Statistics and Computing (2019) 29:335–349
https://doi.org/10.1007/s11222-018-9811-9

Rank aggregation using latent-scale distance-based models

Philip L. H. Yu[1] · Hang Xu[1]

Received: 27 November 2017 / Accepted: 3 April 2018 / Published online: 7 April 2018

Abstract
Rank aggregation aims at combining rankings of a set of items assigned by a sample of rankers to generate a consensus ranking. A typical solution is to adopt a distance-based approach to minimize the sum of the distances to the observed rankings. However, this simple sum may not be appropriate when the quality of rankers varies. This happens when rankers with different backgrounds may have different cognitive levels of examining the items. In this paper, we develop a new distance-based model by allowing different weights for different rankers. Under this model, the weight associated with a ranker is used to measure his/her cognitive level of ranking of the items, and these weights are unobserved and exponentially distributed. Maximum likelihood method is used for model estimation. Extensions to the cases of incomplete rankings and mixture modeling are also discussed. Empirical applications demonstrate that the proposed model produces better rank aggregation than those generated by Borda and the unweighted distance-based models.

Keywords Ranking data · Latent-scale distance-based model · Rank aggregration · Incomplete ranking

1 Introduction

The problem of rank aggregation is to combine a collection of rankings of a set of items to obtain a consensus ranking. It has many applications including combining rankings of sport teams obtained from various sources (Deng et al. 2014), ranking of webpages using meta-search engines (Aslam and Montague 2001) and gene ranking in bioinformatics (DeConde et al. 2006). One particular aspect of these kinds of data is that the number of items being examined is always fairly large, which makes the problem much harder to be tackled.

A popular approach to rank aggregation is to determine the consensus ranking by minimizing the sum of the distances from all the observed rankings to the consensus ranking. This method is actually equivalent to the maximum likelihood (ML) estimation of the consensus ranking under a distance-based model which postulates that the probability of observing a ranking of items decreases exponentially according to its distance to the consensus ranking, or more formally, called the *modal ranking*. Such estimation is acceptable only if the rankings assigned by the rankers are homogeneous, i.e., their rankings are identically distributed. However, rankers with different backgrounds or cognitive levels of examining the items may generate diverse quality of rankings, and hence, the rankings collected are likely non-identically distributed.

Examples of such heterogeneity of ranking abilities are abound. For instance, in ranking of NBA teams studied by Deng et al. (2014), some rankers are professional sport-ranking websites while some others are avid fans or infrequent watchers. In a meta-search study, Dwork et al. (2001) found that some search engines are more powerful than others, and some low-quality search engines referred as spam may provide a very low-quality ranking of webpages because of "paid placement" and "paid inclusion". In biological system study, Lin and Ding (2009) suggested that different omic-scale platform ranking data including DNA variations and RAomics have vast difference of information value regarding the interested problem.

So far, two methods have been available in the literature to tackle the heterogeneity problem due to the diverse quality

The research of Philip L.H. Yu was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 17303515).

✉ Hang Xu
xhang@hku.hk

Philip L. H. Yu
plhyu@hku.hk

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China

of rankings among the rankers. The first method is to allow different pre-specified weights to different rankers suggested by Aslam and Montague (2001) and Lin and Ding (2009). However, it is unclear how to design a suitable weighting scheme in practice. The second method is to allow different latent-scale parameters in the underlying ranking model, see for example Deng et al. (2014) and Lee et al. (2014) but they are not designed for distance-based ranking models. In this paper, we propose a new model called latent-scale distance-based models to tackle the above problem of quality heterogeneity in ranking.

The article is organized as follows. Section 2 introduces the class of distance-based models. In Sect. 3, we propose our latent-scale distance-based models, and adopt an EM algorithm to obtain the ML estimates of the model parameters. Note that when the number of items to be ranked gets large, the determination of the modal ranking and the computation in the E-step may become time consuming. We will devise efficient methods to overcome these problems. Simulation experiments will be conducted in Sect. 3 to demonstrate the performance. Extensions to incomplete rankings and models with multiple modal rankings are discussed in Sect. 4. Several real-world applications are included in Sect. 5. We give some concluding remarks in Sect. 6.

2 Review of distance-based models

In ranking t items, labeled $1, \dots, t$, a ranking π of the t items is a permutation function from $1, \dots, t$ to $1, \dots, t$. For example, $\pi(2) = 3$ means that item 2 ranked third and the inverse $\pi^{-1}(2) = 1$ means that the item ranked second is item 1. To understand people's perception and preference on items, various statistical models for ranking data have been developed over the past few decades. Among them, distance-based models have the advantages of being simple and elegant. Distance-based models (Fligner and Verducci 1986) assume that the probability of observing a ranking π drops exponentially according to its distance from an unknown modal ranking π_0:

$$\Pr(\pi \mid \lambda, \pi_0) = \frac{\exp[-\lambda d(\pi, \pi_0)]}{C(\lambda)}, \tag{1}$$

where $\lambda \geq 0$ is the dispersion parameter, $C(\lambda)$ is the normalizing constant, and $d(\pi, \pi_0)$ is an arbitrary right invariant[1] distance between π and π_0. Table 1 lists some popular distance measures between two rankings. Note that the model with $d = d_K$ (Kendall distance) is called Mallows' ϕ-model (Mallows 1957).

Under the model, the ranking probability is the greatest at the modal ranking π_0 and the probability of a ranking will decay the further it is away from π_0. The rate of decay is governed by the parameter λ. For a large value of λ, the distribution of rankings will be more concentrated around π_0. When λ becomes very small, the distribution of rankings will look more uniform.

It is easy to see that the normalizing constant $C(\lambda)$ in (1) can be written as $C(\lambda) = \sum_{\pi \in \mathcal{P}_t} \exp[-\lambda d(\pi, e)]$, where $\mathcal{P}_t$ be the set of all permutations of $\{1, \dots, t\}$ and $e = (1, 2, \dots, n)$. The closed form expression of $C(\lambda)$ only exists for some distances (Fligner and Verducci 1986), for instance,

$$\text{Kendall distance: } C_K(\lambda) = \prod_{i=1}^{t-1} \frac{1 - \exp[-(t-i+1)\lambda]}{1 - \exp(-\lambda)},$$

$$\text{Cayley distance: } C_C(\lambda) = \prod_{i=1}^{t-1} \left[(t-i)\exp(-\lambda) + 1\right].$$

If $C(\lambda)$ does not have a closed form, evaluating $C(\lambda)$ by summing over the $t!$ possible rankings in $\mathcal{P}_t$ becomes very computationally demanding when t becomes very large, say greater than 10. We will address this problem in Sect. 3.2.

Given a ranking data set $\Pi = \{\pi_k, k = 1, \dots, n\}$, the log-likelihood function of the distance-based model is:

$$\ell(\lambda, \pi_0) = -\lambda \sum_{k=1}^{n} d(\pi_k, \pi_0) - n \ln[C(\lambda)]. \tag{2}$$

The maximum likelihood estimate (MLE) of λ can be easily found because $C(\lambda)$ is deceasing and convex while the MLE $\hat{\pi}_0$ of π_0 is determined by minimizing the sum of distances over $\mathcal{P}_t$:

$$\hat{\pi}_0 = \underset{\pi_0 \in \mathcal{P}_t}{\operatorname{argmin}} \sum_{k=1}^{n} d(\pi, \pi_0). \tag{3}$$

As mentioned before, this is equivalent to the solution of the classical rank aggregation problem. However, such estimation based on sum of equally weighted (weight=1) distances may not be acceptable if the rankers come from different backgrounds and have different cognitive levels of examining the items. This motivates us to develop a new ranking model in the next section to take the diverse quality of rankers into consideration.

3 Latent-scale distance-based model

Suppose that the ranking data Π consists of rankings $\{\pi_k, k = 1, \dots, n\}$ from n rankers. Under the proposed

[1] A distance $d(\pi, \sigma)$ between two rankings π and σ is said to be right invariant if and only if for any ranking τ, $d(\pi, \sigma) = d(\pi \circ \tau, \sigma \circ \tau)$, where $\pi \circ \tau(i) = \pi(\tau(i))$.

Table 1 Some popular distance measures

Distance	Formula
Kendall	$d_K(\pi,\sigma)=\sum_{i<j} I\{[\pi(i)-\sigma(i)][\pi(j)-\sigma(j)]<0\}$
Spearman	$d_S(\pi,\sigma)=\sum_{i=1}^{k}[\pi(i)-\sigma(i)]^2$
Spearman's Footrule	$d_F(\pi,\sigma)=\sum_{i=1}^{k}\lvert\pi(i)-\sigma(i)\rvert$
Hamming	$d_H(\pi,\sigma)=t-\sum_{i=1}^{t}\sum_{j=1}^{t} I(\pi(i)=j)I(\sigma(i)=j)$
Ulam*	$d_U(\pi,\sigma)=t-LIS(\pi\circ\sigma^{-1})$
Cayley	$d_C(\pi,\sigma)=$ minimal number of transpositions needed to transform π to σ

Note that $LIS(\pi)$ is the length of the longest increasing subsequence of π

latent-scale distance-based model, we assume that ranker k has his/her own dispersion parameter λ_k, and conditionally on λ_k, the probability of observing a ranking π_k assigned by ranker k under the new model is given as:

$$\Pr(\pi_k \mid \lambda_k,\pi_0)=\frac{\exp[-\lambda_k d(\pi_k,\pi_0)]}{C(\lambda_k)},\quad k=1,\ldots,n.$$

Introducing different λ_k's can allow that the rankers with higher cognitive levels of examining the items can have larger values of λ_k's so that their rankings are more likely to be closer to the modal ranking π_0 while those rankers with little knowledge about the items can have smaller values of λ_k's and hence their rankings are more likely to be scattered away from π_0. As the background of the rankers are usually unknown, we assume that under the new model, their dispersion parameters λ_k's are randomly drawn from an exponential distribution with an unknown mean λ, denoted by $Exp(\lambda)$. This setting is inherent in Fligner and Verducci (1990), where it is introduced in a Bayesian context as the conjugate prior for the scale.

