Identity testing for Mallows model

Róbert Busa-Fekete
Google Research, New York, USA
busarobi@google.com

Dimitris Fotakis
National Technical University of Athens, Greece
fotakis@cs.ntua.gr

Balázs Szörényi
Yahoo! Research New York, USA
balazs.szorenyi@yahooinc.com

Manolis Zampetakis
Department of Statistics
University of California, Berkeley, USA
mzampet@berkeley.edu

Abstract

In this paper, we devise identity tests for ranking data that is generated from Mallows model both in the *asymptotic* and *non-asymptotic* settings. First we consider the case when the central ranking is known, and devise two algorithms for testing the spread parameter of the Mallows model. The first one is obtained by constructing a Uniformly Most Powerful Unbiased (UMPU) test in the asymptotic setting and then converting it into a sample-optimal non-asymptotic identity test. The resulting test is, however, impractical even for medium sized data, because it requires computing the distribution of the sufficient statistic. The second non-asymptotic test is derived from an optimal learning algorithm for the Mallows model. This test is both easy to compute and is sample-optimal for a wide range of parameters. Next, we consider testing Mallows models for the unknown central ranking case. This case can be tackled in the asymptotic setting by introducing a bias that exponentially decays with the sample size. We support all our findings with extensive numerical experiments and show that the proposed tests scale gracefully with the number of items to be ranked.

1 Introduction

Identity testing of discrete distributions [9, 11] is one of the most fundamental problem which consists of answering a yes-or-no question about the closeness of some explicitly given distribution to an unknown distribution from which random samples are observed. This testing problem have been studied in the classical statistical literature [16], and several tests have been devised in the asymptotic regime which family hypothesis tests are often referred to as goodness-of-fit tests and are routinely used in data analysis [10]. In this work, we extend the identity testing setup to a more general domain that includes rankings over m items which can be viewed as a discrete distribution testing problem over a domain with size $m!$. As a consequence, identity testing of ranking data without any assumption is not feasible, since lower bound is known for identity testing for discrete distribution [26, 11] that is $\Omega(\sqrt{n})$ where n is the domain size. Therefore we study an important subclass of ranking distributions, which is introduced by [20] and also known as the exponential family on rankings. The model has two parameters, the *central ranking* $\pi_0 \in S_m$ and the *spread parameter* $\phi \in [0, 1]$. Based on these, the probability of observing a ranking $\pi \in S_m$ is proportional to $\phi^{d(\pi,\pi_0)}$, where d is a ranking distance, such as the number of discordant pairs, a.k.a Kendall's tau distance. There are many applications of the Mallows model in Machine Learning, to name a few, label ranking [18], online learning [3], recommendation systems [29, 15], clustering [23]. The focus of this study is to devise identity tests for Mallows model that are scalable with the number of items m. Identity testing is a central problem in analysing output of ranking systems where the goal is to

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

decide whether the output ranking data deviates from some expected behaviour, or is biased towards some group of object to be ranked or it is indeed fair ([24, 20, 32] and see Chapter 3-4 of [22]).

The hypothesis testing literature that is related to Mallows model, and more generally to ranking distributions, such as Plackett-Luce [27, 19] or Babington-Smith [22], is not extensive. In our work, we partially fill this gap by devising hypothesis tests for the Mallows model. We focus on identity testing problem, where we assume that the ground truth parameters of a Mallows model are given; this corresponds to the null hypothesis. Our goal is to decide whether the data we observe is generated by the null model or by an alternative one. We show that this kind of goodness-of-fit testing task can be tackled in the traditional asymptotic setting as well as in non-asymptotic scenario in an efficient way regardless whether the central ranking is known. We can summarize our main results as follows:

- We devise a Uniformly Most Powerful Unbiased (UMPU) test for asymptotic case to test spread parameter ϕ with known central ranking π_0.
- We show that, in general, the UMPU test can be converted into a sample-optimal non-asymptotic test which result may be of independent interest. Based on this result, we come up with an optimal non-asymptotic test, however it is hard to apply even for medium sized data.
- Next, we propose a non-asymptotic test that is optimal for small ϕ and easy to compute.
- We also consider the case when the central ranking π_0 is not given, but needs to be estimated, and devise tests for both testing setups.
- We demonstrate the versatility of our algorithm running with large m on synthetic data, and we show that for large $m > 70$ to test the spread parameter, only one single sample is enough.

The paper is organized as follows. Related work is presented in Section 2. Then, we recall the Mallows model and introduce notations. In Section 4, we describe the asymptotic and non-asymptotic testing setups and present our result on the relation of these two settings. In Section 5, we present our algorithms for testing the spread parameter ϕ. In Section 6, we present tests where both parameter of Mallows model are tested. Experiments are in Section 7 and finally conclude the paper in Section 8.

2 Related work

The asymptotic testing has a long history [16] dating back to Pearson's fundamental work. To come up with a asymptotic test for Mallows model is more challenging from a computational point of view than methodological point of view. Testing ranking distribution can be viewed as multinomial testing problem and thus applied, for example, [8], however the domain size is $m!$, so this approach becomes hard to apply even for small m.

The non-asymptotic testing has also long history [9, 6, 5] and references therein, including testing for discrete data. Nevertheless, sample optimal tests for wide range of parameter had been only devised recently [11] which was further strengthened in [31] by showing that this lower bound is instance optimal. Those results imply a sample complexity of order $\Theta(\sqrt{m!}/\varepsilon^2)$ in our ranking setting if we do not exploit the structure induced by Mallows model. In the non-asymptotic setting, testing and learning are related problems. There are several so-called "testing by learning" algorithm which first learns a model, and then decides whether the learnt model is close to the null model or far enough from it which requires a tight lower bound on the total variation distance. Optimal learning, however, does not necessarily imply optimal testing in general, as it is the case for identity testing of discrete distributions (see Subsection 6.6 of [5] for further discussion). Nevertheless, it turned out that for identity testing of Mallows model this is indeed to be the case for a wide range of parameters, since interestingly, upper bound for sample complexity of optimal parameter learning of Mallows model requires asymptotically as many samples as it is needed to show that two models are far from each other in terms of total variation distance in a certain parameter regime. Optimal learning is recently devised for Mallows model by [4] which result we rely on.

