



## An Extended Mallows Model for Ranked Data Aggregation

Han Li, Minxuan Xu, Jun S. Liu & Xiaodan Fan

To cite this article: Han Li, Minxuan Xu, Jun S. Liu & Xiaodan Fan (2020) An Extended Mallows Model for Ranked Data Aggregation, Journal of the American Statistical Association, 115:530, 730-746, DOI: [10.1080/01621459.2019.1573733](https://doi.org/10.1080/01621459.2019.1573733)

To link to this article: <https://doi.org/10.1080/01621459.2019.1573733>



View supplementary material [↗](#)



Published online: 23 Apr 2019.



Submit your article to this journal [↗](#)



Article views: 993



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 4 View citing articles [↗](#)



# An Extended Mallows Model for Ranked Data Aggregation

Han Li<sup>a,b</sup>, Minxuan Xu<sup>b,c</sup>, Jun S. Liu<sup>d</sup>, and Xiaodan Fan<sup>b</sup>

<sup>a</sup>College of Economics, Shenzhen University, Shenzhen, China; <sup>b</sup>Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong; <sup>c</sup>Department of Statistics, University of California, Los Angeles, CA; <sup>d</sup>Department of Statistics, Harvard University, Cambridge, MA

## ABSTRACT

In this article, we study the rank aggregation problem, which aims to find a consensus ranking by aggregating multiple ranking lists. To address the problem probabilistically, we formulate an elaborate ranking model for full and partial rankings by generalizing the Mallows model. Our model assumes that the ranked data are generated through a multistage ranking process that is explicitly governed by parameters that measure the overall quality and stability of the process. The new model is quite flexible and has a closed form expression. Under mild conditions, we can derive a few useful theoretical properties of the model. Furthermore, we propose an efficient statistic called rank coefficient to detect over-correlated rankings and a hierarchical ranking model to fit the data. Through extensive simulation studies and real applications, we evaluate the merits of our models and demonstrate that they outperform the state-of-the-art methods in diverse scenarios. Supplementary materials for this article are available online.

## ARTICLE HISTORY

Received January 2017  
Accepted January 2019

## KEYWORDS

Consensus ranking; Kendall tau distance; Rank aggregation; Rank coefficient.

## 1. Introduction

Many social or scientific problems, such as those considered in political election (de Borda 1781), horse racing (Ali 1998), and oncogene studies (Welsh et al. 2001; True et al. 2006), can be formulated as a ranking of candidates. Given the same test, different graders may come up with different rankings of the candidates. Rank aggregation aims to get a better ranking by aggregating multiple ranking lists.



A simple solution for rank aggregation is to give a final ranking based on the average rank of the candidates, which is also known as the Borda count (de Borda 1781). In addition, Fagin, Kumar, and Sivakumar (2003) suggested using the median rank. These methods perform well when dealing with multiple full ranking lists of similar quality. However, they cannot handle the partial rankings easily (Dwork et al. 2001; Deng et al. 2014).

Instead of using simple statistics, a more principled method for rank aggregation is to use probabilistic models. There has been a substantial amount of work conducted on probabilistic ranking models, including the Thurstone model (Thurstone 1927), the Babington-Smith model (Smith 1950), the Mallows model (Mallows 1957) and its generalizations (Fligner and Verducci 1986), and the Plackett–Luce model (Luce 1959; Plackett 1975); see Critchlow, Fligner, and Verducci (1991) for more details about these models. The distance-based models, such as the Mallows model, have the advantage of being simple and elegant. They assume a central ranking  $\pi_0$ , and the probability of observing a ranking  $\pi$  is inversely proportional to its distance to  $\pi_0$ . Typical examples of distance are the Spearman's footrule distance (Diaconis and Graham 1977), the Kendall tau distance (Kendall 1970), and the Cayley's distance (Diaconis and Graham 1977).


Rank aggregation has also drawn extensive interest in the machine learning field. The classic work by Dwork et al. (2001) proposed Markov chain-based methods, which first construct a transition matrix, whose transition probability measures the pairwise preference, and then rank each item based on the magnitude of its stationary probability. The methods of Dwork et al. (2001) and their variants (DeConde et al. 2006; Lin 2010) have been shown to perform satisfactorily for handling partial rankings and “spam” rankings.

In practice, the problem of rank aggregation can be quite challenging. First, the rankings may not be equally reliable. For example, in a bioinformatics study, some labs can collect and analyze data more efficiently than others. Although Liu et al. (2007), Lin and Ding (2009), Deng et al. (2014), among many others, have proposed weighted rank aggregation approaches, most of them have assigned the weights either in an ad hoc fashion or based on some training data which are often unavailable. Kolde et al. (2012) studied the distribution of the ranks of each item and assigned it a  $p$ -value that described how much better it was ranked than expected by chance. Finally, they ranked all of the items based on their  $p$ -values. The model proposed by Deng et al. (2014) incorporates parameters to measure the reliability of rankings, but it infers only a subset of items to be informative for ranking, without giving their final rank, which is of primary interest in most cases. Badgeley, Sealton, and Chikina (2015) proposed to use Bayes factors to measure the quality of the rankings for genomic dataset aggregation.

Second, a noticeable feature of multiple rankings is that although they may agree with each other for highly ranked items, they may be quite different among lowly ranked items, as the information degrades (Lin and Ding 2009; Hall and Schimek 2012). For instance, in the differential gene expression analysis

**CONTACT** Xiaodan Fan  [xfan@sta.cuhk.edu.hk](mailto:xfan@sta.cuhk.edu.hk)  Lady Shaw Building, Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/r/JASA](http://www.tandfonline.com/r/JASA).

 Supplementary materials for this article are available online. Please go to [www.tandfonline.com/r/JASA](http://www.tandfonline.com/r/JASA).

© 2019 American Statistical Association

in a cancer study, we observe that a few genes are repeatedly highly ranked in different experiments, but the ranks of the remaining genes vary greatly due to noise. Hall and Schimek (2012) proposed a simple approach based on Bernoulli tests to estimate the number of top ranked items that are informative for ranking. Current rank aggregation approaches often treat the disagreement of rankings at different positions equally. In many cases, we are more concerned with the highly ranked items than lowly ranked ones. In practice, according to the final ranking of the genes, researchers will conduct costly follow-up experiments to validate the association of top ranked genes with the disease. This leads to another challenge in ranking modeling: how to characterize the stability of ranking and ensure the ranking accuracy of highly ranked items. Last but not least, we also observe that there may be some “outliers” among the top ranked items. Conventional methods are not robust for handling such outliers.

To address the aforementioned issues, it is desirable to build a robust statistical model that incorporates parameters to measure the quality and stability of the ranking. We consider generalizing the Mallows model by exploiting its stagewise construction, which endows it with nice analytical properties and an efficient implementation. Some authors have extended the Mallows model (Fligner and Verducci 1986) to a more flexible framework, but the model has its own deficiency and cannot address all of the aforementioned problems.

The remainder of this article is organized as follows. We review the Mallows model in Section 2. In Section 3, we develop our ranking model and derive its theoretical properties. In addition, we design a statistic to detect over-correlated rankings and extend our model to a hierarchical ranking model when the independence assumption of rankings is invalid. The inference procedures for the parameters are also presented. In Section 4, we evaluate the performance of our methods via extensive simulation studies. Subsequently, we use our models to analyze prostate cancer studies in Section 5. In Section 6, we show the computation efficiency of our algorithms. We conclude the article with a brief discussion.

## 2. The Mallows Model

Let  $U$  be a set of  $n$  items to be ranked and let a ranking list of the items be  $\mathbf{O} = (y_1 < y_2 < \dots < y_k)$ , where  $y_i \in U, i = 1, \dots, k$ . Here, we assume no ties in the ranking. When  $k = n$ , it is called a full ranking; otherwise, it is called a partial ranking, or more specifically a top- $k$  ranking. Given  $\mathbf{O}$ , we denote  $\pi(y_i)$  as the rank of item  $y_i$ , and  $\pi^{-1}(i)$  as the item assigned rank  $i$ . Here, a smaller rank indicates more preference for the item. Let  $\boldsymbol{\pi}$  and  $\boldsymbol{\pi}^{-1}$  be vectors whose  $i$ th element are  $\pi(y_i)$  and  $\pi^{-1}(i)$ , respectively, and let  $\boldsymbol{\pi}, \boldsymbol{\pi}^k$  denote a full ranking and a top- $k$  ranking, respectively.

The Mallows model was motivated by the Smith (1950) model for paired comparisons. For a full ranking, it can be formulated as follows

$$p(\boldsymbol{\pi}|\boldsymbol{\pi}_0, \phi) = \frac{1}{Z(\phi)} \phi^{d(\boldsymbol{\pi}, \boldsymbol{\pi}_0)}, \quad 0 < \phi < 1,$$

where  $\boldsymbol{\pi}_0$  is the true ranking,  $\phi$  is a dispersion parameter,  $d(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$  is a distance measure between  $\boldsymbol{\pi}$  and  $\boldsymbol{\pi}_0$ , and  $Z(\phi)$

is the normalizing constant. A smaller  $\phi$  indicates that the distribution of rankings is more concentrated around  $\boldsymbol{\pi}_0$ . When  $\phi$  approaches 1, the ranking will become uniformly random. Diaconis (1988) suggested setting  $d(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = d_K(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$ , where  $d_K(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$  is the Kendall tau distance that counts the number of pairwise order disagreements between two rankings, that is,  $d_K(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n I((\pi(y_i) - \pi(y_j))(\pi_0(y_i) - \pi_0(y_j)) < 0)$ , with  $I(\cdot)$  being an indicator function.  $d_K(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$  is also the minimum number of adjacent transpositions that turns  $\boldsymbol{\pi}$  into  $\boldsymbol{\pi}_0$  (Diaconis 1988). Here,  $Z(\phi)$  has a closed form expression with  $Z(\phi) = (1 + \phi)(1 + \phi + \phi^2) \cdots (1 + \phi + \phi^2 + \dots + \phi^{n-1})$ .

The data-generating process of the Mallows model can be described as a stagewise process in which the ranker selects items into  $\boldsymbol{\pi}$  one by one according to decreasing personal preference. Define  $V_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0) \equiv \sum_{j=i+1}^n I(\pi_0(\pi^{-1}(j)) - \pi_0(\pi^{-1}(i)) < 0)$ ,  $i = 1, \dots, n-1$ , which is the number of order disagreement between the  $i$ th item, that is,  $\pi^{-1}(i)$ , and the remaining  $n-i$  items in  $\boldsymbol{\pi}$ . At the first stage, we select item  $\pi^{-1}(1)$ . If the  $(v+1)$ th preferred item in  $\boldsymbol{\pi}_0$  is selected, we have  $V_1(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = v$  which is the number of order disagreement when comparing  $\pi^{-1}(1)$  with the other  $n-1$  items. Thus,  $V_1(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = 0$  indicates that the ranker correctly chose the most preferable item in  $\boldsymbol{\pi}_0$  as the first item in  $\boldsymbol{\pi}$ . The Mallows model indicates that  $V_1(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$  follows the distribution

$$p(V_1(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = v) \propto \phi^v, \quad 0 < \phi < 1, v = 0, 1, \dots, n-1,$$

which is a truncated geometric distribution with finite support  $\{0, \dots, n-1\}$ . We denote it as  $V_1(\boldsymbol{\pi}, \boldsymbol{\pi}_0) \sim G(\phi; 0, n-1)$ . Similarly, at stage  $i$ , we select  $\pi^{-1}(i)$  within the remaining  $n-i+1$  items. We have  $V_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = v$  if the  $(v+1)$ th best item among the remaining is selected and assume  $V_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0) \sim G(\phi; 0, n-i)$ . Note that those  $V_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$ s are independent, and satisfy  $d_K(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_{i=1}^{n-1} V_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$ . As the Mallows model includes only one parameter  $\phi$ , it is also called the Mallows  $\phi$  model.

## 3. An Extended Mallows Model for Rank Aggregation

### 3.1. An Extended Mallows Model

To allow for more flexibility beyond the Mallows model, one can set different  $\phi$  for different rank positions, as the Mallows  $\phi$ -component model (Fligner and Verducci 1986) does. The Mallows  $\phi$ -component model has  $n-1$  free parameters. However, if we have only a few ranking lists, there is no sufficient amount of data to fit the model with so many parameters. For most real examples we have seen, it is desirable to design an extended Mallows model that has only a few parameters.

To modify the Mallows model, we examine its stagewise independent ranking process. At stage  $i$ ,  $p(V_i(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = v) \propto \phi^v, 0 < \phi < 1$ . First, we consider using a weight function  $\alpha(i)$  to measure the stability of the  $i$ th position in a ranking, where  $\alpha(i)$  is assumed to be a non-increasing function of  $i$ , as we assume the top ranked items are more stable. Denote  $\phi_i \equiv \phi(1 - \alpha(i))$ ; thus, it is a nondecreasing function of  $i$ , which indicates that the probability of making mistake or disagreement tends to increase as the rank increases. Based on our empirical evidence,  $\alpha(i) = \alpha^i, 0 \leq \alpha \leq 1$ , is a good choice of the weight function. Figure 1

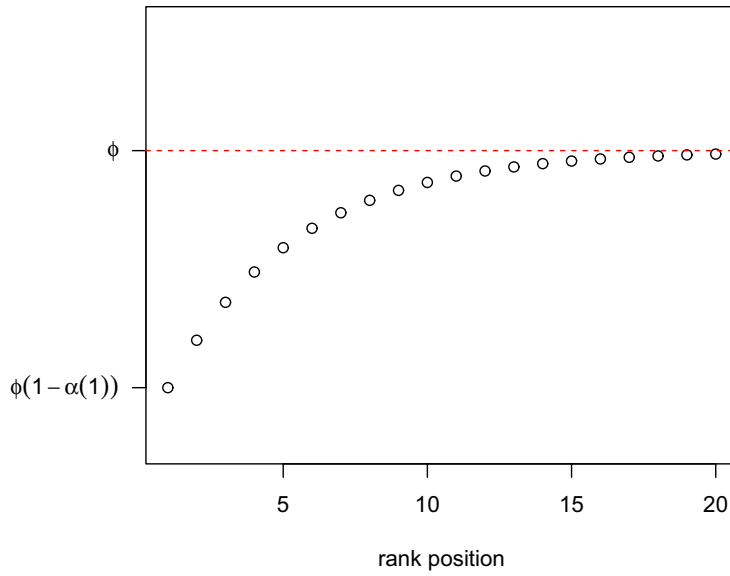


Figure 1. Plot of the weight function  $\phi_i \equiv \phi(1 - \alpha(i)) = \phi(1 - \alpha^i)$ .

shows such an example of  $\phi_i$  by setting  $\phi = 0.7$  and  $\alpha = 0.8$ , that is,  $\phi_i = 0.7(1 - 0.8^i)$ . The following toy example may help illustrate the idea of this specific weight function. Suppose three customers of an ice cream store are asked to guess the top-3 most popular flavors among five flavors (vanilla, chocolate, mango, green tea, strawberry). Those five flavors, ranked as listed above according to a large scale survey, is regarded as the consensus ranking. Given the consensus ranking and  $(\phi_1, \phi_2, \phi_3) = (0.14, 0.25, 0.34)$ , the three customers independently make their rankings. Their ranking lists are (vanilla < chocolate < green tea), (vanilla < mango < chocolate), and (vanilla < chocolate < strawberry), respectively. We can see that they consistently rank vanilla at the top, and rank chocolate either at the second or the third position. Nevertheless their last-ranked flavors are quite different.

