le s_{\beta,1}$. Hence,

$$E[S] = \sum_i \sum_{j<i} x_{ij} y_{ij} \le s_{\beta,1} \sum_i \sum_{j<i} x_{ij} = s_{\beta,1} \mathbb{K}(\sigma).$$

Note that $E[S]$ is the number of inversion on which both π and σ agree. In the following we prove that S is concentrated using Theorem 10. The core intuition is to consider the process P and to show that after each insertion step, S changes by at most $\log n$ w.h.p.

More formally, for each i, the number of inversion with elements $j < i$ on which both σ and π agree is upper bounded by $y_i = i - \pi_i(i)$(recall the definition of $\pi_i(i)$ in process P), that is the number of inversion with elements $j < i$ of π.

Now let the event A be defined as the following: for all $i \in [1, n]$, $|\pi(i) - i| < \min(n - 1, c\log(n)/\beta)$. Using Theorem 1, $\Pr[A] > 1 - n^{-c}$. Let us condition on event A happening. Note that, since the final position of each element change at most of $(c\log n)/\beta$, then y_i is bounded by:

$$|y_i| \le \min\left(n - 1, \frac{2c\log n}{\beta}\right).$$

Thus, after conditioning on A, for each i the number of inversion with elements $j < i$ on which both σ and π agree is upper bounded by $|y_i| \le \min(n - 1, (2c\log n)/\beta)$. Furthermore $\sum_i y_i^2 \le \min(n^3, (4c^2 n \log^2 n)/\beta^2)$.

Now we can apply Theorem 10. Conditioned on A, we get that

$$\begin{aligned}\Pr[\Delta < 0] &\le \Pr[S > \mathbb{K}(\sigma)/2] \\ &\le \exp\left(-\frac{(\frac{1}{2}\mathbb{K}(\sigma) - E[S])^2}{2\min(n^3, \frac{4c^2 n \log^2(n)}{\beta^2})}\right) \\ &\le \exp\left(-\frac{(s'_{\beta,1}\mathbb{K}(\sigma))^2}{2\min(n^3, \frac{4c^2 n \log^2(n)}{\beta^2})}\right).\end{aligned}$$

Hence, when

$$\mathbb{K}(\sigma) \geq \frac{\log(1/\delta)}{s'_{\beta,1}} \min\left(n^{3/2}, \frac{2c\sqrt{n}\log n}{\beta}\right),$$

by applying the union bound, we have the total probability of $\Delta < 0$ to be at most $\delta + n^{-c}$. □

The previous result gives us a bound for two centers and a single sample. It is easy to extend it to more than two centers (up to a polynomial number in n) and more than one sample (up to a polynomial number in n). By appropriately setting δ and c and by using the union bound we can get that for each sample is closer to the center it was sampled from than any other center. Hence, we obtain the following.

COROLLARY 12. *Suppose we have k centers $\{\sigma_1, \ldots, \sigma_k\}$, and the parameters $\beta_1 \leq \cdots \leq \beta_k$. Let $\mathcal{W}$ be an arbitrary set of weights for choosing each center. Let $m = poly(n)$ be the number of samples taken from this mixture. Suppose all pairs σ_i, σ_j satisfy*

$$\mathbb{K}(\sigma_i, \sigma_j) \geq \frac{C\log(m)}{s'_{\beta_1,1}} \min\left(n^{3/2}, \frac{2\sqrt{n}\log n}{\beta_1}\right),$$

for some constant $C > 0$. If the σ_i are known, then all sample points can be assigned to their closest center and we would have the correct clustering with probability $1 - n^{-c}$ for some constant c.

PROOF. For each sample $\pi \sim \mathcal{M}(\sigma_i, \beta_i)$, using the above Lemma 11, we can guarantee that for any other center σ_j, since by assumption $\mathbb{K}(\sigma_i, \sigma_j)$ satisfies the inter-center separation of Lemma 11, $\mathbb{K}(\pi, \sigma_i) < \mathbb{K}(\pi, \sigma_j)$ with probability $1 - n^{-c}$ for some c. By using the union bound, the proof is complete. □

It is useful to note if the centers were known in the Gaussian mixture case, for an analogous claim to Lemma 11, we would only need a separation between the centers that is $\Omega(\log n)$ times the maximum variance.

5. CONCLUSIONS

In this work we initiated the formal study of mixtures of Mallows distributions and prove the impossibility of learning arbitrary mixtures in the most general case when the parameters are different and the learnability of the components are identifiable when the Mallows parameter is the same and known. We point out that the setting where the centers are well-separated is an algorithmically easier setting. Our work suggests that while permutations may not enjoy as nice mathematical properties as Gaussians, they still posses structural characteristics such as embeddability that can still be exploited towards analyzing the corresponding mixture learning problem.

