Bayesian Analysis (2023) **18**, Number 2, pp. 391–414

Informative Priors for the Consensus Ranking in the Bayesian Mallows Model

Marta Crispino[*] and Isadora Antoniano-Villalobos[†,‡]

Abstract. The aim of this work is to study the problem of prior elicitation for the consensus ranking in the Mallows model with Spearman's distance, a popular distance-based model for rankings or permutation data. Previous Bayesian inference for such a model has been limited to the use of the uniform prior over the space of permutations. We present a novel strategy to elicit informative prior beliefs on the location parameter of the model, discussing the interpretation of hyper-parameters and the implication of prior choices for the posterior analysis.

Keywords: Bayesian subjective inference, conjugate priors, Mallows model for rankings, ranking data, permutations, permutohedron.

MSC2020 subject classifications: 62F15, 62F07.

1 Motivation

In recent years, interest in preference data has increased, in part due to internet-related activities. The study of rankings, in particular, has received special attention, since this type of data arise in many fields. Notable examples are electoral systems in which voters are required to rank candidates, as is the case of the Irish general elections (Gormley and Murphy, 2008); automatic recommender systems seeking to aggregate preferences in order to suggest products to the customers (Sun et al., 2012); market research based on surveys in which competing services, or items, are compared or ranked by customers (Dabic and Hatzinger, 2009); medical applications, especially in genomics, in which genes are sometimes ranked according to their expression levels under various experimental conditions (Vitelli et al., 2018), and other data is often transformed into rankings in order minimize the effect of miscalibration error from the measuring devices (Mollica and Tardella, 2014).

The Mallows model (MM) (Mallows, 1957; Diaconis, 1988) is a popular two-parameter distance-based family of models for ranking data, based on the assumption that a modal ranking, which can be interpreted as the consensus ranking of the population, exists. The probability of observing a given ranking is then assumed to decay exponentially fast as its distance from the consensus grows. Individual models with different properties can be obtained depending on the choice of distance on the space of permutations. The scale or precision parameter, controlling the concentration of the distribution, determines the rate of decay of the probability of individual ranks.

[*] Bank of Italy, Via Nazionale, 91 - 00184 Rome, Italy, marta.crispino@bancaditalia.it

[†] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy, isadora.antoniano@unive.it

[‡] Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy

https://doi.org/10.1214/22-BA1307

We focus on the Mallows model with Spearman's distance (MMS), introduced by (Mallows, 1957) with the name of rho-model, since Spearman's distance, when re-scaled to lie between -1 and 1, arises naturally as the correlation between the ranks of two samples. Marden (1995) and Vitelli et al. (2018) have studied Bayesian inference for the MMS, limiting the analysis to the use of a uniform, non-informative prior on the consensus ranking.

Within the Bayesian literature, non-informative and objective priors can be used to provide a sense of neutrality to the analysis by allowing the data to be the only source of information in the estimation procedure. However, when information is available from experts or external sources, it may be argued that a fully Bayesian analysis should include this subjective prior belief. Dawid (1997) clearly stated that "no theory which incorporates non-subjective priors can truly be called Bayesian, and no amount of wishful thinking can alter this reality". While admitting that both approaches may be valid in different situations, in this paper we explore the possibility of including genuine prior information, which might come from a literature review, from an expert or from an earlier data analysis, into the Bayesian Mallows model for ranking data (Vitelli et al., 2018).

Previous proposals to include prior information on the consensus ranking of a MM include Gupta and Damien (2002), who suggest eliciting a prior on the consensus which is constant on conjugacy classes. In other words, they propose a prior that assigns *a priori* equal probability to all permutations with the same cyclic structure. However, the conjugate classes defined by cyclic structures do not coincide with those defined by permutations lying at the same distance (e.g. Spearman's) from the consensus ranking, making this approach impractical for the MMS, as it is difficult to assess a way in which prior information enters the model. Meilă and Bao (2010) and Meilă and Chen (2010) consider the MM with Kendall's distance within the Bayesian paradigm and provide a conjugate prior for the model parameters which is known up to a normalization constant. However, their analysis does not extend to the MMS. Xu et al. (2018) propose an alternative family of models for rankings, based on a mapping of the data to the unit sphere (see also McCullagh, 1993). The location parameter of their model has an interpretation analogous to that of the consensus ranking but it is not limited to be itself a ranking, thus allowing to express a more general form of consensus. The MMS is a particular case of this model, and the authors propose a conjugate Bayesian prior for the consensus parameter. However, the emphasis of the paper is on efficient inference via an approximation of the model's normalizing constant and the use of variational methods; prior elicitation and the inclusion of prior information are not discussed. In a different setting, when data consist of rankings which vary in time, Asfaw et al. (2017) introduce a dynamic version of the Bayesian Mallows model and assume a smoothing prior for modelling the slow time-varying consensus ranking.

In the present work, which stems from Chapter 6 of Crispino (2017), we aim to provide experts using the MMS with a tool to express their beliefs, knowing the effect of prior choices in their analysis, should they wish to do so. With this in mind, by exploiting the notion of permutohedron, also known as permutation polytope, (Thompson, 1993; McCullagh, 1993; Marden, 1995), we propose an explicit form for a conjugate prior on the consensus parameter for the MMS. We then study its properties, presenting some theoretical insights on the prior elicitation problem. Subjective prior information on the consensus ranking can therefore be elicited by choosing appropriate hyper-parameters. The proposed prior density can handle a situation when only partial information is available, which is particularly relevant when the set of items to be ranked is very large. In such cases it is unlikely that a full ranking is *a priori* available, while it could be possible to express some prior belief regarding which are the most (or least) preferred items. An additional advantage of our prior is given by the interpretability of the hyper-parameters in terms of the amount and type of information included.

We initially assume the scale parameter of the MMS to be known, given that in most applications it is considered a nuisance, the interest being focused on the estimation of the consensus ranking (see Vitelli et al., 2018, Section 3). In the more realistic case when the scale parameter is unknown, multiple approaches are possible. For instance, in Vitelli et al. (2018) an exponential prior density is proposed; In Marden (1995), Section 6.4, the conjugate prior for the scale parameter is used, when a uniform prior density for the location is employed. In this manuscript we propose as an alternative a reference prior on the scale parameter, which is a valid option when no prior information on this parameter is available.

The paper is organized as follows. In Section 2 we give an overview of the MMS. In Section 3 we discuss the novel results regarding the conjugate prior for the consensus parameter of the MMS, initially assuming the dispersion parameter to be known (Section 3.1), then (Section 3.2) working with both parameters unknown. In Section 4 we outline the MCMC algorithm used to perform inference on our model, and in Section 5 we illustrate the inference on simple examples, exploiting both simulations and real datasets. We conclude with some final remarks in Section 6.

2 Preliminaries

A (full) ranking of n items, or n-ranking is defined as a map from a finite set, $\{A_1, \dots, A_n\}$, of labeled items to the space $\mathcal{P}_n$ of n-dimensional permutations. A ranking can, therefore, be represented by a vector $\boldsymbol{r} = (r_1, \dots, r_n)$, where r_i is the rank assigned to item A_i according to some criterion. Formally, individual ranks are ordinal numbers, so that $r_i < r_j$ when item A_i is preferred to (ranked lower than) item A_j. Alternatively, rank data may be represented through orderings, which are ordered vectors of labels. Clearly, there is a one-to-one relationship between the two representations, e.g. a possible ranking of the set $A_1, \dots, A_5$ is $\boldsymbol{r} = (1, 3, 4, 5, 2)$, corresponding to the ordering $\boldsymbol{o} = (A_1, A_5, A_2, A_3, A_4)$. Since the ranking vector representation has many advantages in terms of modelling, we will stick to it throughout the paper, and only use the orderings when necessary for illustrative purposes. Given the trivial one-to-one relation between ordinal and cardinal numbers, with a slight abuse of notation, one may consider n-rankings as n-dimensional vectors obtained by permuting the first natural numbers, $\{1, \dots, n\}$. It is then easy to see that $\mathcal{P}_n$ is contained in a $(n-1)$-dimensional affine subspace of $\mathbb{R}^n$. In fact, it is composed by the $n!$ points on the intersection between the hyper-plane with coordinate sums equal to $s_n = n(n+1)/2$ and the surface of an n-dimensional sphere of squared radius $c_n = n(n+1)(2n+1)/6$ centered at the origin. Thus, all the points of $\mathcal{P}_n$ lie on an $(n-1)$-dimensional sphere of squared radius $\phi_n = n(n^2-1)/12$ centered at $\frac{(n+1)}{2}\mathbf{1}_n$, where $\mathbf{1}_n \in \mathbb{R}^n$ denotes the vector with all entries equal to 1 (McCullagh, 1993).

The Mallows model for ranking data (Mallows, 1957) defines the probability that a random n-ranking $\boldsymbol{R}$ takes a value $\boldsymbol{r} \in \mathcal{P}_n$ as

$$\mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \,|\, \boldsymbol{\rho}, \theta, d) = \frac{1}{Z_d(\theta)} \exp\left[-\theta\, d(\boldsymbol{r}, \boldsymbol{\rho})\right], \tag{1}$$

where $\boldsymbol{\rho} \in \mathcal{P}_n$ is a location parameter representing the shared consensus ranking and $\theta \geq 0$ is a scale parameter describing the concentration of the mass around the shared consensus. Different families of models are obtained through different choices of the right-invariant (Diaconis, 1988) distance $d(\cdot,\cdot)$ on $\mathcal{P}_n$. Right-invariance, which ensures that distances are independent of any relabeling of the items, is an important property in this context, as it ensures that the partition function $Z_d(\theta) = \sum_{\boldsymbol{r}\in\mathcal{P}_n} e^{-\theta d(\boldsymbol{r},\boldsymbol{\rho}_I)}$ of the MM does not depend on $\boldsymbol{\rho}$ (Mukherjee, 2016; Vitelli et al., 2018). In the above expression $\boldsymbol{\rho}_I = (1, 2, 3, \ldots, n)$ denotes the identity permutation. Nevertheless, the number of terms in the sum makes direct calculation of this partition function unfeasible for all but very small values of n. As a consequence, the MM is known up to a proportionality constant, except for some particular choices of the distance, for which Z_d has a closed form (Fligner and Verducci, 1986). Different approximation strategies have been proposed (see e.g. McCullagh, 1993; Mukherjee, 2016; Vitelli et al., 2018), allowing inference even with a large number, n, of items. Notice that the distance function induces a partition of $\mathcal{P}_n$ formed by sets of rankings which are equidistant from $\boldsymbol{\rho}$. Within each partition set, the MM assigns equal probability to all rankings. As a consequence, exact computation of the partition function is possible for moderate n, for some choices of $d(\cdot,\cdot)$ for which the cardinalities of the partition sets are known (see e.g. Irurozki et al., 2016; Vitelli et al., 2018). The partitions of $\mathcal{P}_n$ associated to Spearman's distance play a crucial role in understanding the behavior of the prior proposed here for the MMS.

