Informative extended Mallows priors in the Bayesian Mallows model

Marta Crispino\textsuperscript{1} and Isadora Antoniano-Villalobos\textsuperscript{2}

\textsuperscript{1}Mistis team, Inria Grenoble Rhône-Alpes, Inovallée 655 avenue de l’Europe, Montbonnot France.
\textsuperscript{2}Department of Environmental Sciences, Informatics and Statistics, Ca’ Foscari University of Venice, Venice, Italy and Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy.

Abstract

The aim of this work is to study the problem of prior elicitation for the Mallows model with Spearman’s distance, a popular distance-based model for rankings or permutation data. Previous Bayesian inference for such model has been limited to the use of the uniform prior over the space of permutations. We present a novel strategy to elicit subjective 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

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, specially 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). In a coherent analysis of ranking data, the quantification of uncertainty regarding the estimated quantities is a fundamental aspect of decision making. It could, for instance, allow for actions grounded on unreliable estimates to be deferred until more data are available.

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.

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 rescaled to lie between $-1$ and $1$, arises naturally as the correlation between the ranks of two samples. Fligner and Verducci (1990) and Vitelli et al. (2018) have studied Bayesian inference for the MMS, limiting the analysis to the use of a uniform prior on the consensus ranking. As we discuss in Section 3.1, this can be interpreted as a non-informative prior.

Within the Bayesian literature, non-informative and objective priors have attracted much attention in the search for standard go-to procedures when prior information is unavailable. They can also 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 (Mallows, 1957; 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 \textit{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. Meilä and Bao
(2010) and Meilä 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 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 find 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 proper hyper-parameters. In doing this, 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). 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.

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 sketch 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 benchmark 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, \( \{A_1, ..., A_n\} \), of labeled items to the space \( \mathcal{P}_n \) of \( n \)-dimensional permutations. A ranking can, therefore, be represented by a vector \( r = (r_1, ..., r_n) \), 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 \( A_1, ..., A_5 \) is \( r = (1, 3, 4, 5, 2) \), corresponding to the ordering \( o = (A_1, A_5, A_2, A_3, A_4) \). 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, \( \{1, ..., n\} \). 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 \( R \) takes a value \( r \in \mathcal{P}_n \) as

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

where \( \rho \in \mathcal{P}_n \) 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 (see also Definition 3 in the Appendix), 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 \( Z_d(\theta) = \sum_{r \in \mathcal{P}_n} e^{-\theta d(r, \rho_I)} \) of the MM does not depend on \( \rho \) (Mukherjee, 2016; Vitelli et al., 2018). In the above expression \( \rho_I = (1, 2, 3, ..., n) \) 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 \( n \). Therefore, the MM is considered known up to a proportionality constant only, except for some particular choices of the distance, for which \( Z_d \) may have a closed form (Fligner and Verducci, 1986). Different approximation strategies have been proposed (see e.g. McCullagh, 1993; Mukherjee, 2016; Vitelli et al., 2018), al-
lowing 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 \( \rho \). 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 \( d \) 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

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

for \( r, \rho \in \mathcal{P}_n \), which was first introduced with the name rho-model 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 \( R \) follows an MMS distribution, denoted by \( R|\rho, \theta \sim \mathcal{M}(\rho, \theta) \), if its probability mass function is given by

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

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

Given a sample \( R_1, ..., R_N \mid \rho, \theta \sim \mathcal{M}(\rho, \theta) \), the likelihood function takes the form

\[
p(R_1, ..., R_N \mid \rho, \theta) = \frac{1}{Z(\theta)^N} \exp \left[ -\theta \sum_{j=1}^N \| \rho - R_j \|^2 \right].
\]

Therefore, for \( \theta > 0 \), the maximum likelihood estimator (MLE) \( \rho_{MLE} \) is given by

\[
\rho_{MLE} = \arg\min_{\rho \in \mathcal{P}_n} \sum_{j=1}^N \| \rho - R_j \|^2 = \arg\max_{\rho \in \mathcal{P}_n} \rho \cdot \bar{R},
\]

where the dot denotes the scalar product on \( \mathbb{R}^n \), and \( \bar{R} = (\bar{R}_1, ..., \bar{R}_n) \) is the sample mean vector of \( \bar{R}_i = \frac{1}{N} \sum_{j=1}^N R_{ij}, i = 1, \ldots, n \). This is not surprising as the kernel of the MMS distribution coincides with that of an \( n \)-dimensional Gaussian distribution, except it has a finite support. In other words, the MMS is the restriction of the \( n \)-dimensional gaussian to \( \mathcal{P}_n \). Clearly, if \( \bar{R} \notin \mathcal{P}_n \), then the MLE simply coincides with the sample mean. In general, however, \( \bar{R} \notin \mathcal{P}_n \) so a further consideration is required in order to solve the optimization problem.

**Definition 1.** The permutohedron of order \( n \), \( \mathbb{P}_n \), is an \( (n - 1) \)-dimensional polytope embedded in an \( n \)-dimensional space, the vertices of which are formed by
permuting the coordinates of the vector \((1, 2, 3, \ldots, n)\). Equivalently, it is the convex hull of the points \(\rho \in \mathcal{P}_n \subset \mathbb{R}^n\).

The set \(\mathcal{PP}_n\) 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.

**Definition 2.** Let \(\mathbf{r} \in \mathcal{PP}_n\), such that \(r_i \neq r_h\) for all \(i \neq h\). The rank \(\mathbf{Y}(\mathbf{r}) \in \mathcal{P}_n\) defined by \(Y_i = Y_i(\mathbf{r}) = \sum_{h=1}^n \mathbb{1}(r_h \leq r_i)\), \(i = 1, \ldots, n\) is called the rank vector of \(\mathbf{r}\), where \(\mathbb{1}(E)\) denotes the indicator function taking the value 1 if the event \(E\) is true, and 0 otherwise.