The testing literature is very limited for ranking distribution in general. Mallows came up with an approximate solution, in [20, Section 11], to test uniformity against a single parameter Mallows model in the non-asymptotic setting. His approach relies on a normal approximation of the sufficient statistic. Cohen and Mallows [7] presents a goodness of fit test for ranking data in a way the ranking data is handled as a sample from a discrete distribution with $m!$ parameters. That is why their analysis is restricted to $m = 3$ and $m = 4$. In addition to this, the authors presented a goodness of fit test for pairwise marginals of various ranking models for larger m including Mallows model, Placket-Luce

and the Thurstonian model, but their focus was to decide which model fits better to the data in terms of pairwise marginals. Whereas we were interested in testing the parameters of Mallows model with large m values. There are some recent paper for testing ranking data, but these works are not related to parametric ranking models. For example, [21, 17] uses statistics based on permutation kernel functions for testing the equality of two ranking distributions in the asymptotic setting. Finally, there is a very recent paper which testing identity of preference matrices which are marginals of ranking distributions [28].

3 Preliminaries and Notation

Single Parameter Mallows Model. The Mallows model or, more specifically, Mallows ϕ-distribution is a parametrized, distance-based probability distribution that belongs to the family of exponential distributions $\mathcal{R} = \{\mathcal{M}_{\phi,\pi} \mid \phi \in [0,1], \pi \in S_m\}$ with probability mass function $p_{\phi,\pi_0}(\pi) = \phi^{d(\pi,\pi_0)}/Z(\phi,\pi_0)$ where ϕ and π_0 are the parameters of the model: $\pi_0 \in S_m$ is the location parameter also called center ranking and $\phi \in [0,1]$ the spread parameter. Moreover, $d(\cdot,\cdot)$ is a distance metric on permutations, which for our paper will be the Kendall tau distance, that is, the number of discordant item pairs $d_K(\pi,\pi') = \sum_{1\le i<j\le m} \mathbb{I}\{(\pi(i)-\pi(j))(\pi'(i)-\pi'(j)) < 0\}$.

The normalization factor in the definition of the model is equal to $Z(\phi,\pi_0) = \sum_{\pi\in S_m} p_{\phi,\pi_0}(\pi)$. When the distance metric d is the Kendall tau distance we have the identity $Z(\phi,\pi_0) = Z(\phi) = \prod_{i=1}^{m-1}\sum_{j=0}^{i}\phi^j$. Observe that the family of distributions as stated is not an exponential family because of the location parameter π_0. If we fix the permutation parameter then the family $\mathcal{R}(\pi_0) = \{\mathcal{M}_{\phi,\pi_0} \mid \phi \in [0,1]\}$ is an exponential family with natural parameter $\theta = \ln\phi$. From now on, if we use the neutral parametrization, we shall write θ, i.e. $Z(\theta) = \prod_{i=1}^{m-1}\sum_{j=0}^{i} e^{j\theta}$. The log partition function is denoted by $\alpha(\theta) = \log Z(\theta)$. In Appendix A, we review the basic properties of exponential families which we use in our work.

Metrics between distributions. Let p, q be two probability measures in the discrete probability space $(\Omega,\mathcal{A})$ then the total variation distance between p and q is defined as $\mathrm{d_{TV}}(p,q) = \frac{1}{2}\sum_{x\in\Omega}|p(x)-q(x)| = \max_{A\in\mathcal{A}}|p(A)-q(A)|$, and the KL-divergence between p and q is defined as $\mathrm{D_{KL}}(p||q) = \sum_{x\in\Omega} p(x)\ln\left(\frac{p(x)}{q(x)}\right)$.

4 Testing ranking distributions

We shall consider two types of identity tests: *asymptotic* and *non-asymptotic* tests. In the asymptotic case, a set of observations and a significance level $\alpha > 0$ are given in advance, and then our goal is to come up with a test that maximizes the power, i.e. the probability of rejection, if the alternative hypothesis is true, subject to the given level of *significance* α, i.e. the probability of rejection must be below or equal to α if the null hypothesis is true [16]. More concretely, assume that we are given a parametric family of ranking distribution $\mathcal{R} = \{\mathcal{M}_\theta | \theta \in \Omega\}$ where Ω denotes the set of parameters. The observation consists of n rankings $\mathcal{D}_n = \{\pi_1,\dots,\pi_n\}$ from a ranking distribution $\mathcal{M}$. The null hypothesis is $H_0 : \mathcal{M} \in \mathcal{R}_0$ where $\mathcal{R}_0 \subset \mathcal{R}$. As an alternative hypothesis, we consider $H_1 : \mathcal{M} \in \mathcal{R}_1 (\subset \mathcal{R})$ such that $\mathcal{R}_0 \cap \mathcal{R}_1 = \emptyset$. Then the test is a function $f : S_M^n \mapsto \{0,1\}$ where 0 corresponds to the acceptance, and 1 to the rejection, such that probability of rejection $\mathbb{E}[f(\mathcal{D}_n)] \le \alpha$ whenever $\mathcal{M} \in \mathcal{R}_0$ where the expectation is with respect to $\mathcal{M}$. Our goal is to find a test f for which the *power* $\beta_f(\mathcal{M}) = \mathbb{E}[f(\mathcal{D}_n)]$ for all $\mathcal{M} \in \mathcal{R}_1$ is as large as possible. We will deal with randomized test so we consider tests in the form of $f : S_M^n \mapsto [0,1]$ where the output means the probability of rejection.

A test f is called *most powerful* for a given α and $\mathcal{M}' \in \mathcal{R}_1$, if there is no test f' such that $\beta_f(\mathcal{M}') < \beta_{f'}(\mathcal{M}')$. A test is *uniformly most powerful* (UMP) if it is most powerful for any distribution from the alternative hypothesis class. Furthermore, a test is *unbiased*, if $\sup_{\mathcal{M}\in\mathcal{R}_0}\beta_f(\mathcal{M}) \le \alpha \le \inf_{\mathcal{M}\in\mathcal{R}_1}\beta_f(\mathcal{M})$. The unbiased UMP test is referred to as UMPU test.