Note that λ represents the overall mean level of professionalism of the rankers and π_0 represents the consensus ranking common to all the rankers. In the next subsection, we will derive an efficient EM-type algorithm to obtain the ML estimates of the model parameters.

3.1 ML estimation of latent-scale distance-based model

In this subsection, we aim at developing a Expectation-Maximization (EM) algorithm of finding the ML estimates for the latent-scale distance-based model which could work well for any given distance measure and even for a large number (t) of items being ranked, say $t=100$.

Given the observed data $\Pi=\{\pi_k, k=1,\ldots,n\}$, denote $\Lambda=\{\lambda_k, k=1,\ldots,n\}$ be the collection of the latent dispersion variables for the n rankers. Let $\theta=(\lambda,\pi_0)$ be the set of parameters of interest. Augmenting Λ into Π to form the complete data, the complete-data log-likelihood function is found to be:

$$\ell_c(\theta \mid \Pi,\Lambda)=-\sum_{k=1}^{n}[\lambda_k d(\pi_k,\pi_0)+\ln C(\lambda_k)] \\ -n\ln\lambda-\sum_{k=1}^{n}\frac{\lambda_k}{\lambda}. \qquad (4)$$

The E-step here only involves computation of the conditional expectation of the λ_k's given Π and θ. Note that it is not required to compute $E[\ln[C(\lambda_k)]\mid\Pi,\theta]$ as it will not be needed in the M-step. So to compute $E[\lambda_k\mid\pi_k,\theta]$ ($k=1,\ldots,n$), we first use the Metropolis–Hastings (MH) algorithm to obtain random draws of λ_k from the conditional distribution of λ_k given π_k and θ:

$$f(\lambda_k\mid\pi_k,\theta)\propto\frac{\exp\left[-\lambda_k\left(\frac{1}{\lambda}+d(\pi_k,\pi_0)\right)\right]}{C(\lambda_k)}.$$

By implementing the MH algorithm, we choose an independent proposal distribution from $Exp(\lambda)$ with density g, where λ is obtained from θ. Given θ and an initial value $\lambda_k^{(0)}$ (say, drawn at random from g), the MH algorithm sample $\lambda_k^{(s+1)}$ at the $(s+1)$th iteration is as follows:

$$\lambda_k^{(s+1)}=\begin{cases}\lambda_k^* & \text{with probability } \min\left\{1,R(\lambda_k^{(s)},\lambda_k^*)\right\}\\ \lambda_k^{(s)} & \text{otherwise}\end{cases} \qquad (5)$$

where $R(\lambda_k^{(s)},\lambda_k^*)=\frac{f(\lambda_k^*|\pi_k,\theta)g\left(\lambda_i^{(s)}\right)}{f\left(\lambda_k^{(s)}|\pi_k,\theta\right)g(\lambda_i^*)}$. After generating S random draws $\left\{\lambda_k^{(s)},s=1,\ldots,S\right\}$, $E[\lambda_k\mid\pi_k,\theta]$ can be approximated by taking the average of these random draws, $\frac{1}{S}\sum_{s=1}^{S}\lambda_k^{(s)}$.

The M-step is to update the estimate $\hat{\theta}$ by maximizing the conditional expectation of the complete-data log-likelihood $\ell_{\text{com}}(\theta\mid\Pi,\Lambda)$ given Π and $\hat{\theta}^{(s)}$, the estimate of θ obtained at the $(s+1)$th EM iteration. It can be seen that the new estimate $\hat{\theta}^{(s+1)}=(\hat{\lambda}^{(s+1)},\hat{\pi}_0^{(s+1)})$ in M-step is given by:

$$\hat{\lambda}^{(s+1)} = \frac{1}{n}\sum_{k=1}^{n} E\left[\lambda_k \mid \pi_k, \hat{\theta}^{(s)}\right]$$

$$\hat{\pi}_0^{(s+1)} = \operatorname*{argmin}_{\pi_0 \in \mathcal{P}_t} \sum_{k=1}^{n} E\left[\lambda_k \mid \pi_k, \hat{\theta}^{(s)}\right] d(\pi_k, \pi_0). \qquad (6)$$

The new set of $\hat{\theta}$ is then used for calculation of the conditional expectation of the λ_k's in the E-step and the algorithm is iterated until convergence is attained. Unlike (3), our estimate of π_0 minimizes the sum of weighted distances by taking into consideration the individual differences in the conditional expectation of the dispersion parameters λ_k's.

3.2 Computational problems for large number of items

When the number (t) of items to be ranked becomes large, two computational problems will arise. First of all, in calculating Metropolis–Hastings ratio $R(\lambda_k^{(s)}, \lambda_k^*)$ in (5) for large t, it requires evaluating the ratio of two normalizing constants, $C(\lambda_k^*)$ and $C(\lambda_k^{(s)})$, which is computational demanding except for a few distances only (See Sect. 2). Secondly, the exhaustive search algorithm of $\hat{\pi}_0^{(s+1)}$ in the M-step (see (6)) for large t is practically infeasible because the number of possible rankings in $\mathcal{P}_t$ is too large. For example, $\mathcal{P}_{40}$ contains around 10^{47} rankings and enumerating all possible rankings in $\mathcal{P}_{40}$ is certainly unrealistic. Instead of using exhaustive search, Busse et al. (2007) suggested a local neighborhood search algorithm by searching the solution from the permutations within one Cayley distance only. However, our simulation in later section found that this method may cause the $\hat{\pi}_0^{(s+1)}$ stuck at a local minimum and cannot reach the global minimum.

3.2.1 Estimating the ratio of normalizing constants in the Metropolis–Hastings algorithm

To compute the MH ratio $R(\lambda_k^{(s)}, \lambda_k^*)$ in (5), we need to compute the ratio of two normalizing constants evaluated at $\lambda_i^{(s)}$ and λ_i^*:

$$R(\lambda_k^{(s)}, \lambda_k^*) = \frac{C(\lambda_k^{(s)})}{C(\lambda_k^*)} = \frac{\sum_{\pi \in \mathcal{P}_t} \exp\left[-\lambda_k^{(s)} d(\pi, e)\right]}{\sum_{\pi \in \mathcal{P}_t} \exp\left[-\lambda_k^* d(\pi, e)\right]}. \qquad (7)$$

Note that $C(\lambda)/t! = \frac{1}{t!}\sum_{\pi \in \mathcal{P}_t} \exp\left[-\lambda d(\pi, e)\right]$ can be treated as an expectation of $h_\lambda(\pi) = \exp(-\lambda d(\pi, e))$ over a uniform distribution over $\mathcal{P}_t$. Using the idea of importance sampling, $R(\lambda_k^{(s)}, \lambda_k^*)$ is estimated by:

$$\hat{R}(\lambda_k^{(s)}, \lambda_k^*) = \frac{\sum_{\pi \in \mathcal{S}} \exp\left[-\lambda_k^{(s)} d(\pi, e)\right]}{\sum_{\pi \in \mathcal{S}} \exp\left[-\lambda_k^* d(\pi, e)\right]}, \qquad (8)$$

where $\mathcal{S}$ is a set of rankings drawn uniformly from $\mathcal{P}_t$. In order to efficiently obtain a set of rankings uniformly distributed over the huge space $\mathcal{P}_t$ when t is moderate or large, we adopt the quasi-random number generators with low discrepancy (Niederreiter 2010) to generate $t \times 1$ vectors of numbers uniformly distributed in the t-dimensional unit-hypercube, and then order the t numbers in each vector to form a ranking.

3.2.2 Searching $\hat{\pi}_0^{(i+1)}$ over $\mathcal{P}_t$ in the M-step

The minimization problem in (3) and (6) is known to be NP-hard in the literature of combinatorial optimization (see, e.g., Ali and Meilă 2012). To circumvent this difficulty, Aledo et al. (2013) used genetic algorithms (GA) in the case of distance-based model based on Kendall distance and found that GA outperforms existing algorithms such as branch and bound (BB). However, they reported that the CPU time used by GA grows with the increasing of t and is 9.6 higher than that used by the BB algorithm for $t = 50$ and $\lambda = 0.2$. Here, we propose to use a faster algorithm— simulated annealing, to find the global solution of the minimization problem.

Simulated annealing (SA) algorithm proposed by Kirkpatrick et al. (1983) and Černý (1985) is a stochastic search technique. It begins iterating with a high "temperature" so that the algorithm is allowed to explore the solution space so as to move to any candidate position no matter it is better or worse than the current best solution. At the subsequent iterations (i.e., cooling to a lower temperature), the algorithm becomes more restrictive to search and is more likely to accept solutions better than the current best solution. Figure 1 shows our SA algorithm for the minimization problem with objective function shown in (3) or (6).

