Mallows Ranking Models: Maximum Likelihood Estimate and Regeneration

Wenpin Tang [1]

Abstract

This paper is concerned with various Mallows ranking models. We study the statistical properties of the MLE of Mallows' ϕ model. We also make connections of various Mallows ranking models, encompassing recent progress in mathematics. Motivated by the infinite top-t ranking model, we propose an algorithm to select the model size t automatically. The key idea relies on the renewal property of such an infinite random permutation. Our algorithm shows good performance on several data sets.

1. Introduction

Ranked data appear in many problems of social choice, user recommendation and information retrieval. Examples include ranking candidates by a large number of voters in an election (e.g. *instant-runoff voting*), and the document retrieval problem where one aims to design a meta-search engine according to a ranked list of web pages output by various search algorithms. In the sequel, we use the words *ranking* and *permutation* interchangeably. A ranking model is given by a collection of items, and an unknown total ordering of these items.

There is a rich body of literature on probabilistic ranking models. The earliest work dates back to (Thurstone, 1927; 1931) where items are ranked according to the order statistics of a Gaussian random vector. (Bradley & Terry, 1952) introduced an exponential family model by pairwise comparisons, which was extended by (Luce, 1959; Plackett, 1975) with comparisons of multiple items. See also (Hunter, 2004; Cattelan, 2012; Chen & Suh, 2015; Shah & Wainwright, 2017; Hung & Fithian, 2019) for algorithms and statistical analysis of the Bradley-Terry model and its variants.

A more tractable subclass of the Bradley-Terry model was proposed by (Mallows, 1957) as follows. For $n \geq 1$, let

[1]Department of Mathematics, University of California, Los Angeles, USA. Correspondence to: Wenpin Tang <wenpin-tang@math.ucla.edu>.

$\mathfrak{S}_n$ be the set of permutations of $[n] := \{1, \ldots, n\}$. The parametric model

$$\mathbb{P}_{\theta, \pi_0, d}(\pi) = \frac{1}{\Psi(\theta, d)} e^{-\theta d(\pi, \pi_0)} \quad \text{for } \pi \in \mathfrak{S}_n, \quad (1.1)$$

is referred to as the *Mallows model*. Here $\theta > 0$ is the *dispersion parameter*, π_0 is the *central ranking*, $d(\cdot, \cdot) : \mathfrak{S}_n \times \mathfrak{S}_n \to \mathbb{R}_+$ is a discrepancy function which is right invariant: $d(\pi, \sigma) = d(\pi \circ \sigma^{-1}, id)$ for $\pi, \sigma \in \mathfrak{S}_n$, and $\Psi(\theta, d) := \sum_{\pi \in \mathfrak{S}_n} e^{-\theta d(\pi, \pi_0)}$ is the normalizing constant. Mallows primarily considered two special cases of (1.1):

- *Mallows' θ model*, where $d(\pi, \sigma) = \sum_{i=1}^n (\pi(i) - \sigma(i))^2$ is the *Spearman's rho*,

- *Mallows' ϕ model*, where $d(\pi, \sigma) = \text{inv}(\pi \circ \sigma^{-1})$, called the *Kendall's tau*,

where $\text{inv}(\pi) := \#\{(i, j) \in [n]^2 : i < j \text{ and } \pi(i) > \pi(j)\}$ is the number of inversions of π. The general form (1.1) was suggested by (Diaconis, 1988) along with other discrepancy functions. (Diaconis, 1988; 1989) and (Critchlow, 1985) also pioneered the group representation approach to ranked, and partially ranked data.

In this paper we are primarily concerned with the statistical properties of the maximum likelihood estimate (MLE) of the Mallows' ϕ model and its infinite counterpart. Specializing (1.1) with the Kendall's tau, the Mallows' ϕ model is expressed as

$$\mathbb{P}_{\theta, \pi_0}(\pi) = \frac{1}{\Psi(\theta)} e^{-\theta \, \text{inv}(\pi \circ \pi_0^{-1})} \quad \text{for } \pi \in \mathfrak{S}_n. \quad (1.2)$$

It is easily seen that $\mathbb{P}_{\theta, \pi_0}$ has a unique mode π_0 if $\theta > 0$. The Mallows' ϕ model (1.2) is of particular interest, since it is an instance of two large classes of ranking models: distance-based ranking models (Fligner & Verducci, 1986) and multistage ranking models (Fligner & Verducci, 1988). The one parameter model (1.2) also has an $n - 1$ parameter extension where θ is replaced with $\vec{\theta} := (\theta_1, \ldots \theta_{n-1})$ by factorizing the inversion. This $n-1$ parameter model, called the *Generalized Mallows* ($\text{GM}_{\vec{\theta}, \pi_0}$) model, will be discussed in Section 2. See (Critchlow et al., 1991; Marden, 2014) for a review of these ranking models.

(Fligner & Verducci, 1988) showed that if the central ranking π_0 is known, the MLE of θ (or $\vec{\theta}$) can be easily found by

convex optimization. But it is harder to compute the MLE of π_0, and only a few heuristic algorithms are available. As pointed out in (Meilă et al., 2012), the problem of finding the MLE of π_0 for the Mallows' ϕ model is the *Kemeny's consensus ranking problem* which is known to be NP-hard (Young, 1986; Bartholdi et al., 1989). They also gave a branch and bound search algorithm to estimate simultaneously θ (or $\vec{\theta}$) and π_0. See also (Cohen et al., 1998; Ailon et al., 2008; Mandhani & Meilă, 2009) for approximation algorithms for consensus ranking problem, (Busa-Fekete et al., 2014; Irurozki et al., 2016; 2018) for efficient sampling and learning of Mallows models, and (Lebanon & Mao, 2008; Chierichetti et al., 2018) for various generalizations of Mallows ranking models. Though there have been efforts in developing algorithms to estimate θ (or $\vec{\theta}$) and π_0, not much is known about the statistical properties of the MLE $\widehat{\theta}$ and $\widehat{\pi}_0$ even for the simplest model (1.2). We provide statistical analysis to the MLE of the model (1.2), and answer the following questions in Section 3: Are the MLEs $\widehat{\theta}$, $\widehat{\pi}_0$ consistent ? Is the MLE $\widehat{\theta}$ unbiased ? How fast does the MLE $\widehat{\pi}_0$ converge to π_0 ?