In this work we focus on the Mallows model with Spearman's distance, given by $d_S(\boldsymbol{r}, \boldsymbol{\rho}) = ||\boldsymbol{r} - \boldsymbol{\rho}||^2 = \sum_{i=1}^n (\rho_i - r_i)^2$, for $\boldsymbol{r}, \boldsymbol{\rho} \in \mathcal{P}_n$, which was first introduced by Mallows (1957). Notice that Spearman's distance is an unnormalized version of the Spearman's rank correlation, used to measure the statistical correlation between the ranks of two variables, but, when rankings are considered as vectors in $\mathbb{R}^n$, it is simply the squared Euclidean distance, or L_2-norm. Therefore, we say that a random ranking $\boldsymbol{R}$ follows an MMS distribution, denoted by $\boldsymbol{R}|\boldsymbol{\rho}, \theta \sim \mathcal{M}(\boldsymbol{\rho}, \theta)$, if its probability mass function is given by

$$p(\boldsymbol{R} \,|\, \boldsymbol{\rho}, \theta) := \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \,|\, \boldsymbol{\rho}, \theta) = \frac{1}{Z(\theta)} \exp\left[-\theta\, \|\boldsymbol{\rho} - \boldsymbol{R}\|^2\right], \tag{2}$$

where $Z(\theta) := Z_{d_S}(\theta)$ does not have a closed form. Notice that when $\theta = 0$, the MMS reduces to the uniform distribution on $\mathcal{P}_n$.

Given a sample $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N | \boldsymbol{\rho}, \theta \overset{iid}{\sim} \mathcal{M}(\boldsymbol{\rho}, \theta)$, the likelihood function takes the form

$$p(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N | \boldsymbol{\rho}, \theta) = \frac{1}{Z(\theta)^N} \exp \left[-\theta \sum_{j=1}^{N} \|\boldsymbol{\rho} - \boldsymbol{R}_j\|^2 \right]. \tag{3}$$

In most applications the parameter θ is considered a nuisance and the main interest is in the estimation of $\boldsymbol{\rho}$. It can be shown that, for $\theta > 0$, the maximum likelihood estimator (MLE) $\boldsymbol{\rho}_{\text{MLE}}$ is given by

$$\boldsymbol{\rho}_{\text{MLE}} = \underset{\boldsymbol{\rho} \in \mathcal{P}_n}{\operatorname{argmin}} \sum_{j=1}^{N} \|\boldsymbol{\rho} - \boldsymbol{R}_j\|^2 = \underset{\boldsymbol{\rho} \in \mathcal{P}_n}{\operatorname{argmax}}\, \boldsymbol{\rho} \cdot \bar{\boldsymbol{R}} = \boldsymbol{Y}(\bar{\boldsymbol{R}}),$$

where the dot denotes the scalar product on $\mathbb{R}^n$, $\bar{\boldsymbol{R}} = (\bar{R}_1, \ldots, \bar{R}_n)$ is the sample mean vector of $\bar{R}_i = \frac{1}{N} \sum_{j=1}^{N} R_{ij}$, $i = 1, \ldots, n$, and $\boldsymbol{Y}(\boldsymbol{r}) = (Y_1(\boldsymbol{r}), \ldots, Y_n(\boldsymbol{r})) \in \mathcal{P}_n$ is the rank transformation of vector $\boldsymbol{r}$, whose coordinates are defined as $Y_i = Y_i(\boldsymbol{r}) = \sum_{h=1}^{n} \mathbb{1}(r_h \leq r_i)$, $i = 1, \ldots, n$, $\mathbb{1}(E)$ being the indicator function of the event E.

In the remainder, we propose and study an informative prior density for $\boldsymbol{\rho}$, specifically tailored to the MMS, building on the Bayesian Mallows model for ranking data of Vitelli et al. (2018).

3 An informative prior

This section is devoted to the proposal of a prior distribution for the $\boldsymbol{\rho}$ parameter of the MMS. In Section 3.1 we analyze the simpler case in which the precision parameter θ is assumed known. Then, in Section 3.2, we give an intuition on how to deal with the more general and realistic case of unknown θ.

3.1 Known precision parameter

For fixed θ and $\boldsymbol{\rho} \in \mathcal{P}_n$, the likelihood (3) can be simplified as

$$p(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N | \boldsymbol{\rho}, \theta) \propto \exp \left[2\theta \sum_{j=1}^{N} \boldsymbol{\rho} \cdot \boldsymbol{R}_j \right] \propto \exp \left(2\theta N \boldsymbol{\rho} \cdot \bar{\boldsymbol{R}} \right). \tag{4}$$

Notice that the sample mean $\bar{\boldsymbol{R}}$ belongs to the permutohedron of order n, denoted by $\mathbb{pp}_n$, that is the convex hull of the points $\boldsymbol{\rho} \in \mathcal{P}_n \subset \mathbb{R}^n$. The set $\mathbb{pp}_n$ is sometimes called the *permutation polytope* (see e.g. Thompson, 1993; Marden, 1995). This term, however, refers also to a similar polytope whose vertices follow a different order. We, here, use the term permutohedron to avoid ambiguity.

A conjugate prior for $\boldsymbol{\rho} \in \mathcal{P}_n$ is given by

$$\pi(\boldsymbol{\rho} | \boldsymbol{\rho}_0, \eta_0) = \frac{1}{Z^*(\eta_0, \boldsymbol{\rho}_0)} \exp \left[-\eta_0 ||\boldsymbol{\rho}_0 - \boldsymbol{\rho}||^2 \right] \propto \exp \left[2\eta_0\, \boldsymbol{\rho} \cdot \boldsymbol{\rho}_0 \right]. \tag{5}$$

We call this the Extended Mallows Model with Spearman distance (EMMS) and write $\boldsymbol{\rho}|\eta_0, \boldsymbol{\rho}_0 \sim \mathcal{EM}(\boldsymbol{\rho}_0, \eta_0)$. Note that the conjugate prior (5) is analogous to the angle-based model proposed by Xu et al. (2018), originally developed in McCullagh (1993). The two hyper-parameters $\eta_0 \geq 0$ and $\boldsymbol{\rho}_0 \in \mathbb{pp}_n$ can be interpreted as precision and location parameters, respectively, analogous to those of the MMS. In particular, η_0 determines the concentration of the distribution around $\boldsymbol{\rho}_0$, with $\eta_0 = 0$ corresponding to a uniform prior on $\mathbb{pp}_n$ while larger values reflect a stronger prior belief on $\boldsymbol{\rho}_0$. Notice however that, differently from the MMS, the modal parameter $\boldsymbol{\rho}_0$ is not, in general, a permutation, except when it lies on the vertices of the permutohedron $\mathbb{pp}_n$. Recall that Mallows models have the limitation that all rankings which are equidistant (in terms of the distance in (1)) from the consensus ranking have the same probability. For the MMS, this implies in particular that it is not possible to freely assign different masses to different rankings at the same Spearman's distance to the consensus ranking. By allowing the modal parameter of (5) to take any value in the permutohedron $\mathbb{pp}_n$, that is, to be any convex combination of the elements of $\mathcal{P}_n$, such structure can be broken, allowing for a more flexible distribution of the mass. In fact, the prior (5) assigns equal mass to all permutations that lie at the same L_2-norm from $\boldsymbol{\rho}_0$, and greater mass is given to permutations closest to $\boldsymbol{\rho}_0$. For instance, consider the EMMS centered at the barycenter of the permutohedron, that is, with $\boldsymbol{\rho}_0 = \frac{(n+1)}{2}\mathbf{1}_n$. This results in a uniform distribution on rankings for any value of the precision parameter η_0. Small deviations from uniformity can be achieved by letting $\eta_0 > 0$ and $||\boldsymbol{\rho}_0 - \frac{(n+1)}{2}\mathbf{1}_n||^2$ be small. The direction of the vector $\boldsymbol{\rho}_0 - \frac{(n+1)}{2}\mathbf{1}_n$ in $\mathbb{R}^n$ determines the rankings for which the mass increases and those for which it decreases. The case described above, where $\boldsymbol{\rho}_0 = \frac{(n+1)}{2}\mathbf{1}_n$, is therefore equivalent to assigning to $\boldsymbol{\rho}$ the uniform prior on $\mathcal{P}_n$, $\pi(\boldsymbol{\rho}) = \frac{1}{n!}$, like in Marden (1995) and Vitelli et al. (2018).

Note that, since $\boldsymbol{\rho}_0 \in \mathbb{pp}_n$, the partition function in (5),

$$Z^*(\eta_0, \boldsymbol{\rho}_0) = \sum_{\boldsymbol{\rho} \in \mathcal{P}_n} \exp\left[-\eta_0 ||\boldsymbol{\rho}_0 - \boldsymbol{\rho}||^2\right], \tag{6}$$

in general depends on both η_0 and $\boldsymbol{\rho}_0$, unless $\boldsymbol{\rho}_0 \in \mathcal{P}_n \subset \mathbb{pp}_n$, in which case the Z^* is a function only of η_0.[1] This implies that (5) is known up to a normalization constant. However, in the following sections we show that this drawback can be overcome in practice.