By the definition of convex hull, \(\bar{\mathbf{R}} \in \mathcal{PP}_n\) for any set of rankings, \(\mathbf{R}_1, \ldots, \mathbf{R}_N \in \mathcal{P}_n\). The following proposition shows that \(\rho_{\text{MLE}} = \mathbf{Y}(\bar{\mathbf{R}})\) whenever \(\bar{R}_i \neq \bar{R}_h\) for all \(i \neq h\).

**Proposition 1.** Let \(\mathbf{R}_1, \ldots, \mathbf{R}_N|\rho, \theta \sim \mathcal{M}(\rho, \theta)\), and assume that \(\bar{\mathbf{R}} \in \mathcal{PP}_n\) is such that \(\bar{R}_i \neq \bar{R}_h\), for each \(i \neq h\). Then \(\rho_{\text{MLE}} = \mathbf{Y}(\bar{\mathbf{R}})\).

In order to clarify ideas, consider three samples of size \(N = 100\) of 5-dimensional rankings, with sample means \(\bar{\mathbf{R}}_1 = (1, 3, 4, 5, 2), \bar{\mathbf{R}}_2 = (1.1, 3, 4, 4.9, 2)\) and \(\bar{\mathbf{R}}_3 = (1.5, 2.9, 3.8, 4.7, 2.1)\), respectively. All three sample mean vectors have the same rank vector and, consequently, lead to the same MLE of the consensus ranking, \(\rho_{\text{MLE}} = \mathbf{Y}(\bar{\mathbf{R}}_1) = \mathbf{Y}(\bar{\mathbf{R}}_2) = \mathbf{Y}(\bar{\mathbf{R}}_3) = \bar{\mathbf{R}}_1 \in \mathcal{P}_5\). Notice, however, that while the rank vector transformation is formally correct, ensuring that the MLE is a proper ranking, it entails a loss of information. Intuitively, looking at the three sample means, one would attach greater uncertainty to the MLE obtained from the third sample, even if this information is lost when looking at the corresponding rank vector, as it is known that a point estimate alone does not provide an uncertainty assessment. Definition 2 can be generalized to the case of a vector \(\mathbf{r}\) with ties, i.e., for any \(\mathbf{r} \in \mathcal{PP}_n\), by letting \(Y(\mathbf{r})\) be any ranking whose elements satisfy the same ordering relation as those of \(\mathbf{r}\). However, the ranking vector in this case would not be unique and so, a unique \(\rho_{\text{MLE}}\) would not exist. It would be possible, for instance, to obtain a fourth sample with sample mean \(\bar{\mathbf{R}}_4 = (3, 3, 3, 3, 3)\). Then, any ranking in \(\mathcal{P}_5\) would be a rank vector for \(\bar{\mathbf{R}}_4\). This corresponds to a flat likelihood function, for which there is no MLE. Intuitively, any permutation has the same likelihood of being the consensus ranking. This idea is related to the spread or variability of the sample, which in turn is associated to the concentration of the MMS distribution around \(\rho\), in other words, to the precision parameter \(\theta\). This is, again, not surprising, considering the relation of the MMS with the Gaussian distribution, highlighted above.

It follows that, even if in most applications \(\theta\) is considered a nuisance parameter and the main interest is in the estimation of \(\rho\), it is nevertheless necessary to estimate
θ in order to get an idea of the reliability of the ρ estimate. A MLE for θ can be found as the solution to the equation

\[ h(\theta) = \frac{Z'(\theta)}{Z(\theta)} - \frac{1}{N} \sum_{j=1}^{N} \|R_j - \rho_{\text{MLE}}\|^2 = 0, \]

assuming a unique ρ_{MLE} exists. This can be done numerically, for instance, via a Newton-Raphson algorithm (see e.g. Marden, 1995), but the calculations may be cumbersome for all but small values of n.

The Bayesian paradigm is then a natural solution for making inference on the MMS, not only for quantifying uncertainty, but also for including prior information into the statistical analysis. In the remainder, we propose and study an informative prior density, 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 parameters 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, the likelihood (3) can be simplified as

\[ p(R_1, ..., R_N|\rho, \theta) \propto \exp \left[ 2\theta \sum_{j=1}^{N} \rho \cdot R_j \right] \propto \exp (2\theta N \rho \cdot \bar{R}). \]

Therefore, a conjugate prior for ρ ∈ \mathcal{P}_n is given by

\[ \pi(\rho|\eta_0, \rho_0) = \frac{1}{Z^*(\eta_0, \rho_0)} \exp \left[ -\eta_0 \|\rho_0 - \rho\|^2 \right] \propto \exp [2\eta_0 \rho \cdot \rho_0]. \]

We call this, the Extended Mallows Model with Spearman distance (EMMS) and write ρ|η_0, ρ_0 ~ \mathcal{E.M}(\rho_0, \eta_0). The two parameters η_0 ≥ 0 and ρ_0 ∈ \mathbb{R}_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 ρ_0 with η_0 = 0 corresponding to a uniform prior on \mathcal{P}_n, while larger values reflect stronger prior belief on ρ_0. Notice, however, that the modal parameter cannot be interpreted, in general, as a consensus ranking, except when ρ_0 ∈ \mathcal{P}_n, in which case the EMMS simply reduces to a MMS. 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 to take any value in the permutohedron $\mathbb{P}^n$, that is, to be any convex combination of the elements of $P_n$, 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 $\rho_0$, and greater mass is given to permutations closest to $\rho_0$. For instance, consider the EMMS centered at the barycenter of the permutohedron, that is, with $\rho_0 = \frac{(n+1)}{2}1_n$. This results in a uniform distribution on rankings for any value of the precision parameter $\eta_0$. Small deviations from uniformity can be achieved by letting $\eta_0 > 0$ and $||\rho_0 - \frac{(n+1)}{2}1_n||^2$ be small. The direction of the vector $\rho_0 - \frac{(n+1)}{2}1_n$ in $\mathbb{R}^n$ determines the rankings for which the mass increases and those for which it decreases.