When the number of items n is large, learning a complete ranking model becomes impracticable. A line of work by (Fligner & Verducci, 1986; Busse et al., 2007; Meilă & Bao, 2010; Meilă & Chen, 2012) focused on the top-t orderings for the $\text{GM}_{\vec{\theta}, \pi_0}$ model. Among these work, (Meilă & Bao, 2010) proposed a probability model over the top-t orderings of infinite permutations, called the *Infinite Generalized Mallows* ($\text{IGM}_{\vec{\theta}, \pi_0}$) model:

$$\mathbb{P}_{\vec{\theta}, \pi_0}(\pi) = \frac{1}{\Psi(\vec{\theta})} \cdot \exp\left(-\sum_{j=1}^{t} \theta_j s_j(\pi \circ \pi_0^{-1})\right), \quad (1.3)$$

where $s(\pi) := (s_1(\pi), s_2(\pi), \ldots)$ is the *inversion table* of π defined by

$$s_j(\pi) := \pi^{-1}(j) - 1 - \sum_{j' < j} 1_{\{\pi^{-1}(j') < \pi^{-1}(j)\}}, \quad (1.4)$$

and $\Psi(\vec{\theta}) = \prod_{j=1}^{t}(1 - e^{-\theta_j})^{-1}$ is the normalizing constant. By convention, $\theta_j = 0$ for any $j > t$. The integer 't' is referred to as the *model size* of the IGM model. If $\theta_1 = \cdots = \theta_t = \theta$, the $\text{IGM}_{\vec{\theta}, \pi_0}$ model is called the *single parameter* IGM model.

As explained in (Meilă & Bao, 2010), the $\text{IGM}_{\vec{\theta}, \pi_0}$ model (1.3) is the marginal distribution of a random permutation of positive integers. The single parameter IGM was called the *infinite q-shuffle* in (Gnedin & Olshanski, 2010) with parameterization $q = e^{-\theta}$. They provided a nice construction of the infinite q-shuffle, which is reminiscent of *absorption sampling* (Rawlings, 1997; Kemp, 1998) and the *repeated insertion model* (Doignon et al., 2004). The infinite q-shuffle was further extended by (Pitman & Tang, 2017) to

the *p-shifted permutations* of positive integers. But the link between Meilă-Bao's infinite ranking model and infinite q-shuffle or p-shifted permutations does not seem to have been previously noticed. So we point out this connection which will be detailed in Section 2.

One disadvantage of the aforementioned top-t ranking models is that they all require choosing 't' manually. Small model size 't' often leads to poor accuracy, and large model size 't' needs a considerable amount of training time. It was observed in (Pitman & Tang, 2017) that the random infinite limit of the single parameter IGM model has a remarkable renewal property. This suggests a heuristic procedure to select the model size 't' automatically based on Meilă-Bao's search algorithms. We will discuss such an approach in Section 2.

To summarize, the main contributions of this paper are:

- Provide statistical analysis to the MLE of the Mallows' ϕ model as well as the single parameter IGM.

- Propose a selection algorithm for the model size 't' of the top-t Mallows ranking models.

See also (Lebanon & Lafferty, 2002; Lu & Boutilier, 2011; 2014; Awasthi et al., 2014; Vitelli et al., 2017; Liu & Moitra, 2018) for the approximate Baysian inference of Mallows' mixture models for clustering heterogeneous ranked data, and (Huang & Guestrin, 2012; Huang et al., 2012; Meek & Meila, 2014) for hierarchical ranking models.

2. Mallows Models: 't' Selection Algorithm

In this section we provide background on various Mallows ranking models, encompassing the closely related q-shuffles and p-shifted permutations. We also give an algorithm to select the model size 't' for the IGM model (1.3). We follow closely (Fligner & Verducci, 1988; Meilă & Bao, 2010; Pitman & Tang, 2017).

2.1. Finite Mallows Models

Given n items labelled by $[n]$, a ranking $\pi \in \mathfrak{S}_n$ is represented by

- the *word list* $(\pi(1), \pi(2), \ldots, \pi(n))$,

- the *ranked list* $(\pi^{-1}(1)|\pi^{-1}(2)|\ldots|\pi^{-1}(n))$.

Here $\pi(i) = j$ means that the item i has rank j, and conversely $\pi^{-1}(j) = i$ means that the jth most preferred is item i. The idea of multistage ranking is to decompose the ranking procedure into independent stages. The most preferred item is selected at the first stage, the best of the

remaining at the second stage and so on until the least pre-ferred item is selected. The correctness of the choice at any stage is accessed through a central ranking π_0. For example, if $\pi_0 = (3|1|2)$, then the ranking $\pi = (3|2|1)$ gives a correct choice at the first stage, since item 3 is the most preferred in both π and π_0. But at the second stage, among the two remaining items 1 and 2, item 2 is selected by π while the right choice is item 1 according to π_0.

For any ranking $\pi \in \mathfrak{S}_n$ and $j \in [n-1]$, let $(s_1(\pi), \ldots, s_{n-1}(\pi))$ be the inversion table of π defined by (1.4). It is easy to see that $s_j(\pi) \in \{0, \ldots, n-j\}$, and there is a bijection between a ranking π and the inversion table $(s_1(\pi), \ldots, s_{n-1}(\pi))$. The quantity $s_j(\pi \circ \pi_0^{-1})$ measures the correctness of the choice at stage j: $s_j(\pi \circ \pi_0^{-1}) = k$ means that at stage j the $(k+1)^{th}$ best of the remaining items is selected. In the example with $\pi_0 = (3|1|2)$ and $\pi = (3|2|1)$, we have $s_1(\pi \circ \pi_0^{-1}) = 0$ and $s_2(\pi \circ \pi_0^{-1}) = 1$. (Fligner & Verducci, 1986) introduced the multistage rank-ing models of the form:

$$\mathbb{P}_{p, \pi_0}(\pi) = \prod_{j=1}^{n-1} p_j \left(s_j(\pi \circ \pi_0^{-1}) \right), \quad (2.1)$$

where $p_j(\cdot)$ is a probability distribution on $\{0, \ldots, n-j\}$ at stage j. The choice of $p_j(k) = (1 - e^{-\theta})e^{-k\theta}/(1 - e^{-(n-j+1)\theta})$ for $k \in \{0, \ldots, n-j\}$ in (2.1), and the re-markable identity $\sum_{j=1}^{n-1} s_j(\pi) = \text{inv}(\pi)$ for any $\pi \in \mathfrak{S}_n$ yield the Mallows' ϕ model (1.2).

This one parameter model has a natural $n-1$ parameter extension by simply taking $p_j(k) = (1 - e^{-\theta_j})e^{-\theta_j k}/(1 - e^{-(n-j+1)\theta_j})$ for $k \in \{0, \ldots, n-j\}$. The $n-1$ parameter model, called the *Generalized Mallows* ($\text{GM}_{\vec{\theta}, \pi_0}$) model, is then defined by

$$\mathbb{P}_{\vec{\theta}, \pi_0}(\pi) = \frac{1}{\Psi(\vec{\theta})} \exp \left(-\sum_{j=1}^{n-1} \theta_j s_j(\pi \circ \pi_0^{-1}) \right), \quad (2.2)$$

where $\Psi(\vec{\theta}) = \prod_{j=1}^{n-1}(1 - e^{-(n-j+1)\theta_j})(1 - e^{-\theta_j})^{-1}$ is the normalizing constant. The $\text{GM}_{\vec{\theta}, \pi_0}$ model is also called the *Mallows' ϕ-component* model. The model (2.2) can also be expressed in the form $e^{-d_{\vec{\theta}}(\pi, \pi_0)}/\Psi(\vec{\theta})$, with $d_{\vec{\theta}}(\pi, \pi_0) := \sum_{j=1}^{n-1} \theta_j s_j(\pi \circ \pi_0^{-1})$.