In case of non-asymptotic setup, the input is a tolerance parameter $\varepsilon > 0$ and significance parameter $\delta \in (0,1)$, and we assume that the tester has sample access to the unknown distribution $\mathcal{M}$. An (ε,δ) non-asymptotic testing algorithm outputs a sample size n and a test function $f : S_M^n \mapsto \{0,1\}$ such that, generating $\mathcal{D}_n$ from $\mathcal{M}$, we have the following guaranties for f:

1. if the null hypothesis H_0 is true, then it outputs reject ($f(\mathcal{D}_n) = 1$) with probability at most δ, i.e. $\mathbb{E}[f(\mathcal{D}_n)] \leq \delta$
2. if $\mathcal{M} \in \mathcal{R}_1$ such that $d_{\text{TV}}(\mathcal{M}, \mathcal{R}_0) > \varepsilon$, then it outputs reject ($f(\mathcal{D}_n) = 1$) with probability at least $1 - \delta$ where $d_{\text{TV}}(\mathcal{M}, \mathcal{R}_i) = \inf_{\mathcal{M}' \in \mathcal{R}_i} d_{\text{TV}}(\mathcal{M}, \mathcal{M}')$

Given ε and δ, we say that a testing algorithm is sample-optimal for $\mathcal{R}_0$ versus $\mathcal{R}_1$, if no algorithm can have the same confidence guarantee using less samples. If a testing algorithm sample-optimal for any ε and δ, we say that it is uniformly sample-optimal (USO).

One can show that the existence of UMPU test implies the existence of sample-optimal non-asymptotic test as follows with a particular (ε, δ).

Proposition 4.1 *If there exists a UMPU test f_n with a given significance level α for $\mathcal{R}_0$ versus $\mathcal{R}_1$ for sample size n, $n = 1, 2, \ldots$, then one can define a sample-optimal non-asymptotic testing algorithm that outputs n and f on input $\delta = \alpha$ and*

$$\varepsilon_f = \sup_{\mathcal{M} \in \mathcal{R}_1 \setminus \mathcal{R}(f)} d_{TV}(\mathcal{M}, \mathcal{R}_0) \tag{1}$$

where $\mathcal{R}(f) = \{\mathcal{M} \in \mathcal{R}_1 : \beta_f(\mathcal{M}) \geq 1 - \delta\}$.

The proof of Proposition 4.1 is deferred to Appendix B. The opposite direction does not hold, since a non-asymptotic test does not guarantee anything for any $\mathcal{M} \in \mathcal{R}_1$ such that $d_{\text{TV}}(\mathcal{M}, \mathcal{R}_0) \leq \varepsilon$.

5 Testing spread parameter with known central ranking

In this section, we will focus on testing the spread parameter when the central ranking is known.

5.1 Asymptotic case: UMPU test for spread parameter

We will characterize an UMPU test for testing the spread parameter of Mallows model in the asymptotic setting. The UMPU test simply rejects the null hypothesis when the Kendall distance of the data deviates from its expected value under the null hypothesis by a certain margin. This is formalized by Theorem 5.1 whose proof is deferred to Appendix C.

Theorem 5.1 *Let us define a test $f_n(\mathcal{D}_n) = g_n(T_{\pi_0}(\mathcal{D}_n))$ with $T_\pi(\mathcal{D}) = \sum_{\pi' \in \mathcal{D}} d_K(\pi, \pi')$ and $0 \leq t_1 \leq t_2$ where $g_n(\ell) = \mathbb{1}\{\ell \notin [t_1, t_2]\} + c_1 \cdot \mathbb{1}\{\ell = t_1\} + c_2 \cdot \mathbb{1}\{\ell = t_2\}$ with $c_1, c_2 \in [0, 1]$ such that*

$$1 - \frac{\alpha}{2} = \sum_{\ell \in [t_1, \mu]} (1 - g_n(\ell)) \mathbb{P}_{\mathcal{M}_{\phi_0, \pi_0}}(T_{\pi_0}(\mathcal{D}_n) = \ell) = \sum_{\ell \in [\mu, t_2]} (1 - g_n(\ell)) \mathbb{P}_{\mathcal{M}_{\phi_0, \pi_0}}(T_{\pi_0}(\mathcal{D}_n) = \ell).$$

where $\mu = \mathbb{E}_{\mathcal{M}_{\phi_0, \pi_0}}[T_{\pi_0}(\mathcal{D}_n)]$. Then f_n is an UMPU level α test with sample size n for $H_0 : \mathcal{M} \in \mathcal{R}_0 = \{\mathcal{M}_{\phi_0, \pi_0}\}$ versus $H_1 : \mathcal{M} \in \mathcal{R}_1 = \{\mathcal{M}_{\phi, \pi} : \phi \neq \phi_0, \pi = \pi_0\}$.

Even if the UMPU test f_n defined in Theorem 5.1 plays central role in the asymptotic regime, it is hard to compute exactly, as this is discussed next in Remark 1. However it can be approximated by using χ^2 distribution which is presented in Appendix D.

Remark 1 *(Computational complexity) To determine the critical region of the test f_n, one needs to compute the distribution of the sufficient statistic $T_{\pi_0}(\mathcal{D}_n)$ based on n samples under null hypothesis that is defined as*

$$\mathbb{P}_{\mathcal{M}_{\phi, \pi}}(T_\pi(\mathcal{D}_n) = \ell) = \frac{\phi^\ell}{Z(\phi)^n} N(\ell, n) \tag{2}$$

where $N_k = N_k(m) = \#\{\pi \in S_m : d_K(\pi, \pi_{id}) = k\}$ is the Mahonian number [1]. One can compute the Mahonian number in $O(m^2)$ time using the recursion formula. Similarly, $N(\ell, n)$ can be computed using recursion in $O(m^2 n)$ time. From practical point of view, it is more challenging to compute $\mathbb{P}_{\mathcal{M}_{\phi_0, \pi}}(T(\mathcal{D}_n) = \ell)$ for some ℓ that depends on the fraction of large integers. In this regime, the numerical error can get significant for large n or m.

