Efficient and accurate inference for mixtures of Mallows models with Spearman distance

Marta Crispino · Cristina Mollica · Valerio Astuti · Luca Tardella

Received: 20 September 2022 / Accepted: 13 June 2023 / Published online: 5 July 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
The Mallows model (MM) occupies a central role in parametric modelling of ranking data to learn preferences of a population of judges. Despite the wide range of metrics for rankings that can be considered in the model specification, the choice is typically limited to the Kendall, Cayley or Hamming distances, due to the closed-form expression of the related model normalizing constant. This work instead focuses on the Mallows model with Spearman distance (MMS). A novel approximation of the normalizing constant is introduced to allow inference even with a large number of items. This allows us to develop and implement an efficient and accurate EM algorithm for estimating finite mixtures of MMS aimed at i) enlarging the applicability to samples drawn from heterogeneous populations, and ii) dealing with partial rankings affected by diverse forms of censoring. These novelties encompass the critical inferential steps that traditionally limited the use of this distance in practice, and render the MMS comparable (or even preferable) to the MMs with other metrics in terms of computational burden. The inferential ability of the EM scheme and the effectiveness of the approximation are assessed by extensive simulation studies. Finally, we show that the application to three real-world datasets endorses our proposals also in the comparison with competing mixtures of ranking models.

Keywords Ranking data · Distance-based models · Model-based clustering · EM algorithm · Censoring

1 Introduction
Ranking data arise when a sample of \( N \) people is presented a finite set of \( n \) alternatives, called items, and is asked to rank them according to a certain criterion, such as personal preferences or attitudes. Thus, a ranking is the result of a comparative judgment on a set of competing alternatives expressed in the form of order relation.

The increasing interest in ranked data analysis is motivated by several research contexts, among which marketing and political surveys, where items could be consumer goods, political candidates or goals, but also by psychological and behavioral studies consisting, for instance, in ordering of words/topics according to the perceived association with a reference subject (Gormley and Murphy 2008a; Caron et al. 2014). Another typical context for ranking data is sports and, more generally, competitions. Some examples are horse or car races, soccer and basket championships, where the rankings are determined by the relative ability of the competing players or teams (Henery 1981).

However, ranking data are not necessarily the result of a human comparative choice, where typically only a limited number of options can be submitted to the judgment of the rankers. The need of handling large ordered lists of alternatives can be motivated, for example, with the analysis of the results of web-search engines, as well as of programs for matching biological sequences or specific traits with a reference database (DeConde et al. 2006). A further field of

Marta Crispino and Cristina Mollica have contributed equally to this work.

Cristina Mollica
cristina.mollica@uniroma1.it
Marta Crispino
marta.crispino@bancaditalia.it
Valerio Astuti
valerio.astuti@bancaditalia.it
Luca Tardella
luca.tardella@uniroma1.it

1 DG of Economics, Statistics and Research, Bank of Italy, Rome, Italy
2 Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy
application concerns the widespread digital platforms of on demand products, such as movies and songs, where various forms of feedback from the customers (e.g. clicks, ratings, reviews and purchases) are elaborated by the e-commerce companies to capture customers’ preferences and supply them with a personalized ordered list of recommended items (Liu et al. 2019). All these applications are characterized by the challenging task of dealing with a large value $n$ of items which are often not completely ranked, giving rise to partial rankings.

In the wide literature on probability distributions for ranking data, see Marden (1995) and Alvo and Yu (2014) for extensive reviews, the Mallows model (MM) introduced by Mallows (1957) became very popular for its long history, both as a successful approach for preference learning and as a starting point for effective methodological advances. Formally, the MM assumes the existence of a unique modal ranking representing the common comparative judgment on the items, usually referred to as consensus ranking. Moreover, the variability of the population of rankers, that is, the agreement with the consensus ranking, is calibrated by a concentration parameter. A third key ingredient of the MM is the distance between rankings, since the probability of a generic sequence depends on its distance from the consensus; as a consequence the MM is also referred to as distance-based model (DBM).

Even if this model is intuitive and attractive, especially for the fact that it represents an exponential family for random permutations (Diaconis 1988), postulating a MM as a generative process of the observed rankings does not prevent noticeable practical issues. First, one of its peculiarities is the mixed-type parameter space, combining the continuous concentration parameter and the discrete consensus ranking. The latter, in fact, takes values in the space $P_n$ of the $n!$ permutations of the first $n$ integers, whose adequate exploration, needed for making inference on the model, is computationally demanding and challenging as the number of items increases. Second, the estimation task is further complicated in the presence of partial observed rankings and heterogeneity in the sample.

Various directions to generalize the MM and improve its fitting performance have been proposed in the last decades, among which the mixture approach plays a leading role. However, a review of the existing extensions reveals that the version of the MM with Spearman distance (MMS henceforth), one of the historic forms of DBM originally motivated in the pioneering work by Mallows (1957), received very little attention, mainly because it is characterized by an intractable normalizing constant, while the MM with other distances (such as the Kendall, Cayley and Hamming ones) enjoys favourable closed-form expressions (Fligner and Verducci 1986; Irurozki 2014). However, except for the above computational consideration, there is no reason to believe that the Spearman distance would be less powerful or useful than other metrics. As a matter of fact, the choice of the distance should also be driven by interpretative considerations regarding the data being analyzed and the scientific context from which they arise. For example, the Cayley distance is useful in domains such as biology and computer science (Irurozki et al. 2018), while the Hamming is popular in coding theory (Diaconis 1988) or for matchings (Irurozki et al. 2019). On the other hand, the Kendall and Spearman distances represent a natural choice in the preference domain, where data are often collected in the form of rankings, as they both serve as measures of discordance between ordinal observations. While the Kendall counts the number of discordant pairs between two rankings, the Spearman quantifies the discordance by computing the squared deviations between the ranks. This implies that the former is a more robust (Croux and Dehon 2010) but rougher measure of dissimilarity, whereas the latter is more sensitive to outliers, but globally more informative on the degree of discordance (Tarsitano 2009). The higher discriminatory power than the Kendall and, more in general, than the other distances typically used in the MM, is also testified by the wider and grainier range of possible values of the Spearman distance for any number $n$ of items.

Another distinction between the metrics arises when considering the Bayesian perspective concerning the noise process that perturbs the consensus ranking in the data generation process. In this context, the Spearman distance is the preferable choice when there is a prior belief that rankings are mostly perturbed by adjacent transpositions. The foundational basis for this intuition can be found in the original research conducted by Mallows (1957), but the interested readers can also refer to an early version of Vitelli et al. (2018), where a simulation study examines the fitting performance of various metrics based on the stochastic process that generates the rankings.

The aim of this work is twofold: First, to make the MMS accessible and computationally tractable for the analysis of ranking data from a maximum likelihood inferential perspective; Second, to generalize this model within the finite mixture framework, thus enabling model-based clustering of full and partial rankings. The above goals are achieved by means of (i) the introduction of an original approximation of the MMS normalizing constant, which crucially solves the intractability of the model, and (ii) the development of a new Expectation-Maximization (EM) algorithm, which allows to successfully investigate the group structure of rankings with a high number of alternatives, and to handle arbitrary forms of partial rankings.

1 Precisely, they are the unnormalized versions of the corresponding correlation coefficients.

2 This version is available at https://arxiv.org/abs/1405.7945v3. See in particular Section 5.1 and Appendix C.1.
The new approximation, which may be interesting in its own right (e.g. for sampling procedures), is compared with existing approximating approaches for highlighting its relative merits. The new estimation procedure exploits the advantageous property of the MMS (not shared with other forms of MMs) of enjoying a closed-form expression for the maximum likelihood estimate (MLE) of the consensus ranking. Both these novelties encompass the critical inferential steps that traditionally limited the use of this distance in practice, and render the MMS comparable (or even preferable) to the MMs with other distances in terms of computational burden. Finally, the effectiveness and usefulness of our proposals are assessed with extensive simulation studies and applications to real-world ranking data.