2.2. Infinite Mallows Models

Given a countably infinite items labelled by $\mathbb{N}_+ := \{1, 2, \ldots\}$, a ranking π over $\mathbb{N}_+$ is a bijection from $\mathbb{N}_+$ onto itself represented by the word list $(\pi(1), \pi(2), \ldots)$ or the ranked list $(\pi^{-1}(1)|\pi^{-1}(2)|\ldots)$. A top-$t$ order-ing of π is the prefix $(\pi^{-1}(1)|\ldots|\pi^{-1}(t))$. Motivated by the $\text{GM}_{\vec{\theta}, \pi_0}$ model (2.2), (Meilă & Bao, 2010) proposed the *Infinite Generalized Mallow* ($\text{IGM}_{\vec{\theta}, \pi_0}$) model (1.3),

which can also be put in the form $e^{-d_{\vec{\theta}}(\pi, \pi_0)}/\Psi(\vec{\theta})$, with $d_{\vec{\theta}}(\pi, \pi_0) := \sum_{j=1}^{t} \theta_j s_j(\pi \circ \pi_0^{-1})$. In particular, s_j is dis-tributed as $\text{Geo}(1 - e^{-\theta_j})$ on $\{0, 1, \ldots\}$. As explained in (Meilă & Bao, 2010), one can regard π as a top-t ordering, and π_0 as an ordering over $\mathbb{N}_+$. If $\theta_1 = \cdots = \theta_t = \theta$, then the model (1.3) simplifies to

$$\mathbb{P}_{\theta, \pi_0}(\pi) = \frac{1}{\Psi(\theta)} \exp \left(-\theta \sum_{j=1}^{t} s_j(\pi \circ \pi_0^{-1}) \right), \quad (2.3)$$

called the *single parameter* IGM model.

It is easily seen that the single parameter IGM model (2.3) is the marginal distribution of a random permutation of positive integers. Formally, this random infinite permutation is distributed as

$$\mathbb{P}_{\theta, \pi_0}(\pi) = \frac{1}{\Psi(\theta)} \exp \left(-\theta \sum_{j=1}^{\infty} s_j(\pi \circ \pi_0^{-1}) \right). \quad (2.4)$$

In the terminology of (Gnedin & Olshanski, 2010), for π de-fined by (2.4), $\pi \circ \pi_0^{-1}$ is the *infinite $e^{-\theta}$-shuffle*. The latter was generalized by (Pitman & Tang, 2017) to *p-shifted per-mutations*, where $p = (p_1, p_2, \ldots)$ is a discrete distribution on $\mathbb{N}_+$ with $p_1 > 0$. Here we present a further extension of p-shifted permutations.

Let $P = (p_{ij})_{i,j \in \mathbb{N}_+}$ be a stochastic matrix on $\mathbb{N}_+$ with $p^i = (p_{ij})_{j \in \mathbb{N}_+}$ being the ith row of P. Assume that $\lim_{n \to \infty} \prod_{i=1}^{n}(1 - p_{i1}) = 0$. Call a random permutation Π of $\mathbb{N}_+$ a *P-shifted permutation* of $\mathbb{N}_+$ if Π has the distribu-tion defined by the following construction from the indepen-dent sample $(X_i)_{i \geq 1}$, with X_i distributed as p^i. Inductively, let $\Pi_1 := X_1$, and for $i \geq 2$, let $\Pi_i := \psi(X_i)$ where ψ is the increasing bijection from $\mathbb{N}_+$ to $\mathbb{N}_+ \setminus \{\Pi_1, \cdots, \Pi_{i-1}\}$. For example, if $X_1 = 2$, $X_2 = 1$, $X_3 = 2$, $X_4 = 3$, $X_5 = 4$, $X_6 = 1 \ldots$, then the associated permutation is $(2, 1, 4, 6, 8, 3, \ldots)$.

Now the aforementioned infinite ranking models are sub-cases of the P-shifted permutations.

- If $p^i = p$ with $p_1 > 0$ for all i, then we get the p-shifted permutation.

- If $p^i = \text{Geo}(1 - e^{-\theta})$ on $\mathbb{N}_+$ for all i, then we get the single parameter IGM model, or infinite $e^{-\theta}$-shuffle, i.e. $\Pi \stackrel{d}{=} \pi \circ \pi_0^{-1}$ for π distributed according to (2.4).

- If $p^i = \text{Geo}(1 - e^{-\theta_i})$ on $\mathbb{N}_+$ for each i, then $\Pi \stackrel{d}{=} \pi \circ \pi_0^{-1}$ for π an infinite version of the IGM model (1.3).

2.3. 't' Selection Algorithm

We present an algorithm to select the model size 't' automat-ically for the top-t IGM models. The heuristic comes from

the renewal property of the single parameter IGM model (2.3). We need the following vocabulary.

Let Π be a permutation of $\mathbb{N}_+$. Call $n \in \mathbb{N}_+$ a *splitting time* of Π if Π maps $[1, n]$ onto itself. The set of splitting times of Π is the collection of finite right endpoints of some finite or infinite family of *components* of Π, say $\{I_j\}$. So Π does not act as a permutation on any proper subinterval of I_j. These components I_j form a partition of $\mathbb{N}_+$, which is coarser than the partition by cycles of Π. For example, the permutation $\pi = (1)(2, 4)(3) \in \mathfrak{S}_4$ induces the partition by components $[1][2, 3, 4]$.

The idea is to use the single parameter IGM model to preselect 't', which hinges on the renewal property of the latter. Then we proceed to train a top-t ranking model. (Pitman & Tang, 2017) proved that for $p = (p_1, p_2, \ldots)$ a discrete distribution with $p_1 > 0$ and $\sum_{i \geq 1} i p_i < \infty$, a p-shifted permutation Π is a concatenation of independent and identically distributed (i.i.d.) components. That is, Π is characterized by L a distribution on $\mathbb{N}_+$, and $(Q_n)_{n \geq 1}$ a sequence of distributions on indecomposable permutations such that

- the lengths of components $(L_i)_{i \geq 1}$ are i.i.d. as L,

- given the length of a component $L_i = n_i$, the reduced component defined via conjugation of Π by the shift from the component to $[n_i]$ is distributed as Q_{n_i}.

To illustrate, $(\underbrace{2, 3, 4, 1}_{L_1=4}, \underbrace{6, 8, 7, 10, 5, 9}_{L_2=6}, \underbrace{12, 13, 11}_{L_3=3}, \ldots)$
Moreover, the probability generating function of L is given by $F(z) = 1 - \frac{1}{1+\sum_{n=1}^{\infty} u_n z^n}$, with $u_n := \prod_{i=1}^{n} \sum_{j=1}^{i} p_i$. Specializing this renewal construction to the single parameter IGM model, we get the following proposition which is the foundation of Algorithm 1 described right after.