The functions α and β used in the SA algorithm (see Fig. 1) are designed, respectively, to control the speed of temperature cooling and the amount of candidate exploration at each fixed

```
1: Set $m \leftarrow 0$ and $j \leftarrow 0$. Given an initial temperature $\tau_0$
   and an initial solution $\pi^{[0]}$.
2: Simulate a candidate solution $\pi_0^*$ within the neighbor-
   hood of $\pi_0^{[m]}$, according to a proposal density $g^{[m]}(\cdot \mid \pi_0^{[m]})$.
3: Accept $\pi_0^{[m+1]} = \pi_0^*$ with probability:
       $\min\left\{1, \exp\left[\frac{f(\pi_0^{[m]}) - f(\pi_0^*)}{\tau_j}\right]\right\}.$
   Otherwise, set $\pi_0^{[m+1]} = \pi_0^{[m]}$. Set $m \leftarrow m + 1$.
4: Repeat Steps 2 and 3 a total of $m_j$ times.
5: Set $j \leftarrow j + 1$. If $j < j_{\max}$, update $\tau_j = \alpha(\tau_{j-1})$ and
   $m_j = \beta(m_{j-1})$, and go to Step 2.
```

Fig. 1 SA algorithm for (latent-scale) distance-based models with objective function f shown in (3) or (6)

temperature. In this paper, we choose $\tau_j = 0.95\tau_{j-1}$, $m_j = m_{j-1}$ with $\tau_0 = 100$ and $m_0 = 200$.

As seen from the SA algorithm shown in Fig. 1, a new candidate solution is definitely accepted when it is superior to the current solution and possibly accepted even when it is inferior, particularly in the early iterations. Such stochastic search provides SA an opportunity to escape from local minima so as to reach the global minimum.

To apply the SA algorithm, we need to simulate the new candidate solution π_0^* based on a proposal density $g^{[m]}(\cdot \mid \pi_0^{[m]})$. To do so, we swap two randomly selected elements in $\pi_0^{[m]}$. To be sure to reach the global minimum, the cooling process governed by the function α has to be slow. To avoid unnecessary iterations of the SA algorithm, the algorithm ends if $\pi_0^{[m]}$ remains unchanged for five consecutive temperature states. After finishing the SA algorithm, we obtain $\hat{\pi}_0^{(s+1)}$, the $(s+1)$th EM iterate of π_0, and hence the M-step is completed. The EM algorithm continues by iterating the E-step and M-step recursively. It is found that the EM algorithm converges very quickly and it stops in less than 20 iterations in all our simulation experiments and applications.

3.3 Simulation experiments

3.3.1 Simulation study 1: Estimating the ratio of normalizing constants via the importance sampling

In this section, we conduct a simulation study to evaluate the performance of important-sampler estimator (8) for the ratio of normalizing constants in (7). Note that the normalizing constant $C(\lambda)$ has a closed form for Kendall and Cayley distances. We here consider four distance measures: Spearman, Footrule, Hamming and Ulam distances. Due to different scales for these distances, the distances are normalized to have the same maximum value of t: $d^*(\pi, e) = t\, d(\pi, e) / \max d(\pi, e)$ so that we can consider the same set of the values of λ for different distances in the simulation study. More specifically, we study the performance of the ratio estimator

$$\hat{R}(\lambda_1, \lambda_2) = \frac{\hat{C}(\lambda_1)}{\hat{C}(\lambda_2)} = \frac{\sum_{\pi \in \mathcal{S}} \exp\left[-\lambda_1 d^*(\pi, e)\right]}{\sum_{\pi \in \mathcal{S}} \exp\left[-\lambda_2 d^*(\pi, e)\right]},$$

where $\mathcal{S}$ is a set of rankings drawn uniformly from $\mathcal{P}_t$ based on a quasi-random (QR) number method mentioned in Sect. 3.2.1.

In this simulation study, we consider a set of different combinations of (λ_1, λ_2), denoted by Ω_f, such that $\lambda_1 = \lambda_2(1 - f)$, and λ_2 is chosen to be one of the 20 equally spaced values from 0.1 to 2. Here, f represents the gap (or percentage difference) between λ_1 and λ_2 and three different gaps, $f = 0.1, 0.2$ and 0.5, are considered.

Instead of comparing the ratio estimate with its exact value which is computationally expensive to obtain when the number (t) of items becomes large (say $t > 10$), we study its performance based on the maximum relative change of the log-ratio estimates by increasing the number of QR samples in powers of ten. More specifically, between two consecutive QR sample sizes, say $old = 10^{i-1}$ vs $new = 10^i$, we calculate the maximum relative change as

$$\max_{(\lambda_1, \lambda_2) \in \Omega_f} \left\{ \frac{\ln \hat{R}_{\text{new}}(\lambda_1, \lambda_2) - \ln \hat{R}_{\text{old}}(\lambda_1, \lambda_2)}{\ln \hat{R}_{\text{old}}(\lambda_1, \lambda_2)} \right\}.$$

Table 2 shows the maximum relative change of the log-ratio estimates between two normalizing constants for various distances and gaps. It can be seen that our method yields a good estimate of the ratio between two normalizing constants using a number of samples which is much smaller than the number of terms in the exact sum ($20! \approx 10^{18}$ for $t = 20$ and $40! \approx 10^{48}$ for $t = 40$). For the cases of Hamming and Ulam distances, a QR sample size of 10^6 is enough to get an accurate estimate of the ratio. For the cases of Spearman and Footrule distances, we need a larger sample size, say 10^7, in order to have a maximum relative change less than 1%. All the computations were performed on a PC, and the time spent to obtain each value of maximum relative change in Table 2 for $t = 40$ took less than three minutes.

3.3.2 Simulation study 2: Searching over $\mathcal{P}_t$ via the SA algorithm

In this section, we conduct a simulation to compare the efficiency of the simulated annealing (SA) algorithm proposed in Sect. 3.2.2 with the local neighborhood search method in determining the MLE of π_0 in (3).

First of all, ranking data are simulated from a distance-based model with $\lambda^* = 0.3$ and $\pi_0^* = e$. Three distance measures are considered and they are Kendall, Spearman and Footrule. See Appendix for the procedure of simulating ranking data from distance-based model. Then we apply both the SA and local neighborhood search (NS) methods to obtain the MLE $\hat{\pi}_0$ based on the same initial π_0 estimate generated uniformly at random. For each method, we can compute the average log-likelihood difference, $ALD = (\ell(\lambda^*, \hat{\pi}_0) - \ell(\lambda^*, \pi_0^*))/n$, where $\ell(\lambda, \pi_0)$ is defined in (2), and n is the size of data. The more the value of ALD is closer to zero, the better is the searching method.

The simulation is repeated 30 runs and in each run, the MLE estimates of π_0 obtained by the SA and NS methods are recorded and their ALD results are listed in Table 3. It can be seen that our simulated annealing method always performs better than the local neighborhood search method even when the number of items t becomes large. The local

Table 2 Maximum relative change of the log-ratio estimator between two normalizing constants for various distances and gaps

Gap	Spearman distance			Footrule distance			Hamming distance			Ulam distance		
	10%	20%	50%	10%	20%	50%	10%	20%	50%	10%	20%	50%
$t = 20$												
10^3	0.031	0.141	0.067	0.155	0.029	0.044	0.013	0.021	0.010	0.037	0.027	0.026
10^4	0.100	0.048	0.021	0.055	0.064	0.010	0.012	0.024	0.004	0.010	0.002	0.006
10^5	0.011	0.007	0.009	0.003	0.003	0.009	0.007	0.005	0.002	0.004	0.000	0.003
10^6	0.002	0.004	0.001	0.002	0.002	0.003	0.007	0.000	0.000	0.001	0.003	0.001
10^7	0.001	0.001	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.000	0.000
$t = 40$												
10^3	0.033	0.036	0.017	0.087	0.066	0.047	0.012	0.009	0.011	0.039	0.014	0.010
10^4	0.007	0.066	0.013	0.010	0.126	0.013	0.008	0.013	0.004	0.080	0.020	0.005
10^5	0.036	0.037	0.013	0.060	0.039	0.017	0.003	0.007	0.002	0.129	0.008	0.004
10^6	0.011	0.008	0.027	0.016	0.023	0.013	0.008	0.008	0.001	0.008	0.006	0.001
10^7	0.008	0.003	0.009	0.019	0.003	0.008	0.008	0.006	0.000	0.006	0.003	0.010

Note that the row for 10^i represents the maximum relative change of the log-ratio estimator between two normalizing constants estimated based on 10^i and 10^{i-1}

Table 3 Results of ALD for the simulated annealing (SA) algorithm and the local neighborhood search (NS) method for various sample sizes n and the number of items t

Number of items, t	Sample size,	Kendall		Spearman		Footrule	
	n	SA	NS	SA	NS	SA	NS
10	100	0.00	−0.01	0.21	−0.24	0.38	−2.19
	300	0.00	0.00	0.03	−0.53	0.00	−1.32
	500	0.00	0.00	0.01	−0.49	0.00	−2.95
20	100	0.04	−12.61	0.42	−2.96	2.00	−20.94
	300	0.00	0.00	0.33	−1.43	0.57	−20.19
	500	0.00	0.00	0.18	−1.41	1.54	−18.71
50	100	0.01	−107.07	7.65	−9.35	36.41	−165.60
	300	−0.05	−81.62	2.69	−11.00	13.39	−202.12
	500	−0.11	−78.70	1.19	−9.92	10.43	−199.48
100	100	−0.50	−613.21	33.86	−27.44	30.93	−940.31
	300	−0.93	−483.94	16.86	−40.92	26.90	−899.08
	500	−0.91	−22.06	10.27	−44.14	8.67	−924.08

neighborhood search method generally performs satisfactory for small t but its performance deteriorates heavily when t gets large.