Second, to address the outlier problem, we assume that  $\pi^{-1}(i)$  may be drawn at random from the remaining items. More specifically, we assume the following mixture model for  $V_i(\pi, \pi_0)$

$$p(V_i(\pi, \pi_0) = v) = \omega \frac{\phi_i^v}{Z(\phi_i, i)} + (1 - \omega) \frac{1}{n - i + 1}, \quad 0 < \omega \leq 1, \quad (1)$$

where  $\phi_i \equiv \phi(1 - \alpha^i)$  and  $Z(\phi_i, i) = 1 + \phi_i + \dots + \phi_i^{n-i} = \frac{1 - \phi_i^{n-i+1}}{1 - \phi_i}$  is the normalizing constant. Thus, we add a uniform component to the original Mallows model at each stage.

For a partial ranking  $\pi^k$ , we assume its generative process is the same as that of a full ranking, except that the process stops at stage  $k$ . At stage  $i$ , we calculate  $V_i(\pi^k, \pi_0) = \sum_{\{j: \pi^k(y_j) > \pi^k(y_i)\}} I(\pi_0(y_j) - \pi_0(y_i) < 0)$ . Consequently we extend the Mallows model to a ranking model with three parameters  $(\phi, \alpha, \omega)$  for full and partial rankings as

follows

$$\begin{aligned} p(\pi^k | \pi_0, \phi, \alpha, \omega) &= \prod_{i=1}^{\min(k, n-1)} p(V_i(\pi^k, \pi_0) = v_i | \phi, \alpha, \omega) \\ &= \prod_{i=1}^{\min(k, n-1)} \left\{ \omega \frac{[\phi(1 - \alpha^i)]^{v_i}}{Z(\phi(1 - \alpha^i), i)} \right. \\ &\quad \left. + (1 - \omega) \frac{1}{n - i + 1} \right\}, \\ 0 < \phi < 1, 0 \leq \alpha \leq 1, 0 < \omega \leq 1. \quad (2) \end{aligned}$$

Next we show how to connect  $(\phi, \alpha, \omega)$  to the quality of a ranking. Given  $\phi_i = \phi(1 - \alpha^i)$ ,  $\phi$  actually acts as an upper bound of the dispersion parameter of the model. Note that the range of the dispersion parameters in this scenario is  $\phi - \phi_1 = \phi - \phi(1 - \alpha) = \phi\alpha$ . Thus, given  $\phi, \alpha$  controls the stability of the ranking process, with a smaller  $\alpha$  indicating better stability. As for  $\omega$ , it deals with the outlier problem. A larger  $\omega$  implies that the ranking has fewer outliers. Thus,  $(\phi, \alpha, \omega)$  together affect the deviation of a ranking to its consensus ranking.

In addition, note that our model includes two interesting models as special cases. When  $\omega = 1$  and  $\alpha = 0$ , it reduces to the Mallows model. When  $\alpha = 1$ , we have  $\phi_1 = \phi_2 = \dots = \phi_{n-1} = 0$ . In this case, the extended Mallows model (EMM) has a very simple form. It is only when  $v_i = 0$  that the component  $\frac{\phi_i^{v_i}}{Z(\phi_i, i)}$  has positive probability and equals 1. Hence, at stage  $i$ ,  $V_i(\pi, \pi_0) = 0$  has probability  $\omega + (1 - \omega) \frac{1}{n - i + 1}$ , and  $V_i(\pi, \pi_0) = j$  has equal probability  $(1 - \omega) \frac{1}{n - i + 1}$  for all  $j > 0$ . This model was proposed by Fligner and Verducci (1986), who set  $p(V_i(\pi, \pi_0) = v_i) = \frac{\phi_i^{I(v_i > 0)}}{1 + (n - i)\phi_i}$ . Here, we reinvent the model in a different framework.

As our model is composed of independent ranking stages, we can easily derive the expectation and variance of  $d_K(\pi, \pi_0) = \sum_{i=1}^{n-1} V_i(\pi, \pi_0)$  as follows. Similarly, we can derive the expec-

tation and variance of  $d_K(\pi^k, \pi_0)$

$$\begin{aligned} E[d_K(\pi, \pi_0)] &= \sum_{i=1}^{n-1} E[V_i(\pi, \pi_0)] \\ &= \sum_{i=1}^{n-1} \left\{ \omega \left[ \frac{\phi_i}{1-\phi_i} - \frac{(n-i+1)\phi_i^{n-i+1}}{1-\phi_i^{n-i+1}} \right] \right. \\ &\quad \left. + (1-\omega) \frac{n-i}{2} \right\}, \end{aligned} \quad (3)$$

$$\begin{aligned} \text{var}[d_K(\pi, \pi_0)] &= \sum_{i=1}^{n-1} \{E[V_i^2(\pi, \pi_0)] - (E[V_i(\pi, \pi_0)])^2\} \\ &= \sum_{i=1}^{n-1} \left\{ \omega \left[ \frac{\phi_i + \phi_i^2 - 2\phi_i^{n-i+1}}{(1-\phi_i)^2(1-\phi_i^{n-i+1})} \right. \right. \\ &\quad \left. \left. - \frac{[(n-i+1)^2 - 2]\phi_i^{n-i+1} - (n-i)^2\phi_i^{n-i+2}}{(1-\phi_i)(1-\phi_i^{n-i+1})} \right] \right. \\ &\quad \left. + (1-\omega) \frac{(n-i)[2(n-i)+1]}{6} \right\} \\ &\quad - \sum_{i=1}^{n-1} (E[V_i(\pi, \pi_0)])^2. \end{aligned} \quad (4)$$

### 3.2. Theoretical Properties of the New Model

In this section, we show that the EMM, being more flexible, still retains some salient properties of the Mallows model as the following theorems state. Fligner and Verducci (1988) defined a stagewise ranking model as a strongly unimodal model if at each stage  $i$ ,  $p(V_i(\pi, \pi_0) = 0) > p(V_i(\pi, \pi_0) = 1)$ , and  $p(V_i(\pi, \pi_0) = v)$  is a nonincreasing function of  $v$ . By checking Equation (1), we can easily obtain the following theorem.

**Theorem 3.1.** EMM is a strongly unimodal ranking model.

The preceding theorem indicates that EMM has mode  $\pi_0$  and no other maximum. In EMM,  $p(V_i(\pi, \pi_0) = v)$  is a strictly decreasing function of  $v$ . This leads to Theorem 3.2, which states that EMM has the consensus property as defined by Fligner and Verducci (1988). Let  $\tau_{ij}$  be the permutation defined by  $\tau_{ij}(i) = j$ ,  $\tau_{ij}(j) = i$ , and  $\tau_{ij}(g) = g$ , for  $g \neq i, j$ . The consensus property requires that if we transpose  $\pi_0^{-1}(i)$  and  $\pi_0^{-1}(i+1)$  in  $\pi$  (denoting the resulting ranking as  $\pi \circ \tau_{i,i+1}$ ), where originally  $\pi(\pi_0^{-1}(i)) < \pi(\pi_0^{-1}(i+1))$ , we shall have  $p(\pi) \geq p(\pi \circ \tau_{i,i+1})$  with strict inequality for some  $\pi$ , for any  $i = 1, \dots, n-1$ . This property further leads us to expect that the rank of the items is consistent with their true rank as Theorem 3.3 claims.

**Theorem 3.2.** EMM has consensus  $\pi_0$ .

**Proof.** Let  $\tilde{\pi}$  be a general ranking that is either a full ranking or a partial ranking. Denote  $V(\tilde{\pi}, \pi_0) = (V_1(\tilde{\pi}, \pi_0), V_2(\tilde{\pi}, \pi_0), \dots, V_{\min\{k, n-1\}}(\tilde{\pi}, \pi_0))$  and let  $V(\tilde{\pi}, \pi_0)(\mathcal{A})$  be the elements of  $V(\tilde{\pi}, \pi_0)$  with index set

$\mathcal{A}$ ,  $\mathcal{A} \subseteq \{1, \dots, n-1\}$ . Assume  $\tilde{\pi}(\pi_0^{-1}(i)) = a \leq \min(k, n)$ ; otherwise,  $\tilde{\pi} = \tilde{\pi} \circ \tau_{i,i+1}$ . In this case,  $V(\tilde{\pi}, \pi_0)$  differs from  $V(\tilde{\pi} \circ \tau_{i,i+1}, \pi_0)$  only in the  $a$ th component, and

$$\begin{aligned} \frac{p(\tilde{\pi})}{p(\tilde{\pi} \circ \tau_{i,i+1})} &= \frac{p(V_a(\tilde{\pi}, \pi_0) = v)}{p(V_a(\tilde{\pi} \circ \tau_{i,i+1}, \pi_0) = v+1)} \\ &= \frac{\omega \frac{\phi_a^v}{Z(\phi_a, a)} + (1-\omega) \frac{1}{n-a+1}}{\omega \frac{\phi_a^{v+1}}{Z(\phi_a, a)} + (1-\omega) \frac{1}{n-a+1}} > 1. \end{aligned}$$

□

**Theorem 3.3.** For a top- $k$  ranking list  $\pi^k$ , let  $\pi^k(y_i)$  be the rank of item  $y_i$ . In EMM, if  $\pi_0(y_i) < \pi_0(y_j)$ , then  $E[\pi^k(y_i)] < E[\pi^k(y_j)]$ . As a result, given  $\pi_0(y_1) < \pi_0(y_2) < \dots < \pi_0(y_n)$ , we have  $E[\pi^k(y_1)] < E[\pi^k(y_2)] < \dots < E[\pi^k(y_n)]$ .

**Proof.** Without loss of generality, we set  $U = \{1, 2, \dots, n\}$  and  $\pi_0 = (1, 2, \dots, n)$ . Assume the rank of missing item is  $\lambda$ , with  $\lambda > k$ . We first show that for any  $i = 1, 2, \dots, n-1$ ,  $E[\pi^k(i)] < E[\pi^k(i+1)]$ . Let  $j = i+1$ . Given  $p(\pi^k) \geq p(\pi^k \circ \tau_{i,i+1})$ , by assuming  $\pi^k(i) < \pi^k(j)$ , we have  $\pi^k(i)p(\pi^k) + \pi^k(j)p(\pi^k \circ \tau_{ij}) < \pi^k(i)p(\pi^k \circ \tau_{ij}) + \pi^k(j)p(\pi^k)$ .

$$\begin{aligned} E[\pi^k(i)] &= \sum_{\pi^k(i) < \pi^k(j) \leq k} [\pi^k(i)p(\pi^k) + \pi^k(j)p(\pi^k \circ \tau_{ij})] \\ &\quad + \sum_{\pi^k(i) \leq k < \pi^k(j)} [\pi^k(i)p(\pi^k) + \pi^k(j)p(\pi^k \circ \tau_{ij})] \\ &\quad + \sum_{k < \{\pi^k(i), \pi^k(j)\}} \lambda p(\pi^k) \\ &< \sum_{\pi^k(i) < \pi^k(j) \leq k} [\pi^k(i)p(\pi^k \circ \tau_{ij}) + \pi^k(j)p(\pi^k)] \\ &\quad + \sum_{\pi^k(i) \leq k < \pi^k(j)} [\pi^k(i)p(\pi^k \circ \tau_{ij}) + \pi^k(j)p(\pi^k)] \\ &\quad + \sum_{k < \{\pi^k(i), \pi^k(j)\}} \lambda p(\pi^k) \\ &= E[\pi^k(j)]. \end{aligned}$$

Thus, for any  $i < j$ , we have  $E[\pi^k(i)] < \dots < E[\pi^k(j)]$  and  $E[\pi^k(1)] < E[\pi^k(2)] < \dots < E[\pi^k(k)]$ . The result holds automatically for the full ranking list when we set  $k = n$ . □

**Remark 3.1.** Theorem 3.3 implies that, under EMM, if we sort the items according to the mean of their ranks, it will give a consistent estimator of  $\pi_0$  in both the full and partial ranking cases.

Next, we investigate the relationships of different configurations of  $\pi$ . In the original Mallows model, for configurations  $\pi_1, \pi_2$ , if  $d_K(\pi_1, \pi_0) = d_K(\pi_2, \pi_0)$ , then  $p(\pi_1) = p(\pi_2)$ . However, it may not be such a case in EMM, as it assumes non-increasing penalties for order disagreements as the rank increases. We are interested in the case that given  $d_K(\pi_1, \pi_0) = d_K(\pi_2, \pi_0)$ , there are two mismatches in the elements of  $V(\pi_1, \pi_0)$  and  $V(\pi_2, \pi_0)$ . Under mild conditions, we obtain the following theorem.



**Theorem 3.4.** For two configurations  $\pi_1$  and  $\pi_2$ , assume  $V(\pi_1, \pi_0)(a, b) = (v_a, v_b)$ ,  $V(\pi_2, \pi_0)(a, b) = (v'_a, v'_b)$ ,  $a < b$ , and  $V(\pi_1, \pi_0)(g) = V(\pi_2, \pi_0)(g)$ ,  $g \neq a, b$ . Given  $v_a + v_b = v'_a + v'_b$  and  $v_a < v'_a$ , we have  $p(\pi_1) \geq p(\pi_2)$  if either of the following conditions holds: (a)  $\omega = 1$ ; (b)  $0 < \omega < 1$ ,  $\frac{\phi_a^{v_a}/Z(\phi_a, a)}{\phi_b^{v_b}/Z(\phi_b, b)} \geq \frac{n-b+1}{n-a+1}$ .