The posterior density for $\boldsymbol{\rho}$ is given by

$$\pi^N(\boldsymbol{\rho}\,|\theta) \propto \exp\left[2(\eta_0 + \theta N)\ \boldsymbol{\rho} \cdot \left(\frac{\theta N}{\eta_0 + \theta N}\bar{\boldsymbol{R}} + \frac{\eta_0}{\eta_0 + \theta N}\boldsymbol{\rho}_0\right)\right]. \tag{7}$$

The first thing we observe is that the proposed prior is indeed conjugate. In other words, if $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N \,|\, \boldsymbol{\rho}, \theta \overset{iid}{\sim} \mathcal{M}(\boldsymbol{\rho}, \theta)$ and $\boldsymbol{\rho}\,|\,\boldsymbol{\rho}_0, \eta_0 \sim \mathcal{EM}(\boldsymbol{\rho}_0, \eta_0)$, then it holds that

[1] The independence of $Z^*(\eta_0, \boldsymbol{\rho}_0)$ from $\boldsymbol{\rho}_0$ follows from the property of right-invariance of the Spearman's distance, like it happens for $Z(\theta)$.

$\boldsymbol{\rho} \,|\, \theta, \boldsymbol{\rho}_0, \eta_0, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_N \sim \mathcal{EM}(\boldsymbol{\rho}_N, \eta_N)$, with updated parameters:

$$\boldsymbol{\rho}_N = \frac{\theta N}{\eta_0 + \theta N}\bar{\boldsymbol{R}} + \frac{\eta_0}{\eta_0 + \theta N}\boldsymbol{\rho}_0 \in \mathbb{PP}_n, \tag{8}$$

$$\eta_N = \eta_0 + \theta N \geq 0. \tag{9}$$

The above expressions evoke the classical result (Diaconis and Ylvisaker, 1979) that, under regularity conditions, the posterior estimates have the form of a linear combination of the prior belief and the empirical evidence. Furthermore, the reparametrization of (5) obtained by letting $\eta_0 = \theta_0 N_0$ (with the possibility to choose $\theta_0 = \theta$), shows that the mixing weights of the posterior parameters in (8) explicitly depend on N and N_0 which can be thought of as an *a priori* sample size, representing the amount of information on which the expert bases the prior belief about the central tendency of $\boldsymbol{\rho}$. For any finite prior precision, as the sample size increases, the posterior accumulates mass around $\boldsymbol{\rho}_N$, which approaches the sample mean, $\bar{\boldsymbol{R}}$ as N increases. Some insights into the role of the prior hyper-parameters can be obtained by considering limiting situations. An infinite prior precision would express *a priori* certainty, by accumulating all the prior mass on $\boldsymbol{\rho}_0$. The posterior would maintain the infinite precision thus accumulating mass on $\boldsymbol{\rho}_N = \boldsymbol{\rho}_0$. In such hypothetical case, learning would be possible only for infinite sample sizes, with

$$\lim_{N\to\infty} \boldsymbol{\rho}_N = \begin{cases} \boldsymbol{\rho}_0 & \text{if} \quad N_0/N \to \infty, \\ (1-\alpha)\boldsymbol{\rho}_0 + \alpha\bar{\boldsymbol{R}} & \text{if} \quad N_0/N \to (1/\alpha - 1) \in (0,1), \\ \bar{\boldsymbol{R}} & \text{if} \quad N_0/N \to 0. \end{cases}$$

Notice that, if all the coordinates of the vector $\boldsymbol{\rho}_N$ take different values, the maximum *a posteriori* (MAP) of $\boldsymbol{\rho}$ is unique and given by $\boldsymbol{\rho}_{\mathrm{MAP}} = \boldsymbol{Y}(\boldsymbol{\rho}_N) \in \mathcal{P}_n$.

The prior (5) has a shape which is analogous to the one discussed earlier by Gupta and Damien (2002). In their paper, however, the authors propose the use of the Hausdorff distance among subsets (conjugacy classes) of $\mathcal{P}_n$, in place of the squared L_2-norm between a ranking and the location parameter of the prior (5), which is an element of the permutation polytope. This difference implies that the proposal of Gupta and Damien (2002) assigns equal probability to all permutations within a conjugacy class. In particular, all rankings in the modal conjugacy class of the prior are assigned the same mass, even if information may not be available on all such rankings. Furthermore, two permutations in the same class are not necessarily close with respect to the distance used in the MM, which is a crucial element of the model specification. Our proposal, instead, is specifically tailored to the MMS, and gives the possibility to choose whether to give maximum prior weight to a unique permutation, or to more than one. In Section 5.2 we show the inferential differences resulting from using the prior of Gupta and Damien (2002) and our proposal.

To complete this section, we note in the following Result that the findings in Gupta and Damien (2002, Section 3.3) can be extended to our prior (5).

Result 1. *Let* $D(\boldsymbol{\rho}) = \sum_{j=1}^{N} \|\boldsymbol{R}_j - \boldsymbol{\rho}\|^2$, *and* $D^*(\boldsymbol{\rho}) = \|\boldsymbol{\rho}_0 - \boldsymbol{\rho}\|^2$. *Then:*

a) *for each* $\boldsymbol{\rho}_1, \boldsymbol{\rho}_2 \in \mathcal{P}_n$, *and given* θ, η_0, *the ranking* $\boldsymbol{\rho}_1$ *will have higher posterior probability than* $\boldsymbol{\rho}_2$ *if and only if*

$$D(\boldsymbol{\rho}_1) - D(\boldsymbol{\rho}_2) < \gamma[D^*(\boldsymbol{\rho}_2) - D^*(\boldsymbol{\rho}_1)], \tag{10}$$

where $\gamma = \eta_0/\theta$.

b) *for each* $\boldsymbol{\rho} \in \mathcal{P}_n$, *if* $D^*(\boldsymbol{\rho}) \geq D^*(\boldsymbol{\rho}_{MLE})$, $\boldsymbol{\rho}$ *will have lower posterior probability than* $\boldsymbol{\rho}_{MLE}$.

c) *for each* $\boldsymbol{\rho}_1, \boldsymbol{\rho}_2 \in \mathcal{P}_n$, *if* $D^*(\boldsymbol{\rho}_1) = D^*(\boldsymbol{\rho}_2)$, $\boldsymbol{\rho}_1$ *will have higher posterior probability than* $\boldsymbol{\rho}_2$ *if and only if* $D(\boldsymbol{\rho}_1) < D(\boldsymbol{\rho}_2)$.

d) *for each* $\boldsymbol{\rho}_1, \boldsymbol{\rho}_2 \in \mathcal{P}_n$, *if* $D(\boldsymbol{\rho}_1) < D(\boldsymbol{\rho}_2)$ *and* $D^*(\boldsymbol{\rho}_1) < D^*(\boldsymbol{\rho}_2)$, *then* $\boldsymbol{\rho}_1$ *will have higher posterior probability than* $\boldsymbol{\rho}_2$.

The result, analogous to Gupta and Damien's Theorem 2 and corollaries, gives an intuition on the behavior of the posterior density, by providing a relationship between θ and η_0, that determines which rankings receive the highest posterior probabilities. In Section 5.2 we illustrate, through simulated data, some of the consequences of this result on the inference.

Elicitation of the hyper-parameters

The elicitation involves the two hyper-parameters $(\boldsymbol{\rho}_0, \eta_0)$ of (5) which, as mentioned in the previous section, can be interpreted as a location and a precision parameters, analogous to the parameters $(\boldsymbol{\rho}, \theta)$ of the MMS.

An expert would be asked her prior opinion about the modal (also referred to as consensus) ranking $\boldsymbol{\rho}$, and to express it via the vector $\boldsymbol{\rho}_0$. In the simplest case, we request from the expert a prior modal ranking of all the items $\{A_1, \ldots, A_n\}$. If she were able to provide one, this would result in $\boldsymbol{\rho}_0$ being a proper ranking, that is $\boldsymbol{\rho}_0 \in \mathcal{P}_n$. However, particularly in situations when the set of items to be ranked is very large, the expert may only able to express partial information about the consensus ranking. For example, in the field of Genomics, the number n of items, corresponding to genes, is often of the order of thousands, with the geneticists normally knowing barely a few dozens of them, typically, the k most relevant for their analysis. In such a case, the expert would be asked to rank as many items as possible, say the *in-her-opinion* top-k out of n. Then, $\boldsymbol{\rho}_0$ would contain k elicited ranks, and $n-k$ values equal to $(n+k+1)/2$ corresponding to the items that the expert was not able to rank. This corresponds to assigning the same prior mass to all rankings for which the top-k ranks coincide. The uniform distribution on this class represents the lack of prior information on the ranks of the $n-k$ bottom items. Therefore, the vector $\boldsymbol{\rho}_0$ would not be a ranking, $\boldsymbol{\rho}_0 \notin \mathcal{P}_n$. However, being an element of $\mathbb{pp}_n$, it could still be used as hyper-parameter of the EMMS prior, conveying only partial information about the modal ranking $\boldsymbol{\rho}$.

In a second moment, to elicit a value for η_0, we would reason by calibration, in the spirit of Paganin et al. (2021). Consider the prior expectation $f(\eta_0, \boldsymbol{\rho}_0) := \mathbb{E}_{\pi(\boldsymbol{\rho})}\left[\frac{1}{n}||\boldsymbol{\rho} - \boldsymbol{\rho}_0||^2 | \, \eta_0, \boldsymbol{\rho}_0\right]$. This quantity is decreasing in η_0 for each $\boldsymbol{\rho}_0 \in \mathbb{PP}_n$, and can be interpreted as the expected average (per-item) error in the i-th prior rank ρ_{0i} (Vitelli et al., 2018). A value for η_0 may be found by first asking the expert to choose the a priori per-item expected error size, e_0, and then finding the value of η_0 such that $f(\eta_0, \boldsymbol{\rho}_0) = e_0$. We can also guide the expert in the choice of a reasonable value of e_0, for instance by providing the range of possible values of $f(\eta_0, \boldsymbol{\rho}_0)$, and by asking her to express a belief on the per-item expected error size, as a fraction of such range, for instance $e_0 = 0.5(f_{\max} - f_{\min})$. The minimum and maximum values of $f(\eta_0, \boldsymbol{\rho}_0)$ depend on k and n only (that is, on the partial information carried by $\boldsymbol{\rho}_0$) and can be easily computed, for given $\boldsymbol{\rho}_0$, on a grid of η_0 values.[2]

In a different setting, we can imagine a researcher wishing to include covariate information into the analysis. For instance, a certain number $\boldsymbol{x}_h = (x_{h1}, \ldots, x_{hn})$, $h = 1, \ldots, H$, of covariates may be available, describing some features of the items. We could then introduce this information into the prior (5), by choosing a hyper-parameter $\boldsymbol{\rho}_0 = \boldsymbol{\rho}_0(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_H)$ which depends on the relevant covariates. An example of the latter scenario is given in Section 5.3.