Note that the determination of the MLE of π_0 requires minimizing the sum of distances $\sum_{k=1}^{n} d(\pi, \pi_0)$ over $\mathcal{P}_t$. Figure 2 shows the iteration details of the minimization of the average Kendall distance $\frac{1}{n}\sum_{k=1}^{n} d(\pi, \hat{\pi}_0)$ based on two randomly chosen initial estimates of π_0, where the data are simulated from the Kendall distance-based model with $t = 20$ or 50, $n = 100$, and $\pi_0^* = e$. It can be seen that regardless of the choice of initial estimate of π_0, simulated annealing algorithm can achieve the global optimal solution (labeled by the green line) but the local neighborhood search method can sometimes converge to a local optimal solution. It is not surprising to see that the average Kendall distance for the simulated annealing may increase slightly during the iteration as the simulated annealing allows to have some probability to adopt some inferior candidates so that it is more likely to reach the global minimum.

4 Model extensions to incomplete rankings and models with multiple modal rankings

4.1 Incomplete rankings

Incomplete (or partial) ranking data are commonly seen particularly when assessing an item takes much effort and time. Instead of ranking all t items, individuals may be asked to

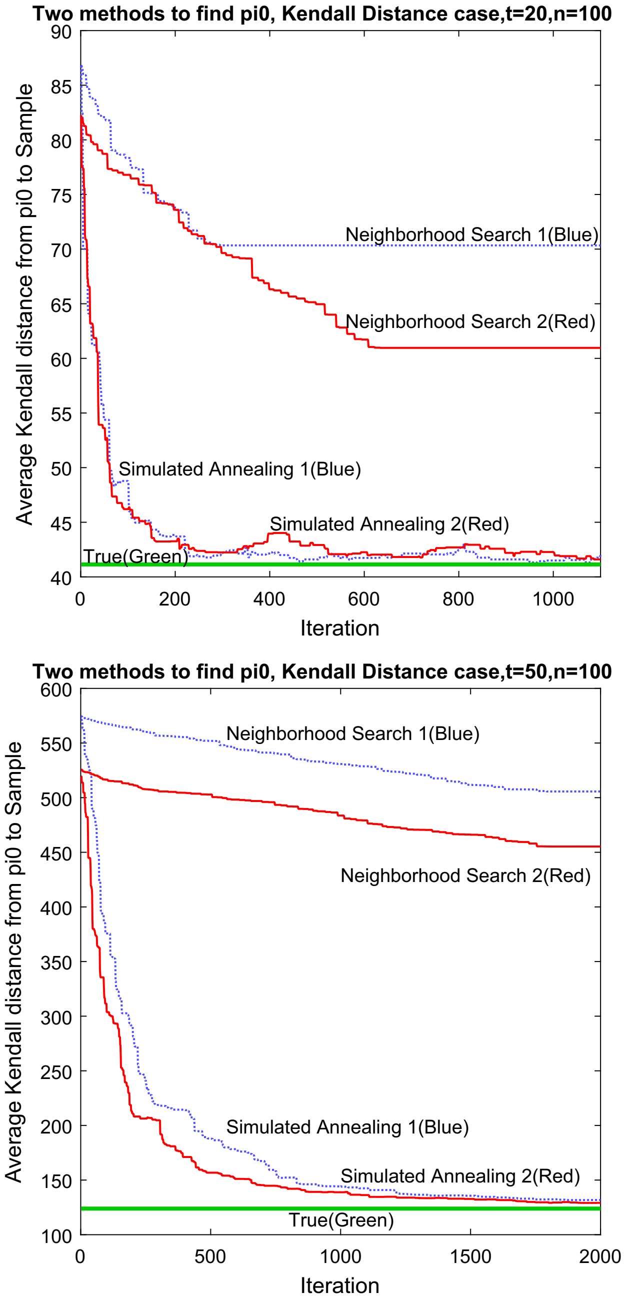

Fig. 2 The iteration details of the minimization of the average Kendall distance based on two randomly chosen initial estimates of π_0 (blue and red lines), where the data are simulated from the Kendall distance-based model with $t = 20, 50, n = 100$, and $\pi_0^* = e$. The minimization is done using the simulated annealing algorithm and the local neighborhood search method. The green line represents the average Kendall distance with $\pi_0 = \pi_{0\,true}$. (Color figure online)

rank the top few items only (called *top-q rankings*) or to rank the items within a subset of the t items only (called *subset rankings*). Note that discrete choices and paired comparisons are special cases of the incomplete ranking data. Let us first extend the notation for incomplete ranking. For instance $\pi^* = (3, 1, 2, 3)$ represents a top two ranking with item 2 ranked first and item 3 ranked second, $\pi^* = (-, 1, 2, -)$ refers to a subset ranking with item 2 more preferred than item 3 and items 1 and 4 unranked, and $\pi^* = (1, 1, 2, -)$ is a combination of these two types.

The incomplete rankings can be treated as a missing data problem, which can be solved by augmenting the missing ranks in Gibbs sampling in the Monte Carlo E-step. Let $\{\pi_1^*, \dots, \pi_n^*\}$ be the observed data of n incomplete rankings, and let $\{\pi_1, \dots, \pi_n\}$ be their corresponding complete rankings. What we need to do is to include one more step in the Gibbs sampling by sampling π_k from its full-conditional distribution $f(\pi_k|\pi_k^*, \lambda_k, \pi_0)$ and all the other steps will be unchanged as if the complete rankings are observed. To sample from $f(\pi_k|\pi_k^*, \lambda_k, \pi_0)$, we have to introduce the concept of compatibility for an incomplete ranking. Let $S(\pi_k^*)$ be the set of complete rankings compatible with π_k^* so that the rank orders are preserved. For example for $\pi^* = (2, -, 3, 4, 1)$, $S(\pi^*) = \{(2, 5, 3, 4, 1), (2, 4, 3, 5, 1), (2, 3, 4, 5, 1), (3, 2, 4, 5, 1), (3, 1, 4, 5, 2)\}$. Notice that

$$f(\pi_k|\pi_k^*, \lambda_k, \pi_0)) \propto f(\pi_k|\lambda_k, \pi_0), \ \pi_k \in S(\pi^*).$$

Obviously, direct sampling from this distribution will be tedious when the size of the compatible set $S(\pi^*)$ becomes large. Instead, we can use the Metropolis–Hastings algorithm to draw samples from this distribution with the proposed candidates generated uniformly from $S(\pi^*)$. The idea of introducing compatible rankings allows us to treat different kinds of incomplete rankings easily. It is easy to sample uniformly from the compatible rankings since we just need to sample the missing ranks under different situations without listing our all members in $S(\pi^*)$. For instance, sampling a complete ranking based on an observed subset ranking of q items can be done by drawing $(t - q)$ integers at random from $\{1, 2, \dots, t\}$ for the missing ranks and then placing the observed subset ranking back to the complete ranking with the order preserved.

When the distance is chosen to be Kendall distance and the data is top-q rankings, Fligner and Verducci (1986) suggests that the Kendall distance can be extended as $d(\pi, \pi_0) = \sum_{i=1}^{t-1} V_i(\pi \circ \pi_0^{-1})$, where $V_i(\pi) = \sum_{j>i} I\{\pi^{-1}(i) - \pi^{-1}(j) > 0\}$ and $\pi \circ \pi_0^{-1} = \pi(\pi_0^{-1})$. For top-$q$ rankings, $V_1, V_2, \dots, V_q$ only depend on π^*. So the probability of observing π^* can be written as

$$\begin{aligned}\Pr(\pi_k^*) &= \exp\left[-\lambda_k \sum_{i=1}^{q} V_i(\pi \circ \pi_0^{-1})\right] \\ &\quad \times \sum_{\pi \in S(\pi_k^*)} \frac{\exp[-\lambda_k \sum_{i>q} V_i(\pi \circ \pi_0^{-1})]}{C(\lambda_k)} \\ &= \frac{\exp[-\lambda_k \sum_{i=1}^{q} V_i(\pi \circ \pi_0^{-1})]}{C_{\text{new}}(\lambda_k, q)}\end{aligned}$$

where $C_{\text{new}}(\lambda_k, q) = \prod_{i=1}^{q}[1-\exp\{-(q-i+1)\lambda_k\}]/[1-\exp(-\lambda_k)]$. For Cayley distance, this rearrangement also exists and similar methods can be applied to the partially ranked data. The EM algorithm of this induced model can be easily derived. Note that the EM algorithm in this special case is much faster, since we don't need to sampling complete rankings in the $S(\pi_k^*)$.

4.2 Models with multiple modal rankings

So far, we assume that there is a single modal ranking π_0. However, it is natural to have different views for different groups of rankers and hence, we need to adopt a mixture modeling framework to allow clusters with distinct modal rankings. Inspired by Murphy and Martin (2003) who extended the use of mixture models to simple distance-based models, we extend our latent-scale models to mixture of latent-scale distance-based models for ranking data.

Given a ranking data set $\Pi = \{\pi_k, k = 1, \ldots, n\}$, denote $\Lambda = \{\lambda_{kg}, k = 1, \ldots, n, g = 1, \ldots, G\}$ be the collection of all latent dispersion variables. The probability of observing ranking data π_k under a mixture of latent-scale distance-based models with G clusters is:

$$P(\pi_k|\Lambda) = \sum_{g=1}^{G} p_g \frac{\exp[-\lambda_{kg} d(\pi_k, \pi_{0g})]}{C(\lambda_{kg})},$$

where p_g is the proportion of cluster g, and the rankings in cluster g follows a latent-scale distance-based model with modal ranking π_{0g} and latent-scale parameters λ_{kg}'s generated independently from $\text{Exp}(\lambda_g)$.