The proof of Theorem 3.4 is given in Appendix A1. Given the condition (a)  $\omega = 1$ , EMM reduces to the Mallows  $\phi$ -component model with  $\phi_1 \leq \phi_2 \leq \dots \leq \phi_{n-1}$ . In this case, it is natural to have  $p(\pi_1) \geq p(\pi_2)$  as we put more penalty on errors in the smaller ranks. Given the condition (b), it is interesting to note that if  $\pi_2^{-1}(a)$  has a larger probability than uniformly selected, which is contrary to  $\pi_2^{-1}(b)$ , that is,  $\frac{\phi_a^{v_a}}{Z(\phi_a, a)} \geq \frac{1}{n-a+1}$ ,  $\frac{\phi_b^{v_b}}{Z(\phi_b, b)} \leq \frac{1}{n-b+1}$ , then let  $v_a = v'_a - i$  and we have  $\frac{\phi_a^{v_a}/Z(\phi_a, a)}{\phi_b^{v_b}/Z(\phi_b, b)} > \frac{\phi_a^{v'_a}/Z(\phi_a, a)}{\phi_b^{v'_b}/Z(\phi_b, b)}$ . Thus, the condition (b) of Theorem 3.4 is satisfied and we obtain  $p(\pi_1) \geq p(\pi_2)$ . Similarly, if we gradually reduce  $v'_a$  by setting  $v_a = v'_a - i$ ,  $v_b = v'_b + i$ ,  $i = 1, \dots, v'_a$ , the probability of the configuration gradually increases. Hence, Theorem 3.4 implies a distinctive property of EMM that it favors the configuration that has fewer mistakes in the highly ranked positions in probability.

### 3.3. A Model for Multiple Ranking Lists and Its Inference

Suppose we have  $m$  ranking lists  $\pi_1, \dots, \pi_m$ , and each is either a full or partial ranking. We assume the lists are conditionally independent and share a common dispersion parameter  $\phi$ . Given  $\phi$ , the  $l$ th ranking has its own reliability measured by  $\omega_l$  and own stability measured by  $\alpha_l$ . Thus, the likelihood function of the whole rankings can be expressed as follows

$$\begin{aligned} p(\pi_1, \dots, \pi_m | \pi_0, \phi, \alpha, \omega) \\ &= \prod_{l=1}^m p(\pi_l | \pi_0, \phi, \alpha, \omega) \\ &= \prod_{l=1}^m \prod_{i=1}^{\min(k_l, n-1)} p(V_i(\pi_l, \pi_0) = v_{li}) \\ &= \prod_{l=1}^m \prod_{i=1}^{\min(k_l, n-1)} \left\{ \omega_l \frac{[\phi(1 - \alpha_l^i)]^{v_{li}}}{Z(\phi(1 - \alpha_l^i), i)} \right. \\ &\quad \left. + (1 - \omega_l) \frac{1}{n - i + 1} \right\}, \end{aligned} \quad (5)$$

where  $\alpha = (\alpha_1, \dots, \alpha_m)$ ,  $\omega = (\omega_1, \dots, \omega_m)$ , and  $k_l$  is the length of the  $l$ th ranking.

Next we propose an expectation conditional maximization (ECM; Meng and Rubin 1993) algorithm to find the MLE of the parameters  $\Psi = (\pi_0, \phi, \alpha, \omega)$ . To simplify the denotations, hereafter we assume the observed ranking lists are full rankings. Note that the ECM algorithm also works in the partial ranking case.

First, we introduce binary latent variables  $\{z_{li}\}$ ,  $i = 1, 2, \dots, n - 1$ ,  $l = 1, 2, \dots, m$ .  $z_{li} = 1$  indicates that the

item at the rank position  $i$  in the  $l$ th ranking is drawn from the distribution  $f_1(v_{li}) = \frac{[\phi(1 - \alpha_l^i)]^{v_{li}}}{Z(\phi(1 - \alpha_l^i), i)}$ , and  $z_{li} = 0$  indicates that it is drawn from the uniform distribution  $f_2(v_{li}) = \frac{1}{n - i + 1}$ . Set  $P(z_{li} = 1) = \omega_l$  and  $P(z_{li} = 0) = 1 - \omega_l$ . Then we apply the routine E-step and M-step of the ECM algorithm to find the MLE of parameters. Let  $\Psi^{(t-1)} = (\pi_0^{(t-1)}, \phi^{(t-1)}, \alpha^{(t-1)}, \omega^{(t-1)})$  denote the parameter value at the  $(t - 1)$ th step of the ECM algorithm. Then the following procedure is performed at the  $t$ th step.

1. In the E-step, we calculate the probability  $p(z_{li} = 1 | \Psi^{(t-1)})$ 

$$= \frac{\omega_l^{(t-1)} \frac{[\phi^{(t-1)}(1 - \alpha_l^{(t-1)i})]^{v_{li}}}{Z(\phi^{(t-1)}(1 - \alpha_l^{(t-1)i}), i)}}{\omega_l^{(t-1)} \frac{[\phi^{(t-1)}(1 - \alpha_l^{(t-1)i})]^{v_{li}}}{Z(\phi^{(t-1)}(1 - \alpha_l^{(t-1)i}), i)} + (1 - \omega_l^{(t-1)}) \frac{1}{n - i + 1}}$$
and  $p(z_{li} = 0 | \Psi^{(t-1)}) = 1 - p(z_{li} = 1 | \Psi^{(t-1)})$ , for  $i = 1, \dots, n - 1$ ,  $l = 1, \dots, m$ .
2. In the M-step, we maximize the following function by iteratively optimize  $\phi$ ,  $\alpha$ ,  $\omega$ ,  $\pi_0$  individually conditional on the updated values of the other parameters

$$\begin{aligned} Q(\Psi | \Psi^{(t-1)}) \\ &= \sum_{l=1}^m \sum_{i=1}^{n-1} \left\{ p(z_{li} = 1 | \Psi^{(t-1)}) \log \left( \omega_l \frac{[\phi(1 - \alpha_l^i)]^{v_{li}}}{Z(\phi(1 - \alpha_l^i), i)} \right) \right. \\ &\quad \left. + p(z_{li} = 0 | \Psi^{(t-1)}) \log \left( \frac{1 - \omega_l}{n - i + 1} \right) \right\}. \end{aligned} \quad (6)$$

- (a) For  $\phi$ , we extract the terms involving  $\phi$  in  $Q(\cdot)$ , and the objective function is

$$\begin{aligned} q_1(\phi) &= \sum_{l=1}^m \sum_{i=1}^{n-1} p(z_{li} = 1 | \Psi^{(t-1)}) \\ &\quad \times \left[ v_{li} \log \phi - \log Z(\phi(1 - (\alpha_l^{(t-1)})^i), i) \right], \\ &\quad 0 < \phi < 1. \end{aligned} \quad (7)$$

As shown in Appendix A2,  $q_1(\phi)$  has a unique global maximum for  $\phi \in [a, b] \subset (0, 1)$ . We use the Newton-Raphson method to find the optimal  $\phi$  and denote it as  $\phi^{(t)}$ .

- (b) For each  $\alpha_l$ ,  $l = 1, 2, \dots, m$ , we extract the terms involving  $\alpha_l$  in  $Q(\cdot)$ , and the objective function is

$$\begin{aligned} q_2(\alpha_l) &= \sum_{i=1}^{n-1} p(z_{li} = 1 | \Psi^{(t-1)}) \\ &\quad \times \left[ v_{li} \log(1 - \alpha_l^i) - \log Z(\phi^{(t)}(1 - \alpha_l^i), i) \right], \\ &\quad 0 \leq \alpha_l < 1. \end{aligned} \quad (8)$$

The function  $q_2(\alpha_l)$  does not have the unique global maximum property as  $q_1(\phi)$ , but empirically we find that  $q_2(\alpha_l)$  has one or two local maximum. We also use the Newton-Raphson method to find the optimal  $\alpha_l$  and denote it as  $\alpha_l^{(t)}$ .

- (c) For each  $\omega_l$ ,  $l = 1, 2, \dots, m$ , we extract the terms involving  $\omega_l$  in  $Q(\cdot)$ , and the objective function is

$$\begin{aligned} q_3(\omega_l) &= \sum_{i=1}^{n-1} \left\{ p(z_{li} = 1 | \Psi^{(t-1)}) \log \omega_l \right. \\ &\quad \left. + p(z_{li} = 0 | \Psi^{(t-1)}) \log(1 - \omega_l) \right\}. \end{aligned} \quad (9)$$

The optimal value of  $\omega_l$  is found to be  $\omega_l^{(t)} = \frac{\sum_{i=1}^{n-1} p(z_{li}=1|\Psi^{(t-1)})}{n-1}$ .

- (d) For  $\pi_0$ , we extract the terms involving  $\{v_{li}\}$  in  $Q(\cdot)$ , and the objective function is

$$q_4(\pi_0) = \sum_{l=1}^m \sum_{i=1}^{n-1} p(z_{li} = 1 | \Psi^{(t-1)}) v_{li} \log[\phi^{(t)}(1 - (\alpha_l^{(t)})^i)]. \quad (10)$$

We adopt the local Kemeny approach (Dwork et al. 2001), which finds the optimal value of  $\pi_0$  by iteratively swapping its neighboring items  $(i, i+1)$ ,  $i = 1, \dots, n-1$ , to check whether the objective function can be increased. If it does, the current  $\pi_0$  is updated by swapping these two neighboring items.

We have observed that the likelihood function of EMM may have multiple local modes. These local modes have almost the same  $\pi_0$  together with similar  $\omega$  and  $\alpha$ . To alleviate the local mode problem, we repeat the ECM algorithm with multiple diverse initial values of the parameters and pick the one resulting in the best likelihood value.

Alternatively, one may formulate EMM in the empirical Bayes framework or the full Bayesian framework, and then implement an MCMC algorithm to infer the parameters. The latter approaches usually take much more time but give the uncertainty measure of the parameters. Here we use a computationally less intensive method called parametric bootstrap (Efron and Tibshirani 1994) to infer the uncertainty of the parameters.

### 3.4. Detecting Over-Correlated Rankings and Remedy

In the preceding model, we assume that the observed multiple ranking lists are mutually independent conditional on the consensus ranking and parameters, which may not always be true. For instance, in political elections, voters from a family tend to have very similar candidate preference lists. It is desirable to check the conditional independence assumption and develop a remedy if any group of over-correlated ranking lists is found. Standard correlation measures, such as the Kendall and Spearman correlation, do not work here because any pair of informative rankings is correlated due to their similarity with the consensus ranking  $\pi_0$ . Ideally, detecting groups of over-correlated rankings needs to check the conditional correlation of rankings, conditional on the consensus ranking and parameters. But since the consensus ranking and parameters are unknown at this stage, an intuitive approach is to check whether the similarity of a ranking pair is significantly greater than the similarities of all ranking pairs in the input data. This approach is not ideal, but it may be powerful if most ranking pairs are conditionally independent or if the goal is to detect rankings with significantly higher correlation as compared with most other rankings. Critchlow and Verducci (1992) proposed a straightforward distance-based method to check whether the rankings are over-close to a particular ranking than another one. However, their method is only computationally feasible in the case when there is a small number of items to be ranked, because it has to search the space of all permutations of the items.

Deng et al. (2014) proposed an approach called the coordinate coefficient (CC) to measure the correlation between rankings. For each item, they studied the distribution of its rescaled ranks, that is,  $\{\frac{\pi_1(y_i)}{n+1}, \dots, \frac{\pi_m(y_i)}{n+1}\}$ . They first used a beta distribution to fit the empirical distribution of the rescaled ranks, then used a Monte Carlo method to find the probability between  $\frac{\pi_s(y_i)}{n+1}$  and  $\frac{\pi_l(y_i)}{n+1}$ , considered as the  $p$ -value measuring the closeness of rankings  $s$  and  $l$  for the item  $i$ . These  $p$ -values for all items are aggregated to construct an association measure of the rankings  $s$  and  $l$ . In their framework, the exact ranks of all items are supposed to be known, and thus the framework is not applicable to partial rankings. Besides, Deng et al. (2014)'s model does not take the position information of the item into consideration. Thus, we are motivated to design a new method to measure the correlation between rankings as follows.

Suppose we have a pair of rankings  $\mathbf{O}_1 = (y_1 < y_2 < \dots < y_{k_1})$  and  $\mathbf{O}_2 = (y'_1 < y'_2 < \dots < y'_{k_2})$ , and denote  $\pi_1$  and  $\pi_2$  as their rank vectors, respectively. We construct our statistic, called the rank coefficient (RC), in a stagewise process. For simplicity, we first assume  $k_1 = k_2 = n$ . At stage 1, we define the rescaled  $V_1(\pi_1, \pi_2)$  as  $RV_1(\pi_1, \pi_2) \equiv \frac{V_1(\pi_1, \pi_2)}{n-1}$ , to measure the closeness of  $\pi_1$  and  $\pi_2$  at rank position 1. Actually,  $RV_1(\pi_1, \pi_2)$  is the ratio of rank disagreement of those pairs  $(y_1 < y_2, \dots, y_1 < y_n)$  in  $\pi_1$  when compared with  $\pi_2$ . Note that  $RV_1(\pi_1, \pi_2)$  is not a symmetric function. Thus, we use  $\overline{RV}_1(\pi_1, \pi_2) \equiv \frac{1}{2} \left( \frac{V_1(\pi_1, \pi_2)}{n-1} + \frac{V_1(\pi_2, \pi_1)}{n-1} \right)$  instead. Similarly, at stage  $i$ , we calculate

$$\overline{RV}_i(\pi_1, \pi_2) \equiv \frac{1}{2} \left( \frac{V_i(\pi_1, \pi_2)}{n-i} + \frac{V_i(\pi_2, \pi_1)}{n-i} \right), \quad i = 2, \dots, n-1.$$

Next, we consider the distribution of  $\overline{RV}_i(\pi_s, \pi_l)$  at stage  $i$  for all pairs of rankings. If these rankings are independent conditional on the consensus ranking and follow a common distribution, the cumulative distribution function  $F(x \leq \overline{RV}_i(\pi_s, \pi_l))$  should follow a uniform distribution on  $(0, 1)$ . A simple way to estimate  $F(x \leq \overline{RV}_i(\pi_s, \pi_l))$  is to use the empirical distribution of the  $\overline{RV}_i$ s. That is, we set

$$p_{sl}(i) \equiv \frac{\sum_{q,t} I(\overline{RV}_i(\pi_q, \pi_t) \leq \overline{RV}_i(\pi_s, \pi_l))}{m(m-1)/2},$$

where  $m$  is the number of rankings. If  $p_{sl}(i)$  is very small, then rankings  $s$  and  $l$  are overly similar at the rank position  $i$ , as compared with all paired rankings. The rank coefficient is defined as  $RC(\pi_s, \pi_l) \equiv -\sum_{i=1}^{n-1} \log(p_{sl}(i))$ . Assuming  $p_{sl}(i) \sim \text{Unif}(0, 1)$ , which shall be true if the rankings are homogenous, we have  $RC(\pi_s, \pi_l) \sim \Gamma(n-1, 1)$ . Thus, we can use the probability to the left of  $RC(\pi_s, \pi_l)$  under  $\Gamma(n-1, 1)$  as the  $p$ -value to measure the correlation of the ranking pair  $(\pi_s, \pi_l)$ . Note that RC is aggregating the relative similarity scores  $-\log(p_{sl}(i))$  of the ranking pair, which are evaluated at each stage separately by comparing with the corresponding empirical distribution. For partial rankings, we need to make a small modification of the preceding method. For instance, at stage 1, not all of the pairs  $y_1 < y_2, \dots, y_1 < y_n$  in  $\pi_1$  are comparable given  $\pi_2$ , as some items are missing in  $\pi_2$ . Thus, we count only those pairs whose orders are comparable and set  $RV_1(\pi_1, \pi_2) =$

$\frac{\sum_{j=2}^n I(\pi_2(y_j) - \pi_2(y_1) < 0)}{\sum_{j=2}^n I(\pi_2(y_j) - \pi_2(y_1) < 0) + \sum_{j=2}^n I(\pi_2(y_j) - \pi_2(y_1) > 0)}$ . Similarly, we calculate any  $RV_i(\pi_s, \pi_t)$  for partial ranking.