[1] https://oeis.org/A008302

5.2 An optimal non-asymptotic test based on UMPU test

To tackle the non-asymptotic case to test $H_0 : \mathcal{M} \in \mathcal{R}_0 = \{\mathcal{M}_{\phi_0,\pi_0}\}$ versus $H_1 : \mathcal{M} \in \mathcal{R}_1 = \{\mathcal{M}_{\phi,\pi_0} : \mathrm{d_{TV}}\left(\mathcal{M}_{\phi,\pi_0}, \mathcal{M}_{\phi_0,\pi_0}\right) > \varepsilon\}$, first we convert the UMPU test f_n into (ε, δ) non-asymptotic test based on Proposition 4.1. For doing this, we pick the smallest n so as the $\mathrm{d_{TV}}(\mathcal{M}_{\phi_0,\pi_0}, \mathcal{M}_{\phi,\pi}) \geq \varepsilon$ for where $\beta_{f_n}(\mathcal{M}_{\phi,\pi}) \geq 1 - \delta$. This algorithm is defined in Algorithm 1.

Algorithm 1 Non-asymptotic test for spread parameter based on UMPU

1: **Input:** ε, δ
2: $n^* = \min\{n \in \mathbb{Z}_+ : s(n) \geq \varepsilon\}$ where $s(n) = \sup_{\mathcal{M} \in \mathcal{R}_1 \setminus \mathcal{R}(f_n)} \mathrm{d_{TV}}\left(\mathcal{M}, \mathcal{M}_{\phi_0,\pi_0}\right)$
3: Take n^* samples $\mathcal{D}_{n^*}$ and ouput $f_{n^*}(\mathcal{D}_{n^*})$

Based on the concentration of the sufficient statistic of exponential family one can compute an upper bound on n^*. The proof of Theorem 5.2 is deferred to Appendix E.

Theorem 5.2 *The sample size n^* required by Algorithm 1 is at most $\frac{1}{2\varepsilon^2} \log \frac{2}{\delta}$.*

According to Proposition 4.1, Algorithm 1 is a uniformly sample optimal non-asymptotic test, since it was derived from an UMPU test. However, the upper bound for the sample complexity given in Theorem 5.2 might not be tight. To show that it is indeed tight, we apply the result of [4] regarding optimal learning of Mallows model in terms of total variation.

Corollary 5.3 *Any non-asymptotic testing algorithm which distinguishes $H_0 : \mathcal{M} \in \mathcal{R}_0 = \{\mathcal{M}_{\phi_0,\pi_0}\}$ from $H_1 : \mathcal{M} \in \mathcal{R}_1 = \{\mathcal{M}_{\phi,\pi_0} : \mathrm{d_{TV}}\left(\mathcal{M}_{\phi,\pi_0}, \mathcal{M}_{\phi_0,\pi_0}\right) > \varepsilon\}$ requires $\Omega(1/\varepsilon^2)$ samples.*

The proof of Corollary 5.3 is presented in Appendix F. The upper bound given in Theorem 5.2, on the one hand is tight up to a logarithmic factor, and on the other hand it does not depend on m in an explicit way, but $\mathcal{R}(f_n)$ indeed does which is used in Line 2 of Algorithm 1 for computing $s(n)$.

Algorithm 2 Non-asymptotic test for spread parameter

1: **Input:** ε, δ
2: $\varepsilon_L = \min\{x \in [0, \phi_0] : h(x) \geq \varepsilon\}$ where $h(x) = \mathrm{d_{TV}}\left(\mathcal{M}_{\phi_0,\pi_0}, \mathcal{M}_{\phi_0 - x,\pi_0}\right)$
3: $\varepsilon_R = \min\{x \in [0, 1 - \phi_0) : h(-x) \geq \varepsilon\}$
4: $\varepsilon_0 = \min\{\varepsilon_L/2, \varepsilon_R/2\}$
5: Take $n \in \Omega\left(\frac{\log(1/\delta)}{m\varepsilon_0^2}\right)$ samples $\mathcal{D}_n$ and solve $\nabla \ln Z(\phi) - \frac{1}{n} T_{\pi_0}(\mathcal{D}_n) = 0$ to obtain $\hat{\phi}$
6: **If** $|\hat{\phi} - \phi_0| \leq \varepsilon_0$ **Then** Output 0 **Else** Output 1 ▷ Based on Theorem 5.4

5.3 Scalable non-asymptotic test based on optimal learning

Algorithm 1 has nice optimality properties, however to solve the optimization in Line 2, the set $\mathcal{R}(f_n)$ has to be computed, that requires to compute the distribution of the sufficient statistic for n samples which is not scalable according to Remark 1. Motivated by this fact, next we devise a more practical algorithm with weaker optimality guaranty. We recall the result of [4] for optimal parameter learning of Mallows models. For a known central ranking π_0, one can compute an estimate for $\theta = \ln \phi$ by finding the root of $\nabla \ln Z(\theta) - \frac{1}{n} T_{\pi_0}(\mathcal{D}) = 0$. Since $\ln Z(\theta)$ is monotone increasing in θ based on Theorem A.1 2), we can find an estimate $\hat{\phi}$ such that $|\hat{\phi} - \phi| < \gamma$ in $O(\log(1/\gamma))$ time based on binary search, therefore we shall neglect the numerical error. If we are given enough sample, the parameter estimate $\hat{\phi}$ determines a Mallows model that is close to the true one. We recall this result from [4] (see Theorem 7 therein for more details).

Theorem 5.4 *For any $\phi < 1$, if $n \geq \Omega\left(\log(1/\delta)/(m\varepsilon_0^2)\right)$, then $\mathbb{P}\left(|\hat{\phi} - \phi| \leq \varepsilon_0\right) \geq 1 - \delta$.*

Theorem 5.4 implies that with $n \geq \Omega\left(\log(1/\delta)/(m\varepsilon_0^2)\right)$ samples, we can estimate the true ϕ with additive error that is of order ε_0. This observation suggests a testing algorithm which is presented

in Algorithm 2. The proposed testing algorithm first computes the neighborhood ϕ_0 for which the total variation is smaller than ε and then test whether the estimated spread parameter is in this neighborhood. Therefore this test falls in the family of "testing by learning". In fact, [4, Corollary 9] implies that if $\varepsilon_0 = \Omega(\sqrt{\log(1/\delta)/m})$, Algorithm 2 requires a single sample.