The outline of the paper is the following. In Sect. 2, after reviewing the MM and its main generalizations, we focus on the inference on the MMS. We then move to detailing the finite mixture framework, also sketching the EM algorithm, and finally we explain the data augmentation strategy to deal with heterogeneous partial rankings. In Sect. 3, we present the new approximation of the model normalizing constant and compare it with available alternatives. An extensive simulation study is performed in Sect. 4, by considering a wide range of scenarios in terms of sample size, number of items, heterogeneity of the sampling distribution, as well as different patterns of data missingness. Section 5 deals with real-world data illustrations, then the paper ends with concluding remarks in Sect. 6.

2 Model and inference

2.1 The Mallows model: a brief review

Let \( r = (r_1, \ldots, r_n) \) be a ranking of \( n \) items with generic entry \( r_i \) indicating the rank assigned to item \( i \). We adopt the usual convention that \( r_i < r_j \) means that item \( i \) is preferred to item \( j \) (the lower the rank is, the more preferred the item is). Both items and ranks are identified with the set \( \{1, \ldots, n\} \), implying that a generic observation \( r \) is a permutation of the first \( n \) integers and, thus, a point in the finite discrete space \( P_n \).

The MM is a well-established parametric class of ranking distributions whose probability is attributed to Mallows (1957). The probability of observing the ranking \( r \) under the MM is

\[
P(r | \rho, \theta) = \frac{e^{-\theta d(r, \rho)}}{Z(\theta, \rho)}, \quad r \in P_n,
\]

where \( \rho \in P_n \) is the consensus ranking, \( \theta \in \mathbb{R}_0^+ \) is the concentration parameter, \( d(\cdot, \cdot) \) is a distance over \( P_n \) and \( Z(\theta, \rho) = \sum_{r \in P_n} e^{-\theta d(r, \rho)} \) is the normalizing constant or partition function. From (1) it follows that, under the MM assumption, the probability of observing a ranking decreases exponentially as its distance from the consensus \( \rho \) increases.

In the attempt to enlarge the flexibility of the MM and improve the goodness-of-fit, various generalizations appeared in the literature. Fligner and Verducci (1986) proposed the Generalized Mallows model (GMM), which considers a \((n-1)\)-dimensional vector of concentration parameters, representing the uncertainty at each stage of the ranking process that differentially affects the attribution of the positions. The GMM, however, can be defined only when the distance is decomposable into the sum of independent components associated to each single stage of the ranking process, such as for the Kendall, Cayley and Hamming metrics. Meilă and Bao (2010) extended the GMM to infinite rankings with the Infinite Generalized Mallows model, to handle those cases where the number of items is very high, or potentially not completely known, such as in the output of search engines. Another proposal to enhance the flexibility of the MM is the incorporation of a varying importance to the ranks through the weighted distance-based model (WDBM) by Lee and Yu (2010), which amounts to use weighted versions of the commonly used distances between rankings, including the Spearman one.

In order to explore the unobserved heterogeneity in the sample, the aforementioned extensions were considered also in the finite mixture modelling approach. For example, Murphy and Martin (2003) described the general EM algorithm to derive the MLE of MM mixtures for an arbitrary metric; Lee and Yu (2012) detailed the same inferential method for the mixture of WDBMs. Differently from the one illustrated by Murphy and Martin (2003), Busse et al. (2007) introduced an EM algorithm for the MM mixtures, working with Kendall distance only but handling both complete and partial rankings of a larger number of items.

However, all the above contributions do not specifically focus on the MMS and, if they handle the Spearman distance, they only work in practice with few items (less than \( n = 10 \), say). This work, instead, is devoted to the MMS and its mixture extension, with the aim of developing a learning strategy able to handle also rankings of a large number of items.

2.2 Likelihood and inference for the MMS

The Spearman distance between two permutations \( r, \rho \in P_n \) is defined as follows

\[
d(r, \rho) = \sum_{i=1}^n (r_i - \rho_i)^2.
\]

It is an unnormalized version of the Spearman rank correlation, used to measure the statistical correlation between the ranks of two variables but, when rankings are consid-
ered as vectors in $\mathbb{R}^n$, it is simply the squared Euclidean distance, or $L_2$-norm. By developing the square in (2) and setting $c_n = n(n + 1)(2n + 1)/6$ and the scalar product $\rho^T r = \sum_{i=1}^n \rho_i r_i$, where the symbol $^T$ denotes the transposition (row vector), one has $d(r, \rho) = 2(c_n - \rho^T r)$. Hence, the MMS can be written as
\[ \mathbb{P}(r | \rho, \theta) = \frac{e^{-2\theta(c_n - \rho^T r)}}{Z(\theta)}, \quad (3) \]
where the normalizing constant $Z(\theta) = \sum_{r \in \mathcal{P}_n} e^{-2\theta(c_n - e^T r)}$ with $e = (1, 2, ..., n)$ does not depend on $\rho$ because of the right-invariance property of the Spearman distance (Diaconis 1988). Note that the number of terms in the sum grows with $n!$, which makes direct calculation of $Z(\theta)$ unfeasible for all but very small values of $n$, typically for at most $n = 11$.

Let $r = (r_1, \ldots, r_N)$ be a random sample of size $N$ drawn from the MMS and $N_l$ be the number of the $l$-th distinct observed ranking $r_l$ for $l = 1, \ldots, L$, such that $\sum_{l=1}^L N_l = N$. The observed-data log-likelihood of the MMS is
\[ \ell(\rho, \theta | r) = -N \left[ \log Z(\theta) + 2\theta \left( c_n - \rho^T \bar{r} \right) \right]. \quad (4) \]
where $\bar{r} = (\bar{r}_1, \ldots, \bar{r}_n)$ is the sample mean rank vector whose $i$-th entry is $\bar{r}_i = 1/N \sum_{l=1}^L N_l r_{il}$. When considering the metric (2) in the MM, it is not surprising that the likelihood of the MMS resembles that of an $n$-variate normal, except for its finite support. It follows that, if $\bar{r} \in \mathcal{P}_n$, then the MLE $\hat{\rho}$ of the consensus ranking simply coincides with the sample mean rank vector. However, in general $\bar{r} \notin \mathcal{P}_n$, so a further consideration is required in order to solve the optimization problem. Specifically, the log-likelihood expression (4) implies that the MLE of the consensus is
\[ \hat{\rho} = \arg \max_{\rho \in \mathcal{P}_n} \rho^T \bar{r}. \quad (5) \]
We observe that the optimization problem (5) is equal to the one discussed in Marden (1995) for estimating the central ranking under uniformity of the random permutations, that the author proved to admit the closed-form solution
\[ \hat{\rho} = (\hat{\rho}_1, \ldots, \hat{\rho}_i, \ldots, \hat{\rho}_n) \quad \text{with} \quad \hat{\rho}_i = \text{rank}(\bar{r}_i) \quad (6) \]
(Theorem 2.2, page 29). This implies that the simple and fast Borda rank aggregation method, which amounts to order the items according to their sample average rank, coincides with the MLE of the consensus ranking of the MMS. Even if this theoretical result, pointed out for the MMS in Feigin and Cohen (1978), efficiently solves the critical step of estimating the discrete parameter $\rho$ regardless of the number $n$ of compared items and the sample size $N$, to the best of our knowledge it has been completely disregarded in those few works dealing with the MMS and extensions thereof. In fact, the optimization (5) has been traditionally addressed as if a closed-form solution did not exist, that is, with computationally intensive approaches, such as global or local search methods whose effectiveness and applicability strongly depend on the values of $n$ and $N$ (see for example Murphy and Martin 2003; Busse et al. 2007). These search methods require the exploration of a discrete space with a fast-growing dimension and, hence, are inefficient both in terms of execution time and convergence achievement as $n$ and $N$ increase, becoming soon infeasible (typically, NP-hard as proven in Bartholdi et al. 1989). The omission of this important result is also testified by the lack of specific and effective software for performing inference on the MMS, that could take advantage of the related computational simplifications. In fact, the use of the Borda method has been so far mainly recommended for the MM with Kendall distance, either as approximation of the MLE of $\rho$ (Ali and Meil˘a 2012) or for initializing the local search (Irurozki et al. 2016).