Our proposed prior naturally handles the case when multiple sources of prior information are available. Notice that, since any $\boldsymbol{\rho}_0 \in \mathbb{PP}_n$ can be expressed as a convex combination of rankings in $\mathcal{P}_n$, it can always be interpreted as arising from multiple (possibly infinite) experts, the calculation of the individually elicited parameters being an exercise in linear algebra.

In the elementary case, two experts may believe, *a priori*, in different modal rankings, say $\boldsymbol{\rho}_{01}$ and $\boldsymbol{\rho}_{02}$. An analyst wishing to express an equally strong prior on such two rankings may simply use the prior (5) with $\boldsymbol{\rho}_0 = (\boldsymbol{\rho}_{01} + \boldsymbol{\rho}_{02})/2 \in \mathbb{PP}_n$. More generally, an analyst may like to aggregate prior opinions from a number L of experts by calculating $\boldsymbol{\rho}_0$ and η_0 as a simple average (see e.g. Burgman et al., 2011) of the individual $\boldsymbol{\rho}_{0,\ell}, \eta_{0,\ell}$ parameters elicited from each expert ℓ. A more robust way of aggregating multiple prior opinions is pooling (O'Hagan et al., 2006), which can also allow to weight unequally the different experts' opinions (Genest et al., 1986). However, when it is not reasonable to think that the experts provide independent observations, these approaches may not be adequate.

[2] If $\boldsymbol{\rho}_0$ is the barycenter of the permutohedron, $\boldsymbol{\rho}_0 = \frac{(n+1)}{2}\mathbf{1}_n$ (which corresponds to the case $k = 0$), the prior is uniform, and the choice of η_0 is not relevant (no prior information is available).

If all the ranks of $\boldsymbol{\rho}_0$ are elicited (which corresponds to the case $\boldsymbol{\rho}_0 \in \mathcal{P}_n$, that is $k = n$), then $f_{\min} \equiv \lim_{\eta_0 \to \infty} f(\eta_0) = 0 < f(\eta_0) \leq \frac{n^2-1}{6} = f_{\max} \equiv f(0)$.

If $k = 1, 2, 3, \ldots, n-2$, then

$$f_{\min} = \frac{1}{12} h(h+1)(h+2), \quad h = n - k - 1;$$

$$f_{\max} = \frac{1}{2n} \sum_{i=1}^{k} (n - 2i + 1)^2 + \frac{1}{2n} \sum_{i=1}^{n-k} \left(\frac{n+k+1}{2} - i \right)^2 + \frac{1}{2n} f_{\min}.$$

An interesting way of dealing with dependent experts is to treat the elicited information as data in the spirit of French (2011) and Albert et al. (2012). This latter approach amounts to performing indirect elicitation, by inferring the hyper-parameters of interest using the posterior of a prior Bayesian analysis. In our framework, this means that, instead of directly using a combination of the experts' opinions $\boldsymbol{\rho}_{0,\ell}$, $\ell \geq 1$, as hyper-parameter of (5), we use the $\boldsymbol{\rho}_{0,\ell}, \ell \geq 1$ as conditionally independent data, and elicit the hyper-parameters $\boldsymbol{\rho}_0$ and η_0 based on a prior analysis. More formally, letting $\boldsymbol{\rho}_{0,1}, \ldots, \boldsymbol{\rho}_{0,L} | \boldsymbol{\rho}_0, \eta_0 \overset{iid}{\sim} \mathcal{M}(\boldsymbol{\rho}_0, \eta_0)$, we can infer, based on the posterior density of such an analysis, $\hat{\eta}_0$ and $\hat{\boldsymbol{\rho}}_0$ and set them equal to the hyper-parameters η_0 and $\boldsymbol{\rho}_0$ of (5) respectively. In Section 5.3, we show how this can be done in a practical example.

3.2 Unknown precision parameter

When θ is unknown, the Bayesian paradigm requires a prior on the pair of parameters $(\boldsymbol{\rho}, \theta)$. We here suggest to choose a joint prior of the form $\pi(\boldsymbol{\rho}, \theta) = \pi(\theta)\pi(\boldsymbol{\rho}|\theta)$, where $\pi(\boldsymbol{\rho}|\theta)$ is the EMMS of (5). Notice that the particular case of prior independence, $\pi(\boldsymbol{\rho}, \theta) = \pi(\theta)\pi(\boldsymbol{\rho})$, is achieved in practice by choosing the hyper-parameter η_0 independent of θ. Regarding the choice of $\pi(\theta)$ some proposals are present in the literature, for instance an exponential density (Vitelli et al., 2018), or the conjugate prior of Marden (1995). Both options can be employed in our framework, if the researcher wishes to put some prior information on the θ parameter. As an alternative, we suggest the use of the Jeffreys prior for θ, which, for small values of n, can be computed exactly and may be an interesting alternative when no information on θ is available *a priori*. The following proposition, proved in Crispino and Antoniano-Villalobos (2022), holds for any MM with a right-invariant distance, and in particular for the MMS.

Proposition 1. *The Jeffreys prior for θ in a MM with right-invariant distance d takes the form*

$$\pi_J(\theta) = \sqrt{\mathbb{V}_{\boldsymbol{R}|\theta}\left[d(\boldsymbol{R}, \boldsymbol{\rho}_I)|\theta\right]}, \tag{11}$$

where $\mathbb{V}_{\boldsymbol{R}|\theta}$ denotes the variance with respect to $\boldsymbol{R} \sim \mathcal{M}(\boldsymbol{\rho}_I, \theta)$, which depends on θ.

The posterior density of the model parameters, with the conjugate prior $\pi(\boldsymbol{\rho}|\theta)$ given in (5) is

$$\pi^N(\boldsymbol{\rho}, \theta) \propto \frac{\pi(\theta)}{Z^N(\theta) Z^*(\eta_0, \boldsymbol{\rho}_0)} \exp\left\{-\theta N \left[\left(\left\|\boldsymbol{\rho} - \bar{\boldsymbol{R}}\right\|^2 + c_n - \left\|\bar{\boldsymbol{R}}\right\|^2\right)\right] - \eta_0 \left\|\boldsymbol{\rho}_0 - \boldsymbol{\rho}\right\|^2\right\}. \tag{12}$$

Equation (12) can be easily evaluated in two cases: when (a) Z^* does not depend on θ, that is, when η_0 is independent of θ (prior independence scenario), or when (b) $\eta_0 = \theta N_0$, and n is small enough, so that Z^* can be calculated exactly, for given prior hyper-parameters $\boldsymbol{\rho}_0$ and N_0 (see also Section 4).

The more problematic case (c) when $\eta_0 = \theta N_0$ and n is too large for computing Z^* exactly, can be handled by using as prior density for θ, $\pi_{\text{large n}}(\theta) \propto Z^*(\theta N_0, \boldsymbol{\rho}_0)$, so that the posterior density (12) can be written as

$$\pi^N(\boldsymbol{\rho},\theta) \propto \frac{1}{Z^N(\theta)} \exp\left\{-\theta\left[N\left(\left\|\boldsymbol{\rho}-\bar{\boldsymbol{R}}\right\|^2 + c_n - \left\|\bar{\boldsymbol{R}}\right\|^2\right) + N_0\left\|\boldsymbol{\rho}_0-\boldsymbol{\rho}\right\|^2\right]\right\}. \tag{13}$$

We believe that the choice of $\pi_{\text{large n}}(\theta)$ motivated by the simplification of the posterior represents, nevertheless, a sensible belief. Indeed, $Z^*(\theta N_0, \boldsymbol{\rho}_0)$ is a decreasing function of θ, and its shape is dominated by an exponential, with rate parameter depending both on N_0, and on $\boldsymbol{\rho}_0$. The larger N_0, the more peaked the density is around $\theta = 0$ (reducing to the improper constant on $\mathbb{R}^+$ when $N_0 = 0$). $\boldsymbol{\rho}_0$ also affects the tightness of the prior (the larger the number of elicited ranks, the more peaked is the density around $\theta = 0$), but its influence is smoother.

In the next section we outline the algorithms developed for inference on the MMS in both cases of known and unknown θ, within the situations (a), (b) and (c) described above.

4 Posterior simulation

Notice that, when $\theta = \theta^*$ is known, the posterior (7) is known up to a normalization constant. Posterior simulation is straightforward in this case and it basically reduces to a visualization problem because of the complexity of the space of permutations. In this simple case, we employ a Metropolis-Hastings (M-H) Markov Chain Monte Carlo (MCMC) scheme for the update of $\boldsymbol{\rho}$. We propose $\boldsymbol{\rho}'$ according to the Leap and Shift distribution of Vitelli et al. (2018), which is an asymmetric proposal centered around the current value of $\boldsymbol{\rho}$. We then accept $\boldsymbol{\rho}'$ with probability $\epsilon = \min\{1, a_{\boldsymbol{\rho}}\}$, where

$$\log a_{\boldsymbol{\rho}} = 2\theta^*(\boldsymbol{\rho}'-\boldsymbol{\rho})\cdot\tilde{\boldsymbol{R}} + \log p_{LS}(\boldsymbol{\rho}'|\boldsymbol{\rho}) - \log p_{LS}(\boldsymbol{\rho}|\boldsymbol{\rho}'), \tag{14}$$

where, $\tilde{\boldsymbol{R}} = N\bar{\boldsymbol{R}} + N_0\boldsymbol{\rho}_0$, and p_{LS} denotes the transition probability of the Leap and Shift distribution. Notice that, for the sake of simplicity, we are considering the case $\eta_0 = \theta^* N_0$, but the results follow trivially for other parametrizations.

When θ is not known, we implement a Metropolis within Gibbs scheme for posterior simulation. However, further considerations must be made for the different cases outlined in Section 3.2. First, we consider case (a), where $\boldsymbol{\rho}$ is assumed *a priori* independent of θ, which amounts to eliciting η_0 of (5) independently of θ; in cases (b) and (c) the precision parameter of the EMMS takes the form $\eta_0 = \theta N_0$.