To show that the sample complexity $O\left(\log(1/\delta)/(m\varepsilon_0^2)\right)$ of Algorithm 2 is optimal, it is enough to show that it matches to the lower bound given in Corollary 5.3. In other words, we have to show that ε_0 is of order $\varepsilon/\sqrt{m}$. This claim is shown next in the Theorem 5.5 for a wide range of parameters.

Theorem 5.5 *There exists a constant $c > 0$ such that, for any $\phi, \phi' \in [0,1]$ and for every $\pi \in S_m$, it holds that*

1. *if $|\phi - \phi'| \leq c/m$ and $\phi, \phi' > c$, then $\mathrm{d_{TV}}(\mathcal{M}_{\phi,\pi}, \mathcal{M}_{\phi',\pi}) \in \Omega(\sqrt{m}|\phi - \phi'|)$,*
2. *if $|\phi - \phi'| \leq c/\sqrt{m}$ and $\phi, \phi' > c$ then $\mathrm{d_{TV}}(\mathcal{M}_{\phi,\pi}, \mathcal{M}_{\phi',\pi}) \in \Omega(\sqrt{m}|\phi - \phi'|^2)$.*

The proof of Theorem 5.5 is deferred to Appendix G. Theorem 5.5 implies that Algorithm 2 is optimal when the difference of parameters of the alternative and the null hypothesis is smaller than c/m, because in this case, $\mathrm{d_{TV}}(\mathcal{M}_{\phi,\pi}, \mathcal{M}_{\phi_0,\pi}) > \varepsilon$ implies that $\varepsilon_0 = \Omega(\varepsilon/\sqrt{m})$. Thus the sample complexity of Algorithm 2 is $O(\log(1/\delta)/\varepsilon^2)$ that matches the lower bound given in Corollary 5.3 up to a logarithmic factor. Albeit to show that Algorithm 2 is optimal for larger parameter values is an interesting open question. Nevertheless, we believe that this algorithm is optimal based on some numerical analysis that is shown in Figure 1.The lower bound that is linear in the difference of the parameters seems to be a very tight approximation even if $|\phi_0 - \phi| \gg 1/m$, whereas the approximation based on the Hellinger distance, i.e. Theorem 5.5 seems to underestimate the total variation distance for small $|\phi_0 - \phi|$. On the other hand, the liner lower bound does not hold for large value of $|\phi_0 - \phi|$, and the lower bound based on Hellinger distances captures the non-linearity of the total variation for larger $|\phi_0 - \phi|$.

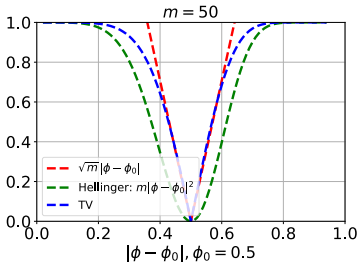

Figure 1: Total variation distance and its approximation based on Theorem 5.5. We set $m = 50$ and $\phi_0 = 0.5$. The x-axis shows the difference in parameters, i.e. $|\phi_0 - \phi|$.

Remark 2 *The Algorithm 1 and the Algorithm 2 are very close to each other when ε is small enough and ϕ_0 is bounded away from 0 and 1. In this case the upper bound $\mathrm{d_{TV}}\left(\mathcal{M}_{\phi_0,\pi_0}, \mathcal{M}_{\phi_0',\pi_0}\right) \leq \sqrt{m}\,|\phi_0 - \phi_0'|$ is very tight and hence both the test and the sample complexity of the two tests are essentially the same. When ε is larger or ϕ_0 is close to either 1 or 0 then the testing problem becomes easier and Algorithm 1 takes advantage of this fact and needs less samples, whereas Algorithm 2 requires the number of samples given by the worst-case bound.*

Remark 3 *(Scalable implementation of Algorithm 2.) The optimization tasks defined in Lines 2 and 3 of Algorithm 2 can be solved by using binary search since the total variation distance is monotone increasing and decreasing function of ε' with $\phi_0 + \varepsilon'$ and $\phi_0 - \varepsilon'$, respectively. Moreover, one can compute the total variation distance in an efficient way as $\mathrm{d_{TV}}(\mathcal{M}_{\phi_0,\pi_0}, \mathcal{M}_{\phi,\pi_0}) = \frac{1}{2}\sum_{\ell=0}^{m(m-1)/2} N_\ell \left|\phi_0^\ell/Z(\phi_0) - \phi^\ell/Z(\phi)\right|$. This implementation works for moderate number of items, i.e. $m \leq 30$. However, based on Theorem 5.5, Algorithm 2 can be implemented in a more scalable way without computing the total variation distance. More detailed, Theorem 5.5 provides an easy way to decide whether the model with estimated parameter is ε-close to the null model in terms of total variation distance.*

6 Testing for spread parameter with unknown central ranking

6.1 Asymptotic case

In this section, we consider that case when the null hypothesis consists of a single model $H_0 : \mathcal{M} \in \mathcal{R}_0 = \{\mathcal{M}_{\phi_0,\pi_0}\}$ and the alternative hypothesis set is $H_1 : \mathcal{M} \in \mathcal{R}_1 = \{\mathcal{M}_{\phi,\pi} : \phi \neq \phi_0 \vee \pi \neq \pi_0\}$. Let us consider a hypothetical test which knows the central ranking π of the underlying model $\mathcal{M}$ and it always applies the $f_n^{\pi}(\mathcal{D}_n) = g_n(T_{\pi}(\mathcal{D}))$ where g_n is defined in Theorem 5.1. The power of this test is clearly an upper bound for the UMPU tests for this testing problem– if there exists UMPU test at all– since we know that f_n^{π} is UMPU if π is known based on Theorem 5.1. However, one can show that in general there is no unbiased test in this case which follows from the following argument.

