Concentric Mixtures of Mallows Models for Top-k Rankings: Sampling and Identifiability

Fabien Collas, Ekhine Irurozki

▶ **To cite this version:**

Fabien Collas, Ekhine Irurozki. Concentric Mixtures of Mallows Models for Top-k Rankings: Sampling and Identifiability. 38th International Conference on Machine Learning, Jul 2021, Online, France. hal-03353642

HAL Id: hal-03353642
https://hal.science/hal-03353642

Submitted on 24 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Concentric Mixtures of Mallows Models for Top-k Rankings: Sampling and Identifiability

Fabien Collas [1] **Ekhine Irurozki** [1 2]

Abstract

In this paper, we study mixtures of two Mallows models for top-k rankings with equal location parameters but with different scale parameters (a mixture of concentric Mallows models). These models arise when we have a heterogeneous population of voters formed by two populations, one of which is a subpopulation of expert voters. We show the identifiability of both components and the learnability of their respective parameters. These results are based upon, first, bounding the sample complexity for the Borda algorithm with top-k rankings. Second, we characterize the distances between rankings, showing that an off-the-shelf clustering algorithm separates the rankings by components with high probability -provided the scales are well-separated.As a by-product, we include an efficient sampling algorithm for Mallows top-k rankings. Finally, since the rank aggregation will suffer from a large amount of noise introduced by the non-expert voters, we adapt the Borda algorithm to be able to recover the ground truth consensus ranking which is especially consistent with the expert rankings.

1. Introdution

Ranked data has been subject to study in different communities starting with Social Choice (Bartholdi et al., 1989), Bioinformatics (Vitelli et al., 2018), and recently in Machine Learning (Busa-Fekete et al., 2014) since rankings arise naturally when ordering items in order of preference. Top-k rankings arise in practice when voters, human or software, see all the items but provide a ranking of their most preferred k items. Examples of top-k rankings are the results displayed in a search engine, which contain just the top 10 most relevant, related search results, out of possibly millions of results.

The Mallows model (MM) (Fligner & Verducci, 1986; Mallows, 1957) is one of the preferred distributions to model rank data. It belongs to the location-scale family since it is parametrized by a location parameter (a.k.a. central ranking) σ_0 and a non-negative scale (a.k.a. dispersion) parameter, θ. The location parameter is the consensus ranking of the distribution. The probability of any other permutation decreases exponentially with its distance to σ_0, where the distance for rankings is, in general, the Kendall's-τ distance. Finally, the dispersion parameter controls the variance of this decay. For other distances see (Irurozki et al., 2019).

Mixtures of MM populations are divided into different subpopulations, each of which can be modeled with a single MM since each is consistent with noisy realizations of their particular consensus ranking. In this paper, we study the particular context of mixtures where all the location parameters are the same σ_0. We denote this situation as *concentric mixture*.

Real world motivation In this work, we consider the following problem. There is a consensus ranking σ_0 representing a complete ranking of a set of n alternatives, i.e., films that are ordered for the preferences, students that are ranked according to their grades in a particular exam or high-quality and low-quality rankings corresponding to different sampling techniques. However, this consensus ranking is unknown and must be estimated from the realization of rankings provided by a collection of m raters or voters. Each voter has ranked his top $k < n$ alternatives.
The population is heterogeneous: a number of them are low-bias-low-variance rankings that will be close to the consensus and the rest will provide low-bias-high-variance rankings, which will be noisier. The population of rankings is modeled as a mixture whose components are 2-concentric MM: Both components will be centered at the same consensus ranking σ_0, but their spread parameter will be different. Our goals are to show identifiability in this scenario by (1) obtaining the consensus ranking with high probability and (2) distinguishing the rankings from each subpopulation with high probability.

This situation is motivated by general noisy settings.

[1]Basque Center for Applied Mathematics, Bilbao, Spain. [2]LTCI, Telecom Paris, Institut Polytechnique de Paris. Correspondence to: Ekhine Irurozki <irurozki@telecom-paris.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Our contributions We study here concentric mixtures of Mallows for top-k rankings. Our contributions are the following:

- We propose efficient algorithms to compute the probability, sample top-k rankings, and sample linear extensions of top-k rankings under the Mallows model.
- We show identifiability of concentric mixtures with the following two results.
  - We analyze the sample complexity that the Borda algorithm needs to return the consensus ranking with high probability for a sample of top-k rankings.
  - We propose an algorithm that separates the rankings from both components of the mixture with high probability. This is the key to estimate the dispersion parameters of each component.
- We propose an improvement for the Borda algorithm for the estimation of the central ranking in concentric mixtures of top-k rankings.

All these results are valid for mixtures of complete rankings.

Related work Partially ranked data and extensions of the Kendall's-τ distance, in particular, have been analyzed extensively. In (Fagin et al., 2003) a family of extensions is studied and the authors show that they are equivalent up to global constants. Based on this work, constant factor approximation algorithms can be found in (Ailon, 2010). Mallows models for top-k rankings under the distance in (Fagin et al., 2003) are given in (Chierichetti et al., 2018), where the authors argue that previous sampling algorithms based on the Repeated Insertion Model (RIM) (Doignon, 2004) cannot be efficiently adapted to sample top-k rankings and propose a $\mathcal{O}(k^2 4^k + k^2 \log n)$ algorithm to sample top-k rankings from a MM. By taking a different approach from theirs, we propose a sampling algorithm of complexity $\mathcal{O}(k \log k)$ for top-k rankings under the MM.

Theoretical identifiability of the parameters of a mixture of MM was first addressed in (Awasthi et al., 2014) after a large number of papers on practical research (D'Elia & Piccolo, 2005; Lee & Yu, 2012; Meila & Chen, 2010). They obtain a polynomial-time algorithm for the case of two mixtures using tensor decomposition. Despite working with arbitrary separation of the centers of the distributions, their algorithm performance drops as the centers of the distributions get closer, being able to correctly identify 10% of the mixtures when both centers are the same (where for identification the authors counted the fraction of times on which their proposed algorithm returned the true rankings that generated the sample).

The problem of learning mixtures of MM was also addressed in (Chierichetti et al., 2015). They propose and analyze algorithms for different mixture settings. They show identifiability when the dispersion parameter is the same and known for all the components and argue that the learnability of the problem can strongly depend on the separation of the consensus rankings of the mixtures' components. In the present work, we focus on this allegedly difficult setting of concentric mixtures of unknown dispersion and show that even concentric mixtures can be identified polynomially provided that the variances in both distributions are different enough.

The first polynomial time algorithm for provably learning the parameters of a mixture of Mallows models with any constant number of components can be found in (Liu & Moitra, 2018). They show that any two mixtures of top-k Mallows models whose components are far from each other and from the uniform distribution in total variation distance are far from each other as mixtures too, provided that $n > 10k^2$. We improve their results showing that for a component close to uniformity and smaller values of n the separation can be done polynomially.

A growing body of recent papers consider simultaneously partial preferences and mixtures of probabilistic models, i.e., Plackett-Luce (Mollica & Tardella, 2017), proposing provably efficient learning algorithms (Liu et al., 2019), sampling linear extensions (Zhao & Xia, 2019), characterizing identifiability (Zhao & Xia, 2020). In this work, we extend the scenario to the Mallows model.

Mallows model belongs to the location-scale family of distributions. The most prominent member of this family is the Gaussian and therefore both Mallows model and Gaussian are usually considered to be analogous. Nonetheless, Mallows model lacks many interesting properties of the Gaussian. In this paper, we show that on the other hand, Mallows model has interesting properties that are not present on the Gaussian such as identifiability of concentric mixtures.

We assume for the theoretical results that the number of alternatives in the rankings is $n > 3$, following traditional assumptions on the computer theory literature. There is often a phase transition between $n \leq 3$ and $n \geq 4$, for example, the NP-hardness result for the Kemeny ranking problem (which Borda approximates) holds provided $n \geq 4$, see (Dwork et al., 2001).

This paper is organized as follows. Section 2 gives background on rankings and distances. Section 3 details the Mallows models for partial permutations and shows efficient ways of dealing with them, sampling, or computing statistics. Section 4.1 shows how to separate both subpopulations of concentric MM. Section 4.2 addresses the problem of the estimation of the consensus ranking. Finally, Sec-

tion 5 details the experimental evaluations and Section 6 concludes the paper.

2. Preliminaries

The group of permutations of n items is denoted S_n.The identity permutation is $e = 1, 2, \ldots, n$, the group operation is the composition $\sigma \cdot \pi$, denoted $\sigma\pi$, and the inverse of σ is denoted σ^{-1}. We consider that permutation σ represents a ranking of items, where $\sigma(i)$ is the rank of item i.