In (a) Z^* is simply constant, so it creates no additional difficulty. Posterior inference can be performed with the efficient scheme of Vitelli et al. (2018, Algorithm 1), by simply modifying the acceptance probabilities of the M-H steps to include the non-uniform prior density on $\boldsymbol{\rho}$.

In cases (b) and (c) we have the additional issue of dealing with Z^*, for which different solutions are possible. In (b), that is for small n, we can compute Z^* on a grid of η_0 values; whenever its evaluation is required within the M-H step for the update of

θ, an approximate value can be obtained via interpolation for values of $\eta_0 = \theta N_0$ not in the grid. In this case we therefore have two steps. First, we update $\boldsymbol{\rho}$ conditional on θ from the posterior full conditional (see (12)),

$$\pi^N(\boldsymbol{\rho}|\theta) \propto \exp\left[2\theta\boldsymbol{\rho}\cdot\tilde{\boldsymbol{R}}\right]. \tag{15}$$

This is done as described above, that is, we propose $\boldsymbol{\rho}'$ according to the Leap and Shift distribution and accept it with probability $\epsilon = \min\{1, a_{\boldsymbol{\rho}}\}$, where $a_{\boldsymbol{\rho}}$ is given in (14), with θ^* equal to the current value of θ. Second, we update θ conditional on $\boldsymbol{\rho}$. Note that the posterior full conditional for θ is

$$\pi^N(\theta|\boldsymbol{\rho}) \propto \pi^N(\boldsymbol{\rho},\theta) \propto \frac{\pi(\theta)}{Z^N(\theta)Z^*(\theta N_0,\boldsymbol{\rho}_0)} \exp\left[-\theta(\tilde{g} - 2\boldsymbol{\rho}\cdot\tilde{\boldsymbol{R}})\right], \tag{16}$$

where $\tilde{g} = (2N + N_0)c_n + N_0\left\|\rho_0\right\|^2$. The proposal θ' is sampled from a log-normal density centered on the current value of θ with a variance tuned in order to obtain a desired acceptance rate.

In (c), that is, for large values of n, only the proposed prior for θ, and therefore its posterior full conditional, changes and it is given by

$$\pi^N(\theta|\boldsymbol{\rho}) \propto \frac{1}{Z^N(\theta)} \exp\left[-\theta(\tilde{g} - 2\boldsymbol{\rho}\cdot\tilde{\boldsymbol{R}})\right]. \tag{17}$$

Posterior simulation is therefore identical to that of case (b), with the obvious difference in the acceptance probability for θ.

5 Illustrative analyses

The examples considered in this section have multiple purposes. First we illustrate the effects of our prior on the inference through very elementary datasets (Sections 5.1 and 5.2). Second, we show an example of how to elicit the hyper-parameters of interest based on covariates (Section 5.3). Finally, in 5.4, we consider an application of data related to the COVID-19 pandemic, where the inclusion of prior information is relevant.

5.1 Simulation study

In this section we illustrate the effect of the prior on the posterior via a small simulated dataset. A small n is used so that all possible permutations can be listed.

We generate a sample of $N = 30$ rankings from $\mathcal{P}_4$ from the MMS with given *true* parameters $\boldsymbol{\rho}^* = (2, 1, 4, 3)$ and $\theta^* = 0.06$. We then set the prior consensus to $\boldsymbol{\rho}_0 = (2, 1, 3, 4)$, and perform inference on the model in different settings corresponding to increasing prior sample size for the prior parametrization $\eta_0 = \theta N_0$, and the Jeffreys prior for θ. The observed sample mean vector is $\bar{\boldsymbol{R}} = (2.33, 2.17, 3, 2.5)$, which leads to $\boldsymbol{\rho}_{\text{MLE}} = \boldsymbol{Y}(\bar{\boldsymbol{R}}) = (2, 1, 4, 3)$. We report in Table 1 the estimated posterior probability (EPP) of each of the rankings in $\mathcal{P}_4$. Notice that $\boldsymbol{\rho}_{\text{MLE}}$ is the ranking with smallest value

$\boldsymbol{\rho}$	$D(\boldsymbol{\rho})$	$D^*(\boldsymbol{\rho})$	$N_0 = 0$	$N_0 = 5$	$N_0 = 10$	$N_0 = 15$	$N_0 = 16$	$N_0 = 20$
(1,2,3,4)	260	2	0.029	0.038	0.050	0.053	0.053	0.050
(1, 2, 4, 3)	230	4	0.172	0.125	0.080	0.052	0.050	0.036
(1, 3, 2, 4)	310	6	0.007	0.003	0.003	0.004	0.004	0.004
(1, 3, 4, 2)	250	10	0.049	0.010	0.005	0.004	0.004	0.003
(1, 4, 2, 3)	330	12	0.004	0.001	0.001	0.002	0.002	0.001
(1, 4, 3, 2)	300	14	0.007	0.002	0.001	0.002	0.001	0.001
(2,1,3,4)	250	0	0.048	0.129	0.257	0.417	0.436	0.546
(2,1,4,3)	**220**	**2**	**0.367**	**0.579**	**0.527**	**0.410**	**0.386**	**0.303**
(2, 3, 1, 4)	350	8	0.003	0.001	0.001	0.002	0.002	0.002
(2, 3, 4, 1)	260	14	0.029	0.005	0.003	0.002	0.002	0.002
(2, 4, 1, 3)	370	14	0.002	0.001	0.001	0.001	0.001	0.001
(2, 4, 3, 1)	310	18	0.006	0.001	0.001	0.001	0.001	0.001
(3,1,2,4)	290	2	0.009	0.010	0.015	0.017	0.023	0.022
(3, 1, 4, 2)	230	6	0.169	0.065	0.032	0.016	0.017	0.012
(3, 2, 1, 4)	340	6	0.003	0.002	0.002	0.003	0.003	0.003
(3, 2, 4, 1)	250	12	0.049	0.007	0.004	0.002	0.002	0.002
(3, 4, 1, 2)	380	18	0.002	0.001	0.001	0.001	0.001	0.001
(3, 4, 2, 1)	350	20	0.003	0.001	0.001	0.001	0.001	0.001
(4, 1, 2, 3)	300	6	0.007	0.005	0.004	0.004	0.004	0.004
(4, 1, 3, 2)	270	8	0.019	0.007	0.006	0.004	0.004	0.004
(4, 2, 1, 3)	350	10	0.003	0.002	0.001	0.001	0.002	0.001
(4, 2, 3, 1)	290	14	0.009	0.002	0.002	0.001	0.001	0.001
(4, 3, 1, 2)	370	16	0.002	0.001	0.001	0.001	0.001	0.001
(4, 3, 2, 1)	340	18	0.003	0.001	0.001	0.001	0.001	0.000

Table 1: Results of the simulation study of Section 5.1. List of the 24 4-rankings (column 1), along with the quantities $D(\boldsymbol{\rho})$ and $D^*(\boldsymbol{\rho})$ defined in Result 1 (columns 2 and 3 respectively). Columns 4 to 9 contain the estimated posterior probabilities of each ranking (rows) and each setting, for increasing values of N_0. Four rows are highlighted: in dark-gray, the prior consensus $\boldsymbol{\rho} = \boldsymbol{\rho}_0$ ($D^*(\boldsymbol{\rho}) = 0$); in light-gray, the rankings nearest $\boldsymbol{\rho}_0$ ($D^*(\boldsymbol{\rho}) = 2$). The MLE (where $D(\boldsymbol{\rho}) = 220$ is minimized) is indicated by bold characters.

of $D(\boldsymbol{\rho})$ (row highlighted in light-gray and with bold characters). Studying this table, we can verify that Result 1 holds. For instance, solving (10) with $\boldsymbol{\rho}_1 = \boldsymbol{\rho}_0$ and $\boldsymbol{\rho}_2 = \boldsymbol{\rho}_{\mathrm{MLE}}$, we obtain that $\boldsymbol{\rho}_0$ has a higher posterior probability than $\boldsymbol{\rho}_{\mathrm{MLE}}$ if and only if $N_0 > 15$, which the empirical results confirm. Also, all rankings $\boldsymbol{\rho}$ with $D^*(\boldsymbol{\rho}) \leq D^*(\boldsymbol{\rho}_{\mathrm{MLE}})$ have lower posterior probabilities than $\boldsymbol{\rho}_{\mathrm{MLE}}$. Furthermore, if $D^*(\boldsymbol{\rho}_1) = D^*(\boldsymbol{\rho}_2)$, then $\boldsymbol{\rho}_1$ has a higher posterior probability than $\boldsymbol{\rho}_2$ if $D(\boldsymbol{\rho}_1) > D(\boldsymbol{\rho}_2)$.

We can also notice the following *sensitivity* behavior of the posterior probabilities: with increasing N_0, the rankings which are closer to $\boldsymbol{\rho}_0$ (in terms of Spearman's distance, or equivalently a smaller $D^*(\boldsymbol{\rho})$) have increasing posterior probabilities, while those that are farthest from $\boldsymbol{\rho}_0$ have decreasing posterior probabilities, even when the distance to the data $D(\boldsymbol{\rho})$ is not too high. An example of this can be seen in the row corresponding to $\boldsymbol{\rho} = (3, 1, 4, 2)$, which has $D(\boldsymbol{\rho}) = 230$ and $D^*(\boldsymbol{\rho}) = 6$ and for which increasing N_0 from 0 to 20 has the effect of decreasing the posterior probability from 0.169 to 0.012. The posterior means of θ in the six settings were 0.068, 0.074, 0.065, 0.06, 0.057, 0.055, while $\theta_{\text{MLE}} = 0.08$.

$\boldsymbol{o}$	Prop.	$N_0 = 0$	$N_0 = 1$	$N_0 = 5$	$N_0 = 10$	$N_0 = 49$	$N_0 = 98$	$D^*(\boldsymbol{\rho})$
ACDEB	0.337	**0.047**	**0.055**	**0.062**	**0.076**	0.105	0.078	4
ADCEB	0.184	0.037	0.044	0.050	0.060	**0.129**	0.129	2
ACDBE	0.122	0.031	0.035	0.041	0.048	0.095	0.103	2
ADCBE	0.082	0.025	0.030	0.034	0.041	0.114	**0.176**	0
ACEDB	0.061	0.022	0.028	0.024	0.025	0.018	0.015	10
CADEB	0.051	0.025	0.028	0.030	0.026	0.023	0.019	8
ADECB	0.051	0.015	0.019	0.021	0.020	0.020	0.020	6

Table 2: Results for the `idea` dataset. List of orderings corresponding to the rankings with the highest observed frequencies in the data (columns 1 and 2 respectively), along with their EPP in different settings, corresponding to values of N_0 between 0 and N (columns 3 to 8). In column 9 we present the Spearman distance between each ranking and the prior mode. The highest EPP of each setting is highlighted in bold characters.