The case described above, where $\rho_0 = \frac{(n+1)}{2}1_n$, is therefore equivalent to assigning to $\rho$ the uniform prior on $P_n$, $\pi(\rho) = \frac{1}{m!}$, like in (Fligner and Verducci, 1990; Vitelli et al., 2018). As Berger et al. (2012) discuss, the natural reference prior for a discrete parameter taking values on a finite support is usually the uniform prior on the parameter space. However, the authors show that the uniform prior, in some cases, may not be objective, which may be a property sought by the analyst. The following result, which holds for the MM with any right-invariant distance, shows that the uniform prior on all rankings, which corresponds to prior (5) with (i) $\eta_0 = 0$ and $\forall \rho_0$, or with (ii) $\rho_0 = \frac{(n+1)}{2}1_n$ and $\forall \eta_0$, is formally the objective prior in the sense of Villa and Walker (2015). The authors propose that an objective prior for a discrete parameter, in this case the mode of the MMS, should assign to each possible value $\rho$ a mass proportional to the minimum Kullback-Leibler divergence between the model with parameter value $\rho$ and the model with any other parameter value, say $\rho^* \neq \rho$. In this way, prior mass is associated to the “worth” of each possible parameter value, defined as the loss in information that would derive from assigning prior probability zero to such value, if it was true.

**Proposition 2.** For any right-invariant distance $d$, the objective prior in the sense of Villa and Walker (2015) for the MM is the uniform prior on the space of permutations, $\rho \sim \text{Unif}(P_n)$.

Note that $\rho_0 \in \mathbb{P}^n$ may not be a permutation, so that the partition function in (5),

$$Z^*(\eta_0, \rho_0) = \sum_{\rho \in P_n} \exp \left[ -\eta_0 ||\rho_0 - \rho||^2 \right],$$

in general depends on $\rho_0$. This implies that (5) is known up to a normalization constant. However, in the following sections we show that this drawback can be overcame in practice. Moreover, the fact that $\rho_0$ does not need to be a permutation
is very convenient for the elicitation problem. The cases where (i) we are interested in including partial information about the consensus ranking, or when (ii) multiple experts’ opinions are available, are naturally handled by this prior. For instance, imagine a simple example of case (i), where we have some information only on the top-$k$ ranked items. We can define $\rho_0$ by fixing the top-$k$ items’ ranks $\{1, ..., k\}$, and giving the same uniform value $(n + k + 1)/2$ to the remaining bottom-$(n - k)$ items’ ranks. An example of case (ii) is to assume that two (or more) experts believe, a priori, in different modal rankings, say $\rho_{01}$ and $\rho_{02}$. An analyst wishing to express an equally strong prior on such two rankings may simply use the prior (5) with $\rho_0 = (\rho_{01} + \rho_{02})/2 \in \mathbb{PP}_n$.

A possible reparametrization of (5) is obtained by letting $\eta_0 = \theta_0 N_0$, which allows to further understand the role of the precision parameter for prior elicitation. One may imagine eliciting prior information from an expert who believes that the consensus ranking is given by some $\rho_0 \in \mathcal{P}_n$ and expresses a degree of uncertainty in this belief through some precision, say $\theta_0$. Within scenario (ii), one may imagine that the analyst, having encountered a number $N_0$ of experts who coincide with this view, wishes to summarize this aggregated information by increasing the prior precision. This is achieved by letting $\eta_0 = \theta_0 N_0$, thus expressing that individual prior belief regarding $\rho_0$ is reinforced by various experts. In the limit, infinite prior precision $\eta_0 \to \infty$ may correspond to a single expert with extremely strong prior belief ($\theta_0 \to \infty$) or to an extremely large number of experts $N_0 \to \infty$ with some prior belief ($\theta_0 > 0$). Intuitively, one may imagine a situation in which the analyst aggregates prior opinions from many experts by calculating $\rho_0$ and $\eta_0$ as convex and linear combinations, respectively, of the individual $\rho_{0,j}, \theta_{0,j}$ parameters elicited from each expert $j$, and considering the number $N_{0,j}$ of experts who agree on both. In order to understand how this could be done, one may consider how the information passes on from the prior to the posterior.

The posterior density for $\rho$ is given by

$$
\pi^N(\rho | \theta) \propto \exp \left[ 2(\eta_0 + \theta N) \rho \cdot \left( \frac{\theta N}{\eta_0 + \theta N} \bar{R} + \frac{\eta_0}{\eta_0 + \theta N} \rho_0 \right) \right] (7)
$$

The first thing we observe is that the proposed prior is indeed conjugate. In other words, if $R_1, ..., R_N \mid \rho, \theta \overset{iid}{\sim} \mathcal{M}(\rho, \theta)$ and $\rho \mid \rho_0, \eta_0 \sim \mathcal{E}\mathcal{M}(\rho_0, \eta_0)$, then it holds that $\rho \mid \theta, \rho_0, \eta_0, R_1, ..., R_N \sim \mathcal{E}\mathcal{M}(\rho_N, \eta_N)$, with updated parameters:

$$
\rho_N = \frac{\theta N}{\eta_0 + \theta N} \bar{R} + \frac{\eta_0}{\eta_0 + \theta N} \rho_0 \in \mathbb{PP}_n, \quad (8)
$$