Every permutation $\sigma \in S_n$ can be uniquely represented by its *inversion vector* $\mathbf{V}(\sigma) = (V_1(\sigma), \ldots, V_{n-1}(\sigma))$,

$$V_j(\sigma) = \sum_{i>j} \mathbb{I}[\sigma(i) < \sigma(j)], \tag{1}$$

where $\mathbb{I}$ is the indicator function and $0 \leq V_j(\sigma) < n - j$. Inversion vectors are also denoted Lehmer codes. There exist two different definitions of the inversion vectors (Fligner & Verducci, 1986; Mandhani & Meila, 2009): this one counts the number of items smaller than $\sigma(j)$ in the tail of σ; the alternative, most popular definition counts the number of items smaller than $\sigma^{-1}(j)$ in the tail of σ^{-1}. When we consider complete permutations both definitions are equivalent. However, with this definition, we can manage naturally partial permutations with consequences that will be clear in subsequent sections, particularly in for its application in sampling algorithms.

The complexity of the bijection between each possible inversion vector and permutations in S_n is $O(n \log n)$ (McClellan et al., 1974).

We consider the Kendall's-τ distance, $d(\sigma, \pi)$, which counts the number of pairwise disagreements between σ and π. We use $d(\sigma)$ to denote $d(\sigma, e)$. Among the properties that this distance satisfies we highlight the following for their implication in the paper. First, the distance to the identity from σ and its inverse is the same $d(\sigma, e) = d(\sigma^{-1}, e)$. Second, the distance is label invariant $d(\sigma, \pi) = d(\sigma\tau, \pi\tau)$ for every triplet $\sigma, \pi, \tau \in S_n$.

The relation between the Kendall's-τ distance and inversion vectors comes from the fact that $d(\sigma) = \sum_j V_j(\sigma)$. For top-$k$ rankings we use a generalization of the Kendall's-τ distance that assumes that items that cannot be compared do not increase the distance. This is equivalent to the generalization of (Fagin et al., 2003) with the p parameter equal to 0.

We consider the Mallows model (MM) to model distributions on S_n. MM expresses the probability of ranking $\sigma \in S_n$ as $p(\sigma) \propto \exp(-\theta d(\sigma, \sigma_0))$. We will make use of the convenient observation made in previous paragraphs that claims that $d(\sigma) = \sum_j V_j(\sigma)$ to rewrite the MM as follows (Meila et al., 2007),

$$p(\sigma) = \frac{\prod_{j=1}^{n-1} \exp(-\theta V_j(\sigma\sigma_0^{-1}))}{\psi_n}$$
$$\text{where} \quad \psi_n = \prod_{j=1}^{n-1} \psi_{n,j} = \prod_{j=1}^{n-1} \frac{1 - \exp(-\theta(n-j+1))}{1 - \exp(-\theta)}. \tag{2}$$

We denote as $M(\sigma_0, \theta)$ a MM with consensus ranking (or location parameter) σ_0 and with dispersion parameter θ. A random permutation distributed according to this model is denoted as $\sigma \sim M(\sigma_0, \theta)$. Since the Kendall's-$\tau$ distance is right invariant, we can assume that $\sigma_0 = e$ w.l.o.g. The base of our sampling algorithm is that the probability distribution of each item in the inversion vector $\mathbf{V}(\sigma\sigma_0^{-1}) = (V_1(\sigma\sigma_0^{-1}), \ldots, V_{n-1}(\sigma\sigma_0^{-1}))$ for $\sigma \sim M(\sigma_0, \theta)$ can be expressed as

$$p(V_j(\sigma\sigma_0^{-1}) = r) = \frac{\exp(-\theta r)}{\psi_{n,j}}. \tag{3}$$

It follows that $p(\sigma)$ can be stated as the product of independent factors, $p(\sigma) = \prod_j p(V_j(\sigma\sigma_0^{-1}))$, (Mandhani & Meila, 2009).

Learning the maximum likelihood estimate (MLE) of a MM given a sample of permutations is done in two stages (Mandhani & Meila, 2009). Firstly, the MLE of the central ranking is the Kemeny ranking of the sample (Ali & Meila, 2012). Since this problem is NP-hard (Dwork et al., 2001), usually the Borda ranking is used. Borda can be computed in quasi-linear time and is an unbiased estimator of the Kemeny ranking of a sample distributed according to Mallows model (Fligner & Verducci, 1988). Secondly, the MLE of the dispersion parameters are obtained numerically (Irurozki et al., 2019).

Mixture models are used to combine different simple probability models to model large, heterogeneous populations. In our motivating problem, we consider a heterogeneous population of two subpopulations: one made of expert voters (low-variance rankings) and another one of non-expert voters (low-variance rankings). This is modeled with a mixture of two concentric components where the probability of each permutation is

$$p(\sigma) = r\frac{\exp(-\theta_g d(\sigma, \sigma_0))}{\psi_n} + (1-r)\frac{\exp(-\theta_b d(\sigma, \sigma_0))}{\psi_n}. \tag{4}$$

where there is one consensus ranking σ_0, a dispersion parameter of the low-variance rankings, θ_g, a dispersion of the high-variance rankings, $\theta_b < \theta_g$ and a proportion of low-variance rankings in the population r, denoted mixture

parameter. We denote mixtures in which all the components have the same central ranking as *concentric*.

3. Top-k Ranking Statistics under the Mallows Model

In this section, we study the problems of sampling top-k rankings from a MM, sampling linear extensions of top-k rankings, and computing the probability of a top-k ranking efficiently.

We start with the primary problem in statistics: Computing the probability of a top-k ranking σ efficiently. The probability of σ is the sum of the probabilities of all its linear extensions, $p(\sigma) = \sum_{\sigma' \in L(\sigma)} p(\sigma')$, so a naive approach computes $p(\sigma)$ in $\mathcal{O}((n \log n)(n-k)!)$. We propose a $\mathcal{O}(n + k \log k)$ expression to compute $p(\sigma)$ in the next lemma.

Lemma 1. *The probability of the top-k ranking σ is*

$$p(\sigma) = \exp(-\theta d(\sigma, \sigma_0)) \frac{\psi_{n-k,\theta}}{\psi_{n,\theta}}.$$

The complexity of the previous expression comes from the normalization constant ($\mathcal{O}(n)$, in Equation (2)) and the computation of the distance ($\mathcal{O}(k \log k)$). A proof can be found in the appendix.

Sampling When we consider the problem of sampling complete rankings (instead of top-k rankings), RIM (Doignon, 2004) offers a convenient alternative. RIM samples a ranking in two steps: First, it samples vector $\mathbf{R}(\sigma) = (R_1, \ldots, R_n)$, where $1 \leq R_i \leq i$ and $p(R_i)$ for $\sigma \sim M(\sigma_0, \theta)$ is known (Doignon, 2004). Second, starting with an empty vector σ and letting i range in $[1, n]$, it inserts item i in position R_i of σ, shifting backwards items if necessary. Due to this shifting strategy, $\sigma(i)$ is not known until the last iteration for every $i \leq n$. This means that the only way of sampling a top-k ranking with RIM is to sample a complete ranking and then to censor it. This is clearly a terrible idea, especially when $k << n$. There exits an improvement for this process with complexity $\mathcal{O}(k^2 4^k + k^2 \log n)$ (Chierichetti et al., 2018).

In the following lines, we introduce a new sampling algorithm for top-k rankings with quasi-linear complexity. Although the theoretical foundation (inversion vectors and its bijection) is known, this algorithm to sample top-k rankings in less than exponential time is novel. Our proposed algorithm can also be used to sample complete rankings by setting $k = n$ and its complexity improves over RIM's. Software implementing the algorithms described here is distributed in `https://github.com/ekhiru/top-k-mallows`.

Algorithm 1 samples a top-k ranking $\sigma \sim M(\sigma_0, \theta)$ in time $\mathcal{O}(k \log k)$ and memory $\mathcal{O}(k)$. It is based on the following results: First, Equation (3) gives the probability of each position of the inversion vector independently $\mathbf{V}(\pi\sigma_0^{-1})$: sampling the first k positions is linear in k. Second, it generates the partial permutation $\pi\sigma_0^{-1}$ from the inversion vector with a quasi-linear time bijection (McClellan et al., 1974). Finally, the top-k ranking σ is distributed as $\sigma \sim M(\sigma_0, \theta)$ where $\sigma = \pi^{-1}$ (Mallows, 1957).

Algorithm 1 Sample top-k in $O(k \log k)$

Data: n, k, θ, σ_0
Result: σ: Top-k ranking of n items distributed according to $M(\sigma_0, \theta)$
for $j \in [1, k]$ **do**
  $V_j(\pi\sigma_0^{-1})$ = random choice in $[n-j]$ with choice probabilities of Eq. (3)
  $\pi\sigma_0^{-1}$ = transform $V(\pi\sigma_0^{-1})$ with the bijection in (McClellan et al., 1974)
  return π^{-1}
end

Sampling linear extensions A similar approach is used to sample a linear extension of a top-k ranking in $\mathcal{O}((n-k)\log(n-k))$, we refer to the supplementary material for the pseudo-code. It follows directly from the previous result. Sampling a linear extension is done by sampling $V_j(\sigma)$ from $j \geq k$. Then, obtaining σ is $\mathcal{O}((n-k)\log(n-k))$ (McClellan et al., 1974).