Proposition 2.1. *Let Π be a random permutation of $\mathbb{N}_+$ distributed as $\mathbb{P}_{\theta, id}$ defined by (2.4). Let L be the common distribution of lengths of components of Π. Then*

$$\mathbb{E}L = \frac{1}{(e^{-\theta}; e^{-\theta})_\infty}, \qquad (2.5)$$

where $(a; q)_\infty := \prod_{k=0}^{\infty} (1 - aq^k)$ is the Q-Pochhammer function.

Pitman-Tang's theory indicates that for the single parameter model (2.4), the first component of $\pi \circ \pi_0^{-1}$ has length L whose expectation is $1/(e^{-\theta}; e^{-\theta})_\infty$. Given the dispersion parameter θ, a complete permutation is expected to occur at some place close to $1/(e^{-\theta}; e^{-\theta})_\infty$. The latter can be regarded as the *effective length* of the random infinite permutation, which suggests a natural candidate for 't' in top-t ranking models. Since θ is unknown, we would like to find the 't' closest to the effective length. Given t in a suitable range $\mathbb{T}$, we fit the single parameter IGM model (2.3) to get
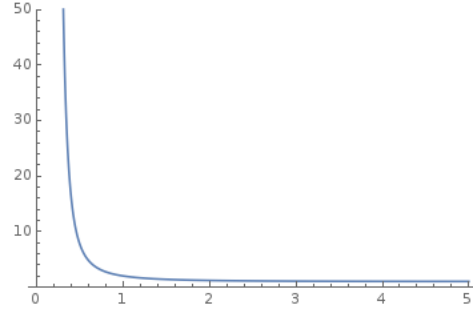

Figure 1. Plot of $\theta \rightarrow 1/(e^{-\theta}; e^{-\theta})_\infty$ for $\theta \in [0, 5]$.

Algorithm 1 't' selection algorithm
$\theta_0 \leftarrow \text{MB}(1)$ (Run Meilă-Bao's algorithm)
Choose $\mathbb{T} \ni 1/(e^{-\theta_0}; e^{-\theta_0})_\infty$ of a small range
Initialize $Err \leftarrow \infty$, $t\_\text{SEL} \leftarrow 0$
for t in $\mathbb{T}$ **do**
    $\theta \leftarrow \text{MB}(t)$ (Run Meilă-Bao's algorithm)
    if $|t - 1/(e^{-\theta}; e^{-\theta})_\infty| < Err$ **then**
        $Err \leftarrow |t - 1/(e^{-\theta}; e^{-\theta})_\infty|$
        $t\_\text{SEL} \leftarrow t$
    end if
end for
return $\min(t\_\text{SEL}, \lambda t_{\max})$

---

the MLE $\widehat{\theta}(t)$ by the algorithms in (Meilă & Bao, 2010). Those algorithms also work for partially ranked data. Then we search for

$$t := \min \left\{ \text{argmin}_{s \in \mathbb{T}} \left| s - \frac{1}{(e^{-\widehat{\theta}(s)}; e^{-\widehat{\theta}(s)})_\infty} \right|, \lambda t_{\max} \right\},$$
$$(2.6)$$

where $\mathbb{T}$ is the range of search for model sizes, $\lambda \in (0, 1)$ is a user-defined cutoff fraction to avoid overfitting, and $t_{\max}$ is the maximum length of permutations in the data. Practically, one starts with 's' of small values to narrow down the choices for the effective length. That way, one only needs to search for a small proportion, not the full range of $[t_{\min}, t_{\max}]$. With 't' selected according to (2.6), we can then fit the IGM model (1.3) by means of MLE.

3. Statistical Properties of the MLE

In this section we provide statistical analysis to the MLE of Mallows models.

3.1. Main Theorems

(Mukherjee, 2016) proved that the MLE $\widehat{\theta}$ for the general Mallows model (1.1) is consistent if π_0 is known. His approach relies on the concept of *permutons* (Hoppen et al.,

2013). The following result shows that the MLE $\widehat{\theta}$ is always biased upwards for the Mallows' ϕ model.

Theorem 3.1 (Bias of $\widehat{\theta}$). *Let $\mathbb{P}_{\theta,\pi_0}$ be defined by* (1.2), *and $\widehat{\theta}$ be the MLE of θ with N samples. Then for each $N \geq 1$,*

$$\mathbb{E}_{\theta,\pi_0}\widehat{\theta} > \theta. \tag{3.1}$$

The analysis of the MLE $\widehat{\pi}_0$ is more subtle since it lives in a discrete space. By general results of (Newey & McFadden, 1994; Choirat & Seri, 2012), one can prove the consistency of $\widehat{\pi}_0$. Here we establish a concentration bound of $\widehat{\pi}_0$ at π_0 from which the consistency is straightforward. This concentration bound also gives a confidence interval for the central ranking in the Mallows' ϕ model.

Theorem 3.2 (Convergence rate of $\widehat{\pi}_0$). *Let $\mathbb{P}_{\theta,\pi_0}$ be defined by* (1.2), *and $\widehat{\pi}_0$ be the MLE of π_0 with N samples. Then for N large enough,*

$$\mathbb{P}_{\theta,\pi_0}(\widehat{\pi}_0 \neq \pi_0) \geq \frac{1}{1-e^{-\theta}}\sqrt{\frac{2}{\pi N}}\left(\cosh\frac{\theta}{2}\right)^{-N}, \tag{3.2}$$

and

$$\mathbb{P}_{\theta,\pi_0}(\widehat{\pi}_0 \neq \pi_0) \leq (n - H_n)n!\left(\cosh\frac{\theta}{2}\right)^{-N}, \tag{3.3}$$

where $H_n := \sum_{i=1}^{n}\frac{1}{i}$ is the harmonic sum.

To the best of our knowledge, Theorems 3.1 and 3.2 are new. Their proof will be given in the next two subsections.

Similar to Theorem 3.1, the MLE of θ is biased upwards for the single parameter IGM model, confirming an observation in (Meilă & Bao, 2010).

Theorem 3.3 (Bias of $\widehat{\theta}$). *Let $\mathbb{P}_{\theta,\pi_0}$ be the distribution of the single parameter IGM model, and $\widehat{\theta}$ be the MLE of θ with N samples. Then for each $N \geq 1$,*

$$\mathbb{E}_{\theta,\pi_0}\widehat{\theta} > \theta. \tag{3.4}$$

Let us mention a few open problems. The rate of $\widehat{\theta}$ is open for both the Mallows' ϕ model and the single parameter IGM model. We believe that the rate is of order $1/\sqrt{N}$ for the Mallows' ϕ model by applying the delta method with a finer variance analysis. We can ask the same questions for the generalized Mallows model ($\text{GM}_{\vec{\theta},\pi_0}$). It is expected that the rate of $\widehat{\pi}$ is of order $e^{-N\beta(\vec{\theta})}$ but the explicit formula of $\beta(\vec{\theta})$ is still missing. Deriving a bound of $\beta(\vec{\theta})$ would also be interesting. The convergence rate of $\widehat{\pi}_0$ for the single parameter IGM model seems to be difficult, since the size of permutations goes to infinity. We leave the analog of Theorem 3.2 open.