Concerning the estimation of the concentration parameter $\theta$, one needs to maximize the profile log-likelihood $\ell(\hat{\rho}, \theta | r)$ to get $\hat{\theta}$ as the solution of the following equation
\[ -Z'(\theta)/Z(\theta) = 2 \left( c_n - \hat{\rho}^T \bar{r} \right). \quad (7) \]
As noticed by Fligner and Verducci (1986), $\hat{\theta}$ is the value equating the expected Spearman distance $E_{\theta}[D]$ under the MMS, that is, the left hand side of (7), to the sample average Spearman distance $d = 1/N \sum_{l=1}^L N_l d(r_l, \hat{\rho})$, corresponding to the right-hand side of (7). Actually, $E_{\theta}[D]$ is a strictly decreasing function of $\theta$, but the root of the above equation is not available in closed-form. So, the MLE can be found numerically, for instance, via a Newton–Raphson algorithm (Marden 1995), although the calculations are demanding for all but small values of $n$, because of the intractability of the model normalizing constant.

Models with intractable partition functions arise frequently in statistics and in machine learning (e.g., exponential family random graph models, Bayesian networks, deep generative models). In the Bayesian framework, when dealing with models that involve intractable partition functions, it leads to what is known as doubly intractable posterior distributions. In such cases, inference has been tackled in the literature either with sampling algorithms aimed at finding consistent estimators for the unknown parameters (see for example Murray et al. 2012; Møller et al. 2006), either with off line approximations of the normalizing constant (Andrieu and Roberts 2009; Vitelli et al. 2018; Xu et al. 2018). The first stream of methods is unfeasible in case of the MMS, because it typically requires the direct simulation from the model likelihood. We here propose to solve the inferential issue by introducing a new analytical approximation of $Z(\theta)$.
2.3 Likelihood and inference for the MMS mixture model

Let \( z_l = (z_{l1}, \ldots, z_{lG}) \) be the latent variable collecting the binary group membership indicators, that is, \( z_{lg} = 1 \) when the \( l \)-th observation belongs to the \( g \)-th group and \( z_{lg} = 0 \) otherwise. The complete-data log-likelihood for the \( G \)-components mixture of MMS is

\[
\ell_c(\rho, \theta, \omega, z|\mathbf{r}) = \sum_{l=1}^{L} \sum_{g=1}^{G} N_l z_{lg} \left[ \log \omega_g + \right.
-2\theta_g \left( c_n - \rho_g^T \mathbf{r}_l \right) - \log Z(\theta_g) \right],
\]

where \( \omega = (\omega_1, \ldots, \omega_G) \) and \( \theta = (\theta_1, \ldots, \theta_G) \) are, respectively, the mixture weights and the component-specific concentration parameters, whereas \( \rho \) is a \( G \times n \) matrix, whose rows indicate the consensus rankings of the mixture components. In order to perform MLE in the presence of latent variables, we employ the EM algorithm (Dempster et al. 1977). In particular, at iteration \( (t+1) \), the proposed EM scheme consists in the following steps:

**E-step:** for \( l = 1, \ldots, L \) and \( g = 1, \ldots, G \),
- compute the posterior membership probabilities
  \[
  \hat{z}_{lg}^{(t+1)} = \frac{\hat{\omega}_g^{(t)} \mathbb{P}(\mathbf{r}_l|\hat{\rho}_g^{(t)} \cdot \hat{\theta}_g^{(t)})}{\sum_{g'=1}^{G} \hat{\omega}_{g'}^{(t)} \mathbb{P}(\mathbf{r}_l|\hat{\rho}_{g'}^{(t)} \cdot \hat{\theta}_{g'}^{(t)})}.
  \]

**M-step:** for \( g = 1, \ldots, G \),
- compute the mixing weights
  \[
  \hat{\omega}_g^{(t+1)} = \frac{\hat{N}_g^{(t+1)}}{N}
  \]
  where \( \hat{N}_g^{(t+1)} = \sum_{l=1}^{L} N_l \hat{z}_{lg}^{(t+1)} \) is the number of sample units allocated in the \( g \)-th component.
- compute the consensus rankings
  \[
  \hat{\rho}_g^{(t+1)} = \left( \hat{\rho}_{g1}^{(t+1)}, \ldots, \hat{\rho}_{gn}^{(t+1)} \right),
  \]
  with \( \hat{\rho}_{gi}^{(t+1)} = \text{rank} \left( \hat{\mathbf{r}}_{gi}^{(t+1)} \right) \) and \( \hat{\mathbf{r}}_{gi}^{(t+1)} = \sum_{l=1}^{L} N_l \hat{z}_{lg}^{(t+1)} \mathbf{r}_{li} / \hat{N}_g^{(t+1)} \).
- determine the concentration parameter estimates \( \hat{\theta}_g^{(t+1)} \) as the solution of
  \[
  E_{\theta_g}[D] = 2 \left( c_n - \rho_g^{(t+1)} \hat{\mathbf{r}}_g^{(t+1)} \right).
  \]

Note that Murphy and Martin (2003) first provided an EM algorithm for the MLE of MM mixtures with a generic metric \( d(\cdot, \cdot) \), including the Spearman. However, their estimation procedure is not specialized for each considered metric implying that, in the case of Spearman distance, it completely disregards the closed-form solution of the M-step for \( \rho \) whereas, for the decomposable metrics, the convenient expression of the normalizing constant is totally ignored. Actually, the applications in their work were limited to full rankings of only \( n = 5 \) items, for which the permutation space can be explored exhaustively and, even with a time-consuming inferential process, the MLE is more likely to be achieved. Instead, our implementation of the EM is completely specialized for the Spearman distance, thus exploiting the shortcuts explained in the previous sections.

Concerning publicly available softwares, the \texttt{R} package \texttt{rankdist} (Qian and Yu 2019) performs MLE for the MM in the presence of full and top-\( q \) ranking data, including the MMS and its finite mixture extension. However, the analytical simplifications associated to the Spearman distance are not considered, and the MLE solution for \( \rho \) is searched locally, slowing down the computational time. In addition, the program generates all the \( n! \) permutations to compute the normalizing constant \( Z(\theta) \) needed to estimate the concentration parameter, and hence crashes for \( n > 11 \) by limiting the range of MMS applications to few items. In case of top-\( q \) rankings the program works with at most \( n = 7 \) items.

The \texttt{pmr R} package (Lee and Yu 2013) performs frequentist inference on the MMS, but again, it generates all the \( n! \) possible permutations, to search the MLE of the consensus ranking globally. As a result, it quickly runs out of memory restricting the MMS applicability to datasets with a maximum of \( n = 11 \) items.