Expressions for the expected distance and variance are included in the appendix for completeness and also appear in (Busa-Fekete et al., 2019).

4. Identifiability of Concentric Mixtures

Identifiability of concentric mixtures has been claimed to be the most difficult scenario in mixture identifiability (Chierichetti et al., 2015). Indeed, the concentric mixtures of Normal components, which belongs to the same family as the MM, are known to be non-identifiable.

Identifiability is guaranteed if we can (1) recover the ground truth ranking and (2) separate the rankings in each population. Each of these points is addressed in the following sections to claim identifiability for concentric mixtures.

4.1. Provably separation of the sub-populations of each component

In this section we consider the problem of separating the rankings of the two components of a concentric mixture of MM. We propose an algorithm that can separate the two sub-populations under mild conditions of the separation of the dispersion parameters among the mixture components

and the mixture parameter r. Our proposed algorithm is based on finding the separation of the mean distance of each top-k ranking σ to all the others in the sample, which is defined as follows,

$$\delta_\sigma = \frac{1}{|S|-1} \sum_{\sigma' \in S \setminus \{\sigma\}} d(\sigma, \sigma'). \tag{5}$$

Recall that the expected distances $\mathbb{E}[d(\sigma, \sigma_0)]$ and $\mathbb{E}[\delta_\sigma]$ for $\sigma \sim M(\sigma_0, \theta)$ only depend on θ. We will use the following observation along this section (Korba et al., 2017): The expected distance between a random Mallows ranking $\sigma \sim M(\sigma_0, \theta)$ and the consensus ranking σ_0, $\mathbb{E}[d(\sigma_0, \sigma)]$, is bounded by the expected pairwise distance of two random Mallows permutations, $\sigma, \sigma' \sim M(\sigma_0, \theta)$ as follows.

$$\frac{1}{2}\mathbb{E}[d(\sigma', \sigma)] \leq \mathbb{E}[d(\sigma_0, \sigma)] \leq \mathbb{E}[d(\sigma', \sigma)]. \tag{6}$$

First, we introduce some notations recalling our motivating example: a concentric mixture represents a population of voters with two homogeneous sub-populations, a sub-populations of low-bias-low-variance rankings distributed according to $M(\sigma_0, \theta_g)$ and another sub-populations of low-bias-high-variance rankings i.i.d. as $M(\sigma_0, \theta_b)$ for $\theta_g > \theta_b$. Let $\beta \sim M(\sigma_0, \theta_b)$ and $\gamma \sim M(\sigma_0, \theta_g)$ be two random top-k rankings, i.e., β belongs to the group of ***bad***, high-variance rankings and γ to the group of ***good***, low-variance rankings. In this section we show that $\mathbb{E}[\delta_\beta]$ and $\mathbb{E}[\delta_\gamma]$ are well separated and give an algorithm that separates the two sub-populations in $\mathcal{O}(m^2)$ time provided that the components are sufficiently far from each other.

Now, we show an auxiliary lemma that bounds the expected distance between random rankings of different components, γ and β.

Lemma 2. *The expected distance between two rankings of different components $\mathbb{E}[d(\beta, \gamma)]$ is bounded as follows*

$$\mathbb{E}[d(\beta, \sigma_0)] \leq \mathbb{E}[d(\beta, \gamma)] \leq \mathbb{E}[d(\beta', \beta)]. \tag{7}$$

The following result shows that the expected mean distances δ_α and δ_β of top-k rankings α, β of different components are well separated.

Theorem 3. *Let $M(\sigma_0, \theta_g)$ and $M(\sigma_0, \theta_b)$ be the two components of a concentric mixture of top-k rankings in which $\mathbb{E}[d(\beta, \sigma_0)] \geq c \cdot \mathbb{E}[d(\gamma, \sigma_0)]$ for $c \geq 2$. Let $r \in [0, 1]$ be the mixture parameter. The expected mean distance between rankings of different components $\beta \sim M(\sigma_0, \theta_b)$ and $\gamma \sim M(\sigma_0, \theta_g)$ is bounded as follows,*

$$\mathbb{E}[\delta_\beta - \delta_\gamma] \geq \mathcal{O}(c \cdot r \cdot \mathbb{E}[d(\gamma, \sigma_0)]). \tag{8}$$

Theorem 3 suggests that a clustering algorithm in δ_σ for every $\sigma \in S$ can segment the population by generating component. We show that, indeed, a single linkage clustering algorithm can separate the sub-populations with proven guarantees.

Theorem 4. *Let $M(\sigma_0, \theta_g)$ and $M(\sigma_0, \theta_b)$ be the two components of a concentric mixture of top-k rankings in which $\mathbb{E}[d(\beta, \sigma_0)] \geq c \cdot \mathbb{E}[d(\gamma, \sigma_0)]$ for $c > 2$. Let $r \in (0, 1]$ be the mixture parameter. There exists an algorithm that separates the samples from both components with probability $1 - \epsilon$ in $\mathcal{O}(m^2)$ when the number of samples is at least*

$$m > \left(\frac{n(n-1)}{(c-2) \cdot r \cdot \mathbb{E}[\gamma, \sigma_0]} \right)^2 \frac{\log(2/\epsilon)}{2}. \tag{9}$$

Proof. First, and based on a direct application of Hoeffding's inequality, note that for $d(\sigma, \sigma')$ $i.i.d.$ random variables with range $\frac{n(n-1)}{2}$ then, it holds that with probability at least $1 - \epsilon$:

$$|\delta_\sigma - \mathbb{E}[d(\sigma, \sigma')]| \leq \frac{n(n-1)}{2} \sqrt{\frac{\log(2/\epsilon)}{2(m-1)}}. \tag{10}$$

Now, we consider the Single-linkage clustering, which separates two components when

$$\mathbb{E}[\delta_\beta - \delta_\gamma] \geq |\delta_\beta - \mathbb{E}[\delta_\beta]| + |\delta_\gamma - \mathbb{E}[\delta_\gamma]|, \tag{11}$$

Thus, the theorem holds if Equation (11) holds. Hence, using Equation (10), the following holds.

$$\begin{aligned} &|\delta_\beta - \mathbb{E}[\delta_\beta]| + |\delta_\gamma - \mathbb{E}[\delta_\gamma]| \\ &\quad \leq 2 \cdot \frac{n(n-1)}{2} \sqrt{\frac{\log(2/\epsilon)}{2m}} \\ &\quad \leq (c-2) \cdot r \cdot \mathbb{E}[\gamma, \sigma_0] \\ &\quad \leq \mathbb{E}[\delta_\beta - \delta_\gamma], \end{aligned} \tag{12}$$

We obtain a bound for m using the second inequality, which concludes the proof.

□

Hence, we have shown that under certain conditions about the expected distances of both populations, we can separate both components with high probability. This would allow us, given that we also know the central permutation, to learn θ_g and θ_b dispersion parameters, using the maximum likelihood estimation as described in (Irurozki, 2014), adapted to the case of top-k rankings here.

Computing δ_σ for every permutation in the sample requires $\mathcal{O}(m^2)$ distance calculations. However, in practice, we

can approximate this value with high probability, reducing the number of distance computations using a result from a corollary of Hoeffding's bound.

Applicability in real-world scenarios is related to the question of whether $c \geq 2$, i.e., "twice the distance in expectation", is a large separation. We need to elaborate on the answer by linking it to the asymptotic of the distance. The maximum Kendal's-τ distance between rankings of n alternatives is at most $n * (n-1)/2$ (grows quadratic with n), while Theorem 4 holds provided that the ratio of the expected distances for two components is linear (the double). Therefore, "twice the distance in expectation" compared to the maximum distance is not a large separation for relatively small rankings ($n = 5$) but it becomes arbitrarily small as we increase n. Moreover, we can deal with cases where one of the distribution is very close to uniformity, which can be a problematic case for some approaches (Liu & Moitra, 2018).

Relation to total variation (TV) distance We have choosen to elaborate the results regarding the expected distance for interpretability on the dataset, but certainly, a discussion on the TV is in order. The TV distance is very easily computed among concentric MM. The TV distance between distributions p_1 and p_2 is

$$d_{TV}(p_1, p_2) = \max_{\sigma \in S_n} |p_1(\sigma) - p_2(\sigma)|. \tag{13}$$

Since in the model $MM(\sigma_0, \theta)$ the permutation σ with the largest (resp. lowest) probability value is the mode σ_0 (resp. the antimode $\bar{\sigma}_0 = (\sigma_0(n), \sigma_0(n-1), \ldots, \sigma_0(1))$, for concentric p_g, p_b ($p_g = M(\sigma_0, \theta_g), p_b = M(\sigma_0, \theta_b)$) it turns out that

$$d_{TV}(p_1, p_2) = p_g(\sigma_0) - p_b(\bar{\sigma}_0) \tag{14}$$

This observation allows bounding the d_{TV} when the separability condition $\mathbb{E}[d(\beta, \sigma_0)] \geq c \cdot \mathbb{E}[d(\gamma, \sigma_0)]$ is satisfied. Let us assume that the condition is satisfied and that $\mathbb{E}[d(\gamma, \sigma_0)]$ and c are in their maximum values (which are a function of n). This is the worst case scenario, since d_{TV} decreases with both $\mathbb{E}[d(\gamma, \sigma_0)]$ and c. For this setting, we can easily obtain the dispersion parameters θ_g and θ_b with numerical methods. To conclude, we plug this values on Eq (14) obtaining the desired bound.