3.2. Proof of Theorem 3.1

We first consider the case where the central ranking π_0 is known. Assume w.l.o.g. that $\pi_0 = id$ by suitably relabelling the items. Then the model (1.2) simplifies to

$$\mathbb{P}_{\theta,id}(\pi) = \exp\left(-\theta\operatorname{inv}(\pi) - \ln\Psi(\theta)\right). \tag{3.5}$$

Given N samples $(\pi_i)_{1\leq i\leq N}$, the MLE $\widehat{\theta}$ is the solution to the following equation:

$$-\frac{\Psi'(\theta)}{\Psi(\theta)} = \frac{1}{N}\sum_{i=1}^{N}\operatorname{inv}(\pi_i). \tag{3.6}$$

Now by writing $\Psi(\theta) = f(e^{-\theta})$ with $f(q) := \sum_{\pi\in\mathfrak{S}_n} q^{\operatorname{inv}(\pi)}$, we have

$$-\frac{\Psi'(\theta)}{\Psi(\theta)} = g(e^{-\theta}), \quad\text{with}\quad g(q) := \frac{qf'(q)}{f(q)}.$$

So $\widehat{\theta} = -\log g^{-1}(\frac{1}{N}\sum_{i=1}^{N}\operatorname{inv}(\pi_i))$. The function $f(q)$ is known as the q-*factorial* (Stanley, 2011). We have

$$g(q) = q\sum_{k=1}^{n-1}\frac{1-(k+1)q^k+kq^{k+1}}{(1-q)(1-q^{k+1})}. \tag{3.7}$$

As observed by (Mallows, 1957; Gnedin & Olshanski, 2010), for π distributed according to (3.5), the number of inversions has the same distribution as a sum of independent *truncated geometric random variables*. That is,

$$\operatorname{inv}(\pi) \stackrel{d}{=} G_{e^{-\theta},1} + \cdots + G_{e^{-\theta},n} \quad\text{for } \pi \sim \mathbb{P}_{\theta,id},$$

where $\mathbb{P}(G_{q,k} = i) = q^i(1-q)/(1-q^k)$ for $i \in \{0,1,\ldots,k-1\}$. Thus, $\mathbb{E}\operatorname{inv}(\pi) = g(e^{-\theta})$. By elementary analysis, $\theta \mapsto g(e^{-\theta})$ is strictly convex and decreasing. So its inverse function $q \mapsto -\log g^{-1}(q)$ is strictly convex. By Jensen's inequality, $-\mathbb{E}\log g^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\operatorname{inv}(\pi_i)\right) > -\log g^{-1}\left(\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\operatorname{inv}(\pi_i)\right)\right)$, which implies that $\mathbb{E}\widehat{\theta} > \theta$.

Now consider the case where the central ranking π_0 is unknown. The MLE $\widehat{\theta}$ is given by

$$\widehat{\theta} = -\log g^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\operatorname{inv}(\pi_i \circ \widehat{\pi}_0^{-1})\right), \tag{3.8}$$

where g is defined as in (3.7), and $\widehat{\pi}_0$ is the MLE of π_0. By the definition of $\widehat{\pi}_0$,

$$\sum_{i=1}^{N}\operatorname{inv}(\pi_i \circ \widehat{\pi}_0^{-1}) \leq \sum_{i=1}^{N}\operatorname{inv}(\pi_i \circ \pi_0^{-1}).$$

Since $q \mapsto -\log g^{-1}(q)$ is strictly convex and decreasing, we get

$$
\begin{aligned}
\mathbb{E}\widehat{\theta} &= -\mathbb{E}\log g^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{inv}(\pi_i \circ \widehat{\pi}_0^{-1})\right) \\
&\geq -\mathbb{E}\log g^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{inv}(\pi_i \circ \pi_0^{-1})\right) \\
&> -\log g^{-1}\left(\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{inv}(\pi_i \circ \pi_0^{-1})\right)\right) = \theta,
\end{aligned}
$$

where the last equality follows from the fact that $\pi_i \circ \pi_0^{-1}$ is distributed according to (3.5) for each i.

3.3. Proof of Theorem 3.2

Assume w.l.o.g. that the true central ranking $\pi_0 = id$. We aim to find the bounds of $\mathbb{P}_{\theta,id}(\widehat{\pi}_0 \neq id)$ given the dispersion parameter θ.

For $\pi, \pi' \in \mathfrak{S}_n$, we say that π is more likely than π', denoted $\pi \succeq \pi'$, if $\sum_{i=1}^{N}\mathrm{inv}(\pi_i \circ \pi^{-1}) \leq \sum_{i=1}^{N}\mathrm{inv}(\pi_i \circ \pi'^{-1})$. For $1 \leq j < k \leq n$, let (j,k) be the transposition of j and k. By the union bound, we get

$$
\mathbb{P}_{\theta,id}((1,2) \succeq id) \leq \mathbb{P}_{\theta,id}(\widehat{\pi}_0 \neq id) \leq \sum_{\pi \in \mathfrak{S}_n}\mathbb{P}_{\theta,id}(\pi \succeq id).
$$

Lower bound: For any permutation $\pi \in \mathfrak{S}_n$, if $\pi(1) > \pi(2)$, then $\mathrm{inv}(\pi \circ (1,2)) = \mathrm{inv}(\pi)-1$, and if $\pi(1) < \pi(2)$, then $\mathrm{inv}(\pi \circ (1,2)) = \mathrm{inv}(\pi)+1$. Thus, $\sum_{i=1}^{N}\mathrm{inv}(\pi_i \circ (1,2))$ equals to

$$
\sum_{i=1}^{N}\mathrm{inv}(\pi_i)+\#\{i : \pi_i(1) < \pi_i(2)\}-\#\{i : \pi_i(1) > \pi_i(2)\}.
$$

Consequently, $\mathbb{P}_{\theta,id}((1,2) \succeq id)$ is given by

$$
\begin{aligned}
&\mathbb{P}_{\theta,id}\left(\#\{i : \pi_i(1) < \pi_i(2)\} \leq \#\{i : \pi_i(1) > \pi_i(2)\}\right) \\
&= \mathbb{P}_{\theta,id}\left(\#\{i : \pi_i(1) > \pi_i(2)\} \geq \frac{N}{2}\right).
\end{aligned} \tag{3.9}
$$

Note that $\mathbb{P}_{\theta,id}(\pi_i(1) > \pi_i(2)) = e^{-\theta}\mathbb{P}_{\theta,id}(\pi_i(1) < \pi_i(2))$ which implies that

$$
\mathbb{P}_{\theta,id}(\pi_i(1) > \pi_i(2)) = \frac{1}{1+e^{\theta}}. \tag{3.10}
$$

Combining (3.9) and (3.10) yields

$$
\begin{aligned}
\mathbb{P}_{\theta,id}((1,2) \succeq id) &= \mathbb{P}\left(\mathrm{Bin}\left(N, \frac{1}{1+e^{\theta}}\right) \geq \frac{N}{2}\right) \\
&\sim \frac{1}{1-e^{-\theta}}\sqrt{\frac{2}{\pi N}}\left(\cosh\frac{\theta}{2}\right)^{-N},
\end{aligned}
$$

where $\mathrm{Bin}(N,p)$ is a binomial random variable with parameters (N,p), and the last estimate follows from the large deviation bound (Arratia & Gordon, 1989) that for $p < a$,

$$
\mathbb{P}(\mathrm{Bin}(N,p) > aN) \sim \frac{(1-p)\sqrt{a}}{(a-p)\sqrt{2\pi(1-a)N}}e^{-NH(a,p)},
$$

where $H(a,p) := a\log\left(\frac{a}{p}\right) + (1-a)\log\left(\frac{1-a}{1-p}\right)$ is the relative entropy between $\mathrm{Bin}(N,p)$ and $\mathrm{Bin}(N,a)$.

