Stat
The ISI's Journal for the Rapid
Dissemination of Statistics Research
(wileyonlinelibrary.com) DOI: 10.100X/sta.0000

Time Varying Rankings with the Bayesian Mallows Model

Derbachew Asfaw[a], Valeria Vitelli[b], Øystein Sørensen[b], Elja Arjas[b,c], Arnoldo Frigessi[b,*]

Received 00 Month 2016; Accepted 00 Month 2016

We present new statistical methodology for analysing rank data, where the rankings are allowed to vary in time. Such data arise, for example, when the assessments are based on a performance measure of the items which varies in time, or if the criteria, according to which the items are ranked, change in time. Items can also be absent when the assessments are made, because of delayed entry or early departure, or purely randomly. In such situations also the dimension of the rank vectors varies in time. Rank data in a time dependent settings thus lead to challenging statistical problems. These problems are further complicated, from the perspective of computation, by the large dimension of the sample space consisting of all permutations of the items. Here we focus on introducing and developing a Bayesian version of the Mallows rank model, suitable for situations in which the ranks vary in time and the assessments can be incomplete. The consequent missing data problems are handled by applying Bayesian data augmentation within MCMC. Our method is also adapted to the task of future rank prediction. The method is illustrated by analysing some aspects of a data set describing the academic performance, measured by a series of tests, of a class of high school students over a period of four years.

Keywords: Bayesian data augmentation; Mallows model; Preference prediction; Footrule distance; MCMC; Incomplete rank data

1. Introduction

Rank data arise in situations where it is desired to order a set of individuals or items in accordance to some criterion. A set of assessors or tests is used to rank the items. Ranking can be complete, when every assessor orders all items, or incomplete, when the assessors give only partial information about their preferences. This can occur when items are rated individually or compared in pairs, or simply if some items are missed in the ranking. Rank data may also arise when

[a]**Hawassa University School of Mathematical and Statistical Sciences, Hawassa, Ethiopia**
[b]**Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway**
[c]**Department of Mathematics and Statistics, University of Helsinki, Finland**

[*]**Email: arnoldo.frigessi@medisin.uio.no**

Prepared using staauth.cls [Version: 2012/05/12 v1.00]

transforming continuous or discrete scores given to the items into ranks, leading to a nonparametric analysis. This is particularly interesting when score scales are difficult to compare across tests or assessors, while ordering is more robust. Examples of ranking problems include aggregating internet search rankings into meta-search results (Dwork et al., 2001), determining winners of competitions and tournaments (Hunter, 2004; Tutz & Schauberger, 2015), voting and elections (Gormley & Murphy, 2006), market research (Dittrich et al., 2000), food preferences (Kamishima & Akaho, 2009), psychology (Regenwetter et al., 2007; Maydeu-Olivares & Bockenholt, 2005), health economics (Krabbe et al., 2007; Ratcliffe et al., 2006), medical treatments (Plumb et al., 2009) and choice of occupation (Yu & Chan, 2001). Rank data lead to interesting and computationally challenging statistical problems, especially due to the dimension of the space of all permutations of the items: it can easily become intractable to enumerate all permutations as would in principle be needed to maximize a posterior probability or a likelihood.

We are interested in preferences and rankings that change in time. For example, the ranking of the preferred social networking sites, or of the preferred political parties, varies in time as they depend on time varying information (Regenwetter et al., 1999). Many games and sports, including races, involve outcomes in which competitors are rank ordered. In some sports, competitors play in multiple events over long periods of time, and it is natural to assume that their abilities change over time (Glickman & Hennessy, 2015). Best selling books as published each week by the New York Times (Caron & Teh, 2012) show time varying preferences, as do the number of annual citations different papers receive (Radicchi et al., 2009).

In this paper we extend the Bayesian framework for inference with the Mallows model (Vitelli et al., 2015), to model the effect of time in rank data. This is particularly challenging when the preferences are incomplete. We propose a new method of data augmentation (Tanner & Wong, 1987) for the Mallows model for rank data.

This paper is organized as follows. Section 2 describes our illustrative example, a student data set collected between 2002 and 2006 in a high school in Italy. Section 3 presents the model for time dependent rank data and our methods of data augmentation, and outlines the MCMC algorithm needed in the computation. Sections 4 and 5 illustrate the use of the method on the school data. Section 6 concludes the paper with a short discussion.

2. Time Dependent Rank Data (TDRD)

Time dependence in rank data can arise when a panel of assessors is asked to rank the same set of items repeatedly over time. Then the preferences can change in time. Sometimes, while preference criteria are stable in table, the characteristics of the items change in time. Here we use a data set of this latter type, in order to illustrate several challenges. We study the case when each assessor is offered only a subset of the items to rank, a subset which changes over time. Furthermore, new items appear and others disappear from the item basket, generating a longer string of consecutive missingness. In this paper we consider a class of students enrolled in a high school program during four years. Each year the students were tested in mathematics, based on several written tests. We are interested in ranking the students, also because the class could be divided into more homogeneous subgroups. For example, the top 5 students can be challenged with more advanced material, while the bottom 5 students should receive special attention. The reason to pass from marks given from each test to ranks is that tests have varying difficulties and ranks are more robust to variation in grading errors.

The class had 18 students, who are here viewed as items to be ranked, and the tests represent the assessors which perform the ranking. The number of tests in the four years was 5, 4, 8, and 8, respectively. The marks were numbers between 0 (worst) and 10 (best). If all 18 students had attended all tests, this would have led to 450 results in total; in the data, however, 69 test results were missing because a varying number of students were absent. The marks from

each test were converted to ranks, also accounting for the fact that the number of students (items) taking different tests was varying. There were many ties in the data, which made this conversion not unique. Here we handle ties by randomizing the ranks involved, repeatedly inside our algorithms, as explained in section 3.

In our data set one student left the class for good by moving to another school at the end of the first year, and another student left at the end of the second year. A third student joined the class at the beginning of the second year. Particularly on these three students there is a lot of potentially useful information missing. Still, it would be of interest to ask questions such as: Is there some systematic way of predicting how these three students would have performed in the (counterfactual) situation in which they had not left the school early, or not arrived late? Or: How would such hypothetical presence of these three students have influenced the ranks of those 15, who in fact worked their way through all four years? More generally, we ask here the question: "What rank would an item have if we would not have excluded it from the basket of all items?" All these questions can be answered in terms of probabilities, based on an assumed model and on the data actually observed.

3. Mallows Model for TDRD

We start by considering the complete data situation. Assume we have n items, present at all time points $t \in 1, 2, ..., T$, and labeled by $A_1, A_2, ..., A_n$. Time is discrete, so that t represents a certain period of time, for example, a year. A ranking is a permutation of the integers $(1, 2, ..., n)$. We denote the set of all permutations by P_n. We assume that a number of assessors, say N_t, ranked the n items at time t. We denote by $R_j^{(t)} = \left(R_{ij}^{(t)}, i = 1, 2, ..., n\right) \in P_n$ the vector of rankings provided by assessor j, where $R_{ij}^{(t)}$ is the rank given to item i. The data collected at time t are then denoted by $R^{(t)} = \left\{R_j^{(t)}, j = 1, 2, ..., N_t\right\}$.

We assume that, for each t, there is a latent ranking $\rho^{(t)} \in P_n$ of the n items, which reflects a consensus of the N_t assessors at time t. The individual assessments are then viewed as perturbations, or imperfect measurements, from that consensus. In the school example, t is a school year, and $\rho^{(t)}$ is the unknown "true" ranking of the students' performance in mathematics in year t.