The \texttt{R} package \texttt{BayesMallows} (Sørensen et al. 2020) implements Bayesian inference for the MMS with any right-invariant distance, including the Spearman, and it also supports partial rankings and mixture models. Even if the procedure scales well with \( n \), the implementation of the MMS does not consider the analytical simplifications of the Spearman distance, which could speed up the algorithm considerably. In \texttt{BayesMallows} the normalizing constant of the MMS is approximated by means of an importance sampling scheme. In Sect. 3 we compare the performance of our novel approximation of \( Z(\theta) \) with different types of approximations, including the one available in Sørensen et al. (2020).

2.4 Mixture model extension for partial rankings

In real applications it could happen that the ranking data are not completely observed. This occurs, for instance, when the assessors rank only a subset of the items, typically their top-\( q \) preferred ones, or there may be missing data either at
random or by design, giving rise to various forms of partial observations.

In this section we extend the mixture analysis previously described for accommodating inference on partial data. In the spirit of Beckett (1993), whose goal was to estimate the parameters of the basic MM from partial observations, we generalize its MLE approach based on the EM algorithm by assuming the mixture of MMS as the sampling distribution. We adopt a data completion argument that suitably combines the two main sets of latent variables: the missing ranks of the partial sequences and the unobserved group memberships.

Suppose that $N_l$ respondents provide a partial sequence $r_l$ where only a subset $I_l \subseteq \{1, 2, \ldots, n\} = \mathcal{I}$ of $n_l = |I_l|$ items are actually ranked. As a consequence, the position $r_{li}$ is known only if $i \in I_l$. We stress that, in our setting, arbitrary types of censoring patterns are allowed, implying that the set of the assigned positions may be any not-empty subset of the first $n$ integers, leading to $n_l = 1, \ldots, n - 1$. Moreover, a full ranking $\tilde{r}$ is defined as the complete sequence obtained by combining the $n_l$ observed entries $r_l$ with the latent unassigned $n - n_l$ ranks for the items $i^l \in \mathcal{I} \setminus I_l$. The latent components of the partially ranked vector are inferred through the EM algorithm, that exploits the information provided by the entire sample to fill the missing entries of the sequence coherently with the observed ones.

Let $\mathcal{C}(r_l) \subset \mathcal{P}_n$ be the set of full rankings which are compatible with the partial sequence $r_l$. This means that $\mathcal{C}(r_l)$ collects all the complete ranked vectors on which the application of a certain censoring pattern yields the partial observation $r_l$. Under the mixture model, the conditional probability of observing the full ranking $\tilde{r}_m \in \mathcal{C}(r_l)$ is

$$p_{r_l}(\tilde{r}_m) := \mathbb{P}(\tilde{r}_m | r_l, \rho, \theta, \omega) = \frac{\sum_{g=1}^{G} \omega_{g} e^{-\beta_{g} (c_{\omega} - \rho_{g}^T \tilde{r}_m)} \exp \{ \log Z(\theta) \}}{\sum_{\tilde{r} \in \mathcal{C}(r_l)} \sum_{g=1}^{G} \omega_{g} e^{-\beta_{g} (c_{\omega} - \rho_{g}^T \tilde{r})} \exp \{ \log Z(\theta) \}}. \tag{8}$$

Equality (8) extends expression (6) in Beckett (1993) to the finite mixture case. The observed frequency $N_l$ of respondents reporting each partial ranking $r_l$ is then distributed over the full rankings that may give rise to $r_l$, in order to obtain the E-step estimate of the frequency of $\tilde{r}_m$ as

$$\hat{N}_m = \sum_{l: \tilde{r}_m \in \mathcal{C}(r_l)} N_l p_{r_l}(\tilde{r}_m). \tag{9}$$

The above sum accounts for the fact that, under varying censoring patterns, the same full ranking $\tilde{r}_m$ can be compatible with different partial rankings; for instance, $\tilde{r} = (1, 2, 4, 3)$ is compatible with both $r_1 = (1, 2, NA, NA)$ and $r_2 = (1, NA, NA, NA)$. Obviously, if the observed data share the same censoring pattern (e.g. they are all top-3 implying $n_l = 3$ for all $l = 1, \ldots, L$), the sum in (9) can be avoided and the computation simplifies greatly.

In practice, the EM scheme detailed in Sect. 2.3 is adapted for handling partial rankings by adding a further E-step, needed to account for the latent ranked entries, and by considering the augmented data $\tilde{r}_m \in \mathcal{C}(r_1) \cup \cdots \cup \mathcal{C}(r_L)$ as the sample to draw inference on.

### 3 A new approximation of the partition function

If, on the one hand, the existence of the closed-form for the MLE of $\rho$ avoids the computationally demanding local search for inferring the consensus ranking, on the other, considering large values of $n$ affects the calculation of $Z(\theta)$ which is needed in the estimation steps of (i) the concentration parameter (see eq. (7)), (ii) the posterior membership probabilities and (iii) the conditional probabilities of observing the full rankings compatible with the partial information (see eq. (8)). In fact, the rapidly-increasing size $n!$ makes the exhaustive enumeration of the sequences in $\mathcal{P}_n$, and hence the direct calculation of $Z(\theta) = \sum_{r \in \mathcal{P}_n} e^{-k_{\beta} (c_{\omega} - e^T r)}$, feasible only for small $n$.

In this section, we propose a novel approximation of the distribution of the Spearman distance which enables inference on rankings of a large number $n$ of items. In fact, this approximation, which may be interesting in its own right (e.g. for sampling procedures), is then used to approximate both $Z(\theta)$ and $E_\theta[D]$.

#### 3.1 The new approximation in practice

Regardless of the metric adopted in the MM, the partition function can be more conveniently expressed as

$$Z(\theta) = \sum_{d_n \in D_n} N_{d_n} e^{-\beta d_n}, \tag{10}$$

where $N_{d_n} = |\{r \in \mathcal{P}_n : d(r, e) = d_n\}|$ is the number of rankings at distance $d_n$ from $e$ and $D_n$ is the set of possible distance values among the $n$-dimensional rankings. In the case of Spearman distance, $D_n = \left\{d_n = 2h : h \in \mathbb{N}_0 \right\}$, implying that the number of terms in (10) is of order $n^3$ rather than $n!$ as in the naive representation of $Z(\theta)$ related to (3). The values $N_{d_n}$ indicate

3 Note that the definition of $Z(\theta)$ corresponds to the canonical partition function of statistical mechanics $Z(T) = \sum_{E_i} \exp \{ - E_i / k_{\beta} T \} = \sum_{E} g(E) \exp \{ - E / k_{\beta} T \}$, where $k_{\beta}$ is the Boltzmann constant, $T$ is the temperature of a system, $E_i$ is the energy of the $i$-th state and $g(E)$ is the density of the states.
the frequency distribution of the Spearman distance under the uniform (null) model and correspond to the sequence A175929 available only up to $n \leq 14$ in the OEIS (Sloane 2017; Vitelli et al. 2018). We here propose a strategy to approximate $N_{dn}$ when $n > 14$.

As proven in Kendall (1970), under the null model the Spearman distance is asymptotically normally distributed (see Fig. 1). However, the convergence is slow and not uniform over the domain of the distribution. In particular, even with $n$ large enough, for the distribution to be reliably approximated in the bulk, the tail behaviour will be different from the one predicted by a normal distribution.