4.2. Estimating the consensus ranking of top-k rankings

In this section we study the following problem: given a sample of top-k rankings distributed as a concentric mixture of MM, find the MLE of the central ranking. Since all the components of a concentric mixture have the same consensus ranking σ_0, this problem boils down to a rank aggregation problem, as in the single-component case: the MLE is exactly given by the Kemeny ranking. Unfortunately, computing this ranking is NP-hard for $n > 4$, (Dwork et al., 2001).

For a sample of complete rankings drawn from a MM, Borda is an approximation to the Kemeny ranking (Fligner & Verducci, 1988). Moreover, Borda is quasi-linear in time and outputs the correct σ_0 w.h.p. with a polynomial number of samples (Caragiannis et al., 2013). However, there is no quality result for the case in which the sample consists of top-k rankings i.i.d. from a MM. In the next lines, we provide a sample complexity for Borda over top-k rankings from a MM and then extend it to the case of concentric mixtures.

A crucial difference between complete and top-k rankings drawn from a MM[1], is that in top-k Mallows rankings σ the probability of observing item i, i.e., $p(\sigma(i) \leq k)$, decreases with i. Intuitively, this means that Borda will need fewer samples to *guess the rank* of smaller i's. We formalize this intuition in the next result: We bound the number of samples that Borda requires to rank items i and $i+1$ in the correct order with probability $1 - \epsilon$.

For this analysis, we first need the following definition.

Definition 5. *Let $i \in [0, n-1]$, $k \in [0, n]$.*

$$\Delta^{ik} = \sum_{\sigma : \sigma(i) \leq k} p(\sigma) - \sum_{\sigma : \sigma(i+1) \leq k} p(\sigma). \tag{15}$$

Despite in general there is no closed-form expression for Δ^{ik}, we can give a convenient expression for the case where $i = 1$, Δ^{1k}.

Lemma 6. *The minimum value for Δ^{ik}, $\Delta^{1k} = \min_i \Delta^{ik}$ can be computed in $O(k^2 + kn)$ as follows:*

$$\begin{aligned} \Delta^{1k} = &\sum_{r_1=0}^{k-1} \frac{\exp(-\theta r_1)}{\psi_{n,1}} \\ &- \sum_{r_1=0}^{k-1} \sum_{r_2=0}^{k-2} \frac{\exp(-\theta r_1)\exp(-\theta r_1)}{\psi_{n,1}\psi_{n,2}} \\ &- \sum_{r_1=k}^{n} \sum_{r_2=0}^{k-1} \frac{\exp(-\theta r_1)\exp(-\theta r_2)}{\psi_{n,1}\psi_{n,2}}. \end{aligned} \tag{16}$$

Based on the above results, we can bound the sample complexity of Borda for giving the correct ordering of items i and $i+1$.

Theorem 7. *Let S be a sample of top-k rankings drawn from a MM with dispersion θ. Borda for sample S orders*

[1] We assume, as in previous sections and w.l.o.g., that $\sigma_0 = e$.

the pair of items i and $i+1$ correctly with probability $1-\epsilon$ when the number of rankings in the sample is at least

$$m \geq \mathcal{O}\left(n^2 \log \epsilon^{-1} \left(\frac{k^2(1-\exp(-\theta))^2}{1-\exp(-\theta n)} - i\Delta^{1k}\right)^{-2}\right). \tag{17}$$

The above theorem formalizes the intuitive idea that the sample complexity of Borda to order correctly items i and $i+1$ on a sample of top-k rankings increases polynomially with i. Therefore, this bound generalizes to the sample complexity of Borda to obtain the correct ranking w.h.p. for a sample of top-k drawn from a mixture of MM with parameters σ_0, θ_g and θ_b.

Corollary 8. *Borda returns the correct central ranking σ_0 when the number of samples in the both component satisfies Equation (17) with $i = n$ and $\theta = \theta_b$.*

Expert Borda In a practical point of view, we can improve Borda quality in the problem of estimating σ_0 for concentric mixtures. The improvement is based on the following observations: (1) The sample complexity is smaller for the component $M(\sigma_0, \theta_g)$ than for $M(\sigma_0, \theta_b)$. (2) we can identify the rankings in $M(\sigma_0, \theta_g)$ w.h.p. as shown in Section 4.1. With this in hand, we propose an improvement for Borda that (1) identifies the low-variance rankings (2) constructs a top-k ranking by aggregating those rankings drawn from $M(\sigma_0, \theta_g)$ and (3) fills the missing $n-k$ positions with the data of the whole sample of rankings. We denote this approach expert Borda and show its improvement empirically in the experimental section.

Summarizing identifiability Let us put all the pieces together: First, we compute δ_σ (Eq (5)) for each ranking. Since this δ_σ will be very different for rankings σ in each component (by Theorem 3), applying the clustering algorithm on δ will separate the rankings in each component. The centers will be the same for both components provided that the number of samples is larger than a known bound (by Theorem 4). Then, identifiability follows, since the spread parameters are computed exactly in linear time, see Section 2 or (Critchlow et al., 1991). Otherwise -if the number of samples is smaller than the statement in Theorem 4- and the centers of both components are not the same, one can use the Expert Borda algorithm to recover the ground true ranking, see Section 4.2.

5. Experiments

In this section, we validate empirically our proposal. The experimental framework is as follows. In the first two experiments, we generate a sample of partial rankings, using Algorithm 1, with parameters $n = 30$ and $k = 10$, from a mixture of concentric MM, both centered at a random σ_0 and with two dispersion parameters, θ_b, θ_g. The mixture parameter is denoted r.

5.1. Voters separability

The goal of this experiment is to separate the two components of rankings based on the mean distance of each of the rankings to all the others as defined in Equation (5). To compute these distances, we used the generalization of Kendall's-τ distance described in Section 2.

The experimental framework is as follows:

- $m_g = 40$ rankings from a $M(\sigma_0, \theta_g)$ such that $\mathbb{E}[d(\gamma, \sigma_0)] \in \{3, 8, 13, \ldots, 48\}$
- $m_b = 60$ rankings from a $M(\sigma_0, \theta_b)$ such that $\mathbb{E}[d(\beta, \sigma_0)] = c \cdot \mathbb{E}[d(\gamma, \sigma_0)]$ with $40 > c \geq 3$ and $\mathbb{E}[d(\gamma, \sigma_0)] \leq 217$ (bound corresponding to the uniform distribution).

For a sample distributed as the model detailed above, we compute each δ_σ and apply a 2-means clustering. The error is measured as the percentage of badly-separated rankings in the sample.

We can observe the results of this experiment in Figure 1 (a). As we can see, as the ratio c increases, the separation gets more accurate. Indeed, for $c \geq 9$ the percentage of wrongly separated rankings is almost always below 10% and is very close to 0 when high-variance rankings are close to the uniform distribution. These results are consistent with Theorem 3.

5.2. Consensus estimate with expert Borda

In these lines, we evaluate *eBorda* (expert Borda) algorithm after Corollary 8. It estimates σ_0 for a sample drawn from a mixture of two concentric Mallows models. We show that eBorda outperforms Borda.

Intuitively, and in the context of the low-variance/high-variance rankings, this aggregation algorithm is based on the following ideas.

- For the top-k items: the low-variance will provide an accurate samples, however
- the probability of observing the items that are not in the top-k is larger for high-variance rankings, although the quality of the provided information might not be very good (the sample complexity is indeed larger, as shown in Section 4.1).

We assume that the number of high-variance rankings is larger than the number of low-variance rankings and our aggregation should *pay more attention* to the low-variance

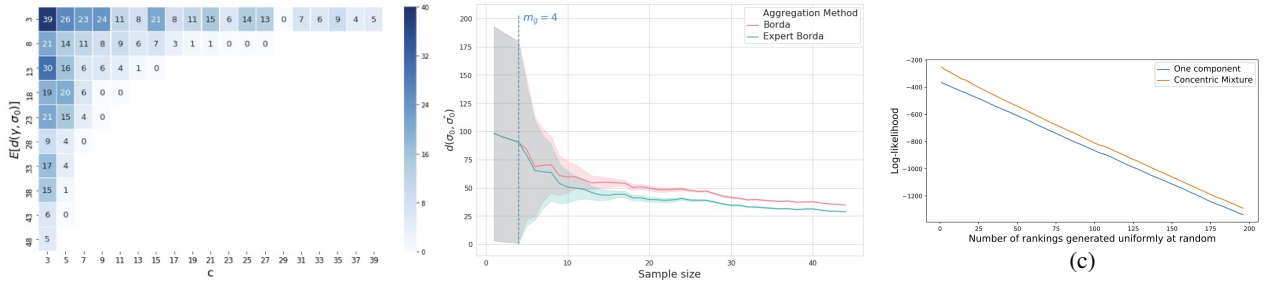

Figure 1. (a) Total error (%) of separation between rankings in each component, using K-means. (b) Distances between the consensus and its estimates, with different sizes of samples and two aggregation methods. (c) Comparison of the log-likelihoods of a distribution of rankings, considering it as a single component of Mallows model or as a concentric mixture of Mallows models with different dispersion parameters

opinion, but also be sure to include the high-variance rankings, specially if there is missing information.

