Model-based learning from preference data

Qinghua Liu, Marta Crispino, Ida Scheel, Valeria Vitelli, Arnoldo Frigessi

Model-based learning from preference data

Qinghua Liu[1,*], Marta Crispino[2,*],
Ida Scheel[1], Valeria Vitelli[3], Arnoldo Frigessi[3]

[*] Jointly first authors
[1]Department of Mathematics, University of Oslo, Oslo, Norway
[2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble, France
[3]OCBE, University of Oslo, Oslo, Norway

Abstract

Preference data occurs when assessors express comparative opinions about a set of items, by rating, ranking, pair comparing, liking or clicking. The purpose of preference learning is to (i) infer on the shared consensus preference of a group of users, sometimes called rank aggregation; or (ii) estimate for each user her individual ranking of the items, when the user indicates only incomplete preferences; this is an important part of recommender systems. We provide an overview of probabilistic approaches to preference learning, including the Mallows, Plackett-Luce, Bradley-Terry models and collaborative filtering, and some of their variations. We illustrate, compare and discuss the use of these models by means of an experiment in which assessors rank potatoes, and with a simulation. The purpose of this paper is not to recommend the use of one best method, but to present a palette of different possibilities for different questions and different types of data.

Keywords— Preference learning with uncertainty, Mallows model, Plackett-Luce model, recommender systems, Bayesian inference, Bradley-Terry model

1 Introduction

Massive Online Open Courses (MOOC) are offered by many universities. Thousands of students sign up for the most popular MOOCs and this has lead to challenges in grading so many written exams. One way to evaluate exams is to let the students themselves check the exams of their peers. All the exams are distributed to students, assigning a small set of exams to each student. Every exam is assigned to several assessors. Every student acts as an assessor and ranks the exams he has been assigned. The task is then to aggregate all these partial rankings into a full ranking of all exams, on the basis of which grades are given: the top 5% receive an A and so on, see for instance Fang et al. (2017), Raman & Joachims (2015). This is an example of rank aggregation, where the aim is to summarize the preferences or opinions about a set of items expressed by a group of assessors. Voting systems, where every individual ranks a list of candidates, are another common application of rank aggregation

(Jacques & Biernacki, 2014). Search result aggregation, as performed for example by metasearch engines (Desarkar et al., 2016), is another form of compiling a consensus ordering of items from independent rankings.

In other situations, the aim is not to produce a consensus ranking of the items, but to estimate the ranking of each individual assessor, when she has only given partial information on her preferences pertaining the items. Recommender systems (Aggarwal, 2016) are algorithms that select, among many, the items which each individual user of a system is likely to be mostly interested in, beyond the ones she has already interacted with. Web-based, video-on-demand companies (like Netflix or Spotify), where customers log-on to their individual accounts, use expressions of preferences made by each customer to predict new videos (items) which she is likely to find attractive. Preference learning is about inferring individual preferences to perform personalized recommendations (Agarwal & Chen, 2016).

This paper is concerned with statistical methods for rank aggregation and individual preference learning. Since these types of problems and the corresponding rank data are very common and important, the literature is vast, see for example the excellent books by Marden (1995), Alvo & Yu (2014), Ricci et al. (2015) and Fürnkranz & Hüllermeier (2010). Many methods are not based on statistical models, and do not allow for quantification of the uncertainty of the predictions. The Borda counts method (Lansdowne & Woodward, 1996) for example, is an easy and classical way of computing rank averages, but is not probabilistic. Machine learning approaches for learning to rank and rank aggregation, are often optimizing some objective functions, as in Kemeny & Snell (1962), but often do not provide uncertainty quantification of the estimates. In this paper we focus on probabilistic methods for rank aggregation and individual preference learning, based on statistical models that allow for uncertainty quantification. We present the most important approaches and illustrate their usage.

There are three main objectives when assessors (users) rate, rank or compare items: (i) aggregate the data coming from a group of homogeneous assessors, and summarize their preferences into a shared consensus ranking of the items; (ii) estimate for each assessor her individual ranking of the items, when the assessors express only incomplete preferences; this amounts to predicting the ranks of (all or some of) the unranked items at the individual level; (iii) cluster the assessors into groups, each sharing a consensus ranking of the items, and classify new assessors. Often objective (iii) appears in combination with the other two, because the assessors are not all sharing a common preference.