3.1. The Complete Data Case

For a fixed time point t, we assume that the observed ranks $R_j^{(t)}, j = 1, 2, ..., N_t$, at time t are conditionally independent, given the corresponding parameters $\rho^{(t)}$ and $\alpha^{(t)}$. Here $\rho^{(t)}$ is the consensus ranking and $\alpha^{(t)}$ a scale parameter. Considering a sequence of rankings $R^{(1:T)} = \left\{R^{(t)}, t = 1, 2, ..., T\right\}$ over T time points, we consider the Mallows likelihood (Mallows, 1957) of the form

$$P\left\{R_1^{(t)}, R_2^{(t)}, .., R_{N_t}^{(t)}, t = 1, ..., T | \alpha^{(1:T)}, \rho^{(1:T)}\right\} = \prod_{t=1}^{T} \prod_{j=1}^{N_t} \left(\frac{1}{Z_n\left\{\alpha^{(t)}\right\}} exp\left[-\frac{\alpha^{(t)}}{n} d\left\{R_j^{(t)}, \rho^{(t)}\right\}\right]\right). \quad (1)$$

It is natural to measure the spread of the rankings $R_{ij}^{(t)}$ around the consensus $\rho^{(t)}$ by means of a distance. Here we consider the particular case of the Footrule distance, i.e. the l_1-distance $d\left(R^{(t)}, \rho^{(t)}\right) = \sum_{j=1}^{N_t} \sum_{i=1}^{n} \left|R_{ij}^{(t)} - \rho_i^{(t)}\right|$. Other distances used in rank models are the Kendall, Spearman, Hamming, Ulam and Cayley distances (Marden, 1995). They are all right invariant, which means they do not depend on the labeling of the items. As discussed in (Vitelli et al., 2015), the normalizing constant $Z_n(\alpha^{(t)})$ does not depend on $\rho^{(t)}$ for right invariant distances.

We assume that the consensus rankings $\rho^{(1:T)} = \left(\rho^{(1)}, \rho^{(2)}, ..., \rho^{(T)}\right)$ do not change too much in consecutive time points, leading to a smoothing prior. We model the transition kernel between latent ranks $\rho^{(t-1)}$ and $\rho^{(t)}$ with a further Mallows model

$$P(\rho^{(t)}|\rho^{(t-1)}, \beta) = \frac{1}{Z_n(\beta)} exp\left[-\frac{\beta}{n} d\left\{\rho^{(t)}, \rho^{(t-1)}\right\}\right] 1_{P_n}\left(\rho^{(t)}\right), \tag{2}$$

where the smoothing parameter $\beta > 0$ describes how strongly the ranks at time t resemble a priori the ranks at time $t-1$. Similarly for the smoothing vector $\alpha^{(1:T)} = \left(\alpha^{(1)}, \alpha^{(2)}, ..., \alpha^{(T)}\right)$ we postulate a Markovian dependence by assuming that conditionally on the value of hyperparameter $\sigma_\alpha > 0$, $P(\alpha^{(t)}|\alpha^{(t-1)}, \sigma_\alpha) \sim N\left(\alpha^{(t-1)}, \sigma_\alpha^2\right) \times 1_{\Re^+}\left\{\alpha^{(t)}\right\}, t = 1, ..., T$. We assume that $\rho^{(t)}$ is independent from $\alpha^{(t)}$ a priori.

Next we consider hyperpriors for all parameters in (1) and (2). We assume that $\rho^{(1)}$ is a priori uniformly distributed in P_n. To specify the prior distribution for $\alpha^{(1)}$ we argue as in (Vitelli et al., 2015): in the Mallows model we have terms of the form $exp\left\{\left(-\alpha^{(t)}/n\right) d\left(R_{ij}^{(t)}, \rho_i^{(t)}\right)\right\}$ contributing multiplicatively to the likelihood. To get some idea of what numerical values of $\alpha^{(1)}$ should have a priori, we can consider how likely it is that the rank $R_{ij}^{(1)}$ given by some assessor j to item i at time 1 deviates from the rank $\rho_i^{(1)}$ by at least n/2. With the footrule distance, this would correspond to $n/2$. We would then have the likelihood contribution $exp\left\{-\alpha^{(1)}/2\right\}$. We thus specify our prior mean for $\alpha^{(1)}$ such that it corresponds to our prior belief that an assessment could be off the mark by n/2. For example, with the prior mean of $\alpha^{(1)}$ equal to 10, the likelihood contribution would be a little less than one percent. We represent this using the exponential distribution $\pi\left(\alpha^{(1)}\right) = \lambda exp\left(-\lambda\alpha^{(1)}\right) 1_{[0,\infty)}\left(\alpha^{(1)}\right)$, with hyperparameter $\lambda = 1/10$. We assume σ_α^2 has an inverse gamma distribution, $P\left(\sigma_\alpha^2\right) = IG\left(a, b\right)$, with shape $a = 1$ and scale $b = 1$.

The joint posterior distribution $P\left(\rho^{(1:T)}, \alpha^{(1:T)}, \beta, \sigma_\alpha | R^{(1:T)}\right)$ of all model parameters, given the observed data $R^{(1:T)}$, can be obtained by applying the chain multiplication rule and the formulas (3) – (6) below. The conditional independence properties assumed in this process are shown in a graphical form in Figure 1. In the final inference, where the main interest is in the consensus rank vector $\rho^{(1:T)}$, the other model parameters are routinely integrated out from this joint posterior.

$$\begin{aligned} P\left(\rho^{(1:T)}|R_{ij}^{(1:T)}, \alpha^{(1:T)}, \beta\right) &\propto \left[\prod_{t=1}^{T} P\left\{R_1^{(t)}, ..., R_{N_t}^{(t)}|\alpha^{(t)}, \rho^{(t)}\right\}\right]\left[P\left\{\rho^{(1)}\right\}\prod_{t=2}^{T} P\left\{\rho^{(t)}|\rho^{(t-1)}, \beta\right\}\right] \\ &= exp\left[-\sum_{t=1}^{T}\frac{\alpha^{(t)}}{n}\sum_{j=1}^{N_t} d\left\{R_j^{(t)}, \rho^{(t)}\right\} - \frac{\beta}{n}\sum_{t=2}^{T} d\left(\rho^{(t)}, \rho^{(t-1)}\right)\right]\left[\prod_{t=1}^{T} 1_{P_n}\left\{\rho^{(t)}\right\}\right]. \end{aligned} \tag{3}$$

The conditional distribution of the scale parameter $\alpha^{(t)}$ is given by

$$\begin{aligned} &P\left(\alpha^{(1:T)}|R_{ij}^{(1:T)}, \rho^{(1:T)}, \sigma_\alpha\right) \\ &\propto \left[\prod_{t=1}^{T} P\left\{R_1^{(t)}, ..., R_{N_t}^{(t)}|\alpha^{(t)}, \rho^{(t)}\right\}\right]\left[P\left\{\alpha^{(1)}\right\}\prod_{t=2}^{T} P\left\{\alpha^{(t)}|\alpha^{(t-1)}, \sigma_\alpha^2\right\}\right] \\ &= \left[\prod_{t=1}^{T}\frac{1}{Z_n\left\{\alpha^{(t)}\right\}^{N_t}}\right] exp\left[-\sum_{t=1}^{T}\frac{\alpha^{(t)}}{n}\sum_{j=1}^{N_t} d\left\{R_j^{(t)}, \rho^{(t)}\right\} - \sum_{t=2}^{T}\frac{1}{2\sigma_\alpha^2}\left\{\alpha^{(t)} - \alpha^{(t-1)}\right\}^2 - \lambda\alpha^{(0)}\right]\left[\prod_{t=1}^{T} 1_{P_n}\left\{\rho^{(t)}\right\}\right]. \end{aligned} \tag{4}$$

Finally, the conditional distribution of β is

$$P\left(\beta|\rho^{(1:T)}\right) \propto \left[\prod_{t=1}^{T}\frac{1}{Z_n\{\beta\}}\right] exp\left[\frac{-\beta}{n}\sum_{t=2}^{T} d\left\{\rho^{(t)}, \rho^{(t-1)}\right\} - \lambda\beta\right]. \tag{5}$$

For σ_α^2 we obtain

$$P\left(\sigma_\alpha^2|\alpha^{(1:T)}\right) \propto P\left(\sigma_\alpha^2\right)\prod_{t=2}^{T} P\left\{\alpha^{(t)}|\alpha^{(t-1)},\sigma_\alpha^2\right\} = \frac{1}{\sigma_\alpha^{2(\alpha+T/2)-1}} exp\left[\frac{1}{\sigma_\alpha^2}\left(-b+1/2\sum_{t=2}^{T}\left\{\alpha^{(t)}-\alpha^{(t-1)}\right\}^2\right)\right]. \quad (6)$$

$P\left(\sigma_\alpha^2|\alpha^{(1:T)}\right)$ is an inverse gamma distribution with scale $b+1/2\sum_{t=2}^{T}\left\{\alpha^{(t)}-\alpha^{(t-1)}\right\}^2$ and shape $\alpha+T/2$.