$$
\eta_N = \eta_0 + \theta N \geq 0. \quad (9)
$$

In particular, the prior hyper-parameters $(\rho_0, \eta_0)$ elicited under scenario (ii) can be interpreted as the posterior parameters obtained by an expert who observes $N = \eta_0/\theta$ instances of $\rho_0 \in \mathcal{P}_n$, where $\theta$ is the known precision of the true data
distribution according to the MMS. Clearly, this interpretation excludes any value of \( \eta_0 \) that is not a multiple of \( \theta \). Nevertheless, such exercise helps to provide an intuition of the role of the prior hyperparameters. Notice that, since any \( \rho_0 \in \mathbb{P}_n \) can be expressed as a convex combination of rankings in \( \mathcal{P}_n \), the prior mode elicited in scenario (ii) can always be interpreted as arising from multiple (possibly infinite) experts, the calculation of the individually elicited parameters being an exercise in linear algebra. The prior precision parameters for which this interpretation is valid, however, are limited. Therefore, in the case in which \( \theta \) is assumed known, an interesting case arises by setting \( \theta_0 = \theta \), i.e. \( \eta_0 = \theta N_0 \). \( N_0 \) can be interpreted as an \textit{a priori} sample size, representing the amount of information on which an expert bases the prior belief about the central tendency of \( \rho \). In this sense, the posterior consensus parameter can be viewed as a weighted average of the prior hyper-parameter \( \rho_0 \) and the observed mean value \( \bar{R} \), with weights proportional to the corresponding sample sizes. For any finite prior precision \( \theta N_0 < \infty \), as the sample size increases, the posterior accumulates mass around \( \rho_N \), which approaches the sample mean, \( \bar{R} \). Some insights into the role of the prior hyper-parameters can be obtained by considering limiting situations. An infinite prior precision \( N_0 = \infty \) would express \textit{a priori} certainty, by accumulating all the prior mass on \( \rho_0 \), a choice that would make sense only for \( \rho_0 \in \mathcal{P}_n \). The posterior would maintain the infinite precision \( \theta N = \infty \) thus accumulating mass on \( \rho_N = \rho_0 \). In such hypothetical case, learning would be possible only for infinite sample sizes, with

\[
\lim_{N \to \infty} \rho_N = \begin{cases} 
\rho_0 & \text{if } N_0/N \to \infty \\
(1-\alpha)\rho_0 + \alpha \bar{R} & \text{if } N_0/N \to (1/\alpha - 1) \in (0,1) \\
\bar{R} & \text{if } N_0/N \to 0 
\end{cases}
\]

Notice that, by Proposition 1, the maximum \textit{a posteriori} (MAP) of \( \rho \) is unique and given by \( \rho_{MAP} = Y(\rho_N) \) provided that all the coordinates of the vector \( \rho_N \) take different values. Furthermore, \( \eta_N \to \infty \) as the sample size \( N \) grows, thus increasing posterior precision.

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. At the same time, we note in the following Theorem that the results in Gupta and Damien (2002, Section 3.3) can be extended to our prior (5).

**Theorem 1.** Let \( D(\rho) = \sum_{j=1}^{N} \|R_j - \rho\|^2 \), and \( D^*(\rho) = \|\rho_0 - \rho\|^2 \). Then:

a) for each \( \rho_1, \rho_2 \in \mathcal{P}_n \), and given \( \theta, \eta_0 \), the ranking \( \rho_1 \) will have higher posterior probability than \( \rho_2 \) if and only if

\[
D(\rho_1) - D(\rho_2) < \gamma[D^*(\rho_2) - D^*(\rho_1)],
\]

(10)

where \( \gamma = \eta_0/\theta \).

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

c) for each \( \rho_1, \rho_2 \in \mathcal{P}_n \), if \( D^*(\rho_1) = D^*(\rho_2) \), \( \rho_1 \) will have higher posterior probability than \( \rho_2 \) if and only if \( D(\rho_1) < D(\rho_2) \).

d) for each \( \rho_1, \rho_2 \in \mathcal{P}_n \), if \( D(\rho_1) < D(\rho_2) \) and \( D^*(\rho_1) < D^*(\rho_2) \), then \( \rho_1 \) will have higher posterior probability than \( \rho_2 \).

The theorem, 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 \( \theta \) and \( \eta_0 \), that determines which rankings receive the highest posterior probabilities. In Section 5 we illustrate, through simulated data, some of the consequences of this theorem on the inference.

### 3.2 Unknown precision parameter

When \( \theta \) is unknown, the Bayesian paradigm requires a prior on the pair of parameters \((\rho, \theta)\). We here suggest to choose a joint prior of the form \( \pi(\rho, \theta) = \pi(\theta)\pi(\rho|\theta) \), where \( \pi(\rho|\theta) \) is the EMMS of eq. (5). Notice that the particular case of prior independence, \( \pi(\rho, \theta) = \pi(\theta)\pi(\rho) \), is achieved in practice by choosing the parameter \( \eta_0 \) independent of \( \theta \). 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 Fligner and Verducci (1990).

Alternatively, we suggest the use of the Jeffreys prior for \( \theta \), which, in some specific cases, has a closed form and may be an interesting alternative when no information on \( \theta \) is available a priori. The following proposition holds for any MM with a right-invariant distance, and in particular for the MMS.

**Proposition 3.** The Jeffreys prior for \( \theta \) in a MM with right-invariant distance \( d \) takes the form

\[
\pi_J(\theta) = \sqrt{\nabla_{\mathbf{R}|\theta} [d(\mathbf{R}, \rho_1)|\theta]},
\]

(11)

where \( \nabla_{\mathbf{R}|\theta} \) denotes the variance with respect to \( \mathbf{R} \sim \mathcal{M}(\rho_1, \theta) \), which depends on \( \theta \).
The posterior density of the model parameters, with the conjugate prior \( \pi(\rho|\theta) \) given in eq. (5) and possibly independent of \( \theta \), and with one of the three prior distributions for \( \theta \) mentioned above is

\[
\pi^N(\rho, \theta) \propto \frac{\pi(\theta)}{Z^N(\theta)Z^*(\eta_0, \rho_0)} \exp \left\{ -\theta N \left[ \| \rho - \bar{R} \|^2 + c_n - \| \bar{R} \|^2 \right] - \eta_0 \| \rho_0 - \rho \|^2 \right\}.
\]  