To usefully approximate the distribution of extreme values, we could in principle use large deviation theory (Cramér 1938; Varadhan 1966) but, as explained in Mukherjee (2016), the computation of explicit results is often not feasible. We can nevertheless exploit general properties of the large deviation rate function, and what we know on the values of $N_{dn}$ for $n \leq 14$, to obtain an approximation. As a matter of fact, in Mukherjee (2016) it is proved the existence and differentiability of the (scaled) cumulant generating function,

$$\lambda (k) := \lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}\left[e^{kn} \right], k \in \mathbb{R},$$

with $x_n = \frac{d_n}{d_{max,n}}$. This result allows us to apply the Gärtner-Ellis theorem (Gärtner 1977; Ellis 1984), to obtain the approximate rate function

$$\xi (x_n) \approx \frac{1}{n} \log \left( \frac{N_{dn}}{n!} \right).$$

with $\xi (x_n)$ only weakly dependent on $n$ for $n \gg 1$. Making use of the known values of $N_{dn}$ for $n \leq 14$, we can derive for $\xi (x_n)$ the form

$$\xi (x_n) = a_0 + a_1 \log [x_n (1-x_n)] + a_2 x_n (1-x_n).$$

The formal derivation of (12) and the value of the coefficients $(a_0, a_1, a_2)$ are detailed in Supplementary Material A. With this result, our approximation of $N_{dn}$ for $8 < d_n < d_{max,n} - 8$ is

$$\hat{N}_{dn} \propto n! \exp [n \xi (x_n)].$$

The above approximation is used for $n > 14$. The values of $N_{dn}$ for $d_n = 0, 2, 4, 6$ (as well as those for their symmetrical $d_n = d_{max,n}, d_{max,n} - 2, d_{max,n} - 4$ and $d_{max,n} - 6$) follow from combinatorial facts and are equal respectively to $1$, $(n - 1)!$, $(\binom{n}{2}^2)$ and $(n - 2)^3/6 - (n - 2)^2 + 23(n - 2)/6 - 1$.

Figure 2 displays the probability mass distribution of the Spearman distance for $n = 14$ (in gray), its limiting normal density (red line) and the density obtained with our new approximate rate function.
3.2 A comparison with other solutions

In this section we compare our new approximation \( \hat{Z}_{\text{New}}(\theta) = \sum_{d \in D_n} \hat{N}_d e^{-\theta \hat{d}_n} \) with two others proposed in the literature, one analytical and the other simulation-based. The former, based on the von Mises-Fisher (vMF) distribution and originally proposed in McCullagh (1993), has been recently exploited by Xu et al. (2018) for inference in their angle-based model for rankings and is given by

\[
Z(\theta) \approx \hat{Z}_{\text{vMF}}(\theta) = \frac{2^{n/2} I_{n-3}(\kappa) \Gamma(n-1)}{\kappa^{n/2} e^{\kappa}},
\]

where \( \kappa = \theta n(n^2 - 1)/6 \), \( I_x(\cdot) \) is the modified Bessel function of the first kind with order \( x \), and \( \Gamma(\cdot) \) is the Gamma function. The reader is referred to Supplementary Material B for the formal derivation of \( \hat{Z}_{\text{vMF}}(\theta) \) in the context of (non-standardized) rankings. The latter \( \hat{Z}_{\text{IS}}(\theta) \), based on the Importance Sampling (IS) technique, is due to Vitelli et al. (2018) to which we refer for details.

In Fig. 3 we plot the difference log \( \hat{Z}(\theta) - \log Z(\theta) \) as a function of \( \theta \) for \( n = 14 \). In the left and middle panels, where the approximations \( \hat{Z}_{\text{New}} \) (in green) and \( \hat{Z}_{\text{vMF}} \) (in orange) are compared over the interval \( \theta \in (0, 1] \) and \( \theta \in (0, 0.02] \) respectively, we notice that for small values of \( \theta \), the vMF approximation and our proposal perform both very good (middle panel). However, the vMF approximation deteriorates fast already for \( \theta > 0.015 \) and its negative bias rapidly becomes prominent (left panel). This happens because the vMF approximation is an integral over the entire \( (n-1) \)-dimensional sphere enclosing all the points of \( P_n \) and, for growing \( \theta \), the permutations are more and more concentrated in a small region of this sphere under the MMS, implying that \( \hat{Z}_{\text{vMF}} \) is strongly biased. On the contrary, our new approximation performs very well for increasing values of \( \theta \), because it exploits the knowledge of the true values of \( N_{d_n} \) available for small \( d_n \), which are the ones that mostly matter in (10).

Instead, the IS approximation has an oscillating behavior around the exact, and its error is generally small even if, when compared to our new approximation, it is closer to the exact for few values of \( \theta \). Note that, even if the performance of the IS is generally good, being based on random sampling, it is computationally very demanding, while our new approximation is immediately available for any value of \( n \).

From the above results we evince that our proposed new approximation performs very well for a wide range of values of \( \theta \), not only for small ones, like the analytical approximation resorting on the vMF distribution. This is particularly relevant when dealing with mixture models, where the cluster-specific concentrations typically grow with the number of clusters detected.

To complete the study of the proposed approximation, in Supplementary Material C we document that its effect on the inference on \( \theta \) is generally negligible (Table 1).

4 Simulation studies

4.1 Experiments with different sample sizes and number of items

The first simulation study considers data drawn from the homogeneous MMS \( (G^* = 1) \) under different parameters settings, and then evaluates the performance of our fitting procedure in terms of model parameters estimates with multiple criteria. Specifically, for each combination of \((N, n)\) with sample size \( N \in \{50, 200, 500, 1000\} \) and number of items \( n \in \{5, 10, 14, 15, 25, 50, 100\} \), we independently simulate 100 datasets from the homogeneous MMS, by randomly choosing the true consensus ranking as \( \rho \sim \text{Unif}(P_n) \).
and sampling uniformly $\theta \sim \text{Unif}(\theta_{\text{min}}, \theta_{\text{max}})$ over different intervals depending on the value of $n$ (see Table 1, which collects the bounds of the uniform sampling intervals). In this way, we have a comparable concentration over the permutation set in the samples with different values of $n$.

The inferential ability of the proposed EM algorithm to recover the actual parameters values has been evaluated by averaging over the 100 datasets the following measures:

- $m_\theta = |\hat{\theta} - \theta|/\theta$, the relative error between the true and estimated concentration;
- $m_\rho = d(\hat{\rho}, \rho)/d_{\text{max},n}$, the relative distance between the true and estimated consensus;
- $\phi_\rho = \mathbb{I}(\hat{\rho} = \rho)$, the matching between the true and estimated consensus;
- $\phi_G = \mathbb{I}(\hat{G} = G^*)$, the matching between the true and estimated number of components. The optimal number $\hat{G}$ of groups is automatically selected by means of the Bayesian Information Criterion, BIC (Schwarz 1978). In particular, we select $\hat{G}$ by applying the elbow rule to the screeplot of the BIC values (Zhao and Hautamäki 2008; Shi et al. 2021), in place of the more popular choice of selecting the minimum BIC, because of the tendency of the MMS to over-fit the data (Murphy and Martin 2003; Qian and Yu 2019).\(^5\)

In Table 2 we report the results of the fit. For all $n$, we see that all performance criteria improve with $N$ and are generally very good, at least for $n \leq 25$. As a matter of fact, the performance of the fitting procedure is linked to the sparsity of the problem: when the data are too sparse (that is, with large $n$ and small $N$), and the concentration parameter is small, recovering the exact true consensus measured by $\phi_\rho$ becomes very difficult (see the results for $n = 50, 100$ in Table 2(b)). However, $m_\rho$ is very small even in these extreme cases, testifying a general good performance of the procedure. In addition, it is enough to sufficiently enlarge the sample size to recover exactly the true $\rho$ in almost all samples considered (see Table 2(b) for $n = 50$ and $N = 10000$).

Finally, from Table 2, we note also that the inferential ability does not seem to be worsened by the introduction of the approximation of $Z(\theta)$. In fact, $m_\theta$ is comparable for the cases $n = 14, 15$, the first estimated with the exact $Z(\theta)$, and the second with $\hat{Z}_{\text{New}}(\theta)$.