5.2 `idea` dataset

For illustrative purposes, in this section we use the benchmark dataset `idea` (see e.g. Fligner and Verducci, 1990; Gupta and Damien, 2002). The data, collected by the Graduate Record Examination (GRE) Board, consist of a sample of $N = 98$ rankings, each of them generated by a college student who was asked to rank $n = 5$ words according to their strength of association with the target word 'idea'. The five words are 'thought' (A), 'play' (B), 'theory' (C), 'dream' (D), and 'attention' (E). Our aim is to show the effect of our informative prior for $\boldsymbol{\rho}$ on the inference. Since n is very small in this example, we can use the exact framework for posterior simulation outlined in Section 4, and choose the Jeffreys prior for the parameter θ, thus reflecting our lack of prior knowledge. In this example, we assume there is reason to believe that $\boldsymbol{o}_0 = (\text{A}, \text{D}, \text{C}, \text{B}, \text{E})$ is the true ordering of association of the five words. We therefore choose the corresponding ranking vector $\boldsymbol{\rho}_0 = (1, 4, 3, 2, 5)$ as the prior mode. The choice of N_0, interpreted as an equivalent sample size, reflects our confidence in $\boldsymbol{\rho}_0$, so we consider different settings, corresponding to increasing values of N_0. Inference is carried out via MCMC posterior simulation, using a sample size of 5×10^4 iterations, after a burn-in of 5×10^3, and the results are shown in Table 2. The orderings corresponding to the most frequently observed rankings in the dataset and their empirical frequencies or sample proportions are shown in columns 1 and 2 respectively, along with their estimated posterior probabilities (EPP) in the different settings (columns 3 to 8). In column 9 we report the Spearman distance between each of the top observed rankings and the prior mode (that is, $D^*(\boldsymbol{\rho})$).

Recall that our prior (5) assigns equal mass to all rankings at the same Spearman distance from $\boldsymbol{\rho}_0$. This behavior has some analogies with the prior of Gupta and Damien (2002). However, while there is always a unique ranking at Spearman's distance 0 from

$\boldsymbol{\rho}_0$, each conjugacy class contains more than one ranking, all of which are assigned the same mass by the prior of Gupta and Damien (2002), henceforth GD. As we show below, this difference has a relevant effect on the posterior inferences based on our prior (5), when compared to the results by GD.

From this table we can notice the following:

- the EPP of (A, D, C, B, E), which corresponds to the prior mode $\boldsymbol{\rho}_0$ (row 4), increases consistently with N_0; when $N_0 = N$, it becomes the posterior modal ranking;
- the ordering (A, C, D, E, B), corresponding to $\boldsymbol{\rho}_{MLE}$ (row 1), remains the ranking with largest EPP provided that the equivalent sample size N_0 is not too large. In other words, if the prior does not assign too much mass to $\boldsymbol{\rho}_0 \neq \boldsymbol{\rho}_{MLE}$;
- the relative ordering of the seven rankings in terms of posterior probability depends on N_0, changing for large values which imply strong prior information.

Comparing our results with the findings of GD (Table 3), we notice that:

1. the posterior distribution of GD places most of the mass (about 0.93) on the top 6 rankings, thus penalizing all other rankings in $\mathcal{P}_5$;
2. the EPP of the prior modal ranking with ordering (A, D, C, B, E), obtained by GD does not increase with the concentration parameter (in their paper denoted by λ^*), but rather decreases (from 0.019 when $\lambda^* = 0$, to 0.0067 when $\lambda^* = 0.1$). This is not in line with the expected behavior of an informative prior.

Our posterior distributions, instead, are generally flatter and, importantly, do not show the contradictory behavior with respect to the concentration parameter exhibited by the results of GD, which is probably a consequence of the complex structure of the conjugacy classes of $\mathcal{P}_5$.

5.3 The prior elicitation problem in practice

In this section we show an example of prior elicitation based on covariates. For the illustration we use the `sushi` benchmark data of Kamishima (2003), which consists of full rankings of $n = 10$ different kinds of sushi items given by $N = 5000$ respondents according to their personal preference. This dataset, available at http://www.kamishima.net/sushi/, has been extensively analyzed (see for instance Lu and Boutilier, 2011; Vitelli et al., 2018; Xu et al., 2018), and exploited in order to show inferential results under different models. We here are not interested in doing inference on this dataset (which requires a mixture model extension, and a deeper analysis), but rather to illustrate a possibility to elicit the hyper-parameters of the proposed prior in a real case study. Indeed, this dataset is particularly interesting because it includes covariates of the sushi item, which we use to build an informative prior over the consensus ranking.

Sushi item	`oil`	`eat`	`price`	`sell`
shrimp	2.73	2.14	1.84	0.84
sea eel	0.93	1.99	1.99	0.88
tuna	1.77	2.35	1.87	0.88
squid	2.69	2.04	1.52	0.92
sea urchin	0.81	1.64	3.29	0.88
salmon roe	1.26	1.98	2.70	0.88
egg	2.37	1.87	1.03	0.84
fatty tuna	0.55	2.06	4.49	0.80
tuna roll	2.25	1.88	1.58	0.44
cucumber roll	3.73	1.46	1.02	0.40

Table 3: Covariate values of interest (columns) for each of the $n = 10$ sushi items (rows).

We begin from the elicitation of the consensus ranking hyper-parameter $\boldsymbol{\rho}_0$ of (5). The following covariates of the sushi items (see Table 3) are likely to have an impact on the personal preference of the respondents:

1. `oil`: the oiliness in taste (measured on a 0–4 continuous scale, where the smaller the value is, the more oily is the sushi item);
2. `eat`: How frequently the sushi item is eaten in sushi shops (measured on a 0–3 continuous scale, where high values correspond to highly frequently sold);
3. `price`: the normalized price of the item;
4. `sell`: the frequency with which the sushi item is sold (measured on a 0–1 continuous scale, where high values correspond to highly frequently eaten).

According to our judgement, we believe that i) the more oily the sushi item is, the more it is preferred; ii) the more eaten, the more it is preferred; iii) price is positively correlated with preference; iv) the more a sushi is sold, the more it is preferred. Clearly, the above assumptions are subjective, and someone else may decide to include these covariates differently (for instance, another judge could let the price play the opposite role). Table 4 shows the rank vectors obtained from the above criteria by applying the rank transformation $\boldsymbol{Y}$ introduced in Section 2 to the covariate vectors of Table 3. Notice that, while the transformation gives rise to proper rankings when applied to the `oil`, `eat` and `price` variables, it does not result in a proper ranking when applied to the `sell` variable (column 5): sea eel, tuna, sea urchin and salmon roe have the same covariate value (0.88 in Table 3), which results in a tied rank (3.5 in Table 4). Similarly, shrimp and egg have the same value (0.84) resulting in the tied rank (6.5). Nonetheless, the transformed vector for the covariate `sell` is an element of the permutation polytope $\mathbb{PP}_{10}$, and is therefore a valid choice for the hyper-parameter $\boldsymbol{\rho}_0$.

An interesting feature of Table 4 is that the rankings induced by the different covariates are not equal but partially agree. The researcher would therefore be interested in combining the prior information coming from these different sources. The simplest possi-

Sushi item	`oil`	`eat`	`price`	`sell`
shrimp	9	2	6	6.5
sea eel	3	5	4	3.5
tuna	5	1	5	3.5
squid	8	4	8	1
sea urchin	2	9	2	3.5
salmon roe	4	6	3	3.5
egg	7	8	9	6.5
fatty tuna	1	3	1	8
tuna roll	6	7	7	9
cucumber roll	10	10	10	10

Table 4: Rank vectors for the sushi items, obtained from the covariates via the rank transformation.

bility in this regard, is to set the prior consensus hyper-parameter equal to the average of the rankings induced by the four covariates, that is, $\boldsymbol{\rho}_{01} = (5.875, 3.875, 3.625, 5.25, 4.125, 4.125, 7.625, 3.25, 7.25, 10) \in \mathbb{PP}_{10}$, or to its rank vector, $\boldsymbol{\rho}_{02} = \boldsymbol{Y}(\boldsymbol{\rho}_{01}) = (7, 3, 2, 6, 4.5, 4.5, 9, 1, 8, 10) \in \mathbb{PP}_{10}$. Alternatively, the four rankings could be given unequal weights, which would amount to calculating a weighted average, in the same spirit of Genest et al. (1986). The elicitation of the precision parameter, η_0, requires a more qualitative reasoning. Considering the parametrization $\eta_0 = \theta_0 N_0$, we may decide to fix $N_0 = 4$, since the consensus hyper-parameter comes from the average of four rankings, which may be interpreted as the opinions of four experts. At the same time, we may choose a relatively large value of θ_0, for instance $\theta_0 = 0.1$ (which is considered large, given the scale of the problem), thus reflecting confidence in $\boldsymbol{\rho}_0$, given the partial agreement of the four rankings used to construct the consensus hyper-parameter.

Another option for the elicitation of η_0, is to reason by calibration, as explained in Section 3. After having elicited the prior modal vector $\boldsymbol{\rho}_{02} = \boldsymbol{Y}(\boldsymbol{\rho}_{01}) = (7, 3, 2, 6, 4.5, 4.5, 9, 1, 8, 10) \in \mathbb{PP}_{10}$, the analyst may elicit the a priori expected per-item error size, e_0, and then solve for η_0. To do so, we first compute the expected error size range with $n = 10$ and $k = 8$, which results in $f_{\max} - f_{\min} = 16.4$. Then, we elicit e_0 as a fraction of the range, say $e_0 = 0.5(f_{\max} - f_{\min}) = 8.2$. Finally, we find the value η_0 such that $f(\eta_0, \boldsymbol{\rho}_0) = e_0$. For instance, if $e_0 = 8.2$ the corresponding value for the prior precision is $\eta_0 = 0.03$, while if $e_0 = 0.1(f_{\max} - f_{\min}) = 1.64$, then $\eta_0 = 0.2$, reflecting the larger a priori confidence of the expert in the elicited modal ranking.