Data is usually classified into two main classes: (i) explicit data, when assessors express a direct opinion about some items, in the form of a rating (by assigning likes, stars), of a ranking (by ordering some items according to their preferences), or of pair comparisons (by indicating which of two items they prefer); (ii) implicit data, when assessors express an indirect preference by choosing some items and not others, in the form of clicking/buying/watching, in general interacting with an item, chosen among several possible ones.

The type of data is application dependent. When there are many items to compare or when the subjective preference is complicated, it is often easier for assessors to compare items in pairs rather than ranking several items at the same time. Another typical situation where pairwise comparison data arise is tournament data, when two players or teams play against each other, see

Glickman (1999). In this paper we will deal with both explicit and implicit data, appearing in three forms: full rankings of all items, partial rankings of some items, and pairwise comparisons of some items.

Once a consensus ranking or an individual ranking is estimated, it can be used in subsequent decisions. In the MOOC example, grades are assigned to each exam; in the recommender system example, a set of items is suggested to each user. These downstream activities can be, and actually often are, challenging and inferential in nature. Where should be the threshold between grade A and B? How many items should be recommended because they are likely to be liked by the user ("exploit") and how many should be suggested among other types of items ("explore")? In this paper we will not discuss these following problems. However, uncertainty quantification of the estimated consensuses and individual rankings, as well as class memberships when clustering is involved, appears to be of fundamental importance. Actions based on unreliable predictions might better be postponed until more data is available and safer predictions can be made. A very natural approach to uncertainty quantification is the Bayesian approach, and in this paper Bayesian inference will play an important role.

This paper is organized as follows. In **Section 2** we introduce the mathematical notation. **Section 3** is dedicated to methods: first for consensus learning (**Section 3.1**) and then for individual preference learning (**Section 3.2**). Then we illustrate several methods in **Section 4**. The purpose is to see various approaches in action and we do this with the help of a compact dataset, namely our potato ranking data. We explain the differences between the various methods, and the contexts when they are most appropriate. **Section 5** is dedicated to a larger, this time simulated, example, where assessors are heterogeneous to meet goal (iii). We conclude with a short discussion of aspects which were not discussed in this paper.

2 Notation

A full ranking of n items is a mapping $\boldsymbol{R} : \mathcal{A} \to \mathcal{P}_n$ from a finite set $\mathcal{A} = \{A_1, ..., A_n\}$, the set of labeled items to be ranked, to the space of n-dimensional permutations $\mathcal{P}_n$. This results from the attribution of a rank $R_i \in \{1, ..., n\}$ to each item A_i, according to some criteria. We denote a full ranking of all items as $\boldsymbol{R} = (R_1, ..., R_n)$. By convention, $R_i < R_k$ means that item A_i is preferred to item A_k, since the rank assigned to A_i is *lower* than the one assigned to A_k (the most preferred item has rank 1), but is read A_i is ranked *higher* than A_k. The full ordering of the n items is an alternative way of representing ranking data, and is denoted by $\boldsymbol{X} = (X_1, ..., X_n)$. Here the components are items in $\mathcal{A}$, ordered from the most preferred to the least, according to $\boldsymbol{R}$. In other words, we have: $X_i = A_k \iff R_k = i$, $\forall i, k = 1, ..., n$. For this we use the shorthand, $\boldsymbol{X} = \boldsymbol{R}^{-1}$. Then $\boldsymbol{X} \in \mathcal{X}_n$ is the set of permutations of the labels in $\mathcal{A}$.

For example, given the following full ranking of the items labelled $\mathcal{A} = \{A_1, ..., A_{10}\}$

$$\begin{array}{rcccccccccc} & A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 & A_8 & A_9 & A_{10} \\ \boldsymbol{R} = & (\,1, & 7, & 8, & 2, & 10, & 4, & 6, & 9, & 3, & 5\,) \end{array},$$

the corresponding ordering vector is the following:

$$\boldsymbol{X} = \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ (A_1, & A_4, & A_9, & A_6, & A_{10}, & A_7, & A_2, & A_3, & A_8, & A_5) \end{matrix}.$$

Some methods are naturally defined on rankings, while others, equivalently, on orderings.