Let $\theta = \{\lambda_g, \pi_{0g}, p_g, g = 1, \ldots, G\}$ be the set of parameters of interest of this mixture model. To obtain the MLE of θ using the EM algorithm, we need to augment into the complete data stated in Sect. 3.1 an additional latent variable $z_k = (z_{k1}, \ldots, z_{kG})$, the membership variable for ranker i which is defined as: $z_{kg} = 1$ if ranker i belongs to cluster g, otherwise $z_{kg} = 0$.

The steps of implementing the EM algorithm for this mixture model is similar to the EM algorithm used in Sect. 3.1. At the E-step of the $(s+1)$th EM iteration, we need to evaluate $E(z_{kg} \mid \pi_k,\ \theta^{(s)})$ and $E(z_{kg}\lambda_{kg} \mid \pi_k,\ \theta^{(s)})$ which can be determined similarly using the Gibbs sampling. First we consider sampling from the full-conditional distribution of $z_k = (z_{k1}, \ldots, z_{kG}), k = 1, \ldots, n$:

$$P(z_{kg} = 1 \mid \Lambda,\ \pi_k,\ \theta^{(s)}) = \frac{\frac{p_g^{(s)} \exp(-\lambda_{kg} d(\pi_k, \pi_{0g}^{(s)}))}{C(\lambda_{kg})}}{\sum_{g'=1}^{G} \frac{p_{g'}^{(s)} \exp(-\lambda_{kg'} d(\pi_k, \pi_{0g'}^{(s)}))}{C(\lambda_{kg'})}}.$$

Then we consider sampling λ_{kg} from its full-conditional distribution. For $z_{kg} = 1$, the full-conditional density of λ_{kg}, $k = 1, \ldots, n, g = 1, \ldots, G$ is

$$f(\lambda_{kg} \mid z_{kg} = 1,\ \pi_k) \propto \frac{\exp\left[-\lambda_{kg}\left(\frac{1}{\lambda_g^{(s)}} + d(\pi_k, \pi_{0g}^{(s)})\right)\right]}{C(\lambda_{kg})}$$

and λ_{kg} can be simulated using similar MH algorithm stated in Sect. 3.1 while for $z_{kg} = 0$, it is easy to see that λ_{kg} can be simulated from $Exp(\lambda_g)$.

At the M-step of the $(s+1)$th EM iteration, we can update $\theta = \{\lambda_g, \pi_{0g}, p_g, g = 1, \ldots, G\}$ as follows: $\hat{\lambda}_g^{(s+1)} = \frac{\sum_{i=1}^{n}((\widehat{z\lambda})_{kg}^{(s+1)})}{\sum_{i=1}^{n} \hat{z}_{ig}^{(s+1)}}$,

$$\hat{\pi}_{0g}^{(s+1)} = \arg\min \sum_{k=1}^{n} (\widehat{z\lambda})_{kg}^{(s+1)} d(\pi_i, \pi_0),$$

and $\hat{p}_g^{(s+1)} = \frac{1}{n}\sum_{k=1}^{n} \hat{z}_{kg}^{(s+1)}$, where $\hat{z}_{kg}^{(s+1)} = E(z_{kg} \mid \pi_k,\ \theta^{(s)})$ and $(\widehat{z\lambda})_{kg}^{(s+1)} = E(z_{kg}\lambda_{kg} \mid \pi_k,\ \theta^{(s)})$ can be computed by the Monte Carlo integration from the above Gibbs sampling. Given the initial parameters for $\theta^{(0)}$, we alternatively run the E-Step and M-Step until the estimates converge. It is found that our EM algorithm converges very quickly within 20 iterations in our gene application.

5 Applications

5.1 Aggregating people's rankings

Consider the data collected in five ranking experiments by Lee et al. (2014) in the University of California Irvine. The participants of these experiments were undergraduates recruited from the human subjects pool of the university. In the experiments, the participants were asked to rank different items according to their knowledge. These items could be US presidents, NFL Superbowl teams, NBA teams, 10 Commandments, etc. The ranking assignments for US Presidents was to put them in chronological order; for NFL Superbowl teams and NBA teams, to put them in the order of their performances of the season; for the Ten Amendments to the Constitution, to put them in the order they appear; and for the Ten Commandments, to put them in the order adhered to by the Jewish and most Protestant religions. Note that for a specific task there are different ranking skills among the rankers in the experiments. All the data are downloaded from http://webfiles.uci.edu/mdlee/LeeSteyversMiller2014Data.zip.

As the ground truth ranking π_0 is finally available or was knowable to the participants when they are assigned the experiment, we can study the wisdom of the crowd effect, i.e., whether our proposed model is able to obtain an aggregated ranking that is close to the ground truth. In each of these experiments, we compare ten rank aggregation methods: Borda count method, Lee's method (2014), the MLE $\hat{\pi}_0$ of simple distance-based models and latent-scale distance-based models based on Kendall, Spearman, Footrule and Ulam distances, estimated using the simulated annealing algorithm in Sect. 3.2.2. Note that the Borda count method (1781) basically computes the average rank of each item assigned by all participants, and the aggregated ranking produced by the Borda count method is then simply generated by ordering the items according to their average ranks. Lee's method is a Bayesian Thurstonian model and the estimated ranking is obtained by ordering the mean ranks of the 1000 posterior utilities simulated by JAGS sampling.

The estimation procedure based on our Monte Carlo EM method converges very fast even for the case of $t = 44$. Usually it takes 5-7 EM iterations to meet the convergence criteria. Table 4 shows the Kendall distance between the ground truth ranking and the estimated aggregated ranking for each rank aggregation method. Of course, other distance measures can also be used for comparison of the methods. We choose Kendall distance here as it can be viewed as a measure of misclassifying the orders of all pairs of items. It can be seen from Table 4 that the latent-scale distance-based models always perform better than their corresponding simple distance-based models and the Borda count method. Our latent-scale models perform the best in four out of the five experiments. Lee's method produces similar performance to our LS model with Spearman distance. This may be because the utilities for the items under the Thurstonian model are independent normally distributed and their log-likelihood involves a squared Euclidean distance of the utilities to their means which resembles the Spearman distance used in our latent-scale distance-based model. Among all the distances considered, latent-scale models based on Kendall and Spearman distances generally perform the best or the second best.

To further illustrate the outstanding performance of the latent-scale Kendall distance-based model in these five experiments, Fig. 3 shows the distribution of the Kendall distances between people's rankings (blue histogram) and the true ranking (green circle). Note that the red circle represents the worst-possible ranking having the largest Kendall distance, and the dotted line shows the distribution of Kendall distance for a ranking generated at random. The aggregated rankings inferred by the latent-scale Kendall distance-based model, the Mallows model (i.e., the standard Kendall distance-based model) and the Borda count method are shown by a black circle labeled "R", a yellow circle labeled "M" and a blue circle labeled "B", respectively.

It can be seen from Fig. 3 that people's rankings indicated by the blue histograms seem not generated at random as the histograms are not close to the distribution for random rankings (dotted line), and there are large individual differences in their rankings. In all these experiments, our latent-scale Kendall distance-based models perform the best as their aggregated rankings are the closest to their ground truths, particularly when people have diverse background on examining the items such as the case of the NBA East 2010 season data. This is because both the Mallows model and Borda count method unrealistically assume the same weights for all rankings.

The diverse background of people is also evidenced by the inserted scatter plots which show the relationship between the conditional mean of λ_k for the kth individual in the final E-step and his/her Kendall distance from the true ranking. Recall that λ_k can be treated as a measure of cognitive level of examining the items inferred by the latent-scale distance-based model. It is found that λ_k is negatively related to the Kendall distance measure, indicating that people having a large value of λ_k are more knowledgeable about the items as they tend to provide better ranking with smaller Kendall distance from the ground truth.

Table 4 Performance comparison in terms of Kendall distance from the true rank

Model	All US presidents $t = 44, n = 26$	NFL 2010 season $t = 32, n = 40$	NBA east 2010 $t = 15, n = 148$	Ten amendments $t = 10, n = 78$	Ten commandments $t = 10, n = 78$
Borda count	78	158	36	6	12
Lee's method	59	156	**23**	2	9
Kendall	69	144	34	4	11
LS-Kendall	**56**	**132**	29	2	**7**
Spearman	68	148	35	3	9
LS-Spearman	66	140	26	1	8
Footrule	68	149	32	4	11
LS-Footrule	66	140	31	2	11
Ulam	117	169	38	2	8
LS-Ulam	91	208	36	**0**	8

LS is the result of latent-scale distance-based model with various distance measures. Number in bold shows the best performance among ten methods