A group of over-correlated rankings is defined as a set of rankings of a minimal size whose pairwise RC  $p$ -values are all smaller than a threshold, such as 0.05. The minimal group size is set as 3 in the simulation and real data studies. We perform such grouping until no other rankings can be grouped. Here we do not adjust the  $p$ -values for multiple comparison, because the  $p$ -values are used jointly instead of individually to call a group of over-correlated rankings as we require all  $p$ -values within the group shall be small than the threshold. This procedure filters out many false positives and automatically decides the number of groups. Note that RC, a nonparametric clustering method in nature, is not meant to detect groups of any clustering structure, which is too ambitious for one paper to achieve. Rather, RC is more suitable for detecting spherically shaped clusters, a situation similar to the scope of  $k$ -means clustering, which we believe is the most common case. If a joint parametric model can be assumed at this stage, a more principled approach is to use BIC (Schwarz 1978) to determine the number of over-correlated ranking groups and find the groups, as suggested by Murphy and Martin (2003).

If any group of over-correlated rankings is detected, we need to modify our model to alleviate the effect of assumption violation in EMM. Since the most common cluster shape may be a sphere, a natural idea is to use a hierarchical model to characterize this type of extra conditional dependence. Assume we have  $M$  blocks of over-correlated rankings, denoted as  $\{G_1, \dots, G_M\}$ , and the remaining individual rankings  $\pi_1, \dots, \pi_m$ . Assume that given the representative ranking  $\kappa_j$ , the rankings within block  $G_j$  are conditionally independent and follow the Mallows  $\phi$  model,  $j = 1, \dots, M$ . Consequently, we formulate the hierarchical extended Mallows model (HEMM) as follows

$$p(G_1, \dots, G_M, \pi_1, \dots, \pi_m | \pi_0, \kappa_1, \dots, \kappa_M, \phi, \omega, \alpha, \tilde{\omega}, \tilde{\alpha}, \tilde{\phi}) \\ = \prod_{i=1}^M p(\kappa_i | \pi_0, \phi, \tilde{\omega}_i, \tilde{\alpha}_i) \prod_{j=1}^m p(\pi_j | \pi_0, \phi, \omega_j, \alpha_j) \\ \prod_{i=1}^M \prod_{\pi_l \in G_i} p(\pi_l | \kappa_i, \tilde{\phi}_i), \quad (11)$$

where  $\tilde{\omega} = (\tilde{\omega}_1, \dots, \tilde{\omega}_M)$ ,  $\tilde{\alpha} = (\tilde{\alpha}_1, \dots, \tilde{\alpha}_M)$ ,  $\tilde{\phi} = (\tilde{\phi}_1, \dots, \tilde{\phi}_M)$ ,  $\omega = (\omega_1, \dots, \omega_m)$ ,  $\alpha = (\alpha_1, \dots, \alpha_m)$ .  $(\tilde{\omega}, \tilde{\alpha}, \tilde{\phi})$  is the counterpart of  $(\omega, \alpha, \phi)$  to model the  $M$  blocks of over-correlated rankings using EMM. We employ an ECM algorithm similar to that of EMM to find the MLEs of those parameters.

## 4. Simulation Studies

In this section, we conduct simulation studies in various contexts to evaluate the performance of EMM, and compare it with 10 established methods: mean rank method (Mean), median rank method (Median), Kendall optimal aggregation (Kendall), Spearman optimal aggregation (Spearman), Markov chain-based methods (MC1, MC2, MC3; Lin 2010; Schimek et al. 2015), robust rank aggregation (RRA; Kolde et al. 2012), Bayesian iterative robust rank aggregation (BIRRA; Badgeley,

Sealfon, and Chikina 2015), and the Mallows model (MM; Mallows 1957). The following is a brief description of some of these methods.

- Mean and Median: they sort the items according to the mean and median ranks of individual items in an ascending order, respectively.
- Kendall and Spearman: they are to find the optimal ranking that minimizes the sum of Kendall tau distance and Spearman's footrule distance between the aggregated ranking and the individual ranking lists, respectively.
- Markov chain-based methods: the basic idea of these methods is to first construct a transition matrix  $P = \{p_{ij}\}_{i,j \in U}$ , where  $p_{ij}$  denotes the transition probability from item  $i$  to item  $j$ , and then find the stationary distribution of  $P$ , and finally sort the items based on their stationary probabilities in a descending order. MC1, MC2, and MC3 differ in how they set their transition rules, that is, the transition probability  $\{p_{ij}\}$ . In MC1, the Markov chain moves to a state with equal probability if the new state is ranked higher than the current state in some rankings, otherwise stays in the current state. In MC2, the Markov chain makes such movement only if the new state is ranked higher than the current state in at least half of the rankings. In MC3, the transition probability is proportional to the percentage of rankings that rank the new state higher than the current state.
- RRA: it uses a  $p$ -value to measure the deviation of order statistics of the ranks of each item from that of a null model, which assumes the items are sampled from a uniform distribution. The aggregated ranking is obtained by sorting the  $p$ -values of items in an ascending order.
- BIRRA: it starts with an initial aggregated ranking by mean ranks, and regards top ranked items as informative items based on their prior probabilities. The items in each ranking are partitioned into multiple bins. The algorithm iteratively calculates the cumulative bin-wise Bayes factors for each ranking by comparing with the current working standard, and updates the aggregated ranking through Bayes rule, till the algorithm converges.

For the top- $k$  ranking, we set the rank of their missing items as  $k + 1$ . To compare the performance of those methods, we use two statistics. The first is the Kendall tau distance up to rank position  $j$ , defined as  $\sum_{i=1}^j V_i(\pi, \pi_0)$ ,  $j = 1, \dots, n - 1$ . The second is true positive rate up to rank position  $j$ , defined as  $\frac{\sum_{i=1}^j I(y_i \in \Lambda)}{v}$ ,  $j = 1, \dots, n - 1$ , where  $\Lambda$  is the set of items that are informative for ranking and  $v$  is the number of elements in  $\Lambda$ .

### 4.1. Simulation Study 1: Simulation Under the EMM Model

We generated multiple ranking lists from our model (Equation (5)). Let  $U = \{1, 2, \dots, 100\}$  and  $\pi_0 = (1, 2, \dots, 100)$ . Set  $\phi = 0.5$  and assume  $\alpha_l \sim \text{Unif}(0, 1)$ ,  $l = 1, \dots, m$ . For each  $\omega_l$ , we sample it from  $\text{Unif}(0.6, 1)$  at a probability  $\eta$ ; otherwise, we sample it from  $\text{Unif}(0, 0.6)$ . Thus, parameter  $\eta$  controls the signal strength of the dataset. A larger  $\eta$  indicates that the qualities of the rankings are more similar and more informative. We generated a couple of top-30 datasets by setting  $\eta =$



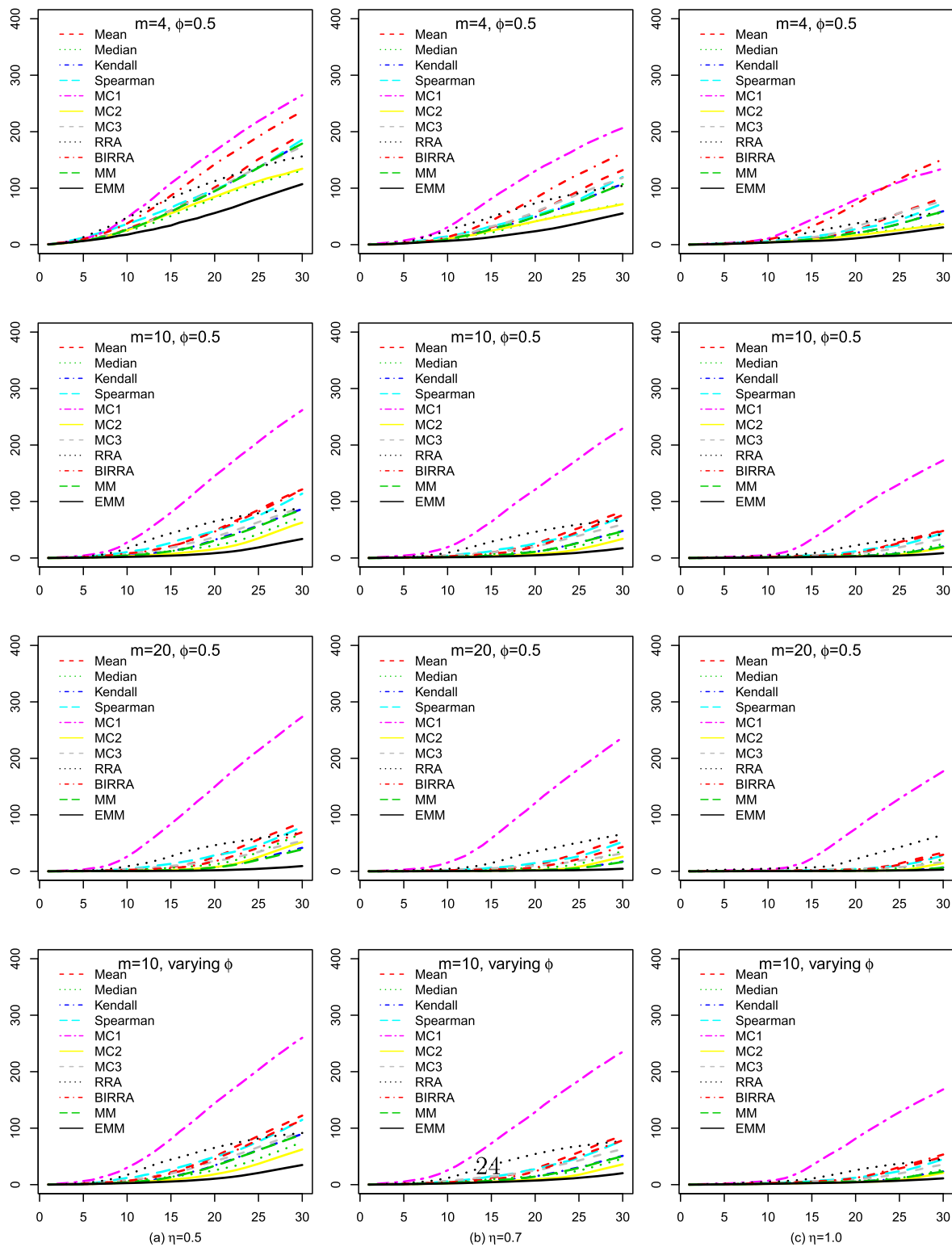
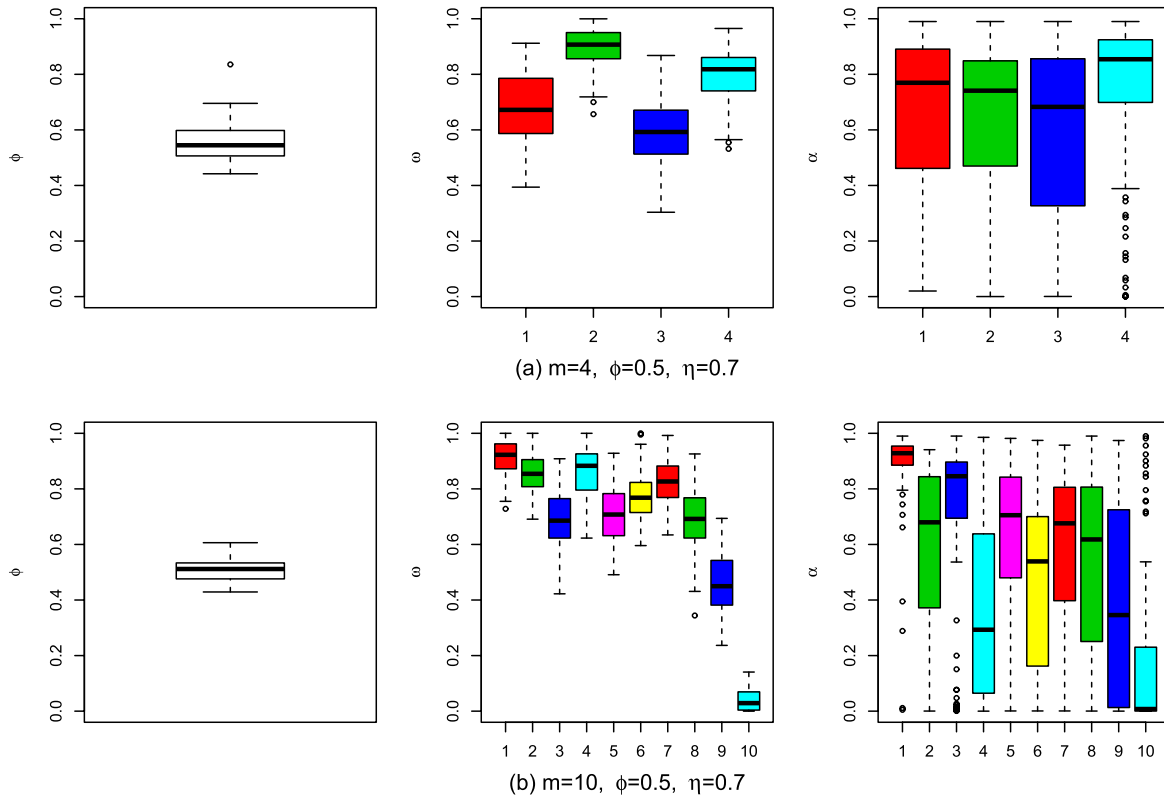


Figure 2. Stagewise Kendall tau distance of consensus rankings in Simulation study 1.