It would be interesting to investigate the feasibility of polynomial time algorithms for learning mixtures when $k = \Theta(1)$, center separation is $\tilde{\Theta}(\sqrt{n}/\beta^c)$ and neither the centers nor the parameters is known. While [5] has settled the question for $k = 2$, it would be interesting to see whether we can develop algorithms requiring polynomial number of samples for arbitrary k, as has been done for other well-behaved distributions [1,10,16]. As we mentioned earlier, it will also be interesting to study the learnability of other choice models in the mixture setting.

6. REFERENCES

[1] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Near-optimal-sample estimators for spherical Gaussian mixtures. In *NIPS*, 2014.

[2] A. Anandkumar, D. Hsu, F. Huang, and S. Kakade. Learning mixtures of tree graphical models. In *NIPS*, pages 1061–1069, 2012.

[3] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. In *COLT*, pages 1–34, 2012.

[4] S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *STOC*, pages 247–257, 2001.

[5] P. Awasthi, A. Blum, O. Sheffet, and A. Vijayaraghavan. Learning mixtures of ranking models. In *NIPS*, 2014.

[6] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

[7] N. Bhatnagar and R. Peled. Lengths of monotone subsequences in a Mallows permutation. *Probability Theory and Related Fields*, To appear.

[8] M. Braverman and E. Mossel. Sorting from noisy information. *CoRR, abs/0910.1191*, 2009.

[9] L. M. Busse, P. Orbanz, and J. M. Buhmann. Cluster analysis of heterogeneous rank data. In *ICML*, pages 113–120, 2007.

[10] S.-O. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014.

[11] K. Chaudhuri and S. Rao. Beyond Gaussians: Spectral methods for learning mixtures of heavy-tailed distributions. In *COLT*, volume 4, page 1, 2008.

[12] F. Chierichetti, A. Dasgupta, R. Kumar, and S. Lattanzi. On reconstructing a hidden permutation. In *RANDOM*, pages 604–617, 2014.

[13] A. Dasgupta, J. Hopcroft, J. Kleinberg, and M. Sandler. On learning mixtures of heavy-tailed distributions. In *FOCS*, pages 491–500, 2005.

[14] S. Dasgupta. Learning mixtures of Gaussians. In *FOCS*, pages 634–644, 1999.

[15] S. Dasgupta and L. J. Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

[16] C. Daskalakis and G. Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In *COLT*, pages 1183–1213, 2014.

[17] J. Doignon, A. Pekec, and M. Regenwetter. The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69(1):33–54, 2004.

[18] D. Dubhashi and A. Panconesi. *Concentration of Measure for the Analysis of Randomised Algorithms.* Cambridge University Press, 2009.

[19] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *COLT*, pages 53–62, 1999.

[20] I. C. Gormley and T. B. Murphy. Exploring voting blocs within the Irish electorate: a mixture modeling approach. *J. Am. Stat. Assoc.*, 103(483):1014–1027, 2008.

[21] I. C. Gormley and T. B. Murphy. A mixture of experts model for rank data with applications in election studies. *Ann. Appl. Stat.*, 2(4):1452–1477, 2008.

[22] M. Hardt and E. Price. Sharp bounds for learning a mixture of two Gaussians. Technical Report 1404.4997v1, ArXiv, 2014.

[23] R. Horn and C. Johnson. *Matrix Analysis*. Matrix Analysis. Cambridge University Press, 2012.

[24] D. Hsu and S. M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *ITCS*, pages 11–20, 2013.

[25] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two Gaussians. In *STOC*, pages 553–562, 2010.

[26] A. T. Kalai, A. Moitra, and G. Valiant. Disentangling Gaussians. *Commun. ACM*, 55(2):113–120, 2012.

[27] R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. *SIAM J. Comput.*, 38(3):1141–1156, 2008.

[28] T. Lu and C. Boutilier. Learning mallows models with pairwise preferences. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 145–152, 2011.

[29] C. L. Mallows. Non-null ranking models I. *Biometrika*, 44(1-2):114–130, 1957.

[30] C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer, 1998.

[31] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.

[32] M. Meila and H. Chen. Dirichlet process mixtures of generalized Mallows models. In *UAI*, pages 358–367, 2010.

[33] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

[34] T. B. Murphy and D. Martin. Mixtures of distance-based models for ranking data. *Computational statistics & data analysis*, 41(3):645–655, 2003.

[35] R. O'Donnell and R. A. Servedio. Learning mixtures of product distributions over discrete domains. *SIAM J. Comput.*, 37(5):1536–1564, 2008.

[36] S. on Chan, I. Diakonikolas, R. A. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.

[37] K. Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[38] Y. Rabani, L. J. Schulman, and C. Swamy. Learning mixtures of arbitrary distributions over large discrete domains. In *ITCS*, pages 207–224, 2014.

[39] D. Titterington and U. Smith, A.; Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.

[40] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci.*, 68(4):841–860, 2004.

[41] H. P. Young. Optimal voting rules. *The Journal of Economic Perspectives*, 9(1):51–64, 1995.