Hence, we sample a population of rankings drawn from a mixture of two Mallows models. We will take $m_g = 4$ rankings from $M(\sigma_0, \theta_g)$, and $m_b = 40$ from $M(\sigma_0, \theta_b)$. θ_g and θ_b are chosen such that $\mathbb{E}[d(\sigma_0, \gamma)] = 10$ and $\mathbb{E}[d(\sigma_0, \beta)] = 75$.

The estimate $\hat{\sigma_0}$ for σ_0 is computed with both the Borda method and with our proposed $eBorda$, using the same growing sample, with size $\{1, 2, 3, ..., 44\}$: first 1 ranking, then 2, ... starting with the low-variance rankings. For each number of rankings, we measure the error of the estimate, $d(\sigma_0, \hat{\sigma_0})$, as the mean between the maximum and the minimum distance the estimate could have to the consensus if all its positions were filled. This is repeated ten times and average values for the distances are given.

The results are given in Figure 1 (b), where the x-axis indicates the number of rankings considered for the estimation, and the y-axis gives the error of estimation. The vertical line marks the step from which high-variance rankings are added to the sample. Finally, the aggregation of top-k rankings results on a single top-k' ranking where $k \leq k' \leq n$. When measuring our top-k' estimate to our complete σ_0 consensus, not all the pairs can be compared. The shadow around the curve is the bound on the distances between any linear extension of the partial estimate and the consensus. Hence the larger k', the smaller the shadow.

As expected, we can observe that $eBorda$ can perform better than Borda. Indeed, the separation of both components allows us to keep a more accurate estimate for the first k positions, using only the low-variance rankings to estimate it. Nevertheless using the high-variance rankings afterward allows us to complete the estimate into a full ranking, making the uncertainty of the error decrease faster as we can see with the shadows around the curves narrowing.

5.3. Semi-synthetic example

To test the identifiability on real data, we used a dataset already used in (Fligner & Verducci, 1986), for which 98 college students were asked to rank five words according to its strength of association with the word "idea". The five words to classify were: (1) thought, (2) play, (3) theory, (4) dream, and (5) attention. These were to be ranked from 1 to 5, 5 being the most strongly associated with the target word. In our present example, $m = 98$ and $n = 5$.

We assumed the dataset to be distributed according to a Mallows Model and estimated its Maximum Likelihood Estimates, $\sigma_0 = 5, 1, 4, 3, 2$ and $\theta_g = 1.43$. We then simulated a sample of $2 \cdot m$ raters generated uniformly at random.

Simulated raters were added, one by one, and at each step, two different models were fitted: (1) a MM and (2) a mixture of concentric MM, for which each component is determined performing a $2 - Means$ clustering, using the same procedure as in Section 5.1. Then, we compute the log-likelihood of each model. The results are represented in Figure 1 (c), we can see that the mixture of concentric MM fits the data better, even in this case in which the number of random is large, larger than $n!$.

6. Conclusions

In this paper, we have studied the allegedly most difficult setting in the learnability of mixtures of location-scale distributions: the case in which the location parameters are the same. We denote this case as concentric and focus on the Mallows model for top-k rankings. This situation arises when we have a low-bias-low-variance population and a low-bias-high-variance subpopulation. For example, when there are two populations of voters (expert/non-expert).

We have proposed a $\mathcal{O}(k \log k)$ sampling algorithm for top-k ranking, which dramatically reduces the requirement of

the samplers in the literature.

We have also proposed an algorithm for the learnability of the parameters of the concentric mixture of top-k rankings with a high probability in polynomial time. It is based on our two following results: We have bounded the sample complexity of the Borda algorithm to recover the ground truth consensus ranking. And second, we have been able to separate the rankings of each component in polynomial time with high probability.

Interesting extensions to our work could be to generalize our results to concentric Mallows mixtures of more than two components, non-concentric mixtures of Mallows model, or other models such as Plackett-Luce's model.

Acknowledgements

We are very grateful to Ioannis Caragiannis for the helpful discussion. This work is also partially funded by the Elkartek program of the Basque Government (2019 KK-2019/00072), by the Industrial Chair "Data science & artificial intelligence for digitalized industry & services" from Télécom Paris, France and by the Spanish Ministry of Innovation (TIN2017-82626-R).

References

Ailon, N. Aggregation of partial rankings, p-ratings and top-m lists. *Algorithmica (New York)*, 2010. ISSN 01784617. doi: 10.1007/s00453-008-9211-1.

Ali, A. and Meila, M. Experiments with Kemeny ranking: What works when? *Mathematical Social Sciences*, 64(1):28–40, 2012. ISSN 0165-4896. doi: http://dx.doi.org/10.1016/j.mathsocsci.2011.08.008. URL http://www.sciencedirect.com/science/article/pii/S0165489611000989.

Awasthi, P., Blum, A., Sheffet, O., and Vijayaraghavan, A. Learning mixtures of ranking models. In *Advances in Neural Information Processing Systems*, pp. 2609–2617, 2014.

Bartholdi, J. J., Tovey, C. A., and Trick, M. A. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 1989. ISSN 01761714. doi: 10.1007/BF00295861.

Busa-Fekete, R., Hüllermeier, E., and Szörényi, B. Preference-Based Rank Elicitation using Statistical Models: The Case of Mallows. In *Proceedings of the 31th International Conference on Machine Learning, (ICML)*, pp. 1071–1079, 2014. URL http://jmlr.org/proceedings/papers/v32/busa-fekete14.html.

Busa-Fekete, R., Fotakis, D., Szörényi, B., and Zampetakis, M. Optimal Learning of Mallows Block Model. *arXiv preprint arXiv:1906.01009*, 2019.

Caragiannis, I., Procaccia, A. D., and Shah, N. When Do Noisy Votes Reveal the Truth? In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, pp. 143–160, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1962-1. doi: 10.1145/2482540.2482570. URL http://doi.acm.org/10.1145/2482540.2482570.

Chierichetti, F., Dasgupta, A., Kumar, R., and Lattanzi, S. On Learning Mixture Models for Permutations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, ITCS '15, pp. 85–92, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3333-7. doi: 10.1145/2688073.2688111. URL http://doi.acm.org/10.1145/2688073.2688111.

Chierichetti, F., Dasgupta, A., Haddadan, S., Kumar, R., and Lattanzi, S. Mallows Models for Top-k Lists. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 4382–4392. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7691-mallows-models-for-top-k-lists.pdf.

Critchlow, D. E., Fligner, M. A., and Verducci, J. S. Probability Models on Rankings. *Journal of Mathematical Psychology*, 35:294–318, 1991.

D'Elia, A. and Piccolo, D. A mixture model for preferences data analysis. *Computational Statistics & Data Analysis*, 49(3):917–934, 2005. ISSN 0167-9473. doi: 10.1016/j.csda.2004.06.012. URL http://www.sciencedirect.com/science/article/pii/S0167947304001987.

Doignon, J.-P. The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69(1):33–54, 2004.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. Rank aggregation methods for the Web. In *International conference on World Wide Web*, WWW '01, pp. 613–622, New York, NY, USA, 2001. ACM. ISBN 1-58113-348-0. doi: 10.1145/371920.372165. URL http://doi.acm.org/10.1145/371920.372165.

Fagin, R., Kumar, R., and Sivakumar, D. Comparing Top k Lists . *SIAM Journal on Discrete Mathematics*, 2003. ISSN 0895-4801. doi: 10.1137/s0895480102412856.

Fligner, M. A. and Verducci, J. S. Distance based ranking models. *Journal of the Royal Statistical Society*, 48(3): 359–369, 1986.

Fligner, M. A. and Verducci, J. S. Multistage Ranking Models. *Journal of the American Statistical Association*, 83(403):892–901, 1988. ISSN 01621459. doi: 10.2307/2289322. URL http://www.jstor.org/stable/2289322?origin=crossref.

Irurozki, E. Sampling and learning distance-based probability models for permutation spaces. pp. 42–44, 2014.

Irurozki, E., Calvo, B., and Lozano, J. A. PerMallows: An R package for mallows and generalized mallows models. *Journal of Statistical Software*, 71, 2019. ISSN 15487660. doi: 10.18637/jss.v071.i12.

Korba, A., Clémençon, S., and Sibony, E. A learning theory of ranking aggregation. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 2017.

Lee, P. H. and Yu, P. L. H. Mixtures of weighted distance-based models for ranking data with applications in political studies. *Computational Statistics & Data Analysis*, 56(8):2486–2500, 2012. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2012.02.002. URL http://www.sciencedirect.com/science/article/pii/S0167947312000679.