0.5, 0.7, 1.0, and  $m = 4, 10, 20$ . Besides, we also considered the case that  $\phi$  is discrepant for different ranking list, in which we set  $\phi \sim \text{Uniform}(0, 1)$ . Under each simulation setting, we generated 100 synthetic datasets and reported the mean of the Kendall tau distance at different rank positions. The results are summarized in Figure 2, which shows the following. First, EMM outperformed the alternatives with a smaller Kendall tau dis-

tance at different positions in all cases. Second, given a specific  $m$ , when  $\eta$  becomes larger, the differences in the performance of those methods become smaller, as the rankings become more informative. Third, given a specific  $\eta$ , when  $m$  becomes large, the performance of all of the methods improved as we have more ranking lists. Finally, given smaller  $m$  and  $\eta$ , EMM significantly outperformed the alternatives.



**Figure 3.** Boxplots of the parameter estimates from parametric bootstrap samples in Simulation study 1. The three columns correspond to the parameters  $\phi$ ,  $\omega$ , and  $\alpha$ , respectively.

To investigate the uncertainty of MLEs, we chose two synthetic data sets in the setting of  $m = 4$ ,  $\phi = 0.5$ ,  $\eta = 0.7$  and  $m = 10$ ,  $\phi = 0.5$ ,  $\eta = 0.7$ , respectively. The boxplots of parameter estimates from parametric bootstrap samples are shown in Figure 3. We can see that as the number of ranking lists increased, the estimation of parameters became more stable. Compared with  $\phi$  and  $\omega$ ,  $\alpha$  exhibited a larger deviation.

#### 4.2. Simulation Study 2: A Synthetic Microarray Data Example

We simulated sets of microarray data in a similar setting as described by Kooperberg et al. (2005) instead of from EMM. Let  $x_{ijgl}$  represent the expression level of the  $i$ th gene from the  $j$ th array of the  $g$ th group of the  $l$ th study for  $i = 1, 2, \dots, 100$ ,  $j = 1, 2, \dots, 10$ ,  $g = 1, 2$ , and  $l = 1, 2, \dots, m$ . Here, the group indicators  $g = 1, 2$  refer to the control and treatment groups, respectively. Generate  $x_{ijgl}$  as  $\mu_i + \delta_{ig} + z_{ijgl}$ , where  $\mu_i \sim \text{Unif}(0, 1)$ ,  $z_{ijgl} \sim N(0, \sigma_{il}^2)$ ,  $\sigma_{il} = (0.3 - 0.02\mu_i)G_iR_l$ ,  $G_i \sim \Gamma(4, 1)$ , and  $R_l \sim \eta \text{Unif}(0.5, 2) + (1 - \eta)\text{Unif}(2, 4)$ . The variation of the gene expression is related to its mean and the study as well. The first 40 genes are set to be differentially expressed, with mean difference  $\delta_{ig} = 0.25(2B_i - 1)Q_i$ , where  $B_i \sim \text{Bern}(0.5)$  and  $Q_i \sim \Gamma(4, 1)$ ,  $i = 1, 2, \dots, 40$ ,  $g = 2$ ; otherwise,  $\delta_{ig} = 0$ . The “true” consensus ranking corresponds to the ordering of the magnitude of  $\frac{\delta_{12}}{(0.3-0.02\mu_1)G_1}, \frac{\delta_{22}}{(0.3-0.02\mu_2)G_2}, \dots, \frac{\delta_{40,2}}{(0.3-0.02\mu_{40})G_{40}}$  from the largest to the smallest. We ranked the genes in each simulated study using the attenuated two-sample  $t$ -test (Tusher, Tibshirani, and Chu 2001).

Similar to the previous simulation study,  $\eta$  controls the signal strength of the rankings. We generated a number of top-40 datasets with different combinations of  $\eta$  and  $m$  by setting  $\eta = 0.5, 0.7, 1.0$ , and  $m = 4, 10, 20$ . In each setting, we generated 100 synthetic datasets and reported the mean of the Kendall tau distance and true positive rate at different positions. The results are summarized in Figures 4 and 5. From those figures, we can see that EMM consistently outperformed the other methods. Overall, the results here are similar to those in the previous simulation study.

#### 4.3. Simulation Study 3: Over-Correlated Rankings

In this simulation study, we test the performance of the RC in detecting the correlation structures among the rankings. Let  $U = \{1, 2, \dots, 100\}$  and  $\pi_0 = (1, 2, \dots, 100)$ . We generated 20 independent full rankings from our model (Equation (5)) by setting  $\phi = 0.6$ ,  $\omega_1 = \omega_2 = \dots = \omega_{20} = 2/3$ , and  $\alpha_1 = \alpha_2 = \dots = \alpha_{20} = 0.5$ . We also considered the following group structure of the 20 rankings

$$G_1 = \{\pi_1, \pi_2, \dots, \pi_5\}, \\ G_2 = \{\pi_6, \pi_7, \dots, \pi_{10}\}, \quad \pi_{11}, \pi_{12}, \dots, \pi_{20}.$$

To generate a group of correlated rankings, we first sampled a representative ranking, and then randomly sampled an item and transposed it with one of the adjacent five items. Let  $s$  denote the number of transposition operations as described previously. To obtain an over-correlated ranking, we set  $s = 10$ , while for a weakly correlated ranking, we set  $s = 30$ . We calculated the Kendall correlation matrix, Spearman correlation matrix, CC

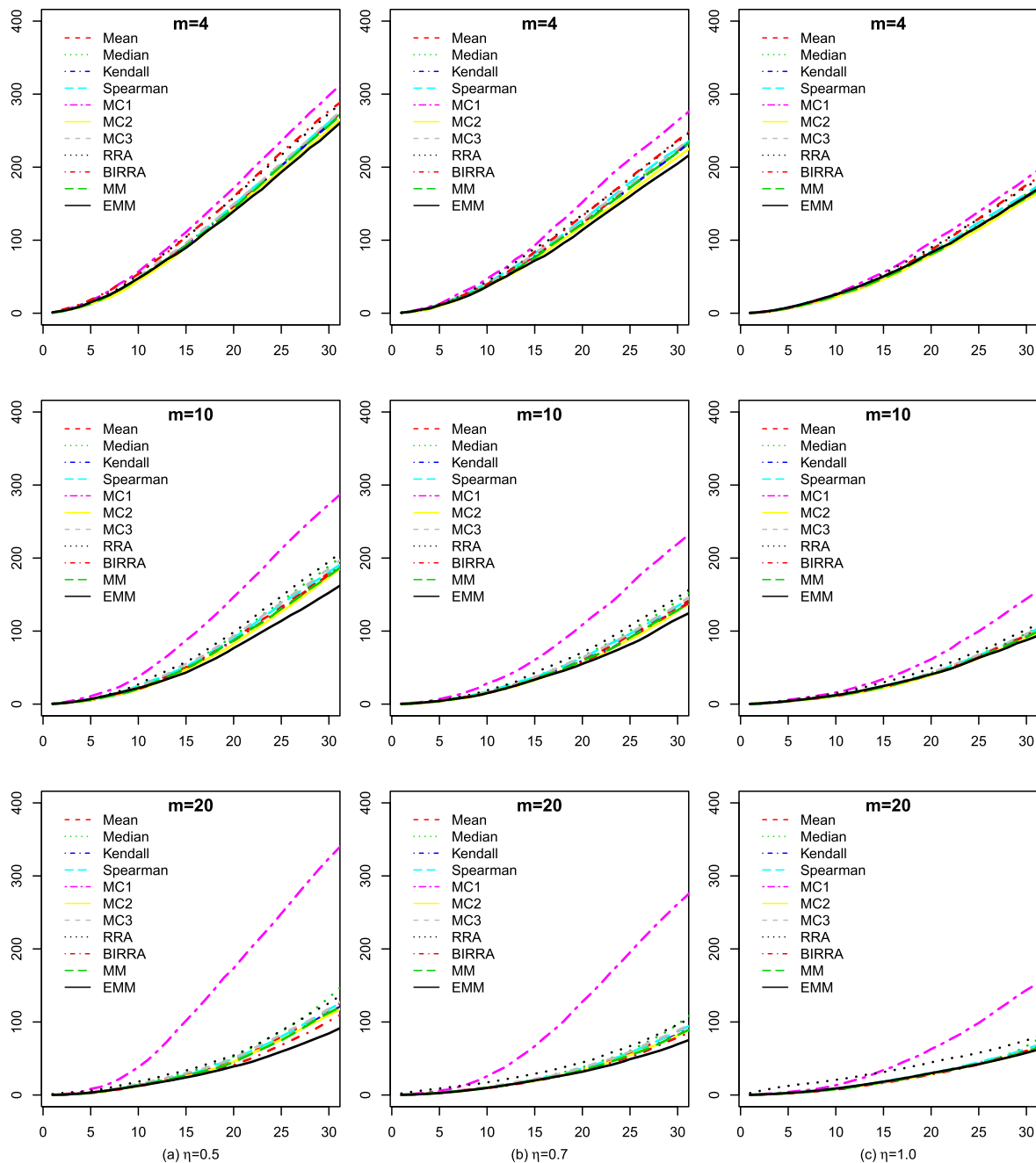


Figure 4. Stagewise Kendall tau distance of consensus rankings in Simulation study 2.

matrix and RC matrix for the independent, strong correlation, and weak correlation cases. The results are shown in Figure 6. In the independent case, no method found any group of correlated rankings. Here, a group of rankings should have at least three rankings. In the strong correlation case, all of the methods found two groups of over-correlated rankings. In the weak correlation case, the Kendall and Spearman correlation methods did not work, while CC and RC still found the two true correlated groups. However, CC reported one false positive group. In all of these cases, both the original and after-thresholding pictures of RC clearly reflected the correlation of rankings, and RC reported a much smaller number of over-correlated pairs than CC. We sampled the 20 rankings repeatedly in each case 100 times and reported the mean of the number of true- and false-positive

paired rankings detected by RC and CC in Table 1. We can see that both RC and CC correctly found the over-correlated groups in all of those runs; however, RC reported a significantly smaller number of false-positive pairs than CC. These results suggest that the RC is indeed quite effective in capturing the correlation structure of rankings.

To evaluate the effect of over-correlated rankings on the aggregation result, we consider two scenarios: (1) rankings of equal quality by fixing  $\phi = 0.6$ ,  $\omega_1 = \omega_2 = \dots = \omega_{20} = 2/3$ , and  $\alpha_1 = \alpha_2 = \dots = \alpha_{20} = 0.5$ ; and (2) rankings of varying quality with  $\phi = 0.6$ ,  $\omega_l \sim \text{Beta}(4, 2)$ ,  $\alpha_l \sim \text{Unif}(0, 1)$ ,  $l = 1, 2, \dots, 20$ . We also used HEMM to fit the data. For each scenario, we simulated 100 synthetic datasets and reported the performance of these methods based on their mean statistics,

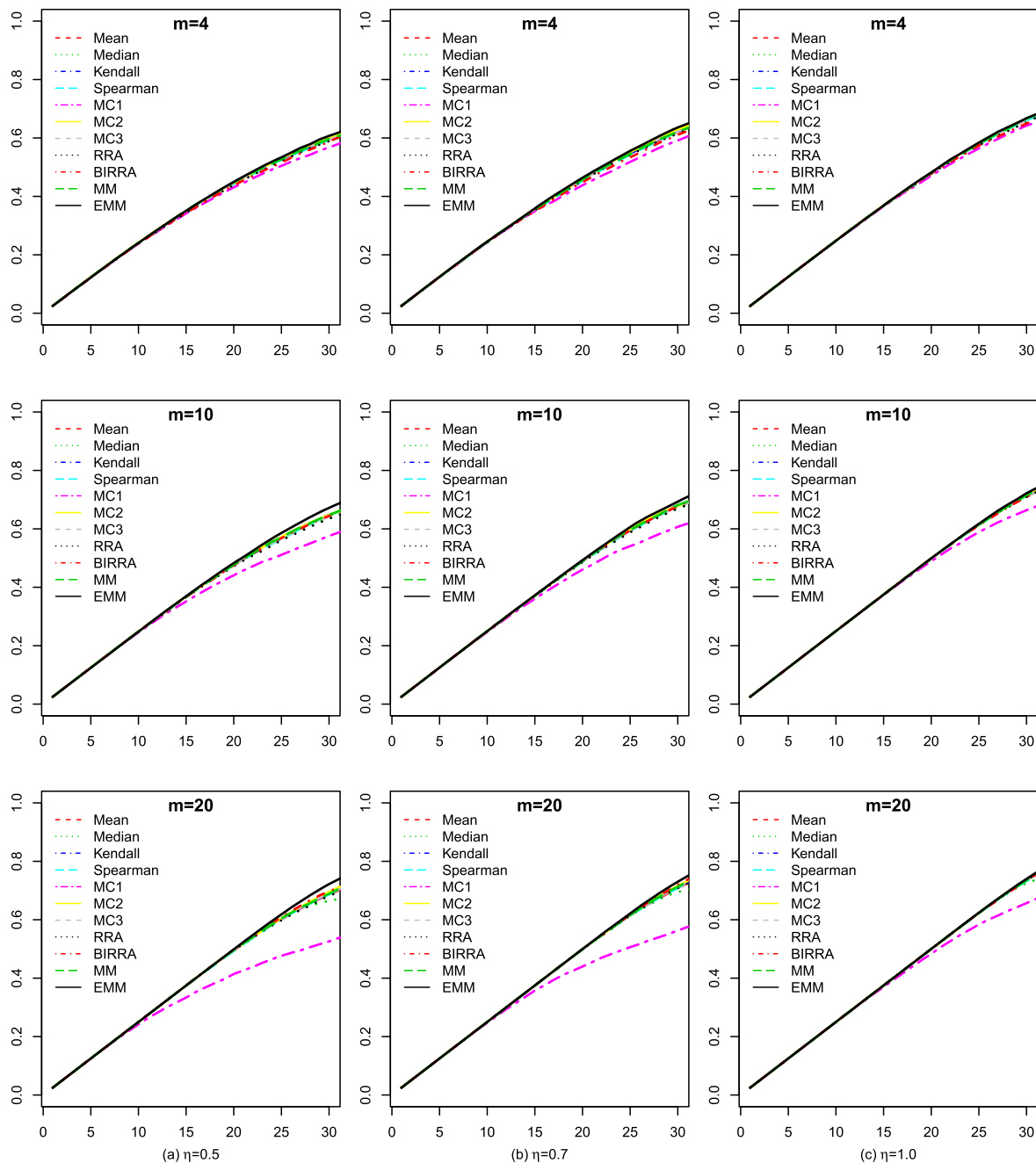


Figure 5. True positive rate of consensus rankings in Simulation study 2.