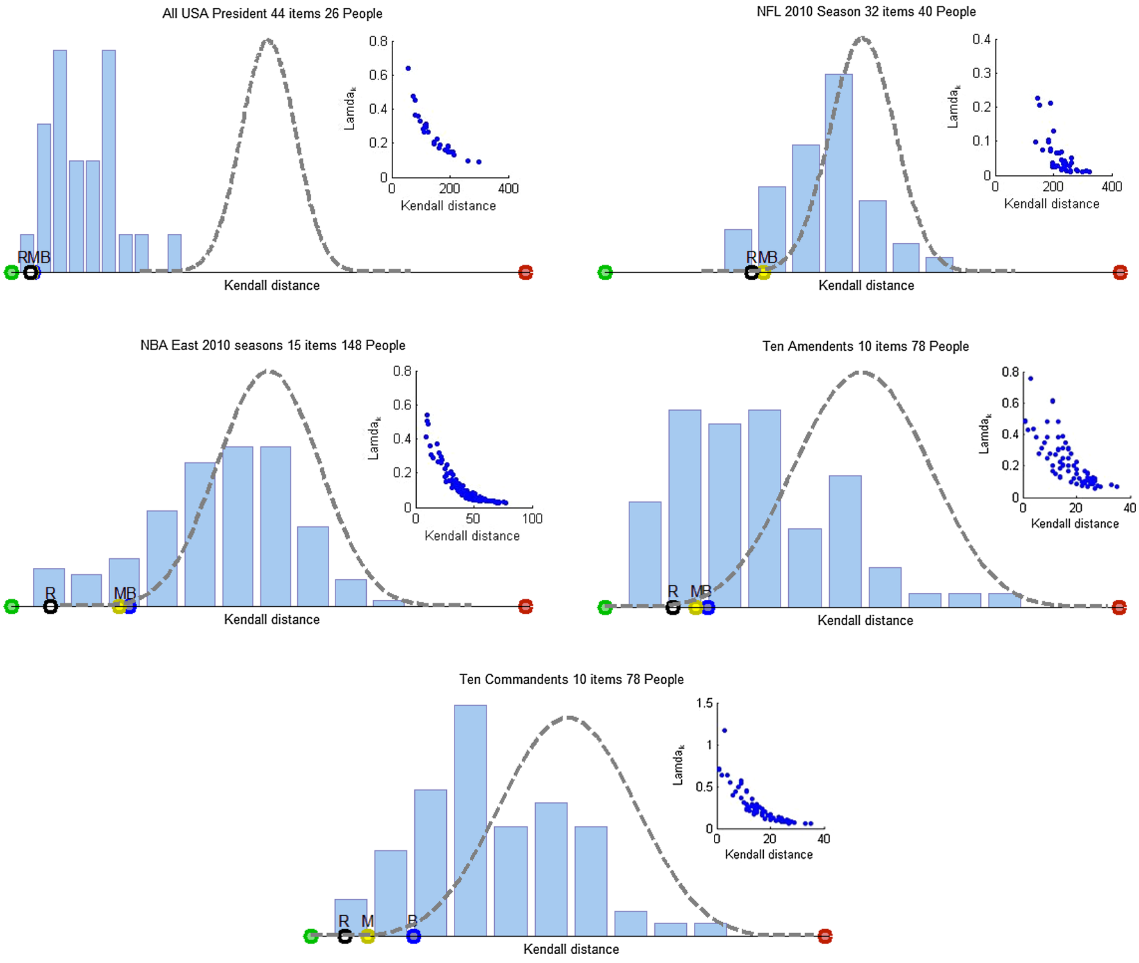

Fig. 3 Each of the above panels corresponds to a different ranking experiment. The distribution of Kendall distances between people' rankings and the true ranking (green circle) is shown by the blue histogram. Note that the red circle represents the worst-possible ranking having the largest Kendall distance, and the dotted line shows the distribution of Kendall distance for a ranking generated at random. The aggregated rankings inferred by the latent-scale Kendall distance-based model, the Mallows model and the Borda count method are shown by a black circle labeled "R" , a yellow circle labeled "M" and a blue circle labeled "B" , respectively. The inserted scatter plots show the relationship between the conditional mean of λ_k for the kth individual and his/her Kendall distance from the true ranking. (Color figure online)

To further explain why our proposed latent-scale Kendall distance-based model could outperform the Borda method and the aggregation based on Mallows model, we apply the multidimensional scaling (MDS) method (see Borg and Groenen 2005) to the Kendall distance matrix obtained from the ranking data on ten commandments. Figure 4 shows the two-dimensional MDS solution. The small blue dots represent the positions of rankings given by 78 rankers. The green circle labeled "True" is the true ranking. The rank aggregation results by our latent-scale model, Borda method, the Mallows model and Lee's Thurstonian Method are denoted by black circle labeled "R", blue circle labeled "B", yellow circle labeled "M" and red circle labeled "T" respectively. Among these methods, the aggregated ranking based on the estimated modal ranking from our latent-scale model is the closest to the true ranking. The aggregated rankings based on Mallows and Borda methods are farther way from the true ranking since they treat the rankings equally so that theiraggregated results are affected by the rankings (blue dots) at the right hand side of the graph.

5.2 Aggregating incomplete ranking data of NBA teams

In this application, we consider the NBA power ranking data studied by Deng et al. (2014) in which 34 judges ranked 30 NBA teams according to the results of 2011–2012 season. Six of them, (P1,…, P6), are complete rankings obtained from professional sports websites, and the other 28 rankings, (S1,…,S28), are collected from a sample of 28 Harvard students in a university survey, in which each student was asked to rank the best 8 NBA teams (top-8 rankings) in the 2011–2012 season based on his/her own knowledge. Each student has classified himself into

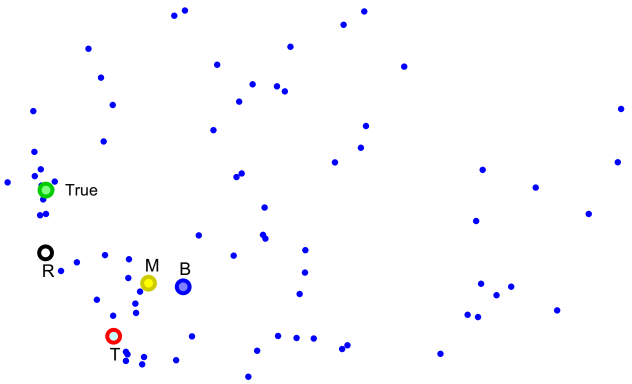

Fig. 4 Two-dimensional multidimensional scaling solution for the ranking data on ten commandments. The aggregated rankings inferred by the latent-scale Kendall distance-based model, the Mallows model, the Borda count method and the Lee's Thurstonian model are shown by a black circle labeled "R" , a yellow circle labeled "M" , a blue circle labeled "B" and a red circle labeled "T" respectively. (Color figure online)

one of the following four groups in the survey: (1) "Avid fans" who never missed NBA games, (2) "Fans" who watched NBA games frequently, (3) "Infrequent watchers" who watched NBA games occasionally, and (4) the "Not interested" who never watched NBA games in the past season. The true ranking of this task is arranged based on their performance of the full season: the top 16 teams reached the playoffs, the top eight survived the first round of the playoffs and so on. The bottom 14 teams and the tied teams in the playoffs are ranked by their winning percentages in the regular season games in 2011–2012.

In this rank aggregation experiment, we consider five rank aggregation methods: traditional Borda count aggregation, Lee's method (2014), Deng's method (2014), Mallows model (distance-based model with Kendall distance) and our latent-scale Kendall distance-based model. After the 2011–2012 NBA season, we can obtain the true ranking of 30 teams (Deng et al. 2014). Table 5 shows the Kendall distance between the ground truth ranking and the estimated aggregated ranking for each rank aggregation method. Among these methods, our latent-scale model with Kendall distance performs the best. Deng's method performs even worse than Borda count because there are many ties in Deng's estimated ranking, whereas Lee's method performs better than Borda count but poorer than our latent-scale model.

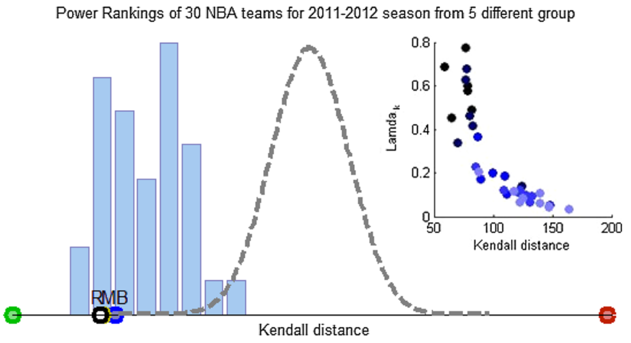

Fig. 5 Distribution of Kendall distances between participants' rankings and the true ranking for the power rankings of 30 NBA teams for 2011–2012 season. In the inserted scatter plot, the darkest points are data from the "professional websites" group and the brightest points are data from "not-interested" group. The darker the point's color is the more professional the group is. (Color figure online)

Figure 5 shows the distribution of Kendall distances between participants' rankings and the true ranking, together with the aggregated rankings based on the Borda count method, Mallows model and LS Kendall model. The Kendall distance here is defined as the average Kendall distance of the rankings in the compatible set from the true ranking. It is clearly seen from Fig. 5 that the observed rankings are not generated at random as the Kendall distribution of the observed rankings (blue histogram) is far away from the Kendall distribution of a random ranking (dotted line). Among the three methods, our latent-scale model (labeled by R) performs the best as it is the closest to the true ranking labeled by a green circle. The inserted scatter plot of Fig. 5 shows the plot of the conditional mean of λ_k for the kth individual against its Kendall distance to the true ranking. Note that the points are colored according to the professionalism of the five groups of judges in the order: Professional websites (darkest), Avid fans, Fans, Infrequent watchers and Not interested (brightest). It can be seen that an individual with a larger conditional mean of λ_k tends to have a darker point and a shorter Kendall distance to the true ranking, indicating that this individual tends to be more professional and understand better the true ranking.