Upper bound: We need the following comparison result.

Proposition 3.4. *For $j, k \in \{1, \ldots, n\}$ and $j < k$, we have*

$$
\mathbb{P}_{\theta,id}((j,k) \succeq id) \leq \left(\cosh\frac{\theta}{2}\right)^{-N}. \tag{3.11}
$$

Now decompose each permutation $\pi \in \mathfrak{S}_n$ into a product of transpositions, say $\pi = \sigma_1 \circ \cdots \circ \sigma_k$. We have

$$
\begin{aligned}
\mathbb{P}_{\theta,id}(\pi \succeq id) &\\
&\leq \mathbb{P}_{\theta,id}(\exists i \in [k] : \sigma_1 \circ \cdots \circ \sigma_i \succeq \sigma_1 \circ \cdots \circ \sigma_{i-1}) \\
&\leq \#_{trans}(\pi)\left(\cosh\frac{\theta}{2}\right)^{-N},
\end{aligned} \tag{3.12}
$$

where $\#_{trans}(\pi)$ is the number of transpositions in π. For any permutation $\pi \in \mathfrak{S}_n$, $\#_{trans}(\pi) = n - \#_{cyc}(\pi)$ with $\#_{cyc}(\pi)$ the number of cycles in π. The upper bound (3.3) follows from (3.12), the union bound and the fact that $\sum_{\pi \in \mathfrak{S}_n}\#_{cyc}(\pi) = n!H_n$.

Proof of Proposition 3.4. A similar argument as before shows that

$$
\mathbb{P}((j,k) \succeq id) = \mathbb{P}((1,2) \succeq id) \quad \text{for } k-j = 1.
$$

Now let $\ell := k - j \in \{1, \ldots, n-1\}$. It is not hard to see that for any permutation $\pi \in \mathfrak{S}_n$, the number of inversions $\mathrm{inv}(\pi \circ (j,k))$ can take 2ℓ values:

$$
\mathrm{inv}(\pi \circ (j,k)) \in \{\mathrm{inv}(\pi) \pm m : m = 1, 3, \ldots, 2\ell - 1)\}.
$$

For $m \in \{-2\ell + 1, -2\ell + 3, \ldots, 2\ell - 3, 2\ell - 1\}$, let

$$
p_m := \mathbb{P}_{\theta,id}\left(\mathrm{inv}(\pi \circ (j,k)) = \mathrm{inv}(\pi) + m\right).
$$

Observe that $\mathbb{P}_{\theta,id}((j,k) \succeq id)$ is given by

$$
\begin{aligned}
&\mathbb{P}_{\theta,id}\left(\sum_{i=1}^{N}\left(\mathrm{inv}(\pi_i \circ (j,k)) - \mathrm{inv}(\pi_i)\right) \leq 0\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{N}Z_i \leq 0\right),
\end{aligned} \tag{3.13}
$$

where Z_i are i.i.d. categorical random variables such that $Z_i = m$ with probability p_m for $m \in \{-2\ell + 1, -2\ell + 3, \ldots, 2\ell - 3, 2\ell - 1\}$. By Cramer's theorem (Dembo & Zeitouni, 2010),

$$\mathbb{P}\left(\sum_{i=1}^{N} Z_i \leq 0\right) \leq \exp\left(-N \sup_{\lambda}\{-\log F(\lambda)\}\right),$$
(3.14)

where

$$F(\lambda) := \sum_{m=-2\ell+1,\, m \text{ odd}}^{2\ell-1} p_m e^{\lambda m},$$

is the moment generating function of Z_i. Note that for $m > 0$, we have $p_{-m} = e^{\theta m} p_m$. Therefore, $\sum_{m=1,\, m \text{ odd}}^{2\ell-1} p_m(1 + e^{\theta m}) = 1$. Moreover,

$$\sum_{m=1,\, m \text{ odd}}^{2\ell-1} p_m(1 + e^{\theta m}) = \sum_{m=1,\, m \text{ odd}}^{2\ell-1} p_m e^{\frac{\theta m}{2}}(e^{\frac{\theta m}{2}} + e^{-\frac{\theta m}{2}})$$

$$\geq \sum_{m=1,\, m \text{ odd}}^{2\ell-1} p_m e^{\frac{\theta m}{2}}(e^{\frac{\theta}{2}} + e^{-\frac{\theta}{2}}),$$

which yields $\sum_{m=1,\, m \text{ odd}}^{2\ell-1} p_m e^{\frac{\theta m}{2}} \leq \frac{1}{2}(\cosh \frac{\theta}{2})^{-1}$. As a consequence,

$$\sup_{\lambda}\{-\log F(\lambda)\} \geq -\log F\left(\frac{\theta}{2}\right)$$

$$= -\log \sum_{m=1,\, m \text{ odd}}^{2\ell-1} 2p_m e^{\frac{\theta m}{2}} \quad (3.15)$$

$$\geq -\log\left(\cosh \frac{\theta}{2}\right)^{-1}.$$

By (3.13), (3.14) and (3.15), we get the estimate (3.11). $\square$

4. Experimental Results

In this section we apply Algorithm 1 to provide experimental results on synthetic data and two real-world data: APA election data (large N, small t_{max}), and University's homepage search (small N, large t_{max}).

4.1. Synthetic Data

We generate 50 sets of $N = 1000$ rankings from the IGM model (1.3) with $\vec{\theta} = (1, 0.9, 0.8, 0.7, 0.6, 0.5, 0, \ldots)$ and $\pi_0 = id$. We restrict all observed rankings to the first $t_{\max} = 6$ components. Table 1 displays the percentage of the model sizes selected by Algorithm 1.

Now we fit the IGM model by MLE with the preselected model size t. The estimate $\widehat{\theta}_1$ lies in the range $[0.94, 1.08]$ with mean 1.0 and standard deviation 0.03, and $\widehat{\theta}_2$ in the range $[0.84, 0.95]$ with mean 0.9 and standard deviation

Table 1. Percentage of model sizes selected for 50 simulated data sets.