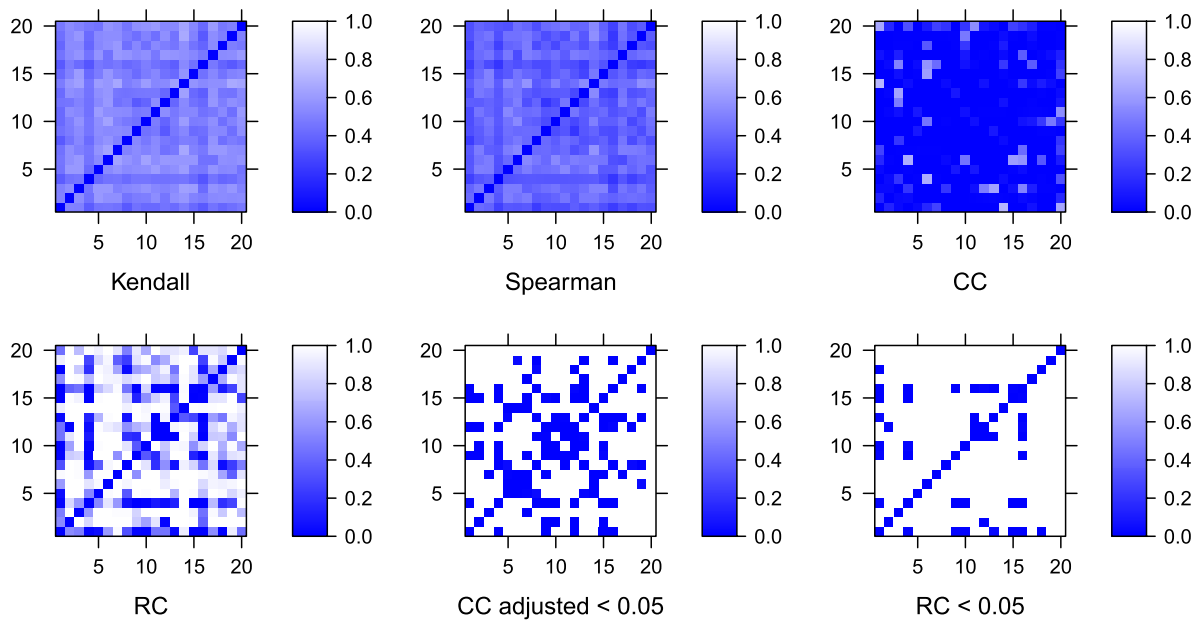
as shown in Figure 7. We can see that HEMM outperformed the other methods in all of the settings and that EMM was the second best method. Compared with other methods, RRA made more mistakes for top ranked items because it uses the distribution of the ranks of items, and a significant portion of the ranks could deviate from its ground truth if the rankings are over-correlated.

## 5. Applications

DeConde et al. (2006) collected ranking lists from five different microarray-based prostate cancer studies (Dhanasekaran et al. 2001; Luo et al. 2001; Welsh et al. 2001; Singh et al. 2002; True

et al. 2006), and the rankings are shown in the first six columns of Table 2. The rankings present the top 25 genes found to be up-regulated in prostate tumors compared with normal prostate tissues. These five studies relied on different technologies, and their results are quite different, although there are indeed quite a number of overlaps.

DeConde et al. (2006) and Lin and Ding (2009) analyzed this dataset, found that the gene list in the study by Luo et al. (2001) was the least common compared with those of the other four studies, and downgraded its weight in their analysis. They also found that the results of Welsh et al. (2001) and Dhanasekaran et al. (2001) were more reliable than the others. We applied EMM to this dataset and observed that the ranking list of Welsh et al. (2001) coincides with the consensus ranking. In the last



(a) Independent rankings

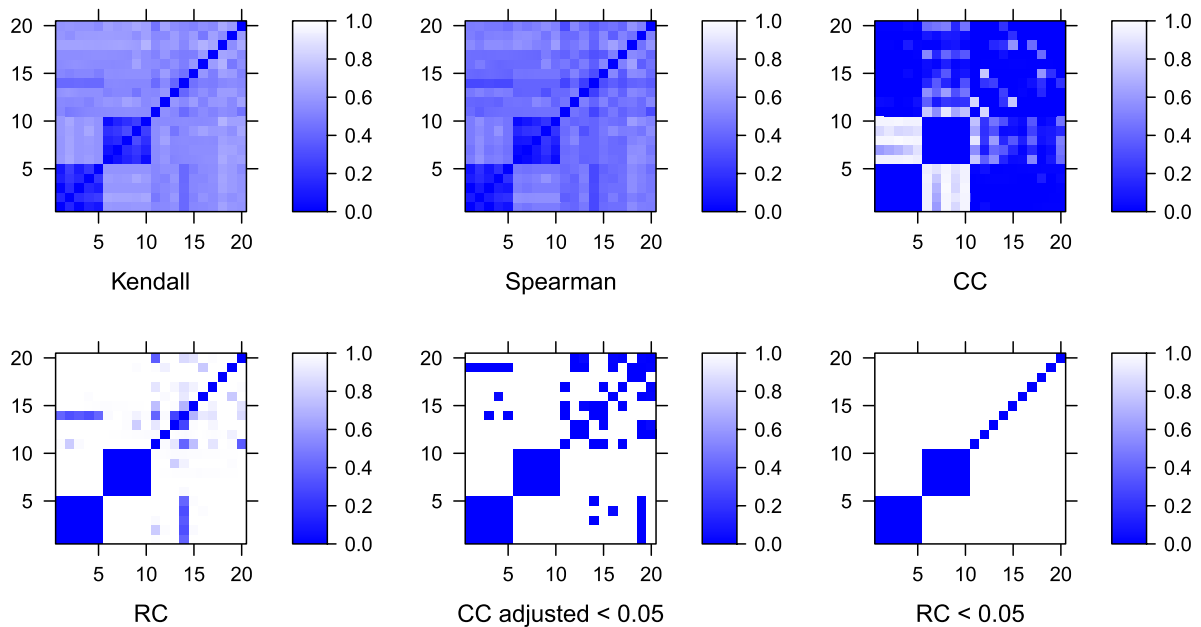
(b) Strongly correlated rankings with  $s = 10$ 

Figure 6. (Continued.)

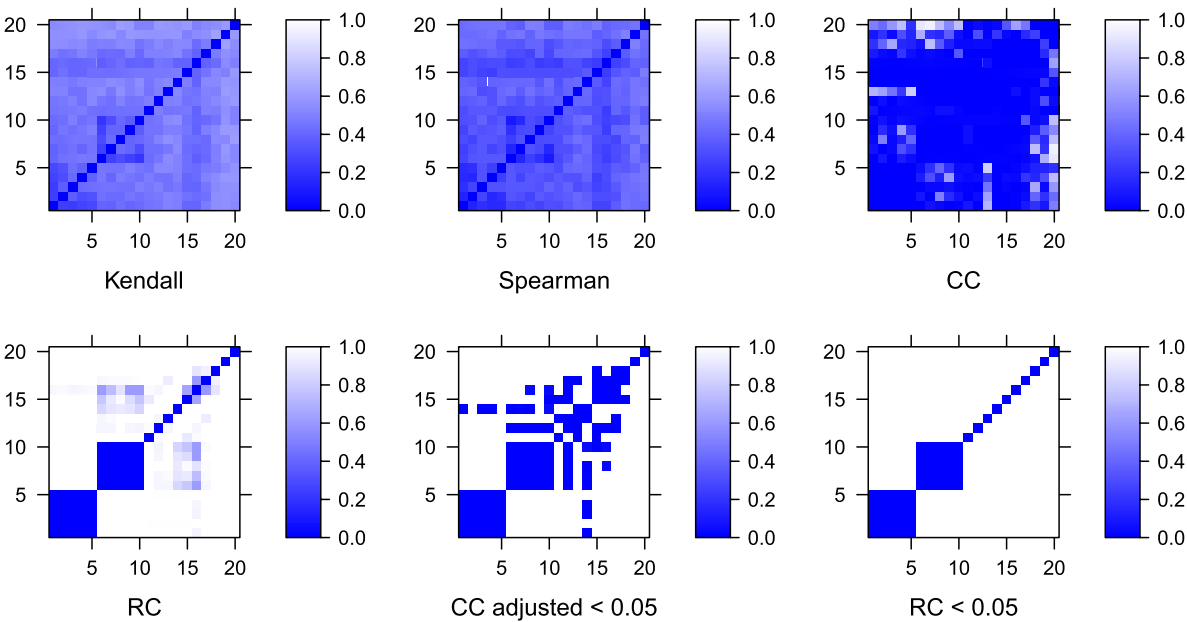
column of Table 2, beside each gene, we indicate the studies in which the gene is present in the top 25 list. The MLEs of the parameters are  $\hat{\phi} = 0.93$ ,  $\hat{\omega} = (0.07, 1.00, 0.90, 0.10, 0.16)$ , and  $\hat{\alpha} = (0.88, 1.00, 0.63, 0.77, 0.87)$ . The studies by Welsh et al. (2001) and Dhanasekaran et al. (2001) have quite a large  $\omega$ , indicating that their qualities are good. The study by Luo et al. (2001) has the smallest  $\omega$  and a large  $\alpha$ , implying that except for a few highly ranked genes, the ranking is not that different from the random ranking. The boxplots of parametric bootstrap samples of the parameters are shown in Figure 8. The MLEs of

$\phi$  and  $\omega$  are more stable than that of  $\alpha$ , suggesting that around the current mode  $(\hat{\phi}, \hat{\omega}, \hat{\alpha})$ , the likelihood function is flat in the direction of  $\hat{\alpha}$ .

## 6. Computation Time

For all the previous simulation studies and the real example in this article, the ECM algorithm converges after a few iterations and it takes less than 4 sec to run. Next we check its computation efficiency in different settings of  $(n, k, m)$ , where  $n$  is the number





(c) Weakly correlated rankings with  $s = 30$

**Figure 6.** Correlation matrix of the rankings using different correlation measures. “CC adjusted” means that the coordinate coefficients have been adjusted by the Bonferroni method as suggested by Deng et al. (2014). For the correlation statistic of Kendall or Spearman, we use 1 minus its value in plotting, such that it resembles the  $p$ -value of RC and CC.

**Table 1.** Performance of the coordinate coefficient and the rank coefficient in detecting correlated ranking pairs.

Case	True positive		False positive	
	CC	RC	CC	RC
Independent case	–	–	38.21(7.86)	13.00(3.57)
$s = 10$	20(0)	20(0)	29.76(9.09)	0.04(0.28)
$s = 30$	20(0)	20(0)	29.59(9.18)	0.09(0.40)

NOTE: The statistic is the mean of the number of detected correlated ranking pairs with  $p$ -value  $< 0.05$  in 100 replicates. The SD is shown in the bracket. “–” means the statistic is not applicable in the corresponding case.

of items to be ranked,  $k$  is the length of the top- $k$  items, and  $m$  is the number of ranking lists. The synthetic data were generated as described in Simulation study 1 by setting  $\phi = 0.5$  and  $\eta = 0.7$ . We also check the computation time of a couple of competing methods for comparison. The mean and SD of the running time from 100 replicates in each setting are summarized in Table 3. We can see that EMM took more computation time than other methods, but for a moderate size of data, its running time was usually less than 1 min, which is practically acceptable.

7. Discussion

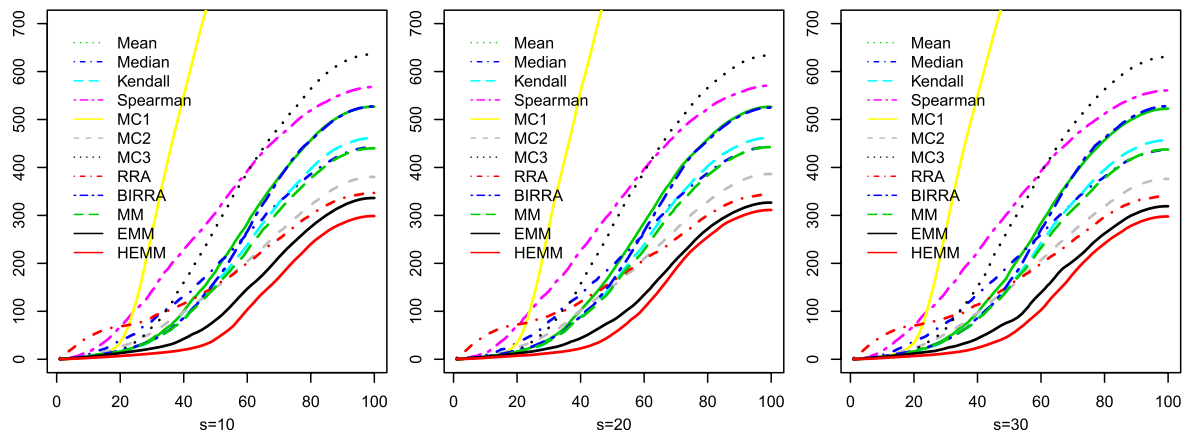
In this article, we proposed an extended Mallows ranking model to aggregate multiple ranking lists of varying quality and stability. Our model is applicable to both full and partial ranking data. The model is quite flexible and has several desirable properties as discussed previously. As some rankings may not be independent, we further designed a statistic called RC to detect over-correlated rankings, and proposed a hierarchical ranking model as a remedy if the independence assumption is violated. We applied our model and its hierarchical version to simulation

studies and real applications in diverse contexts, and showed that they outperformed established approaches. Compared with the alternatives, EMM and HEMM are more elaborate but take more time to infer the parameters.