Liu, A. and Moitra, A. Efficiently learning mixtures of Mallows models. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 627–638. IEEE, 2018.

Liu, A., Zhao, Z., Liao, C., Lu, P., and Xia, L. Learning Plackett-Luce Mixtures from Partial Preferences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33014328.

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

Mandhani, B. and Meila, M. Tractable Search for Learning Exponential Models of Rankings. *Journal of Machine Learning Research*, 5:392–399, 2009.

McClellan, M. T., Minker, J., and Knuth, D. E. The Art of Computer Programming, Vol. 3: Sorting and Searching. *Mathematics of Computation*, 1974. ISSN 00255718. doi: 10.2307/2005383.

Meila, M. and Chen, H. Dirichlet Process Mixtures of Generalized Mallows Models. In *Uncertainty in Artificial Intelligence (UAI)*, pp. 285–294, 2010.

Meila, M., Phadnis, K., Patterson, A., and Bilmes, J. Consensus ranking under the exponential model. In *Uncertainty in Artificial Intelligence (UAI)*, pp. 285–294, Corvallis, Oregon, 2007.

Mollica, C. and Tardella, L. Bayesian Plackett–Luce Mixture Models for Partially Ranked Data. *Psychometrika*, 2017. ISSN 00333123. doi: 10.1007/s11336-016-9530-0.

Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A., and Arjas, E. Probabilistic preference learning with the Mallows rank model. *Journal of Machine Learning Research*, 18 (1), 2018. ISSN 15337928.

Zhao, Z. and Xia, L. Learning Mixtures of Plackett-Luce Models from Structured Partial Orders. In *Advances in Neural Information Processing Systems*, pp. 10143–10153, 2019.

Zhao, Z. and Xia, L. Learning Mixtures of Plackett-Luce Models with Features from Top-$ 1 $ Orders. *arXiv preprint arXiv:2006.03869*, 2020.

Concentric mixtures of Mallows models for top-k rankings: Supplementary Materials

1. Proof of Lemma 1

Proof. Let the linear extensions of σ be $L(\sigma)$. The sum of their probabilities is

$$\begin{aligned} p(\sigma) &= \sum_{\sigma' \in L(\sigma)} p(\sigma') = \sum_{\sigma' \in L(\sigma)} \frac{\prod_{j=1}^{n-1} \exp(-\theta V_j(\sigma'))}{\psi_{n,\theta}} \\ &= \sum_{\sigma' \in L(\sigma)} \frac{\prod_{j=1}^{k} \exp(-\theta V_j(\sigma)) \prod_{j=k+1}^{n-1} \exp(-\theta V_j(\sigma'))}{\psi_{n,\theta}} \\ &= \frac{\exp(-\theta d(\sigma)) \sum_{\sigma' \in L(\sigma)} \prod_{j=k+1}^{n-1} \exp(-\theta V_j(\sigma'))}{\psi_{n,\theta}} \\ &= \frac{\exp(-\theta d(\sigma)) \psi_{n-k,\theta}}{\psi_{n,\theta}}, \end{aligned} \tag{1}$$

where $\psi_{k,\theta}$ is defined in Equation (2). The overall complexity is dominated by $\psi_{n,\theta}$, which is $\mathcal{O}(n)$ □

2. Sampling linear extensions

Algorithm 1 Sampling linear extensions in $\mathcal{O}((n-k)\log(n-k))$

Data: n, k, θ, σ'
Result: σ: Full ranking
$V_j(\sigma) =$ bijection from σ' **for** $j \in [k+1, n]$ **do**
  $V_j(\sigma)$ = random choice in $[n-j]$ with choice probabilities of Eq. (3) σ = transform $V(\sigma)$ with the bijection in (McClellan et al., 1974) **return** σ^{-1}

Along the section, we have made use of the following result.

Lemma 1. *Let $\sigma \in S_n^k$ where $\sigma \sim M(\sigma_0, \theta)$. Then, σ^{-1} is a top-k ranking distributed according to the same distribution, $\sigma^{-1} \sim M(\sigma_0, \theta)$, and $d(\sigma) = d(\sigma^{-1})$.*

Proof. Let $\sigma \sim MM(e, \theta)$ and $\pi = \sigma^{-1}$. Note that for $\sigma \in S_n^k$ then π is a top-k ranking. Moreover, $d(\sigma) = d(\pi)$ and, since the MM is defined upon the distance function, it follows that $p(\sigma) = p(\pi)$ for every σ and therefore $\pi \sim MM(e, \theta)$. In the case that $\sigma_0 \neq e$, taking the right invariance property of the Kendall's-τ distance, it follows that $\pi\sigma_0 \sim MM(\sigma_0, \theta)$. Finally, if $\pi\sigma_0$ is a top-k ranking, then $(\pi\sigma_0)^{-1}$ is a top-k list, which concludes the proof. □

3. Proof of Equation 6 of the paper

Proof. Let p_{ij} be the marginal probability that item i is preferred to item j:

$$p_{ij} = \sum_{\sigma : \sigma(i) < \sigma(j)} p(\sigma). \tag{2}$$

The exact expression can be found in (Busa-Fekete et al., 2014) but for the proof, we just need to highlight that $p_{ij} = 1 - p_{ji}$. This pairwise comparison expression and the assumption that $\sigma_0 = e$ lets us rewrite the expected distance from the mode as follows

$$\mathbb{E}[d(\sigma, \sigma_0)] = \sum_{i<j} p_{ji}. \tag{3}$$

The expected pairwise distance can be written as follows

$$\mathbb{E}[d(\sigma, \sigma')] = 2 * \sum_{i<j} p_{ij} p_{ji} = 2 * \sum_{i<j} p_{ji} - p_{ji}^2. \tag{4}$$

With this restatement, the bound can be easily proved.

$$\sum_{i<j} p_{ji} - p_{ji}^2 \leq \sum_{i<j} p_{ji} \leq 2 * \sum_{i<j} p_{ji} - p_{ji}^2. \tag{5}$$

Note that this result holds for partial permutations as well. □

4. Expected distance $\mathbb{E}[D]$ and variance $\mathbb{V}[D]$ under the Mallows model

Lemma 2. *Let D be a random variable defines as $D = d(\sigma, \sigma_0)$ for a random Mallows ranking σ. The expectation and variance of D are as follows:*

$$\begin{aligned}
\mathbb{E}[D] &= \frac{k \cdot \exp(-\theta)}{1 - \exp(-\theta)} - \sum_{j=n-k+1}^{n} \frac{j \exp(-j\theta)}{1 - \exp(-j\theta)}, \\
\mathbb{V}[D] &= \frac{k \cdot \exp(-\theta)}{(1 - \exp(-\theta))^2} - \sum_{j=n-k+1}^{n} \frac{j^2 \exp(-\theta j)}{(1 - \exp(-\theta j))^2}.
\end{aligned} \tag{6}$$

Proof. The moment generating function $M(t) = \mathbb{E}\left[\exp(tD)\right]$ of the distance $D = d(\sigma_0, \sigma)$ of a random Mallows permutation σ can be factorized in this way (Fligner & Verducci, 1986):

$$M(t) = \prod_j M_j(t) = \prod_j \frac{1 - \exp(t(n - j + 1))}{(n - j + 1)(1 - \exp(t))}. \tag{7}$$

It's derivative, is as follows :

$$\frac{d \ln M_j(t)}{dt} = \frac{\exp(t)}{1 - \exp(t)} - \frac{(n - j + 1)\exp(t(n - j + 1))}{1 - \exp(t(n - j + 1))}. \tag{8}$$

For exponential models as the MM, expected values and variances can be easily written as function of the moment generating function.

$$\begin{aligned}
\mathbb{E}[V_j] &= \frac{d \ln M_j(t)}{dt}\Big|_{t=-\theta} \\
&= \frac{\exp(-\theta)}{1 - \exp(-\theta)} - \frac{(n - j + 1)\exp(-\theta(n - j + 1))}{1 - \exp(-\theta(n - j + 1))}
\end{aligned} \tag{9}$$

and

$$\begin{aligned}\mathbb{V}[V_j] &= \frac{d^2 \ln M_j(t)}{dt^2}\Big|_{t=-\theta} \\ &= \frac{\exp(-\theta)}{(1-\exp(-\theta))^2} - \frac{(n-j+1)^2 \exp(-\theta(n-j+1))}{(1-\exp(-\theta(n-j+1)))^2}.\end{aligned} \tag{10}$$

The proof concludes by noting that $\mathbb{E}[D] = \sum_{j=1}^{k} \mathbb{E}[V_j]$ and $\mathbb{V}[D] = \sum_{j=1}^{k} \mathbb{V}[V_j]$.

□

5. Proof of Lemma 2

Proof. We start by the right hand side of the equation: Let g_{ij} and b_{ij} be the marginal probabilities for good and bad raters respectively, as defined in Equation (2).