Another interesting possibility, mentioned in Section 3.1, is to treat the four rankings as data and perform a prior analysis on these, through a simple Mallows model. The posterior estimates of the parameters resulting from this prior analysis could then be used as the elicited hyper-parameters for the prior. We proceeded as follows. First, we converted the four rankings to pairwise preferences: each item was preferred to all items with strictly higher rank. This preprocessing was done in order to take the `sell` covariate into account, since the Mallows model does not admit rankings with ties (like the `sell` covariate of Table 4) as input. After this transformation, however, the Mallows model for pairwise preferences (Vitelli et al., 2018) can be used. We fit a Mallows model with

Spearman's distance on such data, and a point estimate is obtained from the posterior distribution of the consensus ranking, which is then used as the hyper-parameter in the prior. We choose, as the point estimate, the cumulative probability (CP) consensus ranking, obtained by first assigning rank 1 to the item which has the maximum a posteriori marginal probability of having rank 1, then assigning rank 2 to the item, among the remaining ones, which has the maximum a posteriori marginal posterior probability of having ranks 1 or 2, and so on. As noted in Vitelli et al. (2018), the CP consensus ranking is a robust estimator which can be seen as a sequential MAP estimator. We obtain $\hat{\boldsymbol{\rho}}_{CP} = (7, 3, 2, 6, 4, 5, 9, 1, 8, 10)$[3] and the posterior mean of the scale parameter is $\hat{\theta} = 0.44$. The results of this prior analysis can now be used as hyper-parameters of the prior (5), by setting $\eta_0 = \hat{\theta}$, and $\boldsymbol{\rho}_0 = \hat{\boldsymbol{\rho}}_{CP}$.

5.4 A real-world example: COVID-19 and Italian support policies

In this section we use some data collected by the Bank of Italy in order to show the use of our proposed prior.

The data are part of a special survey (Iseco[4]) carried out between March and May 2020, a period marked by the spread of the COVID-19 pandemic and by the containment measures taken by the Italian Government. The survey contained questions intended to assess how the pandemic was affecting firms' business and how firms were responding to it. In total 3503 firms were interviewed.

We here focus on one particular question (answered by $N = 3462$ firms), which asked the firms which support policies were judged most appropriate to contain the impact of the spread of the Coronavirus on the economy. Each firm was asked to select up to two policies (among $n = 8$ possible options/choices labelled `a1` to `a8`) in order of importance. As such, the answers to the question are an example of top-2 rankings.

Interestingly, the survey was conducted in a 10-week period of time, and in each time point a different sample of firms was interviewed. We then divide the original sample into 10 sub-samples corresponding to the week in which the survey was answered by the firm, and run the model separately in each time period. We denote by $\boldsymbol{R}^{(t)} = \{\boldsymbol{R}_j^{(t)}\}_{j=1}^{N^{(t)}}$ the sample of top-2 rankings provided by the $N^{(t)}$ firms at time point t, and assume that, for each week t, there is a consensus ranking $\boldsymbol{\rho}^{(t)}$ of the eight answers, which reflects the consensus of the $N^{(t)}$ firms at time t. We model this with a Mallows model, thus assuming that, for each t, $\boldsymbol{R}_1^{(t)}, \ldots, \boldsymbol{R}_{N^{(t)}}^{(t)} | \boldsymbol{\rho}^{(t)}, \theta^{(t)} \overset{iid}{\sim} \mathcal{M}(\boldsymbol{\rho}^{(t)}, \theta^{(t)})$. We aim at making inference on $\boldsymbol{\rho}^{(1)}, \ldots, \boldsymbol{\rho}^{(10)}$.

The data, in each time period, are analyzed with an adapted version of the Bayesian Mallows model for partial data (see Vitelli et al., 2018), which can handle top-k rankings

[3] Note that the Maximum a Posteriori (MAP) ranking, in this case would be $\hat{\boldsymbol{\rho}}_{MAP} = (7, 3, 2, 6, 5, 4, 9, 1, 8, 10)$ with an associated EPP of 0.00078. The EPP of $\hat{\boldsymbol{\rho}}_{CP}$ is instead 0.00061. These very low figures imply that the posterior distribution is quite flat, which is expected, since the sample size is only $N = 4$, while the cardinality of the permutation space is 10!.

[4] The data can be processed with Bank of Italy's BIRD system at the following link: https://www.bancaditalia.it/statistiche/basi-dati/rdc/bird/index.html.

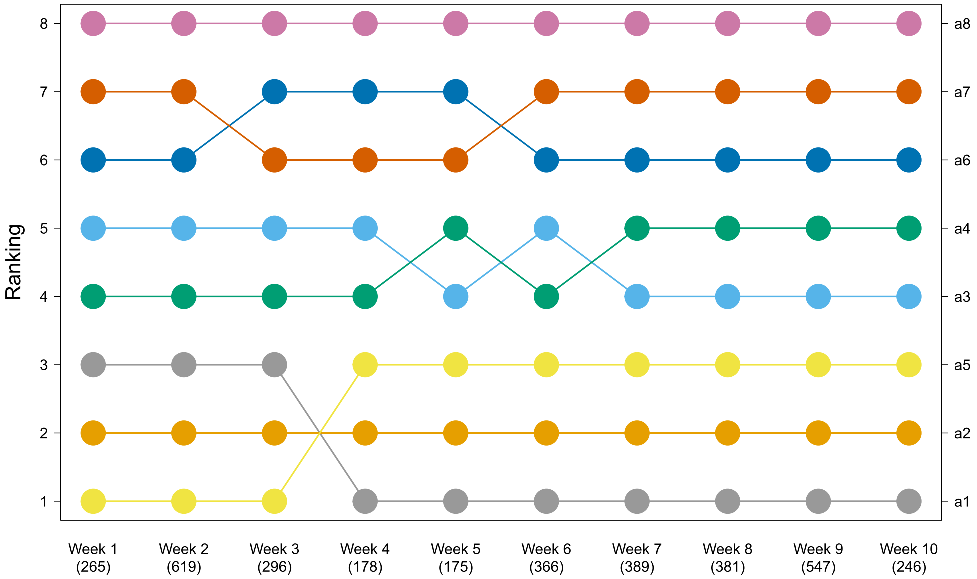

Figure 1: Results of scenario a. The development of the ranks of the considered 8 options (`a1` to `a8`, see right-vertical axis). Ranking obtained on the posterior cumulative probability rankings, computed from the marginals of the posterior $\pi^{N^{(t)}}(\boldsymbol{\rho}^{(t)}|\boldsymbol{R}^{(t)}), t = 1, \dots, 10$. Best option has rank 1.

with the help of data augmentation techniques, in two different settings:

a. with the uniform prior over $\boldsymbol{\rho}^{(t)}, t = 1, \dots, 10$. This amounts to setting the central parameter of (5) equal to the barycenter of the permutohedron, $\boldsymbol{\rho}_0 = \frac{(n+1)}{2}\mathbf{1}_n$, in each time priod t;

b. using a summary of the posterior density of $\boldsymbol{\rho}^{(t)}$ as hyper-parameter for prior (5) in the current week's inference (at time $t + 1$).

Since $n = 8$, we use the exact framework for posterior simulation outlined in Section 4, and choose, for each t, the exponential prior for $\theta^{(t)}$, $\pi(\theta^{(t)}|\lambda) = \lambda \exp(-\lambda\theta^{(t)})$, with $\lambda = 0.1$, after some tuning. In this example we also assume prior independence $\pi(\boldsymbol{\rho}^{(t)}, \theta^{(t)}|\lambda, \eta_0^{(t)}, \boldsymbol{\rho}_0^{(t)}) = \pi(\theta^{(t)}|\lambda)\pi(\boldsymbol{\rho}^{(t)}|\eta_0^{(t)}, \boldsymbol{\rho}_0^{(t)})$, and set the hyper-parameter $\eta_0^{(t+1)}$, $t \geq 1$, equal to the posterior mean of the $\theta^{(t)}$ parameter obtained in the previous period's inference. For each period, we run 100 different chains with 10^4 iterations and discarded the first 10^3 as burn-in.

In Figure 1, obtained under setting a., the CP consensus ranking of the eight options is reported (on the right-vertical axis) for each time period (on the x-axis, in brackets, the number of firms that were interviewed during the corresponding week is also shown).

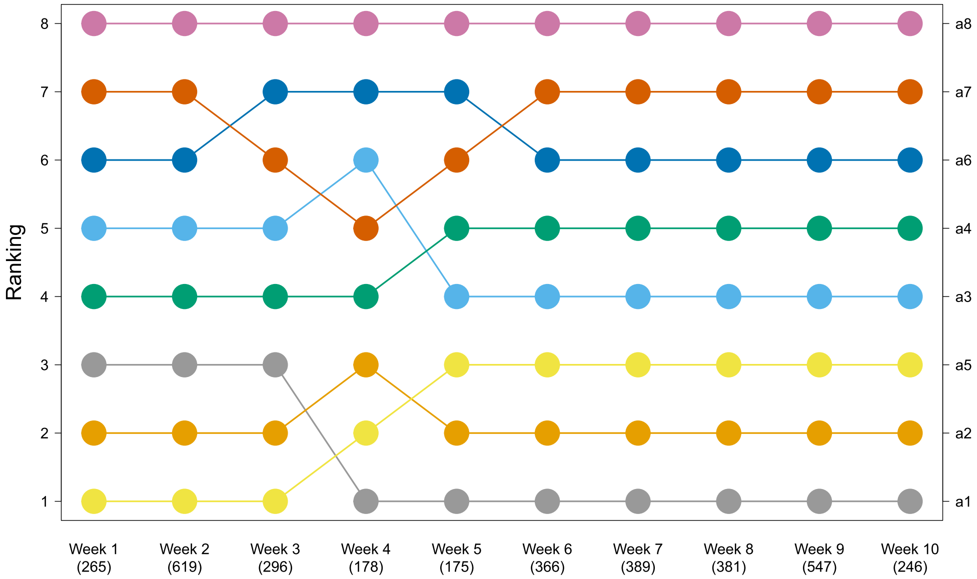

Figure 2: Results of scenario b. The development of the ranks of the considered 8 options (a1 to a8, see right-vertical axis). Ranking obtained on the posterior cumulative probability rankings, computed from the marginals of the posterior $\pi^{N^{(t)}}(\boldsymbol{\rho}^{(t)}|\boldsymbol{R}^{(t)}), t = 1, \ldots, 10$. Best option has rank 1.