3.2. The Case of Missing Data

In our school example, missingness in observed rankings occurs when one or more students miss a test. This can happen sporadically, because of being sick, or more systematically because of longer absence from school. More generally, if one or more items are not available for assessment at a certain point in time, the resulting observed data will consist of the ranks of the items actually present, together with a list of those who were not. In such a situation we consider the collection of all complete data rank vectors which are compatible with observed mutual rankings of the items present. For example, suppose there are 3 items, say A_1, A_2 and A_3, and A_3 is missing. If A_1 is ranked ahead of A_2, we have three possible compatible rankings for $\{A_1, A_2, A_3\}$, namely $(1,2,3), (1,3,2)$ and $(2,3,1)$. We have inserted the item A_3 in all possible positions, which are compatible with the observed ordering of the available items. We call this the set of all allowable fill-ins of the missing ranks, which are compatible with the observation.

We solve the problem of incomplete information in time-dependent rank data using Bayesian data augmentation. We formalize the augmentation as follows. Let $U = \{A_1, A_2, ..., A_n\}$ be the n items. Let $U_j^{(t)}$ be the subset of items ranked by assessor j at time t, and let $n_j^{(t)} = \left|U_j^{(t)}\right|$. The items belonging to the complement set $V_j^{(t)} = U \setminus U_j^{(t)}$ remain then unobserved by this assessor at t. However, here we assume that, in the counterfactual situation in which these items, too would have been observed, they could have been ranked together with those now belonging to $U_j^{(t)}$. Thus we assume that there exist latent ranks for all items $A_i \in U$, denoted here by $\tilde{R}_{ij}^{(t)}$, with values between 1 and n. Let $\tilde{R}_{N_t}^{(t)} = \left\{\tilde{R}_{ij}^{(t)}; 1 \le i \le n, 1 \le j \le N_t\right\}$ and $\tilde{R}^{(1:T)} = \left\{\tilde{R}_{N_t}^{(t)}; t = 1, 2, ..., T\right\}$. We then assume that these latent variables $\tilde{R}^{(1:T)}$ are distributed according to the complete data model specified in eq. (1).

Our next task is to connect, in situations in which some items were missing and therefore not available for ranking, these latent variables to the ranks that were actually provided in the observed data. Supposing that item A_i had been available to assessor j at time t, i.e., $A_i \in U_j^{(t)}$, we denote by $R_{ij}^{(t)}$ its corresponding observed rank. The observed mutual ranking $R_j^{(t)} = \left\{R_{ij}^{(t)}; A_i \in U_j^{(t)}\right\}$ is then assumed to be compatible with the ranking of the same items in the latent and perhaps only partially observed ranking $\left\{\tilde{R}_{ij}^{(t)}; A_i \in U_j^{(t)}\right\}$. Thus, for $A_k, A_h \in U_j^{(t)}, A_k \succ A_h$ holds whenever $\tilde{R}_{kj}^{(t)} < \tilde{R}_{hj}^{(t)}$ is true in the latent ordering of these same items. The observed data are again denoted by $R^{(1:T)} = \left\{R^{(t)}, t = 1, 2, ..., T\right\}$. Note that, as soon as it is known what items are missing and what are available for ranking, the observed ranks are fully determined by the latent ranks. More formally, for each $U_j^{(t)}$ there is a deterministic mapping $r_{U_j^{(t)}}$ from $\{1, 2, ..., n\}$ to $\left\{1, 2, ..., n_j^{(t)}\right\}$ such that $r_{U_j^{(t)}}\left(\tilde{R}_{ij}^{(t)}\right) = R_{ij}^{(t)}$.

Before entering the technical treatment of how the required data augmentation from observed to latent data is performed, we need to consider the issue of whether the possible presence of missing rank information can be considered without biasing the statistical inferences. For this purpose, we postulate that the missingness mechanism, in the sense of specifying the sets $V_j^{(t)}$ of items unavailable to assessor j at time t, or, equivalently, their complements $U_j^{(t)}$, is completely at random (MCAR). More exactly, we assume that the conditional distributions $p\left(U_j^{(t)}|\sigma_\alpha, \alpha^{(1:T)}, \beta, \rho^{(1:T)}, \tilde{R}^{(1:T)}\right)$ are independent from all the conditioning variables, and can therefore be written

simply as $p\left(U_j^{(t)}\right)$. This then implies, by a straightforward application of the chain multiplication rule and the assumed conditional independence properties, that joint distribution of all model variables can be written as

$$p\left(\sigma_\alpha, \alpha^{(1:T)}, \beta, \rho^{(1:T)}, \tilde{R}^{(1:T)}, R^{(1:T)}\right) = p\left(\sigma_\alpha\right) p\left(\alpha^{(1:T)}\right) p\left(\beta\right) p\left(\rho^{(1:T)}|\beta\right) p\left(\tilde{R}^{(1:T)}|\alpha^{(1:T)}, \rho^{(1:T)}\right) p\left(R^{(1:T)}|\tilde{R}^{(1:T)}\right), \tag{7}$$

where the last factor takes the form of the product

$$p\left(R^{(1:T)}|\tilde{R}^{(1:T)}\right) = \prod_{t=1}^{T}\prod_{j=1}^{N_t} p\left(U_j^{(t)}\right) 1\left(r_{U_j^{(t)}}\left(\tilde{R}_j^{(t)}\right) = R_j^{(t)}\right). \tag{8}$$

As a consequence, for given parameter values $\sigma_\alpha, \alpha^{(1:T)}, \beta, \rho^{(1:T)}$, augmented values of $\tilde{R}_j^{(t)}$ matching with observed data $R_j^{(t)}$ can be sampled independently for different t and j, from the corresponding constrained Mallows models proportional to $p\left(\tilde{R}_j^{(t)}|\rho^{(t)}, \alpha^{(t)}\right) 1\left(r_{U_j^{(t)}}\left(\tilde{R}_j^{(t)}\right) = R_j^{(t)}\right)$.

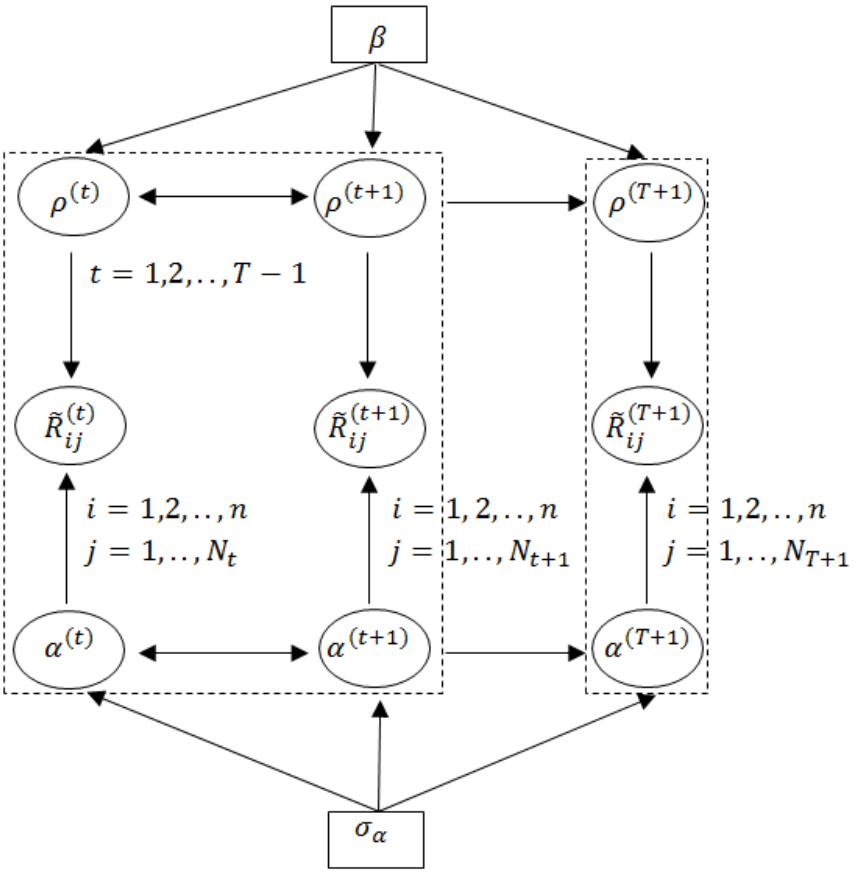

Figure 1. Predictive time dependent rank model represented by a dynamic Bayesian network.