Let us assume that $\forall i < j$, $b_{ij} \geq g_{ij}$ and using Corollary 3 from (Busa-Fekete et al., 2014) we even have that $\forall i < j$, $1 \geq b_{ij} \geq g_{ij} > \frac{1}{2}$. Then $\forall i < j$, $\exists \epsilon_{ij} \in [0, \frac{1}{2}]$ such that $b_{ij} = g_{ij} + \epsilon_{ij}$.

We can rewrite the expected distances as functions of the marginal as follows:

- $\mathbb{E}[d(\beta, \gamma)] = \sum_{i<j} b_{ij} + g_{ij} - 2b_{ij} \cdot g_{ij}$
- $\mathbb{E}[d(\beta, \beta')] = \sum_{i<j} 2b_{ij} - 2b_{ij}^2$

Now we show the expression $\mathbb{E}[d(\beta, \gamma)] \leq \mathbb{E}[d(\beta, \beta')]$ holds, as the following inequalities are equivalent:

$$\begin{aligned}\mathbb{E}[d(\beta,\gamma)] &\leq \mathbb{E}[d(\beta,\beta')] \\ \sum_{i<j} b_{ij} + g_{ij} - 2b_{ij} \cdot g_{ij} &\leq \sum_{i<j} 2b_{ij} - 2b_{ij}^2 \\ \sum_{i<j} g_{ij} - 2b_{ij} \cdot g_{ij} &\leq \sum_{i<j} b_{ij} - 2b_{ij}^2 \\ \sum_{i<j} g_{ij} - 2 \cdot (2g_{ij}^2 + 2\epsilon_{ij} \cdot g_{ij}) &\leq \sum_{i<j} g_{ij} + \epsilon_{ij} - 2(g_{ij}^2 + 2g_{ij} \cdot \epsilon_{ij} + \epsilon_{ij}^2) \\ \sum_{i<j} -2g_{ij}^2 &\leq \sum_{i<j} \epsilon_{ij} - 2\epsilon_{ij}^2 \\ \sum_{i<j} \epsilon_{ij}(\epsilon_{ij} - \frac{1}{2}) &\leq \sum_{i<j} g_{ij}^2.\end{aligned} \tag{11}$$

Which conclude the proof of the right hand side, as the last inequality is always true since $\forall i < j$, $g_{ij}^2 \geq 0$ and $\epsilon_{ij}(\epsilon_{ij} - \frac{1}{2}) \leq 0$.

For the left hand side: Using once again Corollary 3 from (Busa-Fekete et al., 2014) which states that $\forall i < j$, $1 - b_{ij} < \frac{1}{2} < b_{ij}$, we have:

$$\begin{aligned}\mathbb{E}[d(\beta, \sigma_0)] &= \sum_{i<j} (1 - b_{ij}) = \sum_{i<j} g_{ij}(1 - b_{ij}) + (1 - g_{ij})(1 - b_{ij}) \\ &< \sum_{i<j} g_{ij}(1 - b_{ij}) + b_{ij}(1 - g_{ij}) = \mathbb{E}[d(\beta, \gamma)].\end{aligned}$$

□

6. Proof of Theorem 3

Proof. Note that

$$\begin{aligned}\mathbb{E}[\delta_\beta] &= (1-r)\cdot\mathbb{E}[d(\beta,\beta')] + r\cdot\mathbb{E}[d(\beta,\gamma)]\\ \mathbb{E}[\delta_\gamma] &= r\cdot\mathbb{E}[d(\gamma,\gamma')] + (1-r)\cdot\mathbb{E}[d(\beta,\gamma)]\end{aligned} \tag{12}$$

And our goal is to show there is a lower bound for the difference

$$\mathbb{E}[\delta_\beta - \delta_\gamma] = (1-r)\mathbb{E}[d(\beta,\beta')] + (2r-1)\mathbb{E}[d(\gamma,\beta)] - r\mathbb{E}[d(\gamma,\gamma')]. \tag{13}$$

This proof is divided in two parts. First, we show that the following expression holds:

$$c\cdot r\cdot\mathbb{E}[d(\gamma,\sigma_0)] \leq (1-r)\cdot\mathbb{E}[d(\beta,\beta')] + (2r-1)\cdot\mathbb{E}[d(\gamma,\beta)]. \tag{14}$$

In order to show the correctness of the above expression, we will deal with cases where $r < 0.5$ and $r \geq 0.5$ separately.

Case $r < 0.5$:

Starting from the right hand side of Lemma 3 and multiplying by $(2r-1)$ (negative in this case) and adding $(1-r)\mathbb{E}[d(\beta,\beta')]$ on both sides it holds that:

$$\begin{aligned}(1-r)\mathbb{E}[d(\beta,\beta')] + (2r-1)\mathbb{E}[d(\gamma,\beta)] &\geq (1-r)\mathbb{E}[d(\beta,\beta')] + (2r-1)\mathbb{E}[d(\beta,\beta')]\\ =r\cdot\mathbb{E}[d(\beta,\beta')] &\geq r\cdot\mathbb{E}[d(\beta,\sigma_0)] \geq c\cdot r\cdot\mathbb{E}[d(\gamma,\sigma_0)],\end{aligned} \tag{15}$$

where the last two inequalities are obtained from the right hand side of Equation (6) of the original paper and the assumption that $\mathbb{E}[d(\beta,\sigma_0)] \geq c\cdot\mathbb{E}[d(\gamma,\sigma_0)]$ respectively.

Case $r \geq 0.5$:

Starting from the result of the left hand side of Lemma 3 and multiplying by $(2r-1)$ (positive in this case) and adding $(1-r)\mathbb{E}[d(\beta,\beta')]$ on both sides it holds that:

$$\begin{aligned}(1-r)\mathbb{E}[d(\beta,\beta')] + (2r-1)\mathbb{E}[d(\gamma,\beta)] &\geq (1-r)\mathbb{E}[d(\beta,\beta')] + (2r-1)\mathbb{E}[d(\beta,\sigma_0)]\\ \geq(1-r)\mathbb{E}[d(\beta,\sigma_0)] + (2r-1)\mathbb{E}[d(\beta,\sigma_0)] &= r\cdot\mathbb{E}[d(\beta,\sigma_0)]\\ \geq c\cdot r\cdot\mathbb{E}[d(\gamma,\sigma_0)],&\end{aligned} \tag{16}$$

where the last two inequalities are obtained from the right hand side of Equation (6) of the original paper and the assumption that $\mathbb{E}[d(\beta,\sigma_0)] \geq c\cdot\mathbb{E}[d(\gamma,\sigma_0)]$ respectively.

This finishes the first part, in which we show that Equation (14) holds for any value of r. In the second part we will add it to the following result, obtained using the left hand side of Equation (6) of the paper, by multiplying it by $-2r$:

$$-2r\mathbb{E}[d(\gamma,\sigma_0)] \leq -r\mathbb{E}[d(\gamma,\gamma')]. \tag{17}$$

Hence, we finally have, using Equation (13):

$$(c-2)\cdot r\cdot\mathbb{E}[d(\gamma,\sigma_0)] \leq (1-r)\mathbb{E}[d(\beta,\beta')] + (2r-1)\mathbb{E}[d(\gamma,\beta)] - r\mathbb{E}[d(\gamma,\gamma')] = \mathbb{E}[\delta_\beta - \delta_\gamma]. \tag{18}$$

□

7. Preliminaries for sample complexity proofs

Lemma 3. *Let $p(\sigma)$ be the probability of ranking $\sigma \in S_n$ under the Mallows model. For every $1 \leq i, k \leq n$ the following holds.*

$$\sum_{\sigma:\sigma(i)\leq k} p(\sigma) = 1 - \sum_{\sigma:\sigma(n-i)\leq n-k} p(\sigma). \tag{19}$$

Proof. We are going to do a constructive bijection between the permutations in the set $\{\sigma : \sigma(i) \leq k\}$ and those in the set $\{\sigma : \sigma(n-i) \geq n-k\}$ and show that the probabilities are the same. Let us first define these sets.

- $S = \{\sigma : \sigma(i) \leq k\}$
- $S' = \{\sigma : \sigma(i) > n-k\}$
- $S'' = \{\sigma : \sigma(n-i) > n-k\}$

Flip step. First, note that there is bijection between the sets S and S'. Given a permutation $\sigma \in S$ we can construct a permutation $\sigma' \in S'$ by setting $\sigma'(i) = n - \sigma(i) + 1$. Moreover, $d(\sigma') = \binom{n}{2} - d(\sigma)$.

Reverse step. Now, we show a bijection between S' and S''. Given a permutation $\sigma' \in S'$ we can construct a permutation $\sigma'' \in S''$ by setting $\sigma''(i) = \sigma'(n-i)$. Moreover, $d(\sigma'') = \binom{n}{2} - d(\sigma') = d(\sigma)$. This implies that $\sum_{\sigma\in S} p(\sigma) = \sum_{\sigma''\in S''} p(\sigma'')$.