### 4.2 Experiments with heterogeneous data

In this section we study the performance of our procedure in terms of its ability to correctly recover the cluster structure.

We simulate 100 datasets from different mixtures of MMS with varying number of clusters $G^* = 2, 3, 4$, while keeping fixed the number of items $n = 25$, and the sample size $N = 2000$. For each dataset, we sample (i) the mixture weights from a symmetric Dirichlet distribution with large shape

| $n$ | $N$ | $m_\theta$ | $m_\rho$ | $\phi_\rho$ | $\phi_G$ |
|-----|-----|------------|----------|-------------|----------|
| 5   | 50  | .078       | .0000    | 1.00        | 1.00     |
| 5   | 200 | .045       | .0000    | 1.00        | 1.00     |
| 5   | 500 | .034       | .0000    | 1.00        | 1.00     |
| 10  | 1000| .020       | .0000    | 1.00        | 1.00     |
| 10  | 50  | .071       | .0000    | 1.00        | 1.00     |
| 10  | 200 | .034       | .0002    | .97         | .91      |
| 10  | 500 | .024       | .0000    | 1.00        | 1.00     |
| 10  | 1000| .016       | .0000    | 1.00        | 1.00     |
| 14  | 50  | .061       | .0063    | .08         | .99      |
| 14  | 200 | .031       | .0010    | .63         | 1.00     |
| 14  | 500 | .019       | .0001    | .98         | .98      |
| 14  | 1000| .015       | .0000    | 1.00        | 1.00     |

\(^5\) Note that this criterion is useful only for automatizing the simulations: when choosing the number of clusters in applications the researcher typically uses the BIC, as well as other relevant measures, as a guidance, and decides the number of groups based also on other heuristic criteria (such as the size of the clusters, their separation, or prior information depending on the application at hand).
Table 3: Bounds of the intervals for the uniform sampling of \( \theta \) in the three group separation scenarios

| Cluster separation | High | Medium | Low  |
|--------------------|------|--------|------|
| \( \theta_{\min} \) | .004 | .003   | .002 |
| \( \theta_{\max} \) | .006 | .005   | .004 |

Table 4: Inferential performance of the EM algorithm on rankings simulated from alternative MMS-mixtures

| \( G^* \) | Cluster separation | \( \phi_G \) | \( \phi_z \) | \( \phi_G \) | \( \phi_z \) | \( \phi_G \) | \( \phi_z \) |
|-----------|--------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| \( n = 10 \) | | | | | | | |
| full | .0175 | .0000 | 1.00 | .99 | .0175 | .0000 | 1.00 | 1.00 |
| A | .0192 | .0000 | 1.00 | .90 | .0176 | .0000 | 1.00 | .94 |
| B | .0199 | .0002 | .96 | .83 | .0182 | .0002 | .90 | .82 |
| \( n = 15 \) | | | | | | | |

For each sample, we then fit the data with varying number of components \( G = 1, \ldots, 6 \) and select the optimal \( \hat{G} \) by means of the BIC like in the previous section.

In Table 4 we report \( \phi_G = \mathcal{H}(\hat{G}) \) as before, plus the mis-classification rate, defined as the fraction of the sampled units classified in the wrong mixing group, \( \phi_z = \sum_{j=1}^{N} \mathcal{H}(z_j \neq \hat{z}_j) / N \), with the group allocation \( \hat{z}_j \) determined by the maximum a posteriori (MAP) of the posterior membership probabilities. As expected, the ability to correctly reconstruct the clusters decreases as the difficulty of the problem increases. In general, however, we are happy with the results especially because, in the most difficult scenario, the mis-classification rate is always below 15%.

4.3 Simulations for incomplete rankings

Through an additional simulation study, we explore the impact of censoring on the inference, by considering different patterns of missingness of partial rankings drawn from the homogeneous MMS. To this aim, we first simulate 100 datasets of \( N = 500 \) full rankings for the cases \( n = 10 \) and \( n = 15 \). Specifically, the observations are drawn from the MMS with \( \rho \sim \text{Unif}(P_n) \) and \( \theta \sim \text{Unif}(0.1, 0.15) \) for \( n = 10 \) and \( \theta \sim \text{Unif}(0.5, 0.6) \) for \( n = 15 \). Then, partial top-rankings are obtained by randomly censoring all the complete sequences according to two patterns of truncation, labelled respectively as A and B and characterized by an increasing rate of missingness of the items ranked in the bottom positions. In both censoring scenarios, the random number of censored bottom positions varied in \( \{2, 3, 4, 5, 6\} \), but it was sampled by using probability masses \( \{0.1, 0.1, 0.1, 0.1, 0.1\} \) under the censoring scenario A and \( \{0.1, 0.2, 0.4, 0.2, 0.1\} \) under censoring scenario B.

From Table 5 we note that the ability to recover the exact parameter value, as well as the proportion of times that the homogeneous MMS is correctly selected, decreases as the rate of missingness increases. However, for both censoring scenarios, all the criteria exhibit very good values when compared to those obtained in the full ranking analysis, testifying the effectiveness of our EM procedure to draw inference from partial observations.

5 Applications

5.1 Analysis of the sports data

We revisit the well-known sports benchmark data (Marden 1995), available in the R package Rankcluster (Jacques et al. 2014), which collects \( N = 130 \) full rankings. These represent the preferences of students at the University of Illinois towards \( n = 7 \) sports, namely Baseball, Football, Basketball, Tennis, Cycling, Swimming and Jogging. We explored the presence of a cluster structure in the sample through the estimation and comparison of alternative mixture models with \( G = 1, \ldots, 5 \), and components given by: 1) the MM with Spearman distance (MMS); 2) the MM with Kendall distance (MK); 3) the MM with Cayley distance (MMK); 4) the MM with Hamming distance (MMH) and 5) the Plackett-Luce model (PL).

Table 6 shows the best fitting mixture under each parametric scenario, and the corresponding estimated number \( \hat{G} \) of clusters. With the only exception of the MMH that does
Fig. 4  Screeplot of BIC values for alternative number of groups in the MMS-mixture fitted to the reading genres data

not identify the presence of groups (\( \hat{G} = 1 \)), all the other models agree in recognizing the existence of two clusters. However, one can note that our MMS is the model associated with the best fitting performance (BIC=2144.5). The corresponding cluster-specific parameter estimates (mixing weights, concentration and modal orderings) are displayed below.

\[
\hat{\omega}_1 = 0.61 \quad \hat{\theta}_1 = 0.039 \\
\hat{\rho}_1 = \text{Cycling, Novel, Thrillers, Fantasy, Biography, Teenage} \\
\hat{\omega}_2 = 0.39 \quad \hat{\theta}_2 = 0.118 \\
\hat{\rho}_2 = \text{Classic, Novel, Thrillers, Fantasy, Biography, Teenage, Horror, Comics, Poetry, Essay and Humor}
\]

Both groups share the middle position assigned to Cycling, but they are clearly distinguished by top and bottom ranks. In fact, the largest group (\( \hat{\omega}_1 = 0.61 \)) is characterized by the preference towards individual sports, whereas team sports are the most-liked by students of the smallest group (\( \hat{\omega}_2 = 0.39 \)).

5.2 Analysis of the reading genres

In this application we illustrate the effectiveness of our method for handling partial rankings. To this purpose, we consider a brand new dataset that has not been analyzed earlier in the ranking literature. This was obtained from an online survey conducted in Italy in 2019 to investigate reading preferences. It was promoted by the municipal administration of Latina (Latium, Italy), in collaboration with Sapienza University of Rome and the School of Government of the University of Tor Vergata. The dataset is composed of \( N = 507 \) partial top-5 rankings of \( n = 11 \) reading genres ordered by the respondents according to their personal preferences. The genres are Classic, Novel, Thrillers, Fantasy, Biography, Teenage, Horror, Comics, Poetry, Essay and Humor.