(12)

Eq. (12) can be easily evaluated in two cases: when (a) \( Z^* \) does not depend on \( \theta \), that is, when \( \eta_0 \) is independent of \( \theta \), or when (b) \( \eta_0 = \theta N_0 \) and \( n \) is small enough, so that \( Z^* \) can be calculated exactly, for given prior hyperparameters \( \rho_0 \) and \( N_0 \). 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 \( \theta \), \( \pi_{\text{large} \ n}(\theta) \propto Z^*(\theta N_0, \rho_0) \), so that the posterior density (12) can be written as

\[
\pi^N(\rho, \theta) \propto \frac{1}{Z^N(\theta)} \exp \left\{ -\theta \left[ N \left( \| \rho - \bar{R} \|^2 + c_n - \| \bar{R} \|^2 \right) + N_0 \| \rho_0 - \rho \|^2 \right] \right\}. \]  

(13)

In the next section we sketch the algorithms developed for inference on the MMS in both cases of known and unknown \( \theta \), 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 (MH) Markov Chain Monte Carlo (MCMC) scheme for the update of \( \rho \). We propose \( \rho' \) according to the Leap and Shift distribution of Vitelli et al. (2018), which is an asymmetric proposal centered around the current value of \( \rho \). We then accept \( \rho' \) with probability \( \epsilon = \min\{1, a_\rho\} \), where

\[
\log a_\rho = 2\theta^*(\rho' - \rho) \cdot \bar{R} + \log p_{\text{LS}}(\rho'|\rho) - \log p_{\text{LS}}(\rho|\rho'),
\]  

(14)

where, \( \bar{R} = N\bar{R} + N_0\rho_0 \), and \( p_{\text{LS}} \) 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 \( \theta \) 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 \( \rho \) is assumed a priori independent of \( \theta \), which amounts to eliciting \( \eta_0 \) of eq. (5) independently of \( \theta \); 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. Exact posterior inference can be performed when \( n \leq 14 \), that is, when we can compute \( Z \) exactly.
When \( n > 14 \) posterior inference cannot be performed exactly, but we can exploit the efficient scheme of Vitelli et al. (2018, Algorithm 1), which targets an approximation of the posterior density. Only the acceptance probabilities of the two M-H steps are different here, due to the introduction of the non-uniform prior density on \( \rho \).

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 \( \eta_0 \) values; whenever its evaluation is required within the M-H step for the update of \( \theta \), 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 \( \rho \) conditional on \( \theta \) from the posterior full conditional (see eq. (12)),

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

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

\[
\pi^N(\theta|\rho) \propto \frac{\pi^N(\rho, \theta)}{Z^N(\theta)Z^*(\theta N_0, \rho_0)} \exp \left[-\theta(\tilde{g} - 2\rho \cdot \tilde{R})\right],
\]  (16)

where \( \tilde{g} = (2N + N_0)c_n + N_0 \|\rho_0\|^2 \). The proposal \( \theta' \) is sampled from a log-normal density centered on the current value of \( \theta \) 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 \( \theta \), and therefore its posterior full conditional, changes and it is given by

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

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

## 5 Illustrative analyses

### 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 \( P_4 \) from the MMS with given true parameters \( \rho^* = (2, 1, 4, 3) \) and \( \theta^* = 0.06 \). We then set the prior consensus to \( \rho_0 = (2, 1, 3, 4) \), 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 $\theta$. The observed sample mean vector is $\bar{R} = (2.33, 2.17, 3, 2.5)$, which leads to $\rho_{MLE} = Y(\bar{R}) = (2, 1, 4, 3)$. We report in Table 1 the estimated posterior probability (EPP) of each of the rankings in $P_4$. Notice that $\rho_{MLE}$ is the ranking with smallest value of $D(\rho)$ (row highlighted in light-gray and with bold characters). Studying this table, we can verify that Theorem 1 holds. For instance, solving eq. (10) with $\rho_1 = \rho_0$ and $\rho_2 = \rho_{MLE}$, we obtain that $\rho_0$ has a higher posterior probability than $\rho_{MLE}$ if and only if $N_0 > 15$, which the empirical results confirm. Also, all rankings $\rho$ with $D^*(\rho) \leq D^*(\rho_{MLE})$ have lower posterior probabilities than $\rho_{MLE}$. Furthermore, if $D^*(\rho_1) = D^*(\rho_2)$, then $\rho_1$ has a higher posterior probability than $\rho_2$ if $D(\rho_1) > D(\rho_2)$.

We can also notice the following sensitivity behavior of the posterior probabilities: with increasing $N_0$ the rankings which are closer to $\rho_0$ (in terms of Spearman’s distance, or equivalently a smaller $D^*(\rho)$) have increasing posterior probabilities, while those that are farthest, have decreasing posterior probabilities, even if the distance to the data $D(\rho)$ is not so high. An example of this can be seen in the row corresponding to $\rho = (3, 1, 4, 2)$, which has $D(\rho) = 230$ and $D^*(\rho) = 6$ 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 $\theta$ in the six settings were 0.068, 0.074, 0.065, 0.06, 0.057, 0.055, while $\theta_{MLE} = 0.08$.