Figure 6 and Table 6 show the conditional mean of λ_k for the five groups of judges. It is reasonable to see that the conditional means of λ_k for "Professional websites" group tend to the highest while those for "Not interested" group tend to be the smallest. In other words, this conditional mean of λ_k for the kth individual can

Table 5 Performance comparison in terms of Kendall distance from the true rank. LS-Kendall is the result of latent-scale Kendall distance-based model

Aggregating incomplete ranking data of NBA teams task					
Methods:	Borda count	Lee's method	Deng's method	Mallows model	LS-Kendall
Kendall distance from true rank	75	68	114	68	64

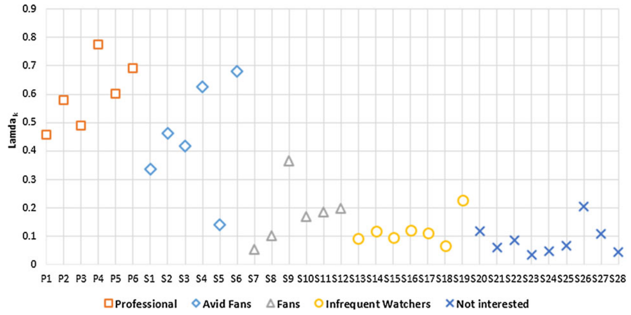

Fig. 6 Estimates of λ_k for five groups of rankers from the power rankings task of 30 NBA teams for 2011–2012 season

be viewed as his/her professional level. We also observe some interesting phenomenons in Fig. 6. The professional level of avid fans S4 and S6 are even higher than some professional websites. Based on their observed rankings, these two students almost successfully picked up all the top-8 teams in the playoffs (top 16). The professional level of avid fan S5 is much lower than the other students in the same group. This precisely reflects the fact that S5 gave high ranks to the two NBA teams, Warriors and Wizards, but these two teams failed to enter the playoffs.

Similar to our latent-scale model, Deng et al. (2014) also defined a kind of γ_k as a quality parameter for each ranker in their Bayesian aggregation approach. In their paper, they find that the quality parameter indicates the basket knowledge level of the rankers. However, we find that their estimated qualities for the NBA power ranking data are fairly different from our estimates of the λ_k's. For example, their estimated qualities of the professional group are much greater than those of the avid fans group. Comparing to our results, our estimates $\hat{\lambda}_k$'s for the professional group are not greatly larger than those for the avid fans group as evidenced from the raw data. Table 6 shows the number of mistakes made in picking up 8 teams survived in the first round of playoffs in NBA data. We can find that some rankers in the professional group (e.g., P2, P3, P4) even made more mistakes than Avid Fans group (e.g., S1, S2, S4, S6).

Comparing the estimated aggregated ranking of our model with the true ranking, we found that our latent-scale model makes only one mistake to pick up 16 teams survived in the playoffs (Top 16): picking Trail Blazers instead of Jazz into the playoffs. However, five out of the six professional websites made two mistakes. Also, our model successfully picks up the champion and the first runner-up team in the season but only two professional websites and 5 of the 28 students successfully picked up these two teams.

5.3 Aggregating and classification of disease subtypes of breast cancer patients based on their gene expressions

Using rank-based methods to solve gene related problems has seen a growing interest in bioinformatics (Naume et al. 2007). In this section, we illustrate the use of our proposed mixture model to analyze a ranked mRNA expression data set with 96 genes from 121 breast cancer patients who are categorized into two disease subtypes according to their ER/PgR-status: estrogen receptor negative (ER−, 41 patients) or positive (ER+, 80 patients). Our aim is to cluster the breast cancer patients into two groups based on their ranked gene expression data and study the classification performance using their actual disease subtypes. The data for 96 genes selected from the KEGG estrogen signaling pathway (Kyoto Encyclopedia of Genes and Genomes: hsa04915) (http://www.genome.jp/kegg/), are obtained from the Stanford Microarray Database (http://genome-www5.stanford.edu/). The ranked normalized log 2 transformed gene expression ratios are retrieved from SMD as our gene ranking data with $t = 96$ and $N = 121$.

We fit a mixture of $G = 2$ latent-scale Kendall distance-based models to the gene ranking data (without using the actual disease subtypes). Table 7 shows the ML estimates of the model parameters. It can be seen that the estimated

Table 6 Numbers of mistakes made in picking up 8 teams survived in the first round of playoffs in NBA data and $\hat{\lambda}_k$ for different groups of rankers

Group	Professional						Avid fans						Fans					
Ranker	P_1	P_2	P_3	P_4	P_5	P_6	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}
Mistakes	2	4	4	4	3	2	2	2	3	2	5	2	6	3	3	3	3	3
$\hat{\lambda}_k$	0.45	0.58	0.48	0.77	0.60	0.69	0.34	0.46	0.42	0.63	0.14	0.68	0.05	0.10	0.36	0.17	0.19	0.20

Group	Infrequent watchers							Not interested individuals								
Ranker	S_{13}	S_{14}	S_{15}	S_{16}	S_{17}	S_{18}	S_{19}	S_{20}	S_{21}	S_{22}	S_{23}	S_{24}	S_{25}	S_{26}	S_{27}	S_{28}
Mistakes	5	5	5	3	4	4	3	5	4	4	7	6	4	3	5	7
$\hat{\lambda}_k$	0.09	0.12	0.10	0.12	0.11	0.07	0.23	0.12	0.06	0.09	0.04	0.05	0.07	0.2	0.11	0.05

Table 7 The ML estimates of the parameters of a mixture of two latent-scale Kendall distance-based models for the gene ranking data

	Group 1	Group 2			Actual disease subtype			
$\hat{p}_g$	0.35	0.65			Type:		ER-	ER+
$\hat{\lambda}_g$	0.074	0.084			Count:		41	80
Patient, k	Kendall distance to		$\hat{\lambda}_{k1}$	$\hat{\lambda}_{k2}$	$\hat{z}_{k1}$	$\hat{z}_{k2}$	Predictive group	Actual Subtype
	$\hat{\pi}_{01}$	$\hat{\pi}_{02}$						
BC-M-003	1089	1037	0.060	0.065	0.01	0.99	2	ER+
BC-M-014	851	721	0.086	0.108	0.00	1.00	2	ER+
BC-M-016	1205	1021	0.052	0.069	0.00	1.00	2	ER+
BC-M-018	1009	1471	0.066	0.035	1.00	0.00	1	ER−
BC-M-020	1061	959	0.072	0.060	1.00	0.00	1	ER−
BC-M-057	1176	1266	0.057	0.048	1.00	0.00	1	ER−
BC-M-066	943	905	0.073	0.080	0.00	1.00	2	ER+
BC-M-080	794	824	0.095	0.088	0.98	0.02	1	ER−
BC-M-100	1649	1569	0.027	0.031	0.05	0.95	2	ER+
BC-M-167	1107	1047	0.062	0.063	0.20	0.80	2	ER−

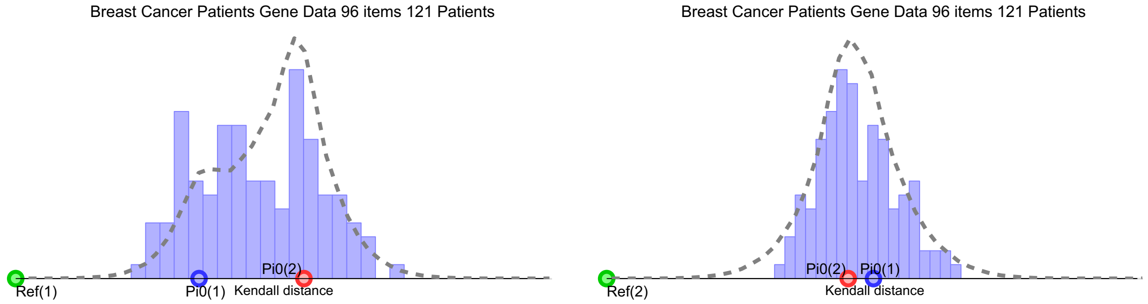

Fig. 7 Distribution of Kendall distances between patients' rankings and a reference ranking (green circle) is shown by the blue histograms. Two reference rankings are randomly picked. The gray dotted line represents the distribution of the Kendall distances implied from the fitted model. The blue circle labeled "Pi0(1)" and the red circle labeled "Pi0(2)" represent the estimated modal rankings $\hat{\pi}_{01}$ and $\hat{\pi}_{02}$ for groups 1 and 2 under the fitted model. (Color figure online)

proportions of two groups ($\hat{p}_1 = 0.35$ and $\hat{p}_2 = 0.65$) are close to those of the actual disease subtypes. As there are 96 genes, we do not list out the estimated modal rankings $\hat{\pi}_{01}$ and $\hat{\pi}_{02}$ of two groups. Instead, ten patients are randomly selected with their Kendall distances to $\hat{\pi}_{0g}$, their predictive estimates of the dispersion variable, $\hat{\lambda}_{kg}$, and the membership indicator, $\hat{z}_{kg}$ ($g = 1, 2$), as shown in Table 7. We also show the predicted group of each selected patient according to the one with higher membership probability. Notice that when the Kendall distance to $\hat{\pi}_{0g}$ is smaller, the $\hat{z}_{kg}$ becomes bigger, meaning that the probability that ranking k belongs to group g is larger. When $\hat{\lambda}_{kg}$ is smaller, the distribution is more concentrated at $\hat{\pi}_{0g}$ and ranking k is more likely to be closer to $\hat{\pi}_{0g}$. From Table 7, patient BC-M-167 is classified to a wrong group since he has similar distances to π_{01} and π_{02}, making him hard to be classified correctly.