Remark 4 (Non-existence of UMPU test.) *If there exists an α level unbiased test f for testing $H_0 : \mathcal{M} \in \mathcal{R}_0 = \{\mathcal{M}_{\phi_0,\pi_0}\}$ vs $H_1 : \mathcal{M} \in \mathcal{R}_1 = \{\mathcal{M}_{\phi,\pi} : \phi \neq \phi_0 \vee \pi \neq \pi_0\}$, based on Theorem 5.1 and the unbiasedness property, it necessarily accepts all sample $\mathcal{D}$ for which $T_{\pi_0}(\mathcal{D}) \in (t_1, t_2)$ where t_1 and t_2 is the border of the critical region of the α level test $f_n^{\pi_0}$. By setting α small enough, the critical region of $f_n^{\pi_0}$ will include $T_{\pi_0}(\{\pi'\})$ for some π' such that $d_K(\pi_0, \pi') \geq 1$. Consequently, $\beta_f(\mathcal{M}_{0,\pi'}) = 0 < \beta_f(\mathcal{M}_{\phi_0,\pi_0}) = \alpha$ which means that the test is biased.*

Nevertheless, we will devise an easy-to-implement test with a small bias that is vanishing exponentially fast in n. To devise such a test, we need to get an estimate of the central ranking with high probability. To find the central ranking π_0 that maximizes likelihood under Mallows model is known to be an NP-hard, since there is a reduction to the *weighted feedback are set problem* [1]. Instead, we shall consider an estimator of central ranking based average ranking which can recover the central ranking with high probability as follows.

Lemma 6.1 *Let $\mathcal{D}_n = \{\pi_1, \dots, \pi_n\}$ is a set of rankings generated i.i.d. from $\mathcal{M}_{\phi,\pi}$ and let $\bar{\pi}$ is the ranking that sorts $[m]$ descending order based on their average rank $s_i = \frac{1}{n}\sum_{i=1}^{n} \pi_j(i)$. Then it holds that $\mathbb{P}(\bar{\pi} \neq \pi) \leq e^{-n(1-\phi)/2+2\log m}$*

The proof is based on [2] and presented in Appendix C. We would like to note that [30] analysed $\mathbb{P}(\bar{\pi} \neq \pi)$ by giving lower bound and upper bound algorithm which approach can be used as well. Based on Lemma 6.1, we can have a simple 2-stage test: 1) take first half of $\mathcal{D}_n$ to compute $\bar{\pi}$, if $\bar{\pi} \neq \pi_0$ then reject else 2) apply $f_n^{\bar{\pi}}$ using the rest of $\mathcal{D}_n$. Let us denote this 2-stage test by $\bar{f}_n$ which works with n samples. Based on Lemma 6.1, one can compute an upper bound on the bias of this test.

Theorem 6.2 *Using the notation above, for any $\alpha \in (0, 1/2)$ and $n > \frac{2}{1-\phi}\log\frac{2}{\alpha m^2}$, it holds that $\bar{f}_{2n}$ is an α level test and its bias is at most $1/2e^{-n(1-\phi)/2+2\log m}$.*

The proof is presented in Appendix I. The power of $\bar{f}_n$ is at least $1-\alpha$ if the the central ranking of the alternative hypothesis does not coincide with π_0.

6.2 Non-asymptotic case

Based on Lemma 6.1 it is easy to see that average ranking algorithm recovers the central ranking with $O\big(\frac{1}{1-\phi}\log\frac{m}{\delta}\big)$ samples with probability at least $1-\delta$, thus it can be applied in the non-asymptotic case. This leads us to a two stage algorithm to test $H_0 : \mathcal{M} \in \mathcal{R}_0 = \{\mathcal{M}_{\phi_0,\pi_0}\}$ versus $H_1 : \{\mathcal{M}_{\phi,\pi} : \mathrm{d_{TV}}\left(\mathcal{M}_{\phi_0,\pi_0}, \mathcal{M}_{\phi,\pi}\right) > \varepsilon\}$ which first estimates the central ranking based on the average ranking algorithm, and then runs Algorithm 1 or 2 with the estimated central ranking. The sample complexity of this two stage algorithm is $O(\max\{\frac{1}{1-\phi}\log(m/\delta), \frac{1}{\varepsilon}\log(1/\delta)\})$ which remains optimal up to a $\log m$ factor since the lower bound presented in Corollary 5.3 applies in this case as well.

7 Experiments

We shall present synthetic experiments to assess the performance asymptotic and non-asymptotic tests. Each result are computed based on 100 repetitions.

7.1 χ^2 approximation of UMPU test

The goal of the first set of experiments is to compare the power of the non-asymptotic test f_n defined in Theorem 5.1 based on exact computation of the distribution of the sufficient statistic and its χ^2 approximation. Figure 2 shows their power for various null and alternative hypothesis. We picked the central ranking to be the identity ranking, since it has no impact on the results. One can see that the approximation slightly deteriorates the power of f_n for small sample size, but the difference is marginal for $n = 50$. We could not compute f_n with $m > 10$ and $n = 50$ in python. However the χ^2 approximation can be run with larger m as well as n. As we can see from this experiment, the χ^2 approximations is accurate already for small sample size and small number of items.

7.2 Testing parameters of Mallows model for $m \leq 30$

In the second experiments we compare Algorithm 1 and 2. To implement Algorithm 1, we need to compute $s(n) = \sup_{\mathcal{M} \in \mathcal{R}_1 \setminus \mathcal{R}(f_n)} \mathrm{d_{TV}}\left(\mathcal{M}, \mathcal{M}_{\phi_0,\pi_0}\right)$, which is based on the distribution of the sufficient statistic. We used χ^2 approximation for the sufficient statistic when we compute $s(n)$. More detailed, we carried out two nested binary searches, one over n and a second nested one over ϕ to compute $s(n)$. To compute $s(n)$, the δ and $1-\delta$ quantile of the distribution of sufficient statistic based on n samples need to be estimated, so the χ^2 approximation has to be good on the tails. With this heuristic, there is no performance guaranty for Algorithm 1, that is why it is interesting to test its performance empirically versus an optimal algorithm. The power of Algorithm 1 and 2 are shown in Figure 3. We run Algorithm 2 with $\log\left(1/\delta\right)/(2m\varepsilon_0^2)$. There are some general trends revealed by these experiments. First, note that the sample sizes are not increasing with m, as one we expected since the sample complexity of both algorithm is $O(1/\varepsilon \log(1/\delta))$. On the other hand, as we pointed out in Remark 2, these two tests are very close to each other when ϕ_0 is not close too neither 0 nor 1. It can be seen that for $\phi_0 = 0.5$, the power of these two tests are very close to each other, and for $\phi_0 = 0.1$ Algorithm 1 has higher power for almost the same number of samples. Note that we used an χ^2 approximation of Algorithm 1, whereas Algorithm 2 is based on exact computation. Hence, we can verify that the χ^2 approximation indeed does work well in our setting even for moderate m.