MODEL SIZE t	1	2	3	4	5	6
PERCENTAGE (%)	0	65	35	0	0	0

Table 2. Accuracy of estimated rank & average training time for 50 simulated data with $t_{max} = 10$ (resp. $t_{max} = 20$, $t_{max} = 40$) and $\vec{\theta} = (1, 0.975, \ldots, 0.775, 0, \ldots)$ (resp. $\vec{\theta} = (1, 0.975, \ldots, 0.525, 0, \ldots)$, $\vec{\theta} = (1, 0.975, \ldots, 0.025, 0, \ldots)$) by the IGM model of model size $t = 1$, $t = 10$ and Algorithm 1.

$t_{max} = 10$	IGM($t=1$)	IGM($t=10$)	ALG 1
ACC. EST. RANK	**100%**	**100%**	**100%**
AVE. TIME	**1.56 s**	14.45 s	2.80 s
$t_{max} = 20$	IGM($t=1$)	IGM($t=10$)	ALG 1
ACC. EST. RANK	94%	**100%**	**100%**
AVE. TIME	**5.73 s**	54.45 s	24.42 s
$t_{max} = 40$	IGM($t=1$)	IGM($t=10$)	ALG 1
ACC. EST. RANK	82%	**100%**	**100%**
AVE. TIME	**70.26 s**	684.65 s	391.20 s

0.02. Moreover, the estimated central rankings restricted to the top 6 ranks are always $(1|2|3|4|5|6)$.

We also generate 50 sets of $N = 1000$ rankings from the IGM model restricted to the first $t_{\max} = 10, 20$ and 40 components. Table 2 shows the accuracy of estimated ranking and average training time by the IGM model of model size $t = 1$, $t = 10$ and Algorithm 1. For rankings of small length, the IGM model of small size gives good accuracy and uses less training time. Algorithm 1 is more appealing for rankings of large length. It has better accuracy than the IGM model of small size, and is less time-consuming than the IGM model of large size.

4.2. APA Data

We consider the problem of ranking with a small number of items and large sample size. The data consists of $N = 15449$ rankings over $t_{max} = 5$ candidates during the American Psychological Association's presidential election in 1980. Among these rankings, there are only 5738 complete rankings and the number of distinct rankings is $n = 207$. See (Coombs et al., 1984; Diaconis, 1989) for further background.

Table 3 displays the estimate of θ for the single parameter IGM with different model sizes. Applying Algorithm 1 yields the selection of $t = 2$. Then by fitting the IGM model with $t = 2$, we get $\widehat{\theta}_1 = 0.46$, $\widehat{\theta}_2 = 0.54$ and $\widehat{\pi}_0 = (3|1|5|4|2)$. By the second order analysis, (Diaconis, 1989) argued that there is a strong effect of choosing candidates $\{1, 3\}$ and $\{4, 5\}$, with candidate $\{2\}$ in the middle. Our

Table 3. Estimates of θ for the single parameter IGM model with $t \in [1, 5]$.

MODEL SIZE t	1	2	3	4	5
ESTIMATED θ	0.47	0.50	0.54	0.62	0.72

Table 4. A list of 10 universities and their estimated ranks.

UNIVERSITY'S NAME	ESTIMATED RANK
OREGON HEALTH SCIENCES UNIV.	2
OUACHITA BAPTIST UNIV.	1
OUR LADY OF THE LAKE UNIV.	6
PACIFIC BAPTIST UNIV.	2
PACIFIC CHRISTIAN COLLEGE	8
PACIFIC LUTHERAN UNIV.	1
WORCESTER STATE COLLEGE	1
YESHIVA UNIV.	19
YORK COLLEGE OF PENNSYLVANIA	6
YOUNG HARRIS COLLEGE	1

result suggests that the pair of candidates $\{1, 3\}$ be in a more favorable position.

4.3. University's Homepage Search

We consider the problem of learning a domain-specific search engine with the data collected by (Cohen et al., 1998). The data consists of 157 universities, the *queries*, and 21 search engines, the *experts*. Each expert search engine outputs a ranked list of up to $t_{max} = 30$ URLs when queried with the university's name. The target output is the university's homepage. There are 10 universities without data, and some search engines return empty list. So there are 147 ranking problems with sample size $N \leq 21$, and each sample has length ranging from 1 to 30.

For each query, we fit the IGM model by MLE with Algorithm 1 to calculate the rank of the university's homepage under the estimated central ranking $\widehat{\pi}_0$. This rank measures the correctness of the model. If the target homepage is not among the URLs returned by the search engines, we put it to the end of the list. The central ranking is estimated by the SORTR heuristic (Meilă & Bao, 2010). Table 4 is an extract of the estimated rank of the target homepages.

Our training model is slightly different from that in (Meilă & Bao, 2010), since they used the parameterization $\vec{\theta} = (\theta_1, \ldots, \theta_{t-1}, \theta_t, \ldots. \theta_t, 0, \ldots)$ for the top-t IGM model. Figure 2 provides the estimates of θ for the single parameter IGM model with different model sizes. By Algorithm 1, 68% select $t = 5$, 20% select $t = 1$, and 11% select $t = 2$ over all 147 queries. We also computed the rank of the homepage for each query and each model size t. Table 5 summarizes the mean and the median of the target rank for these models.

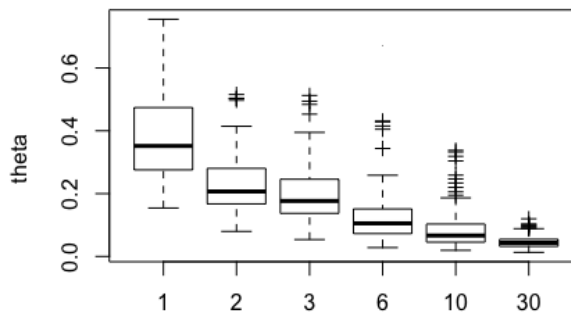(Meilă & Bao, 2010) got a mean rank around 15 and a

Figure 2. Estimates of θ for the single parameter IGM model with $t \in \{1, 2, 3, 6, 10, 30\}$.

Table 5. Mean and median rank of the target homepage under the IGM models.

SIZE t	1	2	3	6	10	30	ALG 1
MEAN R.	3.7	4.2	4.9	8.0	13.2	18.9	6.8
MED. R.	2	2	2	3	5	6	3

median rank around 10. Compared to their results, our experiments give better target rank for small model sizes. This is reasonable because for each query there are only $N \leq 21$ samples and large model sizes may cause overfitting. Small median ranks in Table 5 also supports the validity of the IGM model.

5. Conclusion

In this paper we study various Mallows ranking models. The parameters of interest are the dispersion θ (or $\vec{\theta}$), and the central ranking π_0. Though there have been efficient algorithms to estimate the MLE $(\widehat{\theta}, \widehat{\pi}_0)$, not much is known about the statistical properties of $(\widehat{\theta}, \widehat{\pi}_0)$ except the consistency of $\widehat{\theta}$. Aiming to fill this gap, we prove the biasedness and convergence rate of the MLE of the Mallows' ϕ model and the single parameter IGM model.