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 $\rho$ on 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 $\theta$, thus reflecting our lack of prior knowledge. In this example, we assume there is reason to believe that $o_0=(A,D,C,B,E)$ is the true ordering of association of the five words. We therefore choose the corresponding ranking vector $\rho_0 = (1, 4, 3, 2, 5)$ as the prior mode. The choice of $N_0$, interpreted as an equivalent sample size, reflects our confidence in $\rho_0$, 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 $5 \cdot 10^4$ iterations, after a burn-in of $5 \cdot 10^3$, 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
\[ D(\rho) \times D^*(\rho) = 0 \]

\[ N_0 = 5 \]

\[ N_0 = 10 \]

\[ N_0 = 15 \]

\[ N_0 = 16 \]

\[ N_0 = 20 \]

\[ (2,1,2,3) \]

\[ 250 \]

\[ 0.048 \]

\[ 0.129 \]

\[ 0.257 \]

\[ 0.417 \]

\[ 0.436 \]

\[ 0.546 \]

\[ (2,1,4,3) \]

\[ 220 \]

\[ 0.367 \]

\[ 0.579 \]

\[ 0.527 \]

\[ 0.410 \]

\[ 0.386 \]

\[ 0.303 \]

\[ (2,3,1,4) \]

\[ 350 \]

\[ 0.003 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.002 \]

\[ 0.002 \]

\[ 0.002 \]

\[ (2,3,4,1) \]

\[ 260 \]

\[ 0.054 \]

\[ 0.010 \]

\[ 0.005 \]

\[ 0.004 \]

\[ 0.004 \]

\[ 0.003 \]

\[ (3,1,2,4) \]

\[ 290 \]

\[ 0.009 \]

\[ 0.010 \]

\[ 0.015 \]

\[ 0.017 \]

\[ 0.023 \]

\[ 0.022 \]

\[ (3,1,4,2) \]

\[ 230 \]

\[ 0.169 \]

\[ 0.065 \]

\[ 0.032 \]

\[ 0.016 \]

\[ 0.017 \]

\[ 0.012 \]

\[ (3,2,1,4) \]

\[ 340 \]

\[ 0.003 \]

\[ 0.002 \]

\[ 0.002 \]

\[ 0.003 \]

\[ 0.003 \]

\[ 0.003 \]

\[ (3,2,4,1) \]

\[ 370 \]

\[ 0.002 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ (4,1,2,3) \]

\[ 300 \]

\[ 0.007 \]

\[ 0.005 \]

\[ 0.004 \]

\[ 0.004 \]

\[ 0.004 \]

\[ 0.004 \]

\[ (4,1,3,2) \]

\[ 270 \]

\[ 0.019 \]

\[ 0.007 \]

\[ 0.006 \]

\[ 0.004 \]

\[ 0.004 \]

\[ 0.004 \]

\[ (4,2,1,3) \]

\[ 350 \]

\[ 0.003 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ (4,2,3,1) \]

\[ 290 \]

\[ 0.009 \]

\[ 0.002 \]

\[ 0.002 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ (4,3,1,2) \]

\[ 370 \]

\[ 0.002 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ (4,3,2,1) \]

\[ 340 \]

\[ 0.003 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.001 \]

\[ 0.000 \]

\[ Table 1: \text{Results of the simulation study of Section 5.1. List of the 24 4-rankings (column 1), along with the quantities} \ D(\rho) \text{ and} \ D^*(\rho) \text{ defined in Theorem 1 (columns 2 and 3 respectively). Columns 4 to 9 contain the estimated posterior probabilities of each ranking (rows) and each setting, for increasing values of} \ N_0. \text{ Four rows are highlighted: in dark-gray, the prior consensus} \ \rho = \rho_0 (\ D^*(\rho) = 0); \text{ in light-gray, the rankings nearest} \ \rho_0 (\ D^*(\rho) = 2). \text{ The MLE (where} \ D(\rho) = 220 \text{ is minimized) is indicated by bold characters.}\]

\[ \text{From this table we can notice the following:}\]

\[ \text{• the EPP of} \ (A,D,C,B,E), \text{ which corresponds to the prior mode} \ \rho_0 \text{ (row 4), increases consistently with} \ N_0. \text{ When} \ N_0 = N, \text{ it becomes the posterior modal}\]
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.

| $o$         | Prop. | $N_0 = 0$ | $N_0 = 1$ | $N_0 = 5$ | $N_0 = 10$ | $N_0 = 49$ | $N_0 = 98$ | $D^*(\rho)$ |
|-------------|-------|-----------|-----------|-----------|-----------|-----------|-----------|-------------|
| ACDEB       | 0.337 | 0.047     | 0.055     | 0.062     | 0.076     | 0.105     | 0.078     | 4           |
| ADCBE       | 0.184 | 0.037     | 0.044     | 0.050     | 0.060     | **0.129** | 0.129     | 2           |
| ACDBE       | 0.122 | 0.031     | 0.035     | 0.044     | 0.048     | 0.095     | 0.103     | 2           |
| ADBCE       | 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           |

• the ordering (A,C,D,E,B), corresponding to $\rho_{MLE}$ (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 $\rho_0 \neq \rho_{MLE}$;

• 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 $P_5$;

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 $\lambda^*$), 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 and which is probably a consequence of the complex structure of the conjugacy classes of $P_n$.

5.3 The prior elicitation problem in practice

In this section we exploit covariates to provide an intuition of how to introduce available information in the prior elicitation problem. 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. The data are available at http://www.kamishima.net/sushi/. This dataset is particularly interesting because it includes covariates of the sushi items. We can therefore use this additional information to build an informative prior over the consensus ranking.

We begin from the elicitation of the consensus ranking hyper-parameter \( \rho_0 \) of eq. (5). We believe that the following covariates of the sushi items (see Table 3) may 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 smallest 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).

| 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).