Due to the large sample size and the heterogeneity of the respondents with respect to factors typically impacting on the reading preferences, such as gender, age and education level, we expect the existence of different groups. We explored this hypothesis through the estimation of \( G \)-component MMS-mixtures with \( G = 1, \ldots, 6 \). Figure 4 shows that BIC values remarkably decrease up to 3 components. Moreover, for \( G \geq 5 \) a cluster with a negligible size is always estimated. According to the elbow rule, we considered \( \hat{G} = 3 \) as the optimal number of groups and reported the corresponding cluster-specific parameter estimates in Table 7. The estimates of the modal rankings reveal that more serious and ponderous readings, involving Classic and Essay, are ranked in high positions by respondents of the third cluster, whereas all the genres which are typically the most-liked ones among younger people all occupy bottom positions. This preference pattern is completely reversed in the modal ranking of the second group, with all soft readings (Fantasy, Thrillers, Horror and Comics) ranked in higher positions. The first component represents more hybrid preferences with a mix of the two aforementioned reading types in top and bottom positions. Interestingly, the estimated group structure turned out to be
significantly associated to several individual aspects influencing reading attitudes (see Fig. 5). In fact, the third group is mainly characterized by respondents aged 20 or older, who read many books in the last year and holding a master or doctoral degree. On the other hand, respondents of the second group are mostly teenagers and reported a lower propensity to the reading habit during the last year. Finally, we note that the percentage of males in the second group is remarkably higher than in the other two.

### 5.3 An illustration based on gene expression data

Finally, we apply our methodology on a micro-array gene expression dataset, to illustrate the ability of our method to analyze heterogeneous rankings of a large number of items.

We use data adapted from Khan et al. (2001), which contains micro-array gene expression profiles of four types of small round blue cell tumours (SRBCT) of childhood for $N = 64$ patients. These four different cancers, namely neuroblastoma (NB), Burkitt lymphom (BL), rhabdomyosarcoma (RMS), and the Ewing family of tumours (EWS), present a similar histology of SRBCT and thus often leads to misdiagnosis.

In this illustration we aim at exploiting the MMS-mixture to classify patients into the 4 different cancer types, based on their ranked gene expression profiles.

We select the $n = 60$ most differentially expressed genes throughout the sample and, for each patient, we convert the individual quantitative expression profile into a ranking. The rank transformation is sometimes used in genomics applications, because inference based on rankings is both more robust than working on the actual continuous measurements (Eliseussen et al. 2022), and avoids the arbitrary choice of a numerical pre-processing procedure to handle the raw outcomes (Mollica and Tardella 2014).

We then estimate a 4-component MMS-mixture and check the agreement between the model-based classification with the one based on the true tumour type.

Results are reported in Table 8, where we see that we classify all tumours correctly except one EWS which is erroneously assigned to the RMS group. From Table 9, where we report the classification implied by the simple hierarchical clustering, we note that this method is not able to correctly classify a considerable subset of RMS patients, which are instead erroneously mixed with the BL ones.

In Fig. 6 we plot the tumours based on their gene expression levels in the 2D space of the top-2 principal components of the gene expression levels of the sample. In the figure, the tumours are colored based on their classification obtained with the MMS. It can be seen that two groups classified by the MMS (groups 2 and 3) are highly separable in this space, while groups 1 and 4 appear less separable.

The mixture of MMS also provides the consensus rankings of the four components, which correspond to the estimated modal orderings of genes for the four cancer types. In Table 10, we report the top-3 expressed genes of the different tumours, which are in line with the previous literature.
Fig. 6 Visualization of MMS classification of cancers. The axes are the first two principal components of the original quantitative gene expression levels that identified them as those capable of diagnosing and discerning the considered cancers types (see, for instance, Table 2 in the work by Chen et al. 2007).

6 Conclusions and future work

The specific properties and effective usefulness of the Spearman distance for the analysis of ranking data through the MM has been barely explored in the literature, mainly because of the intractability of the model normalizing constant. Nevertheless, the Spearman distance is a natural candidate in the preference domain, corresponding to the unnormalized version of the popular Spearman rank correlation which measures the concordance between ordinal observations. Motivated by the valuable properties of this metric, including the existence of an analytical solution for the MLE of the modal consensus, in this work we introduce an original approximation of the model normalizing constant. This allows to conveniently encompass the critical inferential steps which have traditionally limited the use of the MMS in practice.

To further enlarge the applicability of the MMS, we also extend the MMS to the finite mixture framework and develop an efficient and accurate EM algorithm for handling both full rankings and partial sequences affected by censoring of arbitrary positions.

Thanks to our contributions, there is no longer computational reason to favor the decomposable metrics over the Spearman in the application of the MM or mixtures thereof.

Moreover, with several applications to real-world data, we prove that our proposals for the MMS have the merit of allowing the use of a distance-based ranking model which combines distinctive meaningful properties and computational advantages.

Of course, our original research can be further extended in several directions. From a computational perspective, an R package to promote a wider use of the MMS and mixtures thereof is currently under construction. From a methodological perspective, model-based clustering could be enhanced by integrating the finite mixture framework with the regression analysis, for identifying significant factors affecting the preferences and better characterizing the groups of rankers. This idea is currently under study (Crispino et al. 2023) by developing a mixture of experts model for the MMS, in the spirit of Gormley and Murphy (2008b). Another valuable methodological contribution could be the extension of the normalizing constant approximation for handling a weighted form of the Spearman distance. This would allow to account for a possible differential uncertainty or importance which typically characterizes the assignment of the single positions in real-word ranking data. Finally, the investigation of the robustness properties of the MMS consensus ranking could be a worthy line of future research. In this regard, some pioneering results on the Borda estimator can be found in Goibert et al. (2022) and Goibert et al. (2023).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-023-10266-8.

Acknowledgements The authors would like to thank an anonymous reviewer for her/his constructive and insightful comments that greatly improved the manuscript. The authors wish also to thank prof. Enrico Casadio Tarabusi for his precious suggestions on the approximation of the Spearman distance distribution. This work was supported by the grant N. RP12117A89FFE8A0 funded by Sapienza University of Rome. The opinions expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy or the Eurosystem.

Author contributions Crispino and Mollica are both first authors as they contributed equally to the development of this work as well as to write the main manuscript text. Astuti developed and described the new approximation in Section 3 and Supplementary Material A. Tardella mainly contributed to implement and computationally optimize the EM algorithm for statistical inference described in Section 2. All authors reviewed the manuscript.