Comlementary step. Since $S'' \cup \{\sigma : \sigma(n-i) \leq n-k\} = S_n$ and both sets are disjoint, to conclude the proof, we just have to note that

$$\sum_{\sigma''\in S''} p(\sigma'') = 1 - \sum_{\sigma:\sigma(n-i)\leq n-k} p(\sigma), \tag{20}$$

which, in turn, implies the results in Equation (19).

□

Lemma 4. *Let Δ^{ik} be defined as is Definition 6 where $p(\sigma)$ is the probability of ranking $\sigma \in S_n$ under the Mallows model. For every $1 \leq i, k \leq n$ the following holds.*

$$\Delta^{ik} = \Delta^{n-i-1,n-k} \tag{21}$$

Proof. Based on the result on Lemma 3, we can rewrite Δ in this way.

$$\begin{aligned}\Delta^{ik} &= \sum_{\sigma:\sigma(i)\leq k} p(\sigma) - \sum_{\sigma:\sigma(i+1)\leq k} p(\sigma) = \Big(1 - \sum_{\sigma:\sigma(n-i)\leq n-k} p(\sigma)\Big) - \Big(1 - \sum_{\sigma:\sigma(n-i-1)\leq n-k} p(\sigma)\Big) \\ &= \sum_{\sigma:\sigma(n-i-1)<n-k} p(\sigma) - \sum_{\sigma:\sigma(n-i)<n-k} p(\sigma) = \Delta^{n-i-1,n-k}.\end{aligned} \tag{22}$$

□

8. Proof of Lemma 6

Proof. The symmetry described in Lemma 4 implies a symmetry in $\arg_i \min \Delta^{ik}$ that allows us focusing on the case $k \geq n/2$. For $k \geq n/2$ $\arg_i \min \Delta^{ik} = 1$. As a summary,

$$\arg\min_i \Delta^{ik} = \begin{cases} n-1 & \text{if } k < n/2, \\ 1 & \text{otherwise} \end{cases} \tag{23}$$

which, in turn implies

$$\min_i \Delta^{ik} = \begin{cases} \Delta^{1,n-k} & \text{if } k < n/2, \\ \Delta^{1,k} & \text{otherwise.} \end{cases} \tag{24}$$

Despite there is not a closed form expression for Δ^{nk}, Δ^{1k} can be computed in $O(k^2)$.

$$\begin{aligned}\Delta^{1k} &= \sum_{\sigma:\sigma(1)\leq k} p(\sigma) - \sum_{\sigma:\sigma(2)\leq k} p(\sigma) \\ &= \sum_{r_1=0}^{k-1} p(V_1 = r_1) - \Big(\sum_{r_1=0}^{k-1}\sum_{r_2=0}^{k-2} p(V_1 = r_1)p(V_2 = r_2) + \sum_{r_1=k}^{n}\sum_{r_2=0}^{k-1} p(V_1 = r_1)p(V_2 = r_2)\Big) \\ &= \sum_{r_1=0}^{k-1} \frac{\exp(-\theta r_1)}{\psi_{n,1}} - \sum_{r_1=0}^{k-1}\sum_{r_2=0}^{k-2} \frac{\exp(-\theta r_1)\exp(-\theta r_1)}{\psi_{n,1}\psi_{n,2}} - \sum_{r_1=k}^{n}\sum_{r_2=0}^{k-1} \frac{\exp(-\theta r_1)\exp(-\theta r_2)}{\psi_{n,1}\psi_{n,2}},\end{aligned} \tag{25}$$

which concludes the proof. □

9. Proof of Theorem 7

Proof. Borda outputs the correct order for the pair of items i and $i+1$ with probability $1-\epsilon$ when the number of permutations is at least

$$m \geq \frac{2n^2 \log \epsilon^{-1}}{(\sum_{j=1}^{n-1} \Delta^{ij})^2}. \tag{26}$$

This expression has been used in sample complexity results that do not consider the spread parameter (Caragiannis et al., 2013). The authors use this expression for Δ^{ij} which can be shown to be equivalent. [1]

$$\sum_{j=1}^{n-1}(\sum_{l=1}^{j}\sum_{\sigma:\sigma(i)=l} p(\sigma) - \sum_{l=1}^{j}\sum_{\sigma:\sigma(i+1)=l} p(\sigma)) = \sum_{j=1}^{n-1}(\sum_{\sigma:\sigma(i)\leq j} p(\sigma) - \sum_{\sigma:\sigma(i+1)\leq j} p(\sigma)) = \sum_{j=1}^{n-1}\Delta^{ij}. \tag{27}$$

In these lines we extend the result by bounding the number of samples (1) w.r.t. the dispersion parameter and (2) considering top-k rankings. To prove these points, we give an upper bound for $\sum_{j=1}^{n-1}\Delta^{ij}$, which is a function of the dispersion parameter θ.

Assume w.l.o.g. that $\sigma_0 = e$, let τ_i be an inversion of positions i and $i+1$, i.e., $\tau_i(i) = i+1$, $\tau_i(i+1) = i$ and $\tau_i(j) = j$ for $j \neq i, i+1$. As for any inversion, the result of the composition $\sigma\tau_i$ swaps positions i and $i+1$ in σ. Therefore,

$$\begin{aligned}\sum_{j=1}^{n-1}\Delta^{ij} &= \sum_{\{\sigma:\sigma(i)<\sigma(i+1)\}} (p(\sigma) - p(\sigma\tau_i)(\sigma(i+1) - \sigma(i)) \leq \sum_{\sigma:\sigma(i)<\sigma(i+1)} (p(\sigma) - p(\sigma\tau_i))k \\ &= k\Big(\sum_{\substack{\sigma:\sigma(i)\leq k \wedge \\ \sigma(i)<\sigma(i+1)}} p(\sigma) - p(\sigma\tau_i) + \sum_{\substack{\sigma:\sigma(i+1)>k \wedge \\ \sigma(i)<\sigma(i+1)}} p(\sigma) - p(\sigma\tau_i)\Big)\end{aligned} \tag{28}$$

Note that for partial permutations of k items, the second sum equals 0. Therefore, the following is equivalent

$$\begin{aligned}&= k \sum_{\substack{\sigma:\sigma(i)\leq k \wedge \\ \sigma(i)<\sigma(i+1)}} p(\sigma) - p(\sigma\tau_i) = k \sum_{\substack{\sigma:\sigma(i)\leq k \wedge \\ \sigma(i)<\sigma(i+1)}} p(\sigma)(1 - \exp(-\theta)) \\ &\leq k(1-\exp(-\theta)) \sum_{\sigma:\sigma(i)\leq k} p(\sigma).\end{aligned} \tag{29}$$

Since $\sum_{\sigma:\sigma(i)\leq k} p(\sigma)$ decreases w.r.t. i the following is equivalent.

[1] Note that the definitions of m and n are interchanged in their paper and that they denote the dispersion in the model as $\phi = \exp(-\theta)$.

$$
\begin{aligned}
&k(1-\exp(-\theta)) \sum_{\sigma:\sigma(i)\leq k} p(\sigma) \leq k(1-\exp(-\theta)) \sum_{\sigma:\sigma(1)\leq k} p(\sigma) - i \min \Delta^{ik} \\
&\leq k(1-\exp(-\theta)) \sum_{r\leq k} p(V_1 = r) - i\Delta^{1k} \leq k^2(1-\exp(-\theta))p(V_1 = 0) - i\Delta^{1k} \\
&= \frac{k^2(1-\exp(-\theta))^2}{1-\exp(-\theta n)} - i\Delta^{1k}.
\end{aligned} \tag{30}
$$

It follows that

$$
\sum_{j=1}^{n-1} \Delta^{ij} \leq \frac{k^2(1-\exp(-\theta))^2}{1-\exp(-\theta n)} - i\Delta^{1k}, \tag{31}
$$

and therefore, Borda outputs the true ranking σ_0 with probability $1-\epsilon$ when the number of samples is at least

$$
m \geq 2n^2 \log \epsilon^{-1} \Big(\frac{k^2(1-\exp(-\theta))^2}{1-\exp(-\theta n)} - i\Delta^{1k} \Big)^{-2}, \tag{32}
$$

which concludes the proof.

□

References

Busa-Fekete, R., Hüllermeier, E., and Szörényi, B. Preference-Based Rank Elicitation using Statistical Models: The Case of Mallows. In *Proceedings of the 31th International Conference on Machine Learning, (ICML)*, pp. 1071–1079, 2014. URL http://jmlr.org/proceedings/papers/v32/busa-fekete14.html.

Caragiannis, I., Procaccia, A. D., and Shah, N. When Do Noisy Votes Reveal the Truth? In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, pp. 143–160, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1962-1. doi: 10.1145/2482540.2482570. URL http://doi.acm.org/10.1145/2482540.2482570.

Fligner, M. A. and Verducci, J. S. Distance based ranking models. *Journal of the Royal Statistical Society*, 48(3):359–369, 1986.

McClellan, M. T., Minker, J., and Knuth, D. E. The Art of Computer Programming, Vol. 3: Sorting and Searching. *Mathematics of Computation*, 1974. ISSN 00255718. doi: 10.2307/2005383.