Therefore, we may include the information contained in these four covariates into the analysis of the ranking data, through the following subjective reasoning: the more oily the sushi item is, the more it is preferred; a sushi item which is frequently eaten, is more likely to be preferred than one eaten less frequently; expensive items are preferred above cheaper ones; finally, items which are sold more, are known more and hence preferred. Clearly, the above assumptions are subjective, and someone else may decide to include these covariates differently (for instance, the price may play the opposite role). Table 4 shows the rank vectors obtained from these criteria by applying the rank transformation of Definition 2 to the covariate vectors of Table 3. Notice that the transformation does not result in a proper ranking for 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). Analogously, shrimp and egg, have the the same value (0.84) resulting a the tied rank (6.5). Nonetheless, the transformed vector for the covariate eat is indeed an element of the permutation polytope $\mathbb{PP}_{10}$, and could therefore be a valid choice for the hyper-parameter $\rho_0$. Another interesting feature of Table 4 is that the rankings induced by the different covariates (columns) are not equal but partially agree. A possible choice for the prior consensus hyper-parameter, which takes into account these four different rankings is to set it equal to the average of the rankings induced by the four covariates, that is, $\rho_{01} = (5.875, 3.875, 3.625, 5.25, 4.125, 4.125, 7.625, 3.25, 7.25, 10) \in \mathbb{PP}_{10}$. Alternatively, one may consider the corresponding rank vector, $\rho_{02} = Y(\rho_{01}) = (7, 3, 2, 6, 4.5, 4.5, 9, 1, 8, 10) \in \mathcal{P}_{10}$.

| 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 of Definition 2.

The elicitation of the precision parameter, $\eta_0$, 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 $\theta_0$, for instance $\theta_0 = 0.1$ (which is considered large, given the scale of the problem), thus reflecting confidence in $\rho_0$, given the partial agreement of the four rankings used to construct the consensus hyper-parameter.

### 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 expressing equally strong prior beliefs on them.

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., 2018).

Appendix

Before stating the proof of Proposition 1, let us introduce the formal notion of right-invariance which will prove useful in the proof.

Definition 3. Right-invariant distance (Diaconis, 1988). A distance function is right-invariant, if $d(\rho, \sigma) = d(\rho \circ \eta, \sigma \circ \eta)$ for all $\eta, \rho, \sigma \in P_n$. With $\rho \circ \eta$ we denote the composition function of two permutations $\rho, \eta \in P_n$, which is defined as $\rho \circ \eta = (\rho \eta_1, ..., \rho \eta_n)$.

Proof of Proposition 1. The following two identities hold by right-invariance (see Definition 3):

$$\sum_{i=1}^{n} \rho_i \bar{R}_i = \sum_{i=1}^{n} i(\bar{R} \circ \rho^{-1})_i \quad (18)$$

$$\sum_{i=1}^{n} \rho_i Y_i(\bar{R}) = \sum_{i=1}^{n} i(Y(\bar{R}) \circ \rho^{-1})_i = \sum_{i=1}^{n} iY_i(\bar{R} \circ \rho^{-1}) \quad (19)$$

Eq. (18) implies that $\hat{\rho}_1 = \arg\max_{\rho \in P_n} \sum_{i=1}^{n} i(\bar{R} \circ \rho^{-1})_i$ is such that $(\bar{R} \circ \hat{\rho}_1^{-1})_1 \leq (\bar{R} \circ \hat{\rho}_1^{-1})_2 \leq \cdots \leq (\bar{R} \circ \hat{\rho}_1^{-1})_n$ (by Lemma 2 in Hüllermeier et al. (2008)).

By (19), it follows that $\hat{\rho}_2 = \arg\max_{\rho \in P_n} \sum_{i=1}^{n} iY_i(\bar{R} \circ \rho^{-1})$, is such that $Y_i(\bar{R} \circ \hat{\rho}_2^{-1})_1 \leq Y_i(\bar{R} \circ \hat{\rho}_2^{-1})_2 \leq \cdots \leq Y_i(\bar{R} \circ \hat{\rho}_2^{-1})_n$. 

Proof of Proposition 2. It is sufficient to prove that
\[
\pi(\rho | \theta) \propto \min_{\rho^* \in \mathcal{P}_n \setminus \{\rho\}} KL[p(\cdot | \rho, \theta) \| p(\cdot | \rho^*, \theta)]
\]
is independent of \( \rho \), for any fixed value of \( \theta \), where
\[
KL[p(\cdot | \rho, \theta) \| p(\cdot | \rho^*, \theta)] = \sum_{r \in \mathcal{P}_n} p(r | \rho, \theta) \log \frac{p(r | \rho, \theta)}{p(r | \rho^*, \theta)}
\]
\[
= \frac{\theta}{Z(\theta)} \sum_{r \in \mathcal{P}_n} \exp \left[ -\theta \| \rho - r \|^2 \right] \left[ \| \rho^* - r \|^2 - \| \rho - r \|^2 \right]
\]
(20)
where the last equality follows from the right-invariance of the Spearman distance. By the same argument, it follows that
\[
\pi(\rho | \theta) \propto \min_{\rho^* \in \mathcal{P}_n \setminus \{\rho\}} KL[p(\cdot | \rho, \theta) \| p(\cdot | \rho^*, \theta)]
\]
\[
= \frac{\theta}{Z(\theta)} \min_{\rho^* \in \mathcal{P}_n \setminus \{\rho\}} \sum_{r \in \mathcal{P}_n} \exp \left[ -\theta \| \rho_I - r \|^2 \right] \left[ \| \rho^* - r \|^2 - \| \rho_I - r \|^2 \right],
\]
(21)
does not depend on \( \rho \), thus ending the proof.

Proof of Theorem 1. The proof is analogous to that of Gupta and Damien (2002, Theorem 2).