3.3. Metropolis-Hastings Algorithm

Numerical estimation of the joint posterior $p\left(\sigma_\alpha, \alpha^{(1:T)}, \beta, \rho^{(1:T)}, \tilde{R}^{(1:T)} | R^{(1:T)}\right)$, based on formula (7), is performed by alternating between (Step 1) sampling augmented values $\tilde{R}^{(1:T)}$ from the conditional distribution

$$p\left(\tilde{R}^{(1:T)} | R^{(1:T)}, \sigma_\alpha, \alpha^{(1:T)}, \beta, \rho^{(1:T)}\right) = p\left(\tilde{R}^{(1:T)} | R^{(1:T)}, \alpha^{(1:T)}, \rho^{(1:T)}\right), \tag{9}$$

and (Step 2) sampling the parameters $\left(\sigma_\alpha, \alpha^{(1:T)}, \beta, \rho^{(1:T)}\right)$ from the conditional distribution

$$p\left(\sigma_\alpha, \alpha^{(1:T)}, \beta, \rho^{(1:T)} | \tilde{R}^{(1:T)}, R^{(1:T)}\right) = p\left(\sigma_\alpha, \alpha^{(1:T)}, \beta, \rho^{(1:T)} | \tilde{R}^{(1:T)}\right). \tag{10}$$

Sampling from (10) can be further divided into Steps 2a-2d, where each sub-step corresponds to sampling from one of the formulas (3) – (6) presented earlier in the complete data situation. In (3) and (4), however, the symbols $R^{(1:T)}$, which now refer to possibly incomplete observed data, need to be replaced by the respective latent variable symbols $\tilde{R}^{(1:T)}$. Finally, as the main interest is generally in inferences on the consensus rankings $\rho^{(1:T)}$, the corresponding marginal posterior can be computed from the Monte Carlo samples taken from the joint posterior $p\left(\sigma_\alpha, \alpha^{(1:T)}, \beta, \rho^{(1:T)}, \tilde{R}^{(1:T)} | R^{(1:T)}\right)$ by restricting to only the corresponding $\rho^{(1:T)}$ coordinates.

Step 1: Given the current $\tilde{R}_j^{(t)}$ and the current values for the parameters, sample the new augmented vectors $\left(\tilde{R}_j^{(t)}\right)'$ separately for each j, $j = 1, 2, ..., N_t$ and t, $t = 1, 2, ..., T$ in $G_j^{(t)}$ (set of all possible fill-ins) from the leap-and-shift proposal distribution (Vitelli et al., 2015) centered at $\tilde{R}_j^{(t)}$, as described in Section 3.2 (by retaining the observed mutual ordering $R_j^{(t)}$ for the items in $U_j^{(t)}$, while then randomly assigning compatible ranks for items in $V_j^{(t)}$ by perturbing $\tilde{R}_j^{(t)}$). The proposed $\left(\tilde{R}_j^{(t)}\right)'$ is then accepted with probability

$$min\left\{1, exp\left[-\frac{\alpha^{(1:T)}}{n}\sum_{j=1}^{N_t} d\left\{\left((\tilde{R}_j^{(1:T)})', \rho^{(1:T)}\right) - d\left(\tilde{R}_j^{(1:T)}, \rho^{(1:T)}\right)\right\}\right]\right\}. \tag{11}$$

Step 2a: Starting at $\alpha^{(1:T)} \geq 0, \rho^{(1:T)} \in P_n$ and given the prior distribution $\pi(\rho^{(1)})$ and $\pi(\alpha^{(1)})$, the proposal $\left(\rho^{(1:T)}\right)'$ is sampled from the symmetric leap and shift distribution and accept it with probability

$$\begin{aligned} &min\left\{1, \frac{\pi((\rho^{(t)})')}{\pi(\rho^{(t)})} exp\left[-\frac{\alpha^{(t)}}{n}\sum_{j=1}^{N_t}\left\{d\left(\tilde{R}_j^{(t)}, (\rho^{(t)})'\right) - d\left(\tilde{R}_j^{(t)}, \rho^{(t)}\right)\right\} - \right.\right. \\ &\left.\left.\frac{\beta}{n}\left\{\left[d\left((\rho^{(t)})', \rho^{(t-1)}\right) + d\left((\rho^{(t)})', \rho^{(t+1)}\right)\right] - \left[d\left(\rho^{(t)}, \rho^{(t-1)}\right) + d\left(\rho^{(t)}, \rho^{(t+1)}\right)\right]\right\}\right], \end{aligned} \tag{12}$$

for time, $t = 2, 3, ..., T-1$.

Step 2b: We sample a proposal $\left(\alpha^{(1:T)}\right)'$ according to $N(\alpha, \sigma_\alpha^2)$ and accept it with probability

$$\begin{aligned} &min\left\{1, \frac{Z_n((\alpha^{(t)})')^{-N_t}\pi((\alpha^{(t)})')}{Z_n(\alpha^{(t)})^{-N_t}\pi(\alpha^{(t)})} exp\left[-\frac{\left((\alpha^{(t)})' - \alpha^{(t)}\right)}{n}\sum_{j=1}^{N_t}\left\{d\left(\tilde{R}_j^{(t)}, \rho^{(t)}\right)\right\} - \right.\right. \\ &\left.\left.\left([\frac{\left((\alpha^{(t)})' - \alpha^{(t-1)}\right)^2}{\sigma_\alpha^2} + \frac{\left((\alpha^{(t)})' - \alpha^{(t+1)}\right)^2}{\sigma_\alpha^2}] - [\frac{\left(\alpha^{(t)} - \alpha^{(t-1)}\right)^2}{\sigma_\alpha^2} + \frac{\left(\alpha^{(t)} - \alpha^{(t+1)}\right)^2}{\sigma_\alpha^2}]\right)\right], \end{aligned} \tag{13}$$

for time, $t = 2, 3, ..., T-1$. We use the obvious simplification of (12) and (13) at $t = 1$ and $t = T$. Note that negative $(\alpha^t)'$s can be proposed, but they are accepted with zero probability because they are outside the prior support. We tested also a lognormal proposal, which worked equivalently in our case.

Step 2c: We sample a proposal β' from $N(\beta, \sigma_\beta^2)$ and the acceptance probability for the Metropolis-Hastings algorithm, is given by

$$min\left\{1, \frac{Z_n(\beta')\pi(\beta')}{Z_n(\beta)\pi(\beta)} exp\left[-\frac{(\beta'-\beta)}{n}\left(d(\rho^{(t)}, \rho^{(t-1)})\right)\right]\right\}. \quad (14)$$

for $t = 2, 3, ..., T$.

Step 2d: Sample σ_α^2 from the inverse gamma distribution with scale $b + 1/2\sum_{t=2}^{T}\left\{\alpha^{(t)} - \alpha^{(t-1)}\right\}^2$ and shape $\alpha + T/2$.

Since $Z_n(\alpha^{(1:T)})$ does not depend on $\rho^{(1:T)}$, it can be computed offline on a grid of $\alpha^{(1:T)}$ values and to yield an estimate over a continuous range, see (Vitelli et al., 2015) for details.

Algorithm: TDRD with Bayesian Mallows Model

Input: $G_1^{(t)}, G_2^{(t)}, ..., G_j^{(t)}$ or $r_{U_j^{(t)}}$, λ, $d(.,.)$, sd_α, sd_β, L, $Z_n(\alpha^{(1:T)})$, T, M.