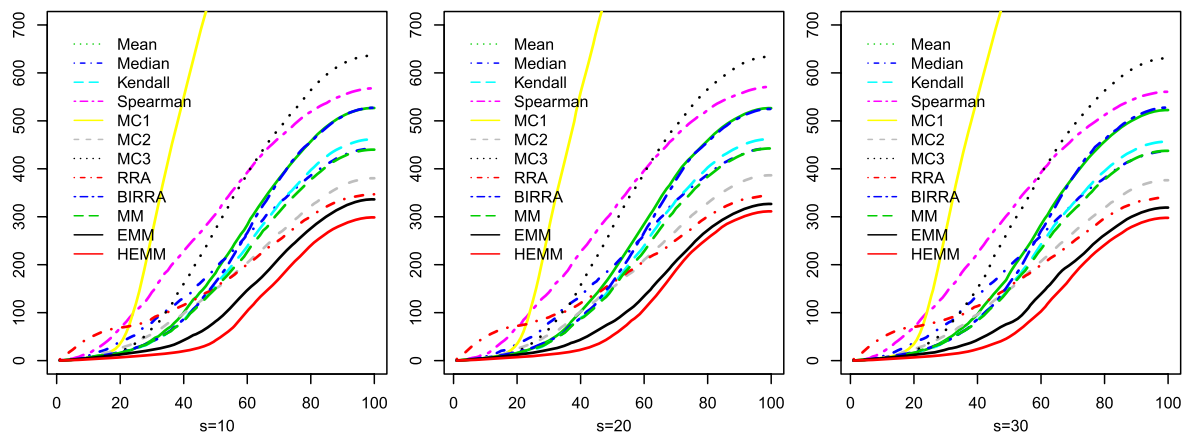
In our empirical studies, we compared the performance of 12 rank aggregation methods. The satisfactory performance of EMM and HEMM encourages their use in practice. Following EMM and HEMM, we showed that MC2 (Lin 2010) was robust

**Table 2.** Rankings of prostate cancer genes and the consensus ranking inferred by EMM.

Rank	Luo(L)	Welsh(W)	Dhana(D)	TRUE(T)	Singh(S)	EMM
1	HPN	HPN	OGT	AMACR	HPN	HPN (LWDTs)
2	AMACR	AMACR	AMACR	HPN	SLC25A6	AMACR (LWDT)
3	CYP1B1	OACT2	FASN	NME2	EEF2	OACT2 (WD)
4	ATF5	GDF15	HPN	CBX3	SAT	GDF15 (WDT)
5	BRCA1	FASN	UAP1	GDF15	NME2	FASN (WDS)
6	LGALS3	ANK3	GUCY1A3	MTHFD2	LDHA	ANK3 (W)
7	MYC	KRT18	OACT2	MRPL3	CANX	KRT18 (WDS)
8	PCDHGC3	UAP1	SLC19A1	SLC25A6	NACA	UAP1 (WDS)
9	WT1	GRP58	KRT18	NME1	FASN	GRP58 (WS)
10	TFF3	PPIB	EEF2	COX6C	SND1	PPIB (WD)
11	MARCKS	KRT7	STRA13	JTV1	KRT18	KRT7 (W)
12	OS-9	NME1	ALCAM	CCNG2	RPL15	NME1 (LWDT)
13	CCND2	STRA13	GDF15	AP3S1	TNFSF10	STRA13 (WD)
14	NME1	DAPK1	NME1	EEF2	SERP1	DAPK1 (W)
15	DRR1A	TMEM4	CALR	RAN	GRP58	TMEM4 (WS)
16	TRAP1	CANX	SND1	PRKACA	ALCAM	CANX (WS)
17	FMO5	TRA1	STAT6	RAD23B	GDF15	TRA1 (W)
18	ZHX2	PRSS8	TCEB3	PSAP	TMEM4	PRSS8 (W)
19	RPL36AL	EMTPD6	EIF4A1	CCT2	CCT2	EMTPD6 (W)
20	ITPR3	PPP1CA	LMAN1	G3BP	SLC39A6	PPP1CA (W)
21	GCSH	ACADSB	MAOA	EPRS	RPL5	ACADSB (W)
22	DDB2	PTPLB	ATP6V0B	CKAP1	RPS13	PTPLB (W)
23	TFCP2	TMEM23	PPIB	LIG3	MTHFD2	TMEM23 (W)
24	TRAM1	MRPL3	FMO5	SNX4	G3BP2	MRPL3 (WT)
25	YTHDF3	SLC19A1	SLC7A5	NSMAF	UAP1	SLC19A1 (WD)



(a) Rankings with equal reliability



(b) Rankings with varying reliability

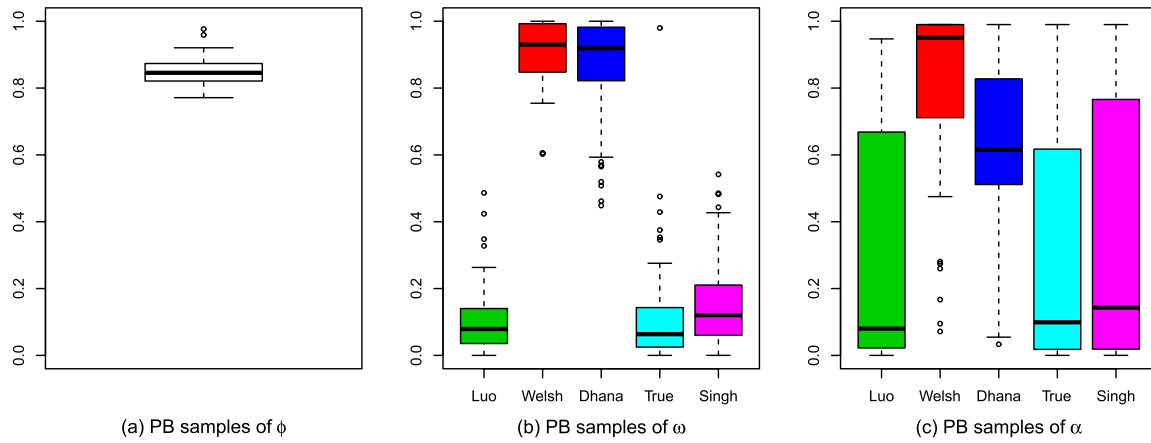
**Figure 7.** Stagewise Kendall tau distance of consensus rankings in Simulation study 3.

as well because it adopts a majority rule to construct the transition matrix. The more recent approaches, RRA and BIRRA, exploit the distribution of the ranks of items and aim to detect those items that are consistently ranked highly in the ranking lists. In the cases of partial and over-correlated rankings, when the rank distribution is conservatively approximated or deviates from the ground truth structurally, we observe that RRA and BIRRA do not perform as well as our methods. They were proposed to analyze datasets with a large number of items for ranking, and only a small proportion of them are informative (Kolde et al. 2012; Badgeley, Sealfon, and Chikina 2015), which is a conventional feature of biology data. We also compared the performance of those methods in such a scenario. To save space, this extra simulation study is given in the supplementary materials.

In our model, we assume the dispersion parameter  $\phi_i = \phi(1 - \alpha(i)) = \phi(1 - \alpha^i)$ , such that the model is well interpretable and has a robust performance as we observed. However, the dispersion may not always be such a case. For example, the dispersion may be constant or increase in a linear way. Besides, people may be quite certain about the rank of the top-ranked and bottom-ranked items, but they are not sure about that of

the middle-ranked items. We check the performance of EMM in these three scenarios and found that EMM still performed robustly well. Details of this extra simulation study are also given in the supplementary materials.

The specification of  $\alpha(i)$ , which controls the reliability of ranking process, is not a simple issue. If we have some prior information how uncertain when people make their rankings, we can set  $\alpha(i)$  accordingly. Note that we can also interpret  $\alpha(i)$  as the function of measuring the importance of different rank positions. In such case, if we are more concern about certain rank positions, we could manually set their  $\alpha(i)$  larger than others. In our studies, we assume no available or specific prior information. Setting  $\alpha(i) = \alpha^i$ , we find EMM works well in the scenarios we investigate. However, one thing we noticed is that the uncertainty of the MLE of  $\alpha$  was large in some cases, as its bootstrap samples manifested, while the MLEs of  $\phi$  and  $\omega$  had smaller deviations. This may due to that the simple form  $\alpha(i) = \alpha^i$  is not sufficient to depict the stability of ranking process in some cases. To address this problem, one may need a more sophisticated function of  $\alpha(i)$ . But for the ranking lists that have a few items, a complicated function  $\alpha(i)$  may lead to over-fitting.



**Figure 8.** Boxplots of parametric bootstrap samples of parameters in prostate cancer study.

**Table 3.** Computation time (unit: second) of different rank aggregation methods.

$(n, k, m)$	(100,30,10)	(100,100,40)	(1000,100,20)	(5000,100,20)	(5000,100,40)
Mean	0.00(0.01)	0.01(0.01)	0.01(0.01)	0.01(0.01)	0.02(0.01)
Kendall	0.09(0.03)	0.24(0.06)	0.52(0.08)	0.71(0.12)	2.19(0.33)
BIRRA	0.07(0.02)	0.14(0.04)	0.38(0.16)	0.41(0.13)	0.87 (0.39)
MM	0.18(0.06)	0.33(0.07)	0.84(0.16)	1.06(0.21)	2.51(0.41)
EMM	0.84(0.28)	5.10(1.03)	19.77(5.15)	25.13(7.22)	92.27(23.00)

NOTE: All computations were done in R 3.4.2 for Windows using a single core of an Intel Core i7-6700 CPU and the RAM of computer is 8GB.

The Mallows model has been generalized by other researchers as well. Murphy and Martin (2003) and Meilă and Chen (2010) studied the clustering structure of the ranking data by proposing the mixture Mallows model and the Dirichlet process mixtures of generalized Mallows model, respectively. Most of other current studies focus on studying the learning and sampling problem of the Mallows model under various distances, including the Kendall tau distance, the Cayley distance, etc. As far as we know, there is little effort on investigating the  $\phi_i$  components of the Mallows model as we did in this article.

Our methods also have some other limitations. First, we do not consider the rankings with ties in the current framework. To handle such case, we may use a modified version of Kendall tau distance as in Fagin et al. (2006) or use the multiple imputation method (Rubin 2004). Second, we do not make use of the covariate information of the items, which may influence their ranks. One speculative idea is to build the item information into the stagewise selection step through a regression framework. Third, in some cases, the ranking lists may be heterogeneous. For instance, in a political election, voters from different parties may have their own candidate preference lists, thus there may be more than one consensus ranking. Murphy and Martin (2003) proposed mixtures of distance-based ranking models and used BIC to determine the number of heterogeneous groups. To extend EMM in this direction is well worth exploration as well. Fourth, when the rankings are over-correlated in a non-spherical fashion, for example, chain correlated, the current RC and HEMM may not be powerful. A possible solution is to set  $p(\pi_i|\pi_{i-1}, \pi_0) \propto \phi^{\omega_1 d_K(\pi_i, \pi_{i-1}) + \omega_2 d_K(\pi_i, \pi_0)}$ , where  $\omega_1, \omega_2$  are penalties for the distance  $d_K(\pi_i, \pi_{i-1})$  and  $d_K(\pi_i, \pi_0)$ , respectively. Further, we may set different  $\phi$  for different rank positions

as EMM does. If  $\frac{\omega_2}{\omega_1} \geq n$ , where  $n$  is the number of items to be ranked, we can follow the proof of Theorem 3.3 to show that the mean rank method gives a consistent estimate of the true ranking. In future work, we will study the properties of such models.

## Appendix

### A1. The proof of Theorem 3.4.

*Proof.*

$$\begin{aligned}
 p(\pi_1) - p(\pi_2) &= \kappa_1 \left\{ \left[ \omega \frac{\phi_a^{v_a}}{Z(\phi_a, a)} + (1 - \omega) \frac{1}{n - a + 1} \right] \right. \\
 &\quad \times \left[ \omega \frac{\phi_b^{v_b}}{Z(\phi_b, b)} + (1 - \omega) \frac{1}{n - b + 1} \right] \\
 &\quad - \left[ \omega \frac{\phi_a^{v'_a}}{Z(\phi_a, a)} + (1 - \omega) \frac{1}{n - a + 1} \right] \\
 &\quad \times \left. \left[ \omega \frac{\phi_b^{v'_b}}{Z(\phi_b, b)} + (1 - \omega) \frac{1}{n - b + 1} \right] \right\} \\
 &= \kappa_1 \kappa_2 \left\{ \omega^2 (n - a + 1)(n - b + 1) (\phi_a^{v_a} \phi_b^{v_b} - \phi_a^{v'_a} \phi_b^{v'_b}) \right. \\
 &\quad + \omega(1 - \omega) Z(\phi_b, b)(n - a + 1) (\phi_a^{v_a} - \phi_a^{v'_a}) \\
 &\quad \left. + \omega(1 - \omega) Z(\phi_a, a)(n - b + 1) (\phi_b^{v_b} - \phi_b^{v'_b}) \right\} \\
 &= \kappa_1 \kappa_2 \left\{ \omega^2 (n - a + 1)(n - b + 1) (\phi_a^{v_a} \phi_b^{v_b} - \phi_a^{v'_a} \phi_b^{v'_b}) \right.
 \end{aligned}$$

$$\begin{aligned}
& + \frac{\omega(1-\omega)Z(\phi_b, b)(n-a+1)}{\phi_b^{v'_b}} (\phi_a^{v_a} \phi_b^{v'_b} - \phi_a^{v'_a} \phi_b^{v_b}) \\
& + \frac{\omega(1-\omega)Z(\phi_a, a)(n-b+1)}{\phi_a^{v'_a}} (\phi_a^{v_a} \phi_b^{v_b} - \phi_a^{v'_a} \phi_b^{v'_b}) \} \\
\end{aligned} \quad (\text{A.1})$$