7.3 Scalable implementation, using single sample

As we observed Algorithm 2 requires only a single sample whenever $\varepsilon_0 = \Omega(\sqrt{\log(1/\delta)/m})$. In addition to this, ε_0 goes to zero as m goes to infinity, since with a fixed ε, the parameter ϕ_ε converges to ϕ when m goes to infinity. Therefore for large m single sample is enough for Algorithm 2. We tested this observation with large m. If we run Algorithm 1 and 2 on a single ranking, they boil down to a very simple algorithm which consists of estimating the ϕ based on a single sample and if the estimate is far from ϕ_0, i.e. $|\hat{\phi} - \phi_0|$ is bigger than a threshold, then it rejects. This threshold is computed by using χ^2 approximation for Algorithm 1 and based on Theorem 5.5 for Algorithm 2. Figure 4 shows the results for $m \in \{100, 1000\}$. As we can see, very similar results can be obtained for large m like before for $m \leq 30$, using only a single ranking which indeed shows that these testing algorithms do scale gracefully with m. For an intuitive explanation, our tests for the spread parameter extract information about ϕ from the outcome of pairwise comparisons between items in π_0, this is how the Kendall-tau distance is computed. The larger m, the more pairwise comparisons available that is why these tests can work based on a single sample intuitively.

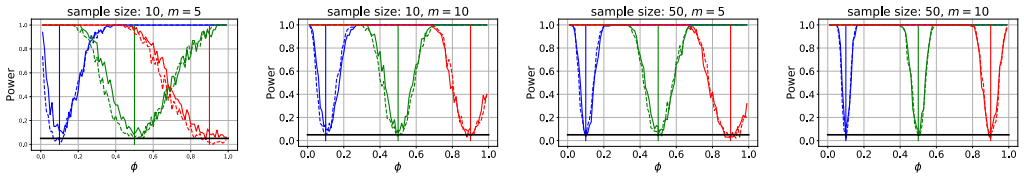

Figure 2: The power of asymptotic tests with $m \in \{5, 10\}$ and $\phi_0 \in \{0.1, 0.5, 0.9\}$, $\alpha = 0.05$ based on $n = \{10, 50\}$ random rankings. The x axis presents parameter ϕ of the underlying model. The solid lines shows the power of the exact test f_n wheres the dashed ones shows the power of the approximate test when the likelihood ratio is approximated by χ^2 distribution.

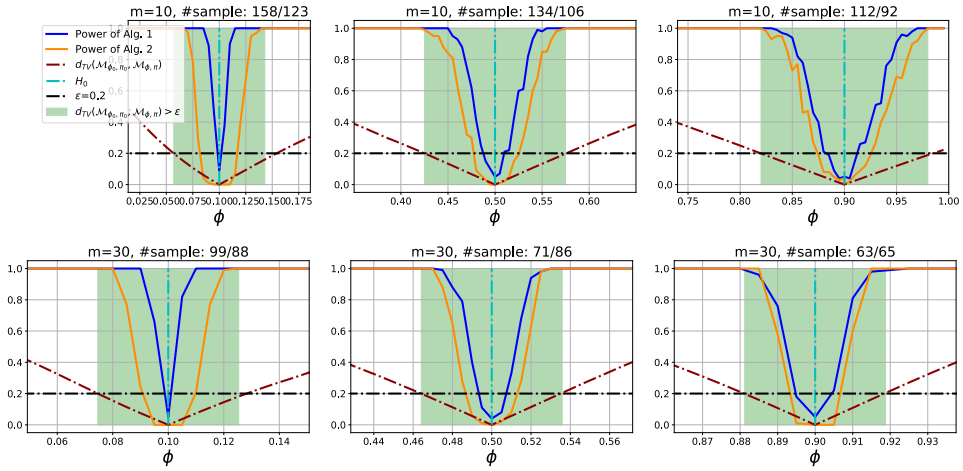

Figure 3: Power function for non-asymptotic tests. The null hypothesis is $\phi_0 \in [0.1, 0.5, 0.9]$ and $\varepsilon = 0.1, \delta = 0.05$. The sample size required by algorithms 2 and 1 is shown in the title of the plots, respectively.

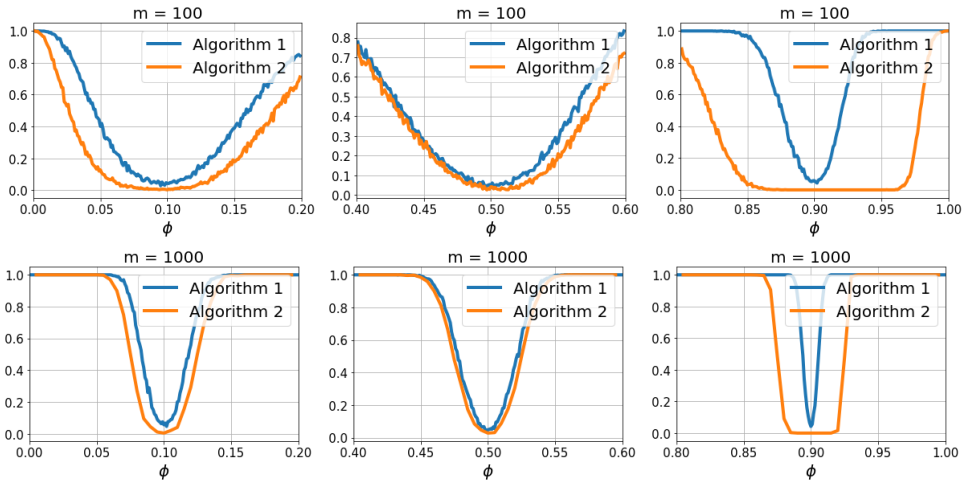

Figure 4: Power function for non-asymptotic tests with $m \in \{100, 1000\}$ by using a single ranking as input. The null hypothesis is $\phi_0 \in [0.1, 0.5, 0.9]$ and $\varepsilon = 0.1, \delta = 0.05$.