Output: Posterior distribution of $\rho^{(1:T)}$, $\alpha^{(1:T)}$, β, σ_α^2, $\tilde{R}_1^{(t)}, ..., \tilde{R}_{N_t}^{(t)}$

Initialize: ρ^1, α^1, β and σ_α

```
if {G_1^(t), G_2^(t), ..., G_j^(t)} are among inputs then
    for t ⟵ 1 to T
        for j ⟵ 1 to N_t
            randomly generate R̃_{j,1}^(t) in G_j^(t)
        end
    end
else
    for t ⟵ 1 to T
        for j ⟵ 1 to N_t
            randomly generate R̃_{j,1}^(t) in G_j^(t) compatible with r_{U_j^(t)}
        end
    end
end
for m ⟵ 1 to M
    for t ⟵ 1 to T
        Update ρ^(t):
            Sample (ρ^(t))' from leap-and-shift distribution centered at ρ_{m-1}^(t)
            Compute: ratio = Equation (12) with ρ^(t) ⟵ ρ_{m-1}^(t) and α^(t) ⟵ α_{m-1}^(t)
            Sample: U ~ U(0,1)
            if U < ratio then
                ρ_m^(t) ⟵ ρ'^(t)
            else
                ρ_m^(t) ⟵ ρ_{m-1}^(t)
            end
        Update α^(t):
```

Sample: $(\alpha^{(t)})' \sim N\left\{\alpha_{m-1}^{(t)}, \sigma_\alpha^2\right\}$
Compute: ratio = Equation (13) with $\rho^{(t)} \longleftarrow \rho_m^{(t)}$ and $\alpha^{(t)} \longleftarrow \alpha_{m-1}^{(t)}$
Sample: $U \sim U(0,1)$
if $U <$ ratio then
$\alpha_m^{(t)} \longleftarrow \alpha'^{(t)}$
else
$\alpha_m^{(t)} \longleftarrow \alpha_{m-1}^{(t)}$
end
Update $\tilde{R}_1^{(t)}, \tilde{R}_2^{(t)}, ..., \tilde{R}_{N_t}^{(t)}$:
for j $\longleftarrow$ 1 to N_t
if $\left\{G_1^{(t)}, ..., G_{N_t}^{(t)}\right\}$ are among inputs then
Sample: $\left(\tilde{R}_j^{(t)}\right)'$ in $G_j^{(t)}$ from leap-and-shift distribution centered at $\tilde{R}_{j,m-1}^{(t)}$
else
Sample: $\left(\tilde{R}_j^{(t)}\right)'$ from leap-and-shift distribution centered at $\tilde{R}_{j,m-1}^{(t)}$ and compatible with $r_{U_j^{(t)}}$
end
Compute: ratio = Equation (11) with $\rho^{(t)} \longleftarrow \rho_m^{(t)}$ and $\alpha^{(t)} \longleftarrow \alpha_m^{(t)}$ and $\tilde{R}_j^{(t)} \longleftarrow \tilde{R}_{j,m-1}^{(t)}$
Sample: $U \sim U(0,1)$
if $U <$ ratio then
$\tilde{R}_{j,m}^{(t)} \longleftarrow \left(\tilde{R}_j^{(t)}\right)'$
else
$\tilde{R}_{j,m}^{(t)} \longleftarrow \tilde{R}_{j,m-1}^{(t)}$
end
end
end
Update β:
Sample: $\beta' \sim N\left(\beta_{m-1}, \sigma_\beta^2\right)$
Compute: ratio = Equation (14) with $\rho^{(t+1)} \longleftarrow \rho^{(t)}$
Sample: $U \sim U(0,1)$
if $U <$ ratio then
$\beta_m \longleftarrow \beta'$
else
$\beta_m \longleftarrow \beta_{m-1}$
end
Update σ_α^2:
Compute: Equation (6) with $\alpha^{(t+1)} \longleftarrow \alpha^{(t)}$
$\sigma_\alpha^2 \sim \text{IG}\left(\alpha + T/2, b + 1/2 \sum_{t=2}^{T} \left\{\alpha^{(t)} - \alpha^{(t-1)}\right\}^2\right)$
end

4. School Data Example

The student data set is available in the supplementary material. In this section we develop new methods needed to represent and understand the results based on our Mallows model using the school data as example.

A first impression of the mutual performance of the 18 students is provided by Figure 2 (a), which shows the ranking of the students at the end of each of the four school years, based on the averages of their grades in the tests during that same year. The coloured letter coding, from A to R, refers to individual students. Note that two or more students can, and in fact do in Figure 2 (a), have the same test average, then leading to a tied rank. We see that student Q joined the class only at the beginning of the second year, while student H left the class after the first year, and student F after two years. In addition the attendance of the students in individual tests varied; for example, of the 16 students in the class during the fourth year, the number of attendees in a test varied from 14 to 16. For missing data we apply the data augmentation methods presented in section 3.2. Estimating the performance of the two students who left school early is very difficult, if at all possible, one wonders.

We consider the posterior distribution of the full 18-dimensional consensus rank vector, given by $p(\rho^{(t)}|R^{(1:t)}), t = 1, 2, .., 4$. The performance of individual students can then be described in terms of the corresponding posterior marginal mode (Figure 2 (b)) or the posterior marginal mean (Figure 2 (c)) of each student. Again, because we look at marginal summaries, some ranks are tied. Note that, unlike the ranks in Fig. 2 (a) directly based on test grade averages in year t, the estimates in Fig. 2 (b) and (c) account for the complete history of test results up to the end of year t, but do so, for $t = 2$, 3 and 4, by progressively discounting the influence of the results from the earlier years. Looking from this perspective, one might anticipate the resulting estimates to be somewhat more stable than those in Figure 2 (a). Moreover, if a student was absent from at most a few tests during a school year, the estimates produced by the data augmentation method, shown in Figures 2 (b) and (c), could be expected to be similar to those that would be obtained in the counterfactual situation in which these students had attended all tests.

But such a conclusion does not seem to hold for the two students, H and F, who had left the school early. They were among the weakest students. Instead, we see in Figures 2 (b) and (c) what looks like a considerable improvement in the performance of these two students over time: in the case of H starting from the second year, at which time he had already left the school, and similarly for F, from the third year onwards. Perhaps unsurprisingly, in such open end situations, where there are no data at all from some time point onwards, the modal and mean estimates obtained by the data augmentation method start shifting towards the center of the range of possible values. This is clearly an artefact produced by the uniform prior in absence of data in an open jaw case.

Given this variation in the point estimates, and particularly in view of the artefact concerning the performance of students H and F, one may ask whether there would be some way to diagnose, and even solve, such potential problems. The Bayesian recipe for solving problems of this kind is to account explicitly for the uncertainties in the estimates, as represented by the full posterior. One way to do this is to consider the marginal posteriors for the ranks of individual students in terms of their cumulative distribution functions (CDFs), shown in Figure 3, and then try to find instances where students could be compared in the sense of stochastic ordering. According to this figure, the same top three students, G, I and K, in the class can be identified with high credibility during all four years of school, with student O being systematically their closest contender. At the low end there is more variability, but in year $t = 4$, a group of three students (N, M and P) can be distinguished quite clearly from the others. (In applications to marketing, such indications can be useful for deciding whether to continue, or to stop, promoting certain items in stock.) In the middle range, the differences between students are somewhat less clear, and also vary somewhat from one year to the next. On the other hand, the ordering between pairs of students in a given year, based in Fig. 2 (b) and (c) on posterior point estimates, can in Fig. 3 often be seen to hold in the sense of stochastic ordering.

While such consideration on marginal posteriors for the consensus rankings, where one CDF dominates another in the sense of stochastic ordering, offers a natural way of comparing student performance, there is another aspect in Figure 3 which catches the eye. The CDF for student H looks almost uniform between ranks 5 and 18 already in year $t = 2$. A similar phenomenon, a major increase in the dispersion of the marginal posterior, can be found in connection of