To compare a large number of items, an infinite ranking model is often useful. A natural infinite generalization of the finite Mallows model appeared in several contexts (Gnedin & Olshanski, 2010; Meilă & Bao, 2010). But neither of them are aware of the other. In this work we make a clear connection between these infinite rankings with further analysis. Relying on a renewal property of the single parameter IGM model, we propose an algorithm to choose the model size 't' automatically. The 't' selection algorithm is tested over synthetic and real data, and shows good performance.

References

Ailon, N., Charikar, M., and Newman, A. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM*, 55(5):23, 2008.

Arratia, R. and Gordon, L. Tutorial on large deviations for the binomial distribution. *Bulletin of mathematical biology*, 51(1):125–131, 1989.

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., Tovey, C., and Trick, M. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6(2):157–165, 1989.

Bradley, R. and Terry, M. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Busa-Fekete, R., Hüllermeier, E., and Szörényi, B. Preference-based rank elicitation using statistical models: The case of Mallows. In *Proceedings of The 31st International Conference on Machine Learning*, volume 32, 2014.

Busse, L., Orbanz, P., and Buhmann, J. Cluster analysis of heterogeneous rank data. In *International Conference on Machine Learning*, pp. 113–120, 2007.

Cattelan, M. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, pp. 412–433, 2012.

Chen, Y. and Suh, C. Spectral MLE: Top-k rank aggregation from pairwise comparisons. In *International Conference on Machine Learning*, pp. 371–380, 2015.

Chierichetti, F., Dasgupta, A., Haddadan, S., Kumar, R., and Lattanzi, S. Mallows models for top-k lists. In *Advances in Neural Information Processing Systems 31*, pp. 4387–4397. 2018.

Choirat, C. and Seri, R. Estimation in discrete parameter models. *Statistical Science*, 27(2):278–293, 2012.

Cohen, W., Schapire, R., and Singer, Y. Learning to order things. In *Advances in Neural Information Processing Systems*, pp. 451–457, 1998.

Coombs, C., Cohen, J., and Chamberlin, J. An empirical study of some election systems. *American Psychologist*, 39(2):140, 1984.

Critchlow, D. *Metric methods for analyzing partially ranked data*, volume 34. Lecture notes in Statistics, Springer, 1985.

Critchlow, D., Fligner, M., and Verducci, J. Probability models on rankings. *Journal of mathematical psychology*, 35(3):294–318, 1991.

Dembo, A. and Zeitouni, O. *Large deviations techniques and applications.* Springer-Verlag, Berlin, 2nd edition, 2010.

Diaconis, P. *Group representations in probability and statistics*, volume 11. Lecture Notes-Monograph Series, 1988.

Diaconis, P. A generalization of spectral analysis with application to ranked data. *Ann. Stat.*, pp. 949–979, 1989.

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

Fligner, M. and Verducci, J. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 359–369, 1986.

Fligner, M. and Verducci, J. Multistage ranking models. *Journal of the American Statistical association*, 83(403):892–901, 1988.

Gnedin, A. and Olshanski, G. q-exchangeability via quasi-invariance. *Ann. Probab.*, 38(6):2103–2135, 2010.

Hoppen, C., Kohayakawa, Y., Moreira, C. G., Ráth, B., and Menezes Sampaio, R. Limits of permutation sequences. *Journal of Combinatorial Theory Series B*, 103(1):93–113, 2013.

Huang, J. and Guestrin, C. Uncovering the riffled independence structure of ranked data. *Electronic Journal of Statistics*, 6:199–230, 2012.

Huang, J., Kapoor, A., and Guestrin, C. Riffled independence for efficient inference with partial rankings. *Journal of Artificial Intelligence Research*, 44:491–532, 2012.

Hung, K. and Fithian, W. Rank verification for exponential families. *Ann. Stat.*, 47(2):758–782, 2019.

Hunter, D. MM algorithms for generalized Bradley-Terry models. *Ann. Stat.*, 32(1):384–406, 2004.

Irurozki, E., Calvo, B., and Lozano, J. PerMallows: An R package for Mallows and generalized Mallows models. *Journal of Statistical Software*, 71(i12), 2016.

Irurozki, E., Calvo, B., and Lozano, J. Sampling and learning Mallows and generalized Mallows models under the Cayley distance. *Methodology and Computing in Applied Probability*, 20(1):1–35, 2018.

Kemp, A. Absorption sampling and the absorption distribution. *J. Appl. Probab.*, 35(2):489–494, 1998. ISSN 0021-9002.

Lebanon, G. and Lafferty, J. Cranking: Combining rankings using conditional probability models on permutations. In *International Conference on Machine Learning*, volume 2, pp. 363–370, 2002.

Lebanon, G. and Mao, Y. Non-parametric modeling of partially ranked data. *Journal of Machine Learning Research*, 9:2401–2429, 2008.

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, 2018.

Lu, T. and Boutilier, C. Learning Mallows models with pairwise preferences. In *International Conference on Machine Learning*, pp. 145–152, 2011.

Lu, T. and Boutilier, C. Effective sampling and learning for Mallows models with pairwise-preference data. *The Journal of Machine Learning Research*, 15(1):3783–3829, 2014.

Luce, R. D. *Individual Choice Behavior a Theoretical Analysis*. John Wiley & Sons Inc., 1959.

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

Mandhani, B. and Meilă, M. Tractable search for learning exponential models of rankings. In *International Conference on Artificial Intelligence and Statistics*, pp. 392–399, 2009.

Marden, J. *Analyzing and modeling rank data*. Chapman and Hall/CRC, 2014.

Meek, C. and Meila, M. Recursive inversion models for permutations. In *Advances in Neural Information Processing Systems*, pp. 631–639, 2014.

Meilă, M. and Bao, L. An exponential family model over infinite rankings. *Journal of Machine Learning Research*, 10:3481–3518, 2010.

Meilă, M. and Chen, H. Dirichlet process mixtures of generalized Mallows models, 2012. arXiv:1203.3496.

Meilă, M., Phadnis, K., Patterson, A., and Bilmes, J. Consensus ranking under the exponential model. *arXiv:1206.5265*, 2012.

Mukherjee, S. Estimation in exponential families on permutations. *Ann. Stat.*, 44(2):853–875, 2016.

Newey, W. and McFadden, D. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4: 2111–2245, 1994.

Pitman, J. and Tang, W. Regenerative random permutations of integers. *arXiv:1704.01166. To appear Ann. Probab.*, 2017.

Plackett, R. The analysis of permutations. *Applied Statistics*, pp. 193–202, 1975.

Rawlings, D. Absorption processes: models for q-identities. *Adv. in Appl. Math.*, 18(2):133–148, 1997. ISSN 0196-8858. doi: 10.1006/aama.1996.0504.

Shah, N. and Wainwright, M. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

Stanley, R. P. *Enumerative Combinatorics: Volume 1*. Cambridge University Press, 2nd edition, 2011.

Thurstone, L. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

Thurstone, L. Rank order as a psycho-physical method. *Journal of Experimental Psychology*, 14(3):187, 1931.

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: 158–1, 2017.

Young, H. P. Optimal ranking and choice from pairwise comparisons. *Information pooling and group decision making*, pp. 113–122, 1986.