$$\text{where } \kappa_1 = \prod_{i \neq a, b} \left[ \omega \frac{\phi_i^{v_i}}{Z(\phi_i, i)} + (1-\omega) \frac{1}{n-i+1} \right],$$

$$\kappa_2 = Z(\phi_a, a)Z(\phi_b, b)(n-a+1)(n-b+1).$$

First, given  $v_a + v_b = v'_a + v'_b$ ,  $v_a < v'_a$  and  $\phi_a \leq \phi_b$ , we can easily derive that  $\phi_a^{v_a} \phi_b^{v_b} - \phi_a^{v'_a} \phi_b^{v'_b} \geq 0$ . Thus, the first term in Equation (A.1) is nonnegative. Second, given  $v_a < v'_a$  and  $v_b > v'_b$ , we have  $\phi_a^{v_a} \phi_b^{v'_b} - \phi_a^{v'_a} \phi_b^{v_b} > 0$  and  $\phi_a^{v_a} \phi_b^{v_b} - \phi_a^{v'_a} \phi_b^{v'_b} < 0$ . If  $\frac{\omega(1-\omega)Z(\phi_b, b)(n-a+1)}{\phi_b^{v'_b}} \geq \frac{\omega(1-\omega)Z(\phi_a, a)(n-b+1)}{\phi_a^{v'_a}}$ , which is equivalent to  $\frac{\phi_a^{v'_a}/Z(\phi_a, a)}{\phi_b^{v'_b}/Z(\phi_b, b)} \geq \frac{n-b+1}{n-a+1}$ , the sum of the last two terms in Equation (A.1) is no less than  $\frac{\omega(1-\omega)Z(\phi_a, a)(n-b+1)}{\phi_a^{v'_a}} (\phi_a^{v_a} \phi_b^{v'_b} - \phi_a^{v'_a} \phi_b^{v_b} + \phi_a^{v_a} \phi_b^{v_b} - \phi_a^{v'_a} \phi_b^{v'_b}) = \frac{\omega(1-\omega)Z(\phi_a, a)(n-b+1)}{\phi_a^{v'_a}} (\phi_a^{v_a} \phi_b^{v_b} - \phi_a^{v'_a} \phi_b^{v'_b}) \geq 0$ . When condition (a) satisfies,  $\omega = 1$ , we need only to consider the first term, and it is nonnegative. When condition (b) satisfies, the sum of the three terms is nonnegative. Thus, the whole term  $p(\pi_1) - p(\pi_2) \geq 0$  if either condition satisfies.  $\square$

A2. The proof for that

$$\begin{aligned}
q_1(\phi) &= \sum_{l=1}^m \sum_{i=1}^{n-1} p(z_{li} = 1 | \Psi^{(t-1)}) \\
&\quad \times \left[ v_{li} \log \phi - \log Z(\phi(1 - (\alpha_l^{(t-1)})^i), i) \right]
\end{aligned}$$

has a unique global maximum when  $\phi \in [a, b] \subset (0, 1)$ .

**Proof.** We first consider the term  $f(\phi) \equiv v_{li} \log \phi - \log Z(\phi(1 - (\alpha_l^{(t-1)})^i), i)$ . Set  $e^{-\theta} \equiv \phi$  ( $-\theta \geq 0$ ) and  $e^{-c} \equiv 1 - (\alpha_l^{(t-1)})^i$  ( $c \geq 0$ ). Thus,  $f(\phi)$  can be reformulated as the following function of  $\theta$

$$\tilde{f}(\theta) = v_{li}(-\theta) - \log Z(e^{-(\theta+c)}, i).$$

Then  $\tilde{f}'(\theta) = -v_{li} - \frac{Z'_{\theta}(e^{-(\theta+c)}, i)}{Z(e^{-(\theta+c)}, i)}$ . Next we show  $g(\theta) \equiv \frac{Z'_{\theta}(e^{-\theta}, i)}{Z(e^{-\theta}, i)}$  is a strictly increasing function of  $\theta$  for  $\theta > 0$

$$\begin{aligned}
Z(e^{-\theta}, i) &= 1 + e^{-\theta} + \dots + e^{-(n-i)\theta} \\
&= \frac{1 - e^{-(\gamma+1)\theta}}{1 - e^{-\theta}} \quad (\text{where } \gamma \equiv n-i) \\
Z'_{\theta}(e^{-\theta}, i) &= \frac{(\gamma+1)e^{-(\gamma+1)\theta} - \gamma e^{-(\gamma+2)\theta} - e^{-\theta}}{(1 - e^{-\theta})^2}
\end{aligned}$$

$$\begin{aligned}
g(\theta) &\equiv \frac{Z'_{\theta}(e^{-\theta}, i)}{Z(e^{-\theta}, i)} \\
&= \frac{(\gamma+1)e^{-(\gamma+1)\theta}}{1 - e^{-(\gamma+1)\theta}} - \frac{e^{-\theta}}{1 - e^{-\theta}} \\
&= \frac{\gamma+1}{e^{(\gamma+1)\theta} - 1} - \frac{1}{e^{\theta} - 1} \\
g'(\theta) &= -\frac{(\gamma+1)^2 e^{(\gamma+1)\theta}}{(e^{(\gamma+1)\theta} - 1)^2} + \frac{e^{\theta}}{(e^{\theta} - 1)^2}.
\end{aligned}$$

To determine the sign of  $g'(\theta)$ , we consider the following ratio of the two parts in  $g'(\theta)$

$$\begin{aligned}
\varphi(\theta) &\equiv \frac{e^{\theta}(e^{(\gamma+1)\theta} - 1)^2}{(\gamma+1)^2 e^{(\gamma+1)\theta} (e^{\theta} - 1)^2} \\
&= \frac{1}{(\gamma+1)^2 e^{\gamma\theta}} \cdot \left( \frac{e^{(\gamma+1)\theta} - 1}{e^{\theta} - 1} \right)^2 \\
&= \frac{1}{(\gamma+1)^2 e^{\gamma\theta}} \cdot (1 + e^{\theta} + \dots + e^{(\gamma-1)\theta})^2 \\
&> \frac{1}{(\gamma+1)^2 e^{\gamma\theta}} \cdot ((\gamma+1)e^{\gamma\theta/2})^2 = 1 \\
&\quad (\text{applying the formula } e^{a\theta} + e^{(\gamma-a)\theta} \geq 2e^{\gamma\theta/2}).
\end{aligned}$$

Thus for  $\theta > 0$ ,  $\varphi(\theta) > 1$ ,  $g'(\theta) > 0$ . Then  $g(\theta)$  is a strictly increasing function of  $\theta$  and  $\tilde{f}'(\theta) = -v_{li} - g(\theta + c)$  is a strictly decreasing function of  $\theta$ . As a result,  $\tilde{f}(\theta)$  is a concave function of  $\theta$ ,  $\theta \in [-\log a, -\log b]$ . For the objective function  $q_1(\phi) = q_1(e^{-\theta})$  is a weighted sum of those concave functions, it is a concave function of  $\theta$  as well. Consequently  $q_1(\phi)$  has a unique global maximum for  $\phi \in [a, b] \subset (0, 1)$ .  $\square$

## Supplementary Materials

The supplementary materials include (1) two extra simulation studies and one extra real application on the NBA team data from Deng et al. (2014); (2) a R package called "ExtMallows" and its instruction manual for using MM, EMM, and HEMM, and the RC method to check the independence of the rankings; (3) a Windows GUI toolkit for conducting rank aggregation using either EMM or HEMM. The instructions are illustrated with datasets and examples.

## Acknowledgments

We thank Professor Persi Diaconis for his helpful comments and suggestions in investigating rank aggregation problem.

## Funding

The research is supported in part by a grant from the Research Grants Council of the Hong Kong SAR (No. CUHK14203915) and a Mathematical Tianyuan Fund of the National Natural Science Foundation of China (No. 11626160).

## References

- Ali, M. M. (1998), "Probability Models on Horse-Race Outcomes," *Journal of Applied Statistics*, 25, 221–229. [730]
- Badgley, M. A., Sealfon, S. C., and Chikina, M. D. (2015), "Hybrid Bayesian-Rank Integration Approach Improves the Predictive Power of Genomic Dataset Aggregation," *Bioinformatics*, 31, 209–215. [730, 736, 743]

- Critchlow, D. E., Fligner, M. A., and Verducci, J. S. (1991), "Probability Models on Rankings," *Journal of Mathematical Psychology*, 35, 294–318. [730]
- Critchlow, D. E., and Verducci, J. S. (1992), "Detecting a Trend in Paired Rankings," *Journal of the Royal Statistical Society, Series C*, 41, 17–29. [735]
- de Borda, J. C. (1781), *Mémoire sur les Élections au Scrutiny*, Histoire de l'Académie Royale des Sciences. [730]
- DeConde, R. P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., and Etzioni, R. (2006), "Combining Results of Microarray Experiments: A Rank Aggregation Approach," *Statistical Applications in Genetics and Molecular Biology*, 5, 1–23. [730,740]
- Deng, K., Han, S., Li, K. J., and Liu, J. S. (2014), "Bayesian Aggregation of Order-Based Rank Data," *Journal of the American Statistical Association*, 109, 1023–1039. [730,735,742,745]
- Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001), "Delineation of Prognostic Biomarkers in Prostate Cancer," *Nature*, 412, 822–826. [740,741]
- Diaconis, P. (1988), "Group Representations in Probability and Statistics," *Lecture Notes-Monograph Series*, 11, i–192. [731]
- Diaconis, P., and Graham, R. L. (1977), "Spearman's Footrule as a Measure of Disarray," *Journal of the Royal Statistical Society, Series B*, 39, 262–268. [730]
- Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001), "Rank Aggregation Methods for the Web," in *Proceedings of the 10th International Conference on World Wide Web*, ACM, pp. 613–622. [730,735]
- Efron, B., and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press. [735]
- Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. (2006), "Comparing Partial Rankings," *SIAM Journal on Discrete Mathematics*, 20, 628–648. [744]
- Fagin, R., Kumar, R., and Sivakumar, D. (2003), "Efficient Similarity Search and Classification via Rank Aggregation," in *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, ACM, pp. 301–312. [730]
- Fligner, M. A., and Verducci, J. S. (1986), "Distance Based Ranking Models," *Journal of the Royal Statistical Society, Series B*, 48, 359–369. [730,731,732]
- (1988), "Multistage Ranking Models," *Journal of the American Statistical Association*, 83, 892–901. [733]
- Hall, P., and Schimek, M. G. (2012), "Moderate-Deviation-Based Inference for Random Degeneration in Paired Rank Lists," *Journal of the American Statistical Association*, 107, 661–672. [730,731]
- Kendall, M. G. (1970), *Rank Correlation Methods* (4th ed.), New York: Hafner. [730]
- Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012), "Robust Rank Aggregation for Gene List Integration and Meta-analysis," *Bioinformatics*, 28, 573–580. [730,736,743]
- Kooperberg, C., Aragaki, A., Strand, A., and Olson, J. (2005), "Significance Testing for Small Microarray Experiments," *Statistics in Medicine*, 24, 2281–2298. [738]
- Lin, S. (2010), "Space Oriented Rank-Based Data Integration," *Statistical Applications in Genetics and Molecular Biology*, 9, 1–23. [730,736,742]
- Lin, S., and Ding, J. (2009), "Integration of Ranked Lists via Cross Entropy Monte Carlo With Applications to mRNA and microRNA Studies," *Biometrics*, 65, 9–18. [730,740]
- Liu, Y.-T., Liu, T.-Y., Qin, T., Ma, Z.-M., and Li, H. (2007), "Supervised Rank Aggregation," in *Proceedings of the 16th International Conference on World Wide Web*, ACM, pp. 481–490. [730]
- Luce, R. D. (1959), *Individual Choice Behavior: A Theoretical Analysis*, New York: Wiley. [730]
- Luo, J., Duggan, D. J., Chen, Y., Sauvageot, J., Ewing, C. M., Bittner, M. L., Trent, J. M., and Isaacs, W. B. (2001), "Human Prostate Cancer and Benign Prostatic Hyperplasia Molecular Dissection by Gene Expression Profiling," *Cancer Research*, 61, 4683–4688. [740,741]
- Mallows, C. L. (1957), "Non-null Ranking Models," *Biometrika*, 44, 114–130. [730,736]
- Meilă, M., and Chen, H. (2010), "Dirichlet Process Mixtures of Generalized Mallows Models," in *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI'10*, Arlington, VA: AUAI Press, pp. 358–367. [744]
- Meng, X.-L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278. [734]
- Murphy, T. B., and Martin, D. (2003), "Mixtures of Distance-Based Models for Ranking Data," *Computational Statistics & Data Analysis*, 41, 209–215. [736,744]
- Plackett, R. L. (1975), "The Analysis of Permutations," *Applied Statistics*, 24, 193–202. [730]
- Rubin, D. B. (2004), *Multiple Imputation for Nonresponse in Surveys* (Vol. 81), New York: Wiley. [744]
- Schimek, M. G., Budinska, E., Kugler, K. G., Svendova, V., Ding, J., and Lin, S. (2015), "TopKLists: A Comprehensive R Package for Statistical Inference, Stochastic Aggregation, and Visualization of Multiple Omics Ranked Lists," *Statistical Applications in Genetics and Molecular Biology*, 14, 311–316. [736]
- Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [736]
- Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., and Lander, E. S. (2002), "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, 1, 203–209. [740]
- Smith, B. B. (1950), "Discussion of Professor Ross's Paper," *Journal of the Royal Statistical Society, Series B*, 12, 53–56. [730,731]
- Thurstone, L. L. (1927), "A Law of Comparative Judgment," *Psychological Review*, 34, 273. [730]
- True, L., Coleman, I., Hawley, S., Huang, C.-Y., Gifford, D., Coleman, R., Beer, T. M., Gelmann, E., Datta, M., Mostaghel, E., Knudsen, B., Lange, P., Vessella, R., Lin, D., Hood, L., and Nelson, P. S. (2006), "A Molecular Correlate to the Gleason Grading System for Prostate Adenocarcinoma," *Proceedings of the National Academy of Sciences of the United States of America*, 103, 10991–10996. [730,740]
- Tusher, V., Tibshirani, R., and Chu, G. (2001), "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proceedings of the National Academy Science of the United States of America*, 98, 5116–21. [738]
- Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F., and Hampton, G. M. (2001), "Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer," *Cancer Research*, 61, 5974–5978. [730,740,741]