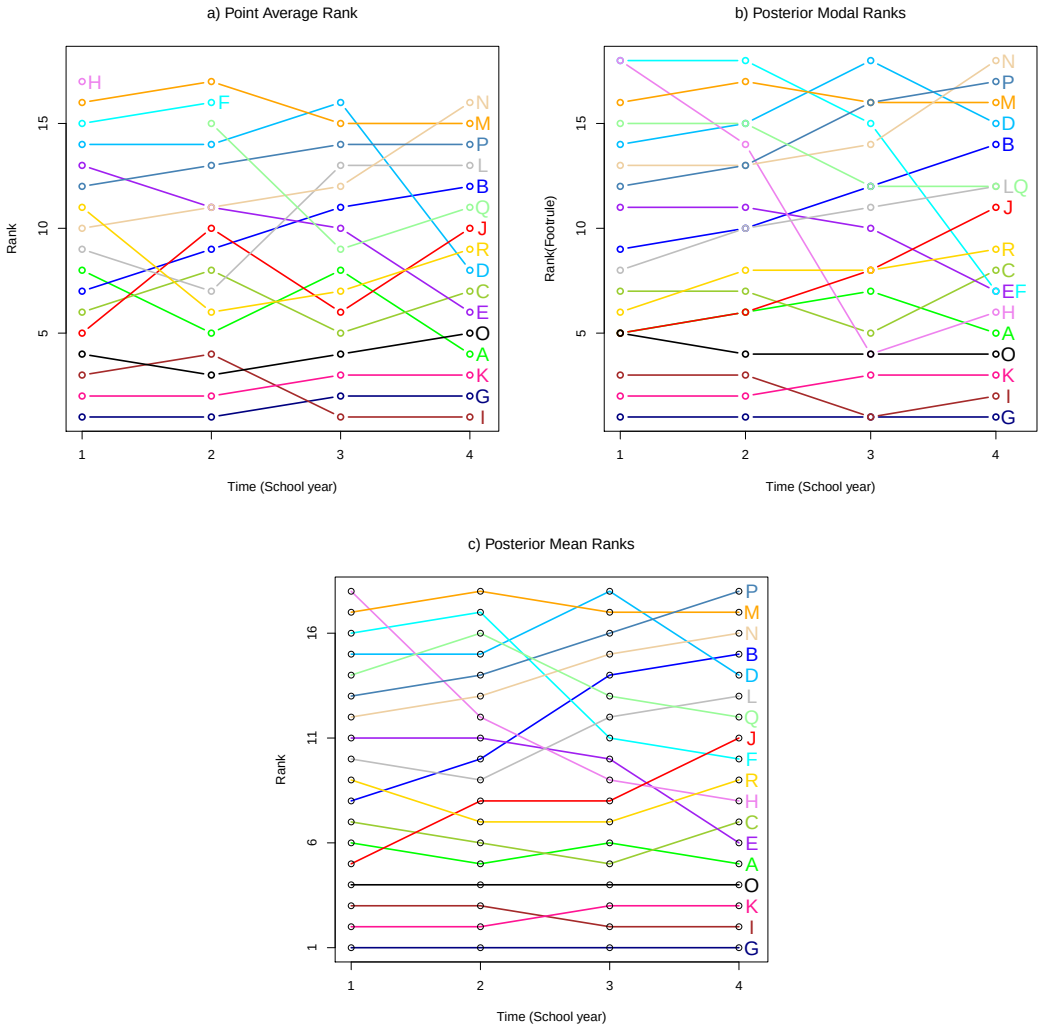

Figure 2. The development of the ranks of the considered 18 students, measured at the end of each of the four school years: ranking based on (a) the grade averages from the tests in that same year; (b) the posterior marginal modal rank, and (c) the posterior marginal means, computed from the marginals of the posterior $p(\rho^{(t)}|R^{(1:t)})$, for $t = 1, ..., 4$. Best student has rank 1.

student F when comparing years $t = 2$ and $t = 3$. But we already have a natural explanation for these observations: students H and F had not attended a single test in those years, and in this open end situation the posteriors became, in fact, predictive distributions. In the consequent model based prediction, the memory from earlier test results is seen to progressively fade out (with the strength of the memory being represented in the model by the discount factor $e^{-\beta}$), and the dispersion of the posterior distributions increases accordingly. Our model thus gives clear indications on the posterior uncertainty.

The marginal posteriors $p(\alpha^{(t)}|R^{(1:t)}), t = 1, ..., 4$, and $p(\beta|R^{(1:4)})$ for the model parameters are shown in Figure 4. The stochastically smaller values of the scale parameter $\alpha^{(1)}$ in comparison to the other $\alpha^{(t)}$ reveals that the precision

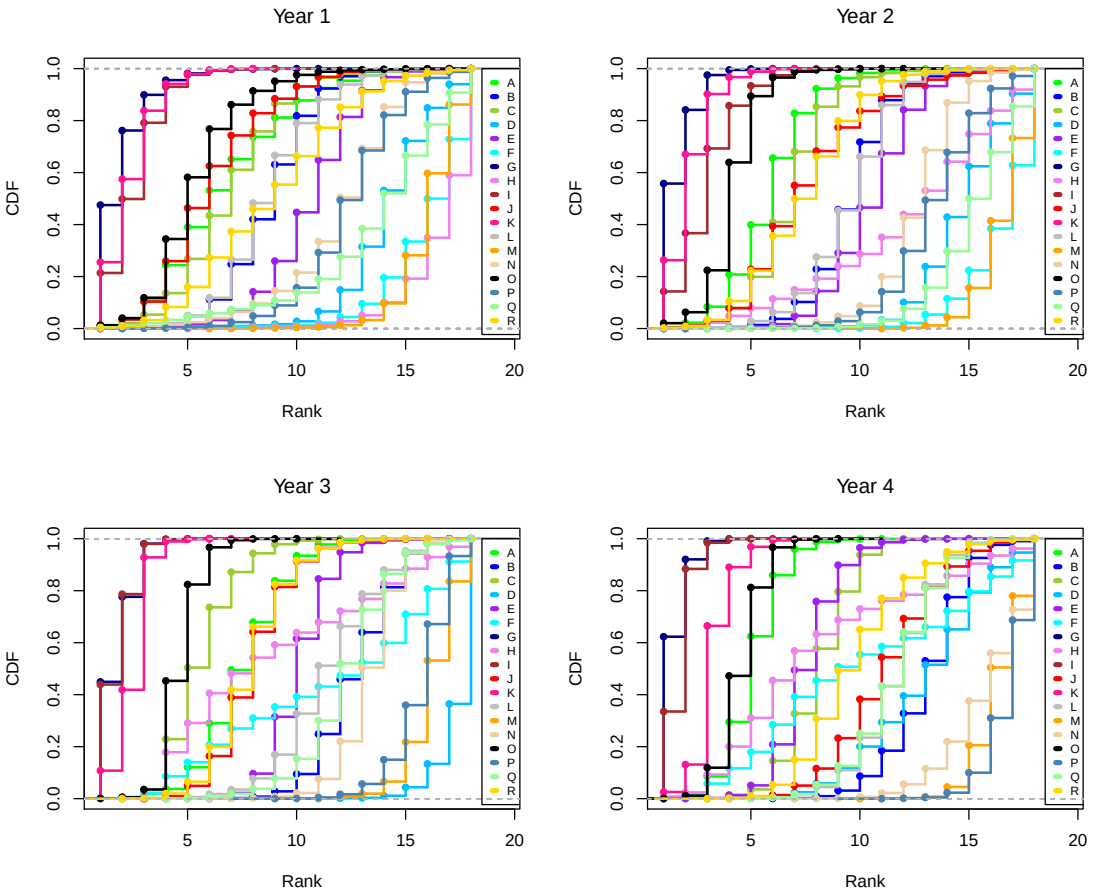

Figure 3. Marginal posterior CDFs for the consensus ranks of the individual students at the end of each of the four school years.

of the consensus ranking after only the first year of data was less precise than those from the second year onwards.

5. Prediction of Individual Test Results

The goal in section 4 was to estimate the consensus ranking of all 18 students at the end of each of the four school years. The attendance of the students in individual tests varied in time, and three students missed one or more complete years at school. This missing data problem was handled by applying, within the considered MCMC, a Bayesian data augmentation method described in section 3.2. All computations and comparisons were then systematically performed in terms of full 18-dimensional rank vectors. The situation changes if, instead of attempting to estimate the consensus rank of the students, the goal is to predict his or her ranking in an individual test, based on results from earlier tests. Then, if it is known which students are going to be absent from the considered test, it is no longer relevant to include them in the ranking. Another obvious difference, is that prediction of the correct outcome of a game involves a far higher degree of uncertainty than the assessment of the strength of a player. The same holds true for predicting the results from an individual test.

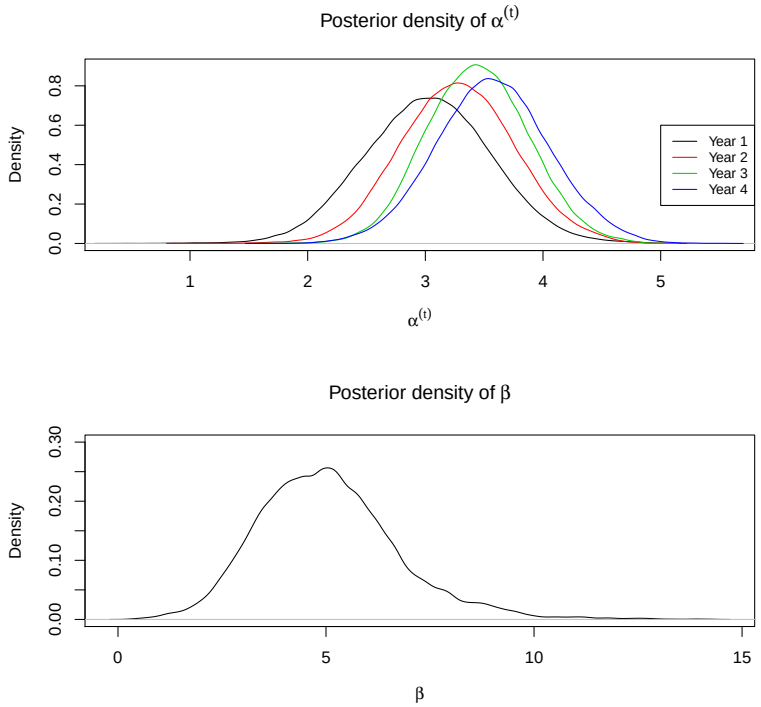

Figure 4. Posterior densities for model parameters $\alpha^{(t)}$ and β when applying the Footrule distances.