Proof of Proposition 3. The Jeffreys prior for a parameter \( \theta \) is defined as \( \pi(\theta) \propto \sqrt{\mathcal{I}(\theta)} \), where \( \mathcal{I}(\theta) \) is the Fisher information function of the statistical model
\[
\mathcal{I}(\theta) = -\mathbb{E}_{R|\theta} \left[ \frac{d^2}{d\theta^2} \ln p(R|\theta, \rho) \right].
\]
Recall that for the MM it holds
\[
\ln p(R|\theta, \rho) = -\theta d(R, \rho) - \ln Z_d(\theta),
\]
where \( Z_d(\theta) = \sum_{r \in \mathcal{P}_n} e^{-\theta d(r, \rho)} \). Let us simplify the notation here and set \( Z_d(\theta) := Z(\theta) \). Then,
\[
\frac{d^2}{d\theta^2} \ln p(R|\theta, \rho) = -\frac{Z''(\theta) Z(\theta) - [Z'(\theta)]^2}{[Z(\theta)]^2},
\]
20
is independent of $\rho$. Notice also that

$$Z'(\theta) = - \sum_{r \in P_n} d(r, \rho_I) e^{-\theta d(r, \rho_I)} = -Z(\theta) \sum_{r \in P_n} d(r, \rho_I)p(r|\theta, \rho = \rho_I)$$

$$= - \mathbb{E}_{R|\theta}[d(R, \rho_I)|\theta]Z(\theta),$$

and

$$Z''(\theta) = \sum_{r \in P_n} d^2(r, \rho_I) e^{-\theta d(r, \rho_I)} = Z(\theta) \sum_{r \in P_n} d^2(r, \rho_I)p(r|\theta, \rho = \rho_I)$$

$$= \mathbb{E}_{R|\theta}[d^2(R, \rho_I)|\theta]Z(\theta).$$

Then,

$$\frac{d^2}{d\theta^2} \ln p(R|\theta, \rho) = - \frac{\mathbb{E}_{R|\theta}[d^2(R, \rho_I)|\theta][Z(\theta)]^2 - [\mathbb{E}_{R|\theta}[d(R, \rho_I)|\theta]]^2[Z(\theta)]^2}{[Z(\theta)]^2}$$

$$= -\nabla_{R|\theta}[d(R, \rho_I)|\theta].$$

From the previous equations, it finally holds the result:

$$\pi(\theta) = \sqrt{I(\theta)} = \sqrt{\mathbb{E}_{R|\theta} \left[ \nabla_{R|\theta}[d(R, \rho_I)|\theta]|\theta \right]} = \sqrt{\nabla_{R|\theta}[d(R, \rho_I)|\theta]|\theta}. \quad \square$$

**References**

Berger, J. O., Bernardo, J. M., and Sun, D. (2012). “Objective priors for discrete parameter spaces.” *Journal of the American Statistical Association*, 107(498): 636–648.

Crispino, M. (2017). “Bayesian Learning of Ranking data.” Unpublished doctoral dissertation, Bocconi University, Milan, Italy.  
URL https://drive.google.com/file/d/1Sb2UlJBlcALYcdQ5WiShNKVn5QgXt/view

Dabic, M. and Hatzinger, R. (2009). “Zielgruppenadäquate Abläufe in Konfigurationssystemen - Eine empirische Studie im Automobilmarkt - Partial Rankings.” In Hatzinger, R., Dittrich, R., and Salzberger, T., editors, *Präferenzanalyse mit R: Anwendungen aus Marketing, Behavioural Finance und Human Resource Management*. 1

Dawid, A. (1997). “Comments on ‘non-informative priors do not exist’.” *Journal of Statistical Planning and Inference*, 65(1): 178–180. 2
Diaconis, P. (1988). *Group representations in probability and statistics*, volume 11 of *Lecture Notes - Monograph Series*. Hayward, CA, USA: Institute of Mathematical Statistics. 2, 4, 19

Fligner, M. A. and Verducci, J. S. (1986). “Distance based Ranking Models.” *Journal of the Royal Statistical Society B*, 48(3): 359–369. 4

— (1990). “Posterior probabilities for a consensus ordering.” *Psychometrika*, 55(1): 53–63. 2, 8, 11, 14

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. 1

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. 2, 10, 11, 14, 15, 20

Hüllermeier, E., Fürnkranz, J., Cheng, W., and Brinker, K. (2008). “Label ranking by learning pairwise preferences.” *Artificial Intelligence*, 172(16): 1897–1916. 19

Irurozki, E., Calvo, B., and Lozano, A. (2016). “PerMallows: An R Package for Mallows and Generalized Mallows models.” *Journal of Statistical Software*, 71. 5

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. 17

Mallows, C. L. (1957). “Non-Null Ranking Models. I.” *Biometrika*, 44(1/2): 114–130. 2, 4, 5

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*, volume 64 of *Monographs on Statistics and Applied Probability*. Cambridge, MA, USA: Chapman & Hall. 3, 6, 7

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. 3, 4

Meilă, M. and Bao, L. (2010). “An Exponential Model for Infinite Rankings.” *Journal of Machine Learning Research*, 11: 3481–3518. 2

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. 3
Mollica, C. and Tardella, L. (2014). “Epitope profiling via mixture modeling for ranked data.” *Statistics in Medicine*, 33(21): 3738–3758.

Mukherjee, S. (2016). “Estimation in exponential families on permutations.” *The Annals of Statistics*, 44(2): 853–875.

Sørensen, Ø., Crispino, M., Liu, Q., and Vitelli, V. (2018). *BayesMallows: Bayesian Preference Learning with the Mallows Rank Model*. R package version 0.1.1. URL https://github.com/osorensen/BayesMallows

Sun, M., Lebanon, G., and Kidwell, P. (2012). “Estimating probabilities in recommendation systems.” *Journal of the Royal Statistical Society C*, 61(3): 471–492.

Thompson, G. (1993). “Generalized permutation polytopes and exploratory graphical methods for ranked data.” *The Annals of Statistics*, 1401–1430.

Villa, C. and Walker, S. (2015). “An objective approach to prior mass functions for discrete parameter spaces.” *Journal of the American Statistical Association*, 110(511): 1072–1082.

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.

Xu, H., Alvo, M., and Yu, P. L. (2018). “Angle-based models for ranking data.” *Computational Statistics & Data Analysis*, 121: 113 – 136.

**Acknowledgement**

The authors would like to thank Sonia Petrone, Elja Arjas and Arnoldo Frigessi for their insightful comments.