We see that answer a1 (gray line) acquires popularity over time, passing from being ranked as the third option (first three weeks), to being ranked first in the remaining weeks. Answer a5 (yellow line) instead goes from being the first choice (first three weeks) to be classified as third in the following weeks.

We then repeat the analysis using setting b. and inspect the differences in the inference. Under scenario b., where the inference in time $t + 1$ is enriched with some prior information gained in period t, one might expect the resulting estimates to be more stable than those in scenario a. Indeed, in Figure 2 we see that answer a5 (yellow line) loses popularity over time as in scenario a., but it passes from being ranked first to being ranked third more smoothly, by being ranked second in week 4. Note that, indirectly, this may also affect other answers' ranks because the rankings are mutually exclusive (see how a2, orange line, makes room for a5 in week 4, to allow the smoother adjustment). The increased stability in the time development of some ranks is also apparent when looking at answers a4 and a3 (green and light blue lines respectively).

6 Conclusion

In this paper we have proposed an informative prior distribution for the consensus ranking of the Mallows model with Spearman's distance. The peculiarity of the proposed prior is that it is a location-scale family for which the location parameter does not need to be a ranking. This is convenient for the elicitation problem, since the prior can naturally handle the case when it is difficult to indicate a full ranking which is *a priori* the most likely. For instance, when the total number of items in the application considered is very large, it may be unlikely that an expert is able to elicit a prior ranking over all the items. On the contrary, it may be possible to put some prior information only over the top-ranked items. This is often the case in genomics applications, where thousands of genes are considered in the statistical analysis, but only few of them are known to be related to some disease. Another case which is naturally handled by our prior, is when multiple competing rankings are available prior to the analysis, and we are interested in including all of them into the analysis.

A limitation, discussed in Section 4, arises from the intractability of the normalizing constant Z^* of (5) when the location parameter is not itself a ranking. Possible directions for future work include exploring tractable approximations for this quantity, perhaps in the spirit of Mukherjee (2016). In general, more efficient methods for posterior simulations might be developed, but these developments fall outside of the scope of the present work. We do hope, however, that some of the ideas presented here can shed light on potentialities and limitations of the Mallows model with Spearman's distance, and encourage further developments in constructing more flexible priors.

All the simulation algorithms are implemented in R with the cpp package, and will soon be integrated into the **BayesMallows** R package (Sørensen et al., 2020).

Supplementary Material

Additional proofs (DOI: 10.1214/22-BA1307SUPP; .pdf). Proof of Result 1 and of Proposition 1.

References

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., Rousseau, J., et al. (2012). "Combining expert opinions in prior elicitation." *Bayesian Analysis*, 7(3): 503–532. MR2981623. doi: https://doi.org/10.1214/12-BA717. 400

Asfaw, D., Vitelli, V., Sørensen, Ø., Arjas, E., and Frigessi, A. (2017). "Time-varying rankings with the Bayesian Mallows model." *Stat*, 6(1): 14–30. MR3599889. doi: https://doi.org/10.1002/sta4.132. 392

Burgman, M. A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L., and Twardy, C. (2011). "Expert Status and Performance." *PLOS ONE*, 6(7): 1–7. 399

Crispino, M. (2017). "Bayesian Learning of Ranking data." Unpublished doctoral dissertation, Bocconi University, Milan, Italy. URL https://drive.google.com/file/d/1Sb2UlJBIcjAlEYcdQ5WiShNKVn5QYgXt/view 392

Crispino, M., and Antoniano-Villalobos, I. (2022). "Additional proofs." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1307SUPP. 400

Dabic, M. and Hatzinger, R. (2009). "Zielgruppenadäquate Abläufe in Konfigurationssystemen – eine empirische Studie im Automobilmarkt: Das Paarvergleichs-Pattern-Modell für Partial Rankings." In Hatzinger, R., Dittrich, R., and Salzberger, T. (eds.), *Praeferenzanalyse mit R: Anwendungen aus Marketing, Behavioural Finance und Human Resource Management*. 391

Dawid, A. (1997). "Comments on 'non-informative priors do not exist'." *Journal of Statistical Planning and Inference*, 65(1): 178–180. 392

Diaconis, P. (1988). *Group representations in probability and statistics*, volume 11 of *Lecture Notes – Monograph Series*. Hayward, CA, USA: Institute of Mathematical Statistics. MR0964069. 391, 394

Diaconis, P. and Ylvisaker, D. (1979). "Conjugate priors for exponential families." *The Annals of statistics*, 269–281. MR0520238. 397

Fligner, M. A. and Verducci, J. S. (1986). "Distance based Ranking Models." *Journal of the Royal Statistical Society B*, 48(3): 359–369. MR0876847. 394

Fligner, M. A. and Verducci, J. S. (1990). "Posterior probabilities for a consensus ordering." *Psychometrika*, 55(1): 53–63. MR1060264. doi: https://doi.org/10.1007/BF02294743. 404

French, S. (2011). "Aggregating expert judgement." *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 105(1): 181–206. MR2783806. doi: https://doi.org/10.1007/s13398-011-0018-6. 400

Genest, C., Zidek, J. V., et al. (1986). "Combining probability distributions: A critique and an annotated bibliography." *Statistical Science*, 1(1): 114–135. MR0833278. 399, 407

Gormley, I. C. and Murphy, T. B. (2008). "A mixture of experts model for rank data with applications in election studies." *The Annals of Applied Statistics*, 2(4): 1452–1477. MR2655667. doi: https://doi.org/10.1214/08-AOAS178. 391

Gupta, J. and Damien, P. (2002). "Conjugacy class prior distributions on metric-based ranking models." *Journal of the Royal Statistical Society B*, 64(3): 433–445. MR1924299. doi: https://doi.org/10.1111/1467-9868.00343. 392, 397, 404, 405

Irurozki, E., Calvo, B., and Lozano, A. (2016). "PerMallows: An R Package for Mallows and Generalized Mallows models." *Journal of Statistical Software*, 71. 394

Kamishima, T. (2003). "Nantonac Collaborative Filtering: Recommendation Based on Order Responses." In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 583–588. New York, NY, USA: ACM. 405

Lu, T. and Boutilier, C. (2011). "Learning Mallows Models with Pairwise Preferences." In Getoor, L. and Scheffer, T. (eds.), *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, 145–152. New York, NY, USA: ACM. 405

Mallows, C. L. (1957). "Non-Null Ranking Models. I." *Biometrika*, 44(1/2): 114–130. MR0087267. doi: https://doi.org/10.1093/biomet/44.1-2.114. 391, 392, 394

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*, volume 64 of *Monographs on Statistics and Applied Probability*. Cambridge, MA, USA: Chapman & Hall. MR1346107. 392, 393, 395, 396, 400

McCullagh, P. (1993). "Models on Spheres and Models for Permutations." In Fligner, M. A. and Verducci, J. S. (eds.), *Probability Models and Statistical Analyses for Ranking Data*, 278–283. New York, NY: Springer New York. MR1237210. doi: https://doi.org/10.1007/978-1-4612-2738-0_14. 392, 394, 396

Meilă, M. and Bao, L. (2010). "An Exponential Model for Infinite Rankings." *Journal of Machine Learning Research*, 11: 3481–3518. MR2756191. 392

Meilă, M. and Chen, H. (2010). "Dirichlet Process Mixtures of Generalized Mallows Models." In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, 358–367. Corvallis, OR, USA: AUAI Press. 392

Mollica, C. and Tardella, L. (2014). "Epitope profiling via mixture modeling for ranked data." *Statistics in Medicine*, 33(21): 3738–3758. MR3260657. doi: https://doi.org/10.1002/sim.6224. 391

Mukherjee, S. (2016). "Estimation in exponential families on permutations." *The Annals of Statistics*, 44(2): 853–875. MR3476619. doi: https://doi.org/10.1214/15-AOS1389. 394, 411

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Expert Probabilities*. Wiley. 399

Paganin, S., Herring, A. H., Olshan, A. F., and Dunson, D. B. (2021). "Centered Partition Processes: Informative Priors for Clustering (with Discussion)." *Bayesian Analysis*, 16(1): 301 – 370. MR4255332. doi: https://doi.org/10.1214/20-ba1197. 399

Sørensen, Ø., Crispino, M., Liu, Q., and Vitelli, V. (2020). "**BayesMallows**: Bayesian Preference Learning with the Mallows Rank Model." *R journal*. 411

Sun, M., Lebanon, G., and Kidwell, P. (2012). "Estimating probabilities in recommendation systems." *Journal of the Royal Statistical Society C*, 61(3): 471–492. MR2914522. doi: https://doi.org/10.1111/j.1467-9876.2011.01027.x. 391

Thompson, G. (1993). "Generalized permutation polytopes and exploratory graphical methods for ranked data." *The Annals of Statistics*, 1401–1430. MR1241272. doi: https://doi.org/10.1214/aos/1176349265. 392, 395

Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A., and Arjas, E. (2018). "Probabilistic preference learning with the Mallows rank model." *Journal of Machine Learning Research*, 18(158): 1–49. MR3813807. 391, 392, 393, 394, 395, 396, 399, 400, 401, 405, 407, 408

Xu, H., Alvo, M., and Yu, P. L. (2018). "Angle-based models for ranking data." *Computational Statistics & Data Analysis*, 121: 113–136. MR3759202. doi: https://doi.org/10.1016/j.csda.2017.12.004. 392, 396, 405

Acknowledgments

The authors would like to thank Sonia Petrone, Elja Arjas and Arnoldo Frigessi for their insightful comments. We are also grateful to an anonymous associate editor and two anonymous referees for their helpful comments that greatly improved a previous version of the paper.

The views expressed in this paper are those of the authors and do not necessarily reflect those of the Bank of Italy.