For the Bayesian Mallows model, the prediction task can be handled by considering the corresponding predictive distributions. However, the varying attendance of students in the tests complicates the issue. One possibility would be to perform all computations, both for drawing statistical inferences and for making the predictions, separately for each collection of students corresponding to true attendance, thus eliminating from the data those absent. Such elimination of part of data seems wasteful, however, and it also forces one to repeat closely similar computations for each predicted test. Here we follow a different path, and compute the predictive distributions in the fixed setting of 18-dimensional rank vectors. The prediction concerning the correct ranking of the students, who in fact took the test, is then handled within the MCMC by performing the obvious mapping from the ranks of all 18 students to the ranks of the actual attendees.

This idea is illustrated by considering the final test (No. 8) at the end of the fourth school year. The number of students taking that test was 15 (with students H and F having left the school already earlier, but also student N missing this test) so that the range of ranks was from 1 to 15. The prediction of the student rankings was carried out in three different ways: by using, as background data, the test results from (a) the two first school years; (b) the three first school years; and (c) when combining (b)with results from the seven tests preceding the considered test No. 8 in year four.

Figure 5 shows the posterior predictive CDFs for the ranks in test No. 8 computed in ways (a) on the top left , (b) on the top right and (c) on the bottom. The true ranks derived from the observed test grades are indicated with small balls of the same colour, on the x-axis. Note that there were some ties: students A, E and K were all ranked 4^{th}, because they got the same grade in this last test, and, for the same reason, B and L were ranked 11^{th}. Table 1 gives

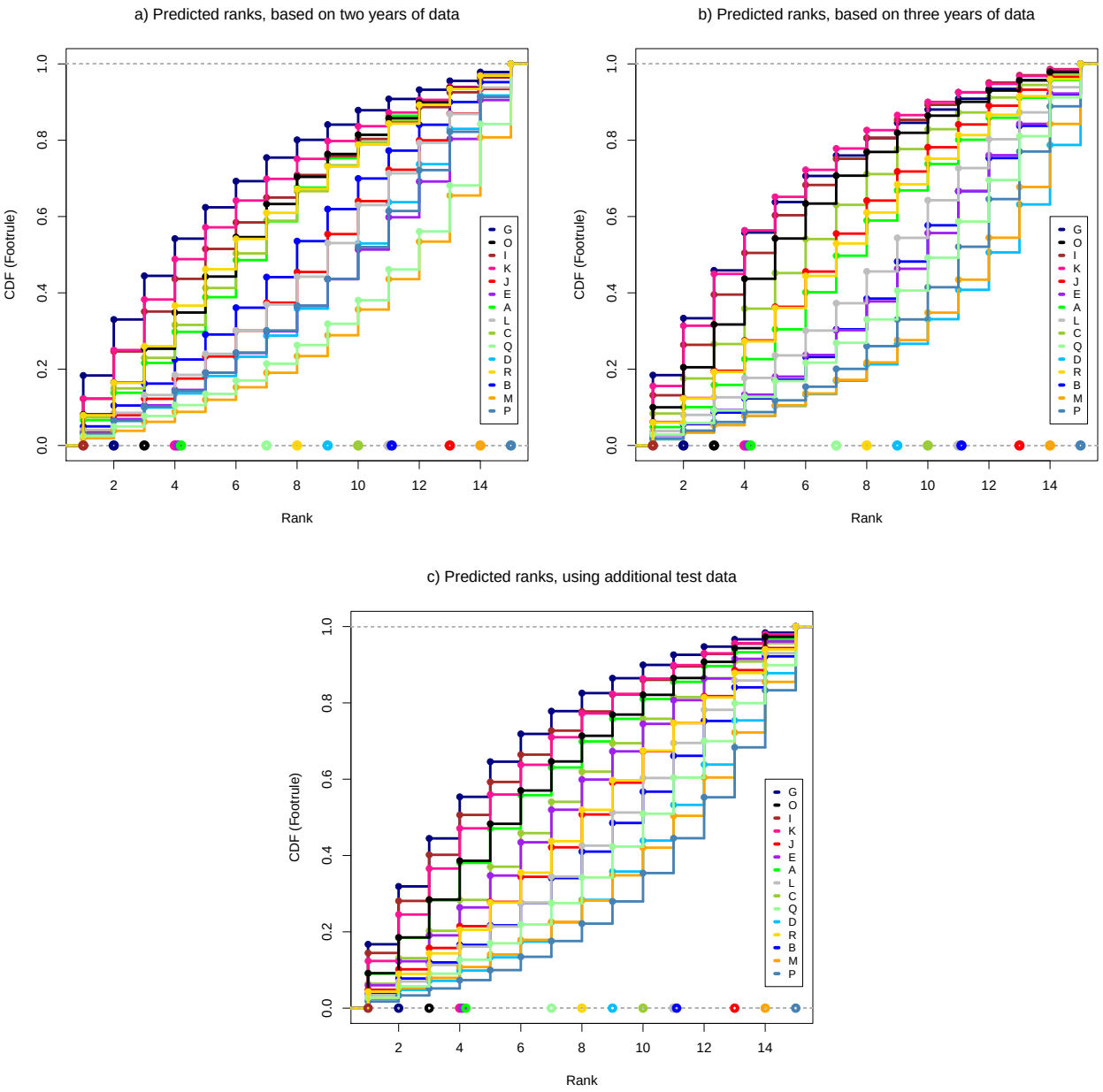

Figure 5. The posterior predictive CDFs for the ranks of individual students in the final test (No.8) in year four, when based on test results from the first two school years (a), the first three school years (b), and when (b) is combined with results from the seven preceding tests in year four (c). The coloured balls on the x-axis show the observed true ranks.

some numerical values for the predictive probabilities appearing in Figure 5.

More generally, when comparing the three predictions (a), (b) and (c) to each other, it would seem plausible that,

Based on background data from				
Student	True rank	Two years	Three years	Three years + Seven preceding tests
I	1	0.5	0.6	0.61
G	2	0.61	0.64	0.65
O	3	0.43	0.53	0.48
K	4	0.56	0.66	0.57
A	4	0.39	0.31	0.47
E	4	0.21	0.19	0.34
Q	7	0.15	0.17	0.17
R	8	0.44	0.36	0.27
D	9	0.2	0.11	0.12
C	10	0.42	0.44	0.37
B	11	0.3	0.18	0.21
L	11	0.23	0.24	0.2
J	13	0.23	0.36	0.29
M	14	0.12	0.11	0.15
P	15	0.2	0.12	0.11

Table 1. Posterior predictive probabilities of being ranked among the true top-5 in test No. 8, when using three different sets of background data as explained in the text. The results correspond directly to the values of the posterior CDFs in Figure 5 when read at rank 5, arranged in the order of the true outcome from that test.

when progressing from situation (a) via (b) to (c), the prediction should in both steps become more accurate. This is because the predictions made later are based on more background information on the earlier performance of the students, and also since such additional information is more recent and therefore closer in time to the predicted event itself.

We have done such a comparison in Figure 6, by computing the distribution of the prediction error $(d\,(R_{8,pred}, R_{8,obs})\,|data)$, where $R_{8,obs}$ and $R_{8,pred}$ are the observed and the predicted rank vector for test No. 8 for the 15 students who took that test, d is the Footrule distance, and data represents the background data in each of the three situations (a), (b) and (c). The figure (CDFs on the left, densities on the right) shows that, at least in the considered context, having more background information was indeed useful in the sense that it made the prediction error stochastically smaller.