Declarations

Conflict of interest The authors declare no conflict of interest.
References

Ali, A., Meilă, M.: Experiments with Kemeny ranking: What works when? Math. Soc. Sci. 64(1), 28–40 (2012)
Alvo, M., Yu, P.L.H.: Statistical methods for ranking data. Frontiers in Probability and the Statistical Sciences. Springer, New York, USA (2014)
Andrieu, C., Roberts, G.O.: The pseudo-marginal approach for efficient Monte Carlo computations. Ann. Stat. 37(2), 697–725 (2009)
Bartholdi, J.J., Tovey, C.A., Trick, M.A.: The computational difficulty of manipulating an election. Soc. Choice Welf. 6(3), 227–241 (1989)
Beckett, L.A.: Maximum likelihood estimation in mallows’ model using partially ranked data. In: Fligner, M.A., Verducci, J.S. (eds.) Probability Models and Statistical Analyses for Ranking Data, pp. 92–107. Springer, New York (1993)
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)
Caron, F., Teh, Y.W., Murphy, T.B.: Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. Ann. Appl. Stat. 8(2), 1145–1181 (2014)
Chen, Q.R., Vansant, G., Oades, K., Pickering, M., Wei, J.S., Song, Y.K., Monforte, J., Khan, J.: Diagnosis of the small round blue cell tumors using multiplex polymerase chain reaction. J. Mol. Diagn. 9(1), 80–88 (2007)
Cramer, H.: Sur un nouveau théoreme-limite de la théorie des probabilités. Actualités scientifiques et industrielles 736, 2–23 (1938)
Crispino, M., Modugno, L., Mollica, C.: The Mallows model with respondents’ covariates. Preliminary draft available upon request. (2023)
Croux, C., Dehon, C.: Influence functions of the Spearman and Kendall correlation measures. Stat. Methods Appl. 19, 497–515 (2010)
Deheuvels, P., Vovelle, P.: Boundary consistent estimation of the extremal index. Ann. Stat. 21, 101–110 (1993)
Diaconis, P.: Group representations in probability and statistics. Institute of Mathematical Statistics. (1988)
Eliseussen, E., Fleischer, T., Vitelli, V.: Rank-based Bayesian variable selection for genome-wide transcriptomic analyses. Stat. Med. 41(23), 4532–4553 (2022)
Ellis, R.S.: Large deviations for a general class of random vectors. Ann. Prob. 12(1), 1–12 (1984)
Feigin, P., Cohen, A.: On a model for concordance between judges. J. R. Stat. Soc. Ser. B Method. 40(2), 203–213 (1978)
Fligner, M.A., Verducci, J.S.: Distance based ranking models. J. R. Stat. Soc. B 48(3), 359–369 (1986)
Gärtner, J.: On large deviations from the invariant measure. Theory Prob. Appl. 22(1), 24–39 (1977)
Goibert, M., Calauzènes, C., Irurozki, E., Clémençon, S. (2023). Robust consensus in ranking data analysis: Definitions, properties and computational issues. arXiv preprint arXiv:2303.12878
Goibert, M., Clémençon, S., Irurozki, E., Mozharovskyi, P. (2022). Statistical depth functions for ranking distributions: definitions, statistical learning and applications. arXiv preprint arXiv:2201.08105
Gormley, I.C., Murphy, T.B.: Exploring voting blocs within the Irish electorate: a mixture modeling approach. J. Am. Stat. Assoc. 103(483), 1014–1027 (2008)
Gormley, I.C., Murphy, T.B.: A mixture of experts model for rank data with applications in election studies. Ann. Appl. Stat. 2(4), 1452–1477 (2008)
Henery, R.J.: Permutation probabilities as models for horse races. J. R. Stat. Soc. Ser. B Stat. Method. 43(1), 86–91 (1981)
Irurozki, E.: Sampling and learning distance-based probability models for permutation spaces. Ph. D. thesis, Department of Computer Science and Artificial Intelligence of the University of the Basque Country. (2014)
Irurozki, E., Calvo, B., Lozano, A.: PerMallows: an R package for Mallows and generalized Mallows models. J. Stat. Softw. 71(12), 1–30 (2016)
Irurozki, E., Calvo, B., Lozano, A.: Sampling and learning the Mallows and generalized Mallows models under the Cayley distance. Method. Comput. Appl. Prob. 20(1), 1–35 (2018)
Irurozki, E., Calvo, B., Lozano, J.A.: Mallows and generalized Mallows model for matchings. Bernoulli 25(2), 1160–1188 (2019)
Jacques, J., Grimondrez, Q., Biernacki, C.: Rankcluster: An R package for clustering multivariate partial rankings. J. R. Stat. Soc. 61(2), 101–110 (2014)
Kendall, M.G.: Rank correlation methods, 4th edn. Griffin, London (1970)
Khan, J., Wei, J.S., Ringnér, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7(6), 673–679 (2001)
Lee, P.H., Yu, P.L.H.: Distance-based tree models for ranking data. Comput. Stat. Data Anal. 54(6), 1672–1682 (2010)
Lee, P.H., Yu, P.L.H.: Mixtures of weighted distance-based models for ranking data with applications in political studies. Comput. Stat. Data Anal. 56(8), 2486–2500 (2012)
Lee, P.H., Yu, P.L.H.: An R package for analyzing and modeling ranking data. BMC Med. Res. Method. 3(1), 1–11 (2013)
Liu, Q., Reiner, A., Frigessi, A., Scheel, I.: Diverse personalized recommendations with uncertainty from implicit preference data with the Bayesian Mallows model. Knowl. Based Syst. 186, 104960 (2019)
Mallows, C.L.: Non-null ranking models. I. Biometrika 44(1/2), 114–130 (1957)
Marden, J.I.: Analyzing and Modeling Ranking Data, Volume 64 of Monographs on Statistics and Applied Probability. Cambridge, USA: Chapman & Hall. (1995)
McCullagh, P.: Models on spheres and models for permutations. In: Fligner, M.A., Verducci, J.S. (eds.) Probability Models and Statistical Analyses for Ranking Data, pp. 278–283. Springer, New York (1993)
Meilă, M., Bao, L.: An exponential model for infinite rankings. J. Mach. Learn. Res. 11, 3481–3518 (2010)
Møller, J., Pettitt, A.N., Reeves, R., Berthelsen, K.K.: An efficient Markov Chain Monte Carlo method for distributions with intractable normalising constants. Biometrika 93(2), 451–458 (2006)
Mollica, C., Tardella, L.: Epitope profiling via mixture modeling for ranked data. Stat. Med. 33(21), 3738–3758 (2014)
Mukherjee, S.: Estimation in exponential families on permutations. Ann. Stat. 44(2), 853–876 (2016)
Murphy, T.B., Martin, D.: Mixtures of distance-based models for ranking data. Comput. Stat. Data Anal. 41(3–4), 645–655 (2003)
Murray, I., Ghahramani, Z., MacKay, D.: MCMC for doubly-intractable distributions. (2012) arXiv preprint arXiv:1206.6848
Qian, Z., Yu, P.L.H.: Weighted distance-based models for ranking data using the R package rankdist. J. Stat. Softw. 90(5), 1–31 (2019)
Schwarz, G.: Estimating the dimension of a model. Ann. Stat. 6(2), 461–464 (1978)
Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., Liu, J.: A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP J. Wireless Commun. Network. 2021(1), 31 (2021)
Sloane, N. J. A.: The Encyclopedia of Integer Sequences. (2017)
Sørensen, Ø., Crispino, M., Liu, Q., Vitelli, V.: BayesMallows: an R Package for the Bayesian Mallows Model. R J. 12(1), 324–342 (2020)

Tarsitano, A. et al. Comparing the effectiveness of rank correlation statistics. P: Dip. di Economia e Statistica, University of della Calabria. (2009)

Varadhan, S.S.: Asymptotic probabilities and differential equations. Commun. Pure Appl. Math. 19(3), 261–286 (1966)

Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A., Arjas, E.: Probabilistic preference learning with the Mallows rank model. J. Mach. Learn. Res. 18(158), 1–49 (2018)

Xu, H., Alvo, M., Yu, P.L.H.: Angle-based models for ranking data. Comput. Stati. Data Anal. 121, 113–136 (2018)

Zhao, Q., Hautamäki, V.: Knee Point Detection in BIC for Detecting the Number of Clusters. In Blanc-Talon, J., Bourouanne, S., Philips, W., Popescu, D., Scheunders, P. (eds) Advanced Concepts for Intelligent Vision Systems. ACIVS 2008. Lecture Notes in Computer Science, vol 5259. Springer, Berlin, Heidelberg, pp. 664–673. (2008)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.