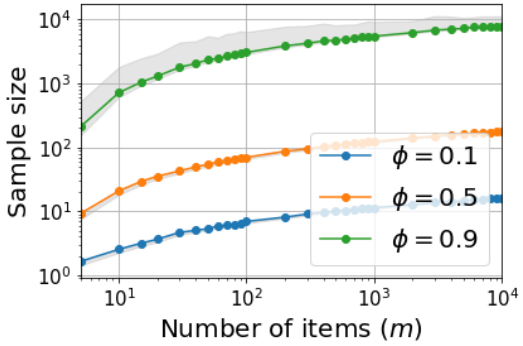

Figure 5: The sample size with error bar that average rank algorithms requires for recovering the central ranking of Mallows model with various number of items ranged from 5 to 10000.

7.4 Experiments with unknown central ranking

In the next set of experiments, we test the average ranking algorithm to recover the central ranking. We run this algorithms with different input sample size and estimate their accuracy, i.e. how many times they can recover the central ranking. The results are shown in Figure 5. Note that recovering the central ranking can be done based on relatively small number of examples, that is we need only ≈ 100 to recover the central ranking with high probability, even if it is shown that this is an NP-hard problem.

8 Conclusion and Future Work

We introduced several identity tests for Mallows model. We had found that the χ^2 approximation does not deteriorate the performance of the exact UMPU test f_n by a significant margin. Based on the asymptotic exact test, we devised a non-asymptotic test for which we applied the same χ^2 approximation and we had found that this test achieves similar result to the non-asymptotic test based on optimal learning. Our results clearly show that scalable identity test for Mallows model is indeed feasible even for $m = 1000$, since the spread parameter can be tested based on a single ranking when m is large and, moreover, the central ranking can be also recovered based on a small number of samples when $\phi_0 \leq 0.5$. This suggest a very simple and scalable approach for identity testing of ranking data that generated from Mallows model: take ≈ 100 rankings, estimate the central ranking on 100 rankings, and use only a single sample to test the spread parameter by using either Algorithm 1 with χ^2 approximation or Algorithm 2.

One interesting future direction is to explore hypothesis testing for the Generalized Mallows Model [12]. Apart from testing for the vector of the spread parameters, it would be interesting if we can test whether the samples come from a simple Mallows model, with the alternative that the samples come from a Generalized Mallows Model.

Acknowledgments and Disclosure of Funding

Dimitris Fotakis is supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant", project BALSAM, HFRI-FM17-1424.

References

[1] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, pages 684–693, 2005.

[2] Mark Braverman and Elchanan Mossel. Sorting from noisy information. *CoRR*, abs/0910.1191, 2009.

[3] Róbert Busa-Fekete, Eyke Hüllermeier, and Balázs Szörényi. Preference-based rank elicitation using statistical models: The case of Mallows. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1071–1079. JMLR.org, 2014.

[4] Róbert Busa-Fekete, Balázs Szörényi, Dimitris Fotakis, and Manolis Zampetakis. Optimal learning for mallows block model. In *International Conference on Computational Learning Theory (COLT)*, pages ??–??, 2019.

[5] Clément L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library, 2020.

[6] Siu-on Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. *CoRR*, abs/1308.3946, 2013.

[7] Ayala Cohen and C. L. Mallows. Assessing goodness of fit of ranking models to data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(4):361–374, 1983.

[8] Noel Cressie and Timothy R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):440–464, 1984.

[9] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A Servedio. Learning poisson binomial distributions. *Algorithmica*, 72(1):316–357, 2015.

[10] Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26(1):363–397, 02 1998.

[11] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. *Electron. Colloquium Comput. Complex.*, 24:133, 2017.

[12] Michael A Fligner and Joseph S Verducci. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 359–369, 1986.

[13] P.D. Grünwald. *The minimum description length principle*. The MIT Press, 2007.

[14] Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer, 2011.

[15] Guy Lebanon and Yi Mao. Non-parametric modeling of partially ranked data. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 857–864, 2007.

[16] E.L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 2005.

[17] Bruce G. Lindsay, Marianthi Markatou, and Surajit Ray. Kernels, degrees of freedom, and power properties of quadratic distance goodness-of-fit tests. *Journal of the American Statistical Association*, 109(505):395–410, 2014.

[18] Tyler Lu and Craig Boutilier. Effective sampling and learning for mallows models with pairwise-preference data. *J. Mach. Learn. Res.*, 15(1):3783–3829, 2014.

[19] R. Luce and P. Suppes. *Handbook of Mathematical Psychology*, chapter Preference, Utility and Subjective Probability, pages 249–410. Wiley, 1965.

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

[21] Horia Mania, Aaditya Ramdas, Martin J. Wainwright, Michael I. Jordan, and Benjamin Recht. On kernel methods for covariates that are rankings. *Electron. J. Statist.*, 12(2):2537–2577, 2018.

[22] John I. Marden. *Analyzing and Modeling Rank Data*. Chapman & Hall, 1995.

[23] Marina Meila and Le Bao. An exponential model for infinite rankings. *Journal of Machine Learning Research*, 11:3481–3518, 2010.

[24] Sumit Mukherjee. Estimation in exponential families on permutations. *The Annals of Statistics*, 44(2):853–875, 2016.

[25] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[26] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inf. Theory*, 54(10):4750–4755, 2008.

[27] R. Plackett. The analysis of permutations. *Applied Statistics*, 24:193–202, 1975.

[28] Charvi Rastogi, Sivaraman Balakrishnan, Nihar Shah, and Aarti Singh. Two-sample testing on pairwise comparison data and the role of modeling assumptions. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1271–1276, 2020.

[29] Mingxuan Sun, Guy Lebanon, and Paul Kidwell. Estimating probabilities in recommendation systems. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 734–742, 2011.

[30] Wenpin Tang. Mallows ranking models: maximum likelihood estimate and regeneration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6125–6134. PMLR, 09–15 Jun 2019.

[31] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[32] L. H. Philip Yu, Jiaqi Gu, and Hang Xu. Analysis of ranking data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] We hope there is none.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We used synthetic data.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported the repetitions from the error bar is clear for binary error. Otherwise we reported error bar.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]