6. Conclusion

In this paper we developed a new method based on the Mallows rank model, for preferences which vary in time and involve missing observations. Data of this type arise in marketing, where customers/users are asked from time to time to rank, rate or compare a number of products contained in a basket. Opinions of customers change in time, possibly smoothly. Also, the content of the basket does not remain fixed, as some items may be dropped and others added. For example, Shi, et al. (2012) presented a novel visualization method to help users explore the changes in value and ranking in large time series data. Some rankings, such as best-seller lists or trends measured by Twitter and Wikipedia, are intrinsically volatile, changing daily (Blumm et al., 2012). These types of data involve massive missingness, due to permanent changes in the basket and limited expression of interests of customers. Here the missing data problem was

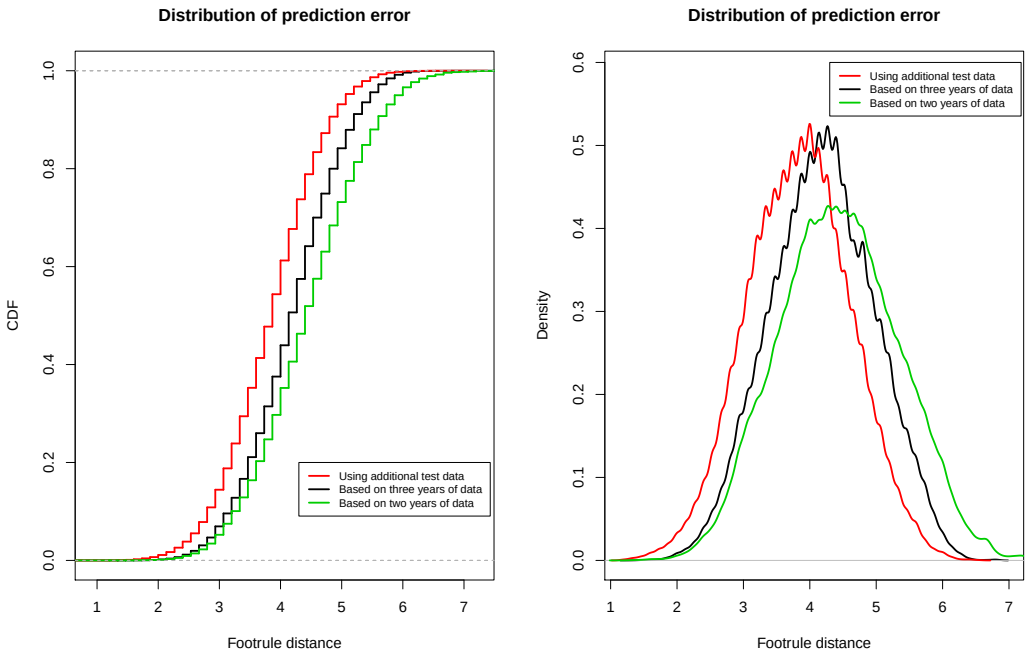

Figure 6. Distribution of prediction error in the form of cummulative distribution function (left) and in the form of density function (right).

handled by applying Bayesian data augmentation within the Mallows model with Footrule distance.

Prediction can be seen as a missing data problem, solved here by sampling from the posterior distribution using Gibbs sampling within Metropolis-Hasting algorithm. Our method describes uncertainty in predictions, which can be used to understand the reliability of the forecasts themselves. Goodness-of-fit could be investigated by a systematic leave-one-out approach and Bayesian prediction. We discussed several ways to learn from the joint posterior distribution over all permutations of n items, producing interpretable summaries. We used the Footrule distance in the Mallows model, though other right invariant distance could be used instead. We leave it to future work to compare various distances, to select the most appropriate ones for missingness imputation and prediction.

Acknowledgement

We would like to thank Roberta Michelini for the student data, the NORHED project at Hawassa University for funding, and Zeytu Gashaw for discussions.

References

Blumm, N, Ghoshal, G, Forro, Z, Schich, M, Bianconi, G, Bouchaud, JP Barabasi, AL (2012), Dynamics of ranking processes in complex systems. *Physical Review Letters*, **109**(12), 128701.

Caron, F, and Teh, YW (2012). Bayesian nonparametric models for ranked data. *In Advances in Neural Information Processing Systems* (pp. 1520-1528).

Craig, BM, Busschbach, JJV & Salomon, JA (2009), Modeling ranking, time trade-off, and visual analog scale values for eq-5d health states: A review and comparison of methods. *Medical Care*, **47**(6), 634-641.

Dittrich, R, Katzenbeisser, W & Reisinger, H (2000), The analysis of rank ordered preference data based on Bradley-Terry type models. *OR Spektrum*, **22**, 117-134.

Dwork, C, Kumar, R, Naor, M & Sivakumar, D (2001), Rank aggregation methods for the web. *In Proceedings of the* 10^{th} *international conference on World Wide Web* (pp. 613-622). ACM.

Glickman, ME & Hennessy, J (2015), A stochastic rank ordered logit model for rating multi-competitor games and sports. *Journal of Quantitative Analysis in Sports*, **11**(3), 131-144.

Golder, SA & Macy, MW (2011), Diurnal and seasonal mood vary with work, sleep, and day length across diverse cultures. *Science*, **333**(6051), 1878-1881.

Gormley, IC & Murphy, TB (2006), Analysis of Irish thirdâĂŘlevel college applications data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**(2), 361-379.

Hidalgo, CA, Blumm, N, BarabÃąsi, AL, & Christakis, NA (2009), A dynamic network approach for the study of human phenotypes. *PLoS Comput Biol*, **5**(4), e1000353.

Hunter, DR (2004), MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, 384-406.

Kamishima, T & Akaho, S (2009), Efficient clustering for orders. *In Mining complex data* (pp. 261-279). Springer Berlin Heidelberg.

Krabbe, PFM, Salomon, JA & Murray, CJL (2007), Quantification of health states with rank-based nonmetric multidimensional scaling. *Medical Decision Making*, **27**, 395-405.

Mallows, CL (1957), Non-null ranking models. I. *Biometrika*, **44**(1/2), 114-130.

Mantegna, RN & Stanley, HE (1995), Scaling behaviour in the dynamics of an economic index. *Nature*, **376**(6535), 46-49.

Marden, JI (1995), Analyzing and modeling rank data, volume 64 of Monographs on Statistics and Applied Probability.

Maydeu-Olivares, A & Bockenholt, U (2005), Structural equation modeling of pairedcomparison and ranking data. *Psychological Methods*, **10**(3), 285-304.

Plumb, AAO, Grieve, FM, and Khan, SH (2009), Survey of hospital cliniciansâĂŹ preferences regarding the format of radiology reports. *Clinical Radiology*, **64**, 386-394.

Radicchi, F, Fortunato, S, Markines, B, and Vespignani, A (2009), Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, **80**(5), 056103.

Ratcliffe, J, Brazaier, J, Tsuchiya, A, Symonds, T & Brown, M (2006), Estimation of a preference based single index from the sexual quality of life questionnaire (SQOL) using ordinal data. *Discussion Paper Series, Health Economics and Decision Science, The University of Sheffield*, **06**, 6.

Regenwetter, M, Falmagne, JC, & Grofman, B (1999), A stochastic model of preference change and its application to 1992 presidential election panel data. *Psychological Review*, **106**(2), 362.

Regenwetter, M, Ho, MHR and Tsetlin, I (2007), Sophisticated approval voting, ignorance priors, and plurality heuristics: A behavioral social choice analysis in a Thurstonian framework. *Psychological Review, 114*(4), 994-1014.

Shi, C, Cui, W, Liu, S, Xu, P, Chen, W, & Qu, H (2012), RankExplorer: Visualization of ranking changes in large time series data. *Visualization and Computer Graphics, IEEE Transactions on*, **18**(12), 2669-2678.

Tanner, MA, & Wong, WH (1987), The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, **82**(398), 528-540.

Tutz, G & Schauberger, G (2015), Extended ordered paired comparison models with application to football data from German Bundesliga. *AStA Advances in Statistical Analysis*, **99**(2), 209-227.

Vitelli, V, Srensen, Ø, Frigessi, A, & Arjas, E (2015), Probabilistic preference learning with the Mallows rank model. *arXiv preprint arXiv:1405.7945*.

Yu, PLH & Chan, LKY (2001), Bayesian analysis of wandering vector models for displaying ranking data. *Statistica Sinica*, **11**, 445-461.