To better illustrate our fitting result, Fig. 7 shows the Kendall distribution of the observed ranking data and the fitted model with reference to two randomly selected rankings. From the figure, we can find that the mixture model actually fits the data quite well since the fitted Kendall distribution fairly fits the observed data. To access the performance of our model classification, Fig. 8 shows the ROC curve of the disease subtype classification based on the rule $\hat{z}_{k1} > c$ for group 1. For $c = 0.5$, our model correctly classifies 85.12% of the patients. Since this is an unsupervised learning, we may think the proposed mixture of latent-scale distance-based models has a relatively powerful classification of disease subtypes for breast cancer patients based on the gene ranking data.

6 Concluding remarks

In this paper, we proposed a new class of latent-scale distance-based models which accounts for the heterogene-

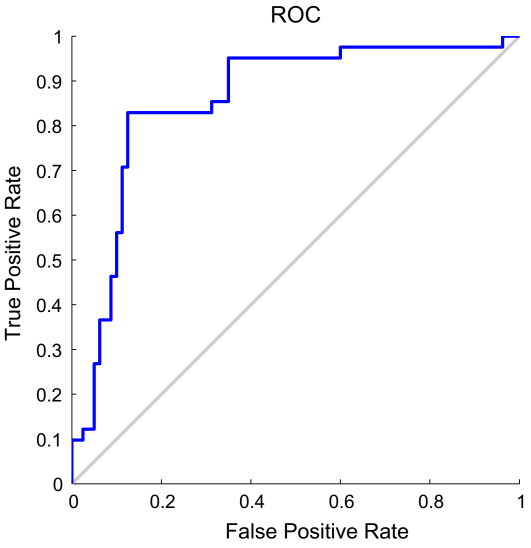

Fig. 8 ROC curve of the classification of the disease subtypes for breast cancer patients

ity in the background or expertise among the rankers. Our simulation experiments demonstrated that our proposed EM algorithm is computational efficient for any distance measure and even for a large number of items. Our real-world applications in Sect. 5 revealed that the proposed latent-scale distance-based model outperforms existing rank aggregation methods including Borda and those based on unweighted distance-based models.

Note that for some existing Bayesian rank aggregation methods such as Lee et al. (2014), the output of the MCMC algorithm is a list of rankings sampled from the posterior distribution. It is still necessary to find a proper rank aggregation method to combine these rankings. In Deng et al. (2014)'s method, its rank aggregation result is highly dependent on the hyper-parameter p, the proportion of relevant entities in the rankings which is often unknown in practice. Different choice of p will then end up with different results of rank aggregation. An improper choice of p might even lead to many ties in the aggregated ranking. Unlike these methods, our proposed latent-scale distance-based models can directly estimate the aggregated ranking while taking into account the diverse quality of rankers.

When applying our rank aggregation method, we need to choose a distance measure for our proposed model. Although there is no universal best rule of choosing a particular distance, we recommend using Kendall distance or Spearman distance as these two distances show consistently good results in our simulation experiments and applications. Hamming and Ulam distances usually perform poorly since they tend to be less sensitive to a small change in the number of items, particularly for large number of items.

Appendix: Sampling from a general distance-based Model

It is straightforward to simulate ranking data from the Kendall distance-based model, because of the nice decomposition of Kendall distance into a set of independent variables $V_i\left(\pi_s \circ \pi_0^{-1}\right)$ (see Sect. 4.1) which can be sampled easily. The detailed algorithm can be found in Ceberio et al. (2014). Recently Irurozki et al. (2018, 2014) developed two different methods to sample from models using Cayley distance and Ulam distance. Their simulation methods require the knowledge of special properties of the distance measures and they may not be able to be generalized to other distances. Here we introduce a general method to sample from any distance-based model.

We need to sample from $f(x)$ here. The method begins at $s = 0$ with the selection of $X^{(0)} = x^{(0)}$ drawn at random from some stating distribution g, with the requirement that $f\left(x^{(0)}\right) > 0$. Given $X^{(s)} = x^{(s)}$ the algorithm generates $X^{(s+1)}$ as follows:

1. Sample a candidate value X^* from a proposal distribution $g\left(\cdot \mid x^{(s)}\right)$.
2. Compute the Metropolis–Hastings ratio $R_{MH}\left(x^{(s)}, X^*\right)$, where

$$R_{MH}\left(x^{(s)}, X^*\right) = \frac{f\left(X^*\right) g\left(x^{(s)} \mid X^*\right)}{f\left(x^{(t)}\right) g\left(X^* \mid x^{(t)}\right)}.$$

3. Sample a random value for $X^{(s+1)}$ as follow:

$$X^{(s+1)} = \begin{cases} X^* & \text{with probability } \min\left\{1, R_{MH}\left(x^{(s)}, X^*\right)\right\} \\ x^{(s)} & \text{otherwise} \end{cases}$$

4. Assign $s := s + 1$ and go to step 1

Back to our problem, given π_0, λ, we need to sample from the distance-based model: $f(\pi) = \frac{\exp[-\lambda d(\pi,\pi_0)]}{C(\lambda)}$ is the distance function we choose. Note that we don't need to calculate the normalizing constant $C(\lambda)$ in the $R_{MH}\left(x^{(s)}, X^*\right)$ function because the $C(\lambda)$ terms cancel. For the proposal distribution $g(\cdot|.)$ here, we will introduce two proposal distribution and compare their difference.

The proposal distribution for the algorithm can be chosen as a symmetric distribution so that $g(x^* \mid x^{(s)}) = g(x^{(s)} \mid x^*)$. In our case, our choice of $g(x^* \mid x^{(s)})$ imposes small perturbation of the elements of $x^{(s)}$. Given $x^{(s)}$, we uniformly picked two elements of $x^{(s)}$ and swap their positions as our proposal π^*. Since this swap distribution is symmetric, so the Hasting corrections in the Metropolis–Hastings ratio cancel, and the ratio is given as:

$$R_{MH}\left(x^{(s)}, X^{*}\right) = \frac{\exp\left(-\lambda d\left(\pi^{*}, \pi_0\right)\right)}{\exp\left(-\lambda d\left(\pi^{(s)}, \pi_0\right)\right)}$$

When we need to sample from the distance-based Latent-scale Models, we first sample λ_k from $exp(\lambda)$. Then we sample π_k from simple distance-based model with λ_k using the method above.

References

Aledo, J.A., Gámez, J.A., Molina, D.: Tackling the rank aggregation problem with evolutionary algorithms. Appl. Math. Comput. **222**, 632–644 (2013)

Ali, A., Meilă, M.: Experiments with Kemeny ranking: what works when? Math. Soc. Sci. **64**(1), 28–40 (2012)

Aslam, J.A., Montague, M.: Models for metasearch. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 276–284 (2001)

Borg, I., Groenen, P.J.: Modern Multidimensional Scaling: Theory and Applications. Springer, Berlin (2005)

Busse, L.M., Orbanz, P., Buhmann, J.M.: Cluster analysis of heterogeneous rank data. In: Proceedings of the 24th International Conference on Machine Learning, pp. 113–120 (2007)

Ceberio, J., Irurozki, E., Mendiburu, A., Lozano, J.A.: Extending distance-based ranking models in estimation of distribution algorithms. In: Evolutionary Computation (CEC), 2014 IEEE Congress on, pp. 2459–2466. IEEE (2014)

de Borda, J.: Mémoire sur les élections au scrutin. Mémoires de l'Académie Royale des Sciences Année, pp. 657–664 (1781)

DeConde, R.P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., Etzioni, R.: Combining results of microarray experiments: a rank aggregation approach. Stat. Appl. Genet. Mol. Biol. 5(1):Article 15 (2006)

Deng, K., Han, S., Li, K.J., Liu, J.S.: Bayesian aggregation of order-based rank data. J. Am. Stat. Assoc. **109**(507), 1023–1039 (2014)

Dwork, C., Kumar, R., Naor, M., Sivakumar, D.: Rank aggregation methods for the web. In: Proceedings of the 10th International Conference on World Wide Web, pp. 613–622 (2001)

Fligner, M.A., Verducci, J.S.: Distance-based ranking models. J. R. Stat. Soc. B **48**(3), 359–369 (1986)

Fligner, M.A., Verducci, J.S.: Posterior probabilities for a consensus ordering. Psychometrika **55**(1), 53–63 (1990)

Irurozki, E., Calvo, B., Lozano, J.A.: Sampling and learning the Mallows model under the Ulam distance. Technical report, Department of Computer Science and Artificial Intelligence, University of the Basque Country (2014)

Irurozki, E., Calvo, B., Lozano, J.A.: Sampling and learning the Mallows and generalized Mallows models under the Cayley distance. Methodol. Comput. Appl. Probab. **20**(1), 1–35 (2018)

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., et al.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983)

Lee, M.D., Steyvers, M., Miller, B.: A cognitive model for aggregating people's rankings. PLoS One **9**(5), e96431 (2014)

Lin, S., Ding, J.: Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies. Biometrics **65**(1), 9–18 (2009)

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

Murphy, T.B., Martin, D.: Mixtures of distance-based models for ranking data. Comput. Stat. Data Anal. **41**(3), 645–655 (2003)

Naume, B., Zhao, X., Synnestvedt, M., Borgen, E., Russnes, H.G., Lingjærde, O.C., Strømberg, M., Wiedswang, G., Kvalheim, G., Kåresen, R., et al.: Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. Mol. Oncol. **1**(2), 160–171 (2007)

Niederreiter, H.: Quasi-Monte Carlo Methods. Wiley Online Library (2010)

Černý, V.: Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. J. Optim. Theory Appl. **45**(1), 41–51 (1985)