Preferences are expressed by N users or assessors, indexed by $j \in \{1, ..., N\}$. The full ranking given by user j is denoted by $\boldsymbol{R}_j$, with components $R_{ji}, i \in \{1, ..., n\}$.

There is an important relation between a ranking and pairwise preferences. For an unordered pair of items $\{A_i, A_k\}$, we denote a pairwise preference between the two items as $(A_i \prec A_k)$, if item A_i is preferred to item A_k. Given a full ranking $\boldsymbol{R} \in \mathcal{P}_n$, it is immediate to list all the possible $n(n-1)/2$ pairwise orderings among the items, according to the rule

$$(A_{t_1} \prec A_{t_2}) \iff R_{t_1} < R_{t_2}, \qquad t_1, t_2 \in \{1, ..., n\}, \quad t_1 \neq t_2. \tag{1}$$

We refer to the pairwise preferences obtained above as *derived pairwise preferences* (DPP), to distinguish them from *real pairwise preferences* (RPP), which arise when assessors compare items in pairs directly, rather than providing a full ranking. The main difference between DPP and RPP is that DPP are always complete and transitive, while RPP can be incomplete (if an assessor does not compare all the possible pairs), and non-transitive (if an assessor happens to contradict herself).

3 Methods

In this paper we group preference learning methods according to their tasks, namely (1) methods for learning the consensus ranking of a set of items by a group of assessors, or rank aggregation methods, and (2) methods for learning the individual rankings, when they are not readily available from the data, in order to perform personalized recommendations. In most practical situations, it is also necessary to partition the assessors into homogeneous groups, each one sharing a consensus ranking. In this section we assume that the group of users is homogeneous, while in **Section 5** we will show an example requiring clustering of the users.

Most of the methods in this paper assume that items are characterized by a real-valued score (or utility). The score can be shared by a group of assessors, or each assessor can have her own score. The assessors' preferences depend on this score such that the item with a higher score at the moment of comparison has a tendency to be preferred to an item with lower score. When the task is rank aggregation, the purpose is to infer the shared latent score. In the case of learning individual ranking, the focus is on the assessors' individual scores, even though many methods also provide an estimate of the shared score. Other methods are based on a latent-ranking of the items, with a parameter varying in $\mathcal{P}_n$.

3.1 Methods for rank aggregation, inference on the consensus

We start with the class of so-called order statistics models, originating from the work of Thurstone (1927), who introduced the *Law of Comparative Judgment*.

Formally, the order statistics model assumes that in a ranking task involving n items, there exist n random score variables $Y_1, ..., Y_n$, assumed independent and each distributed according to its own distribution F_i. The model then assigns a ranking $\boldsymbol{R} = (R_1, ..., R_n) \in \mathcal{P}_n$ to the items according to the probability

$$P(\boldsymbol{R}) = P(Y_{X_1} < Y_{X_2} < \cdots < Y_{X_n} | \boldsymbol{X} = \boldsymbol{R}^{-1}). \tag{2}$$

Under the order statistics model, the generative process of a ranking of n items is determined by the relative ordering of the n random scores. The most common simplification of (2) is to assume an additive parametric model for the random scores, that is, $Y_i = u_i + \epsilon_i$, $i = 1, ..., n$, where the u_i is the mean score associated with item A_i and ϵ_i captures its variability. Such models are known as Thurstone order statistics models. Depending on the choice of $F_i(y) = F(y - u_i)$ different models arise: the Gaussian assumption on F gives rise to the Thurstone model (Thurstone, 1927) and the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959) assumes that F is Gumbel. The probability in (2) has a closed form under the BTL model, therefore many works on order statistics ranking models are based on the BTL model. Based on data $\boldsymbol{R}_1, ..., \boldsymbol{R}_N$ from N users, the purpose is to estimate the parameter vector $\boldsymbol{u} = (u_1, ..., u_n)$, which can be converted to a consensus ranking by applying the *rank* transformation $\text{rank}(\boldsymbol{u})$, defined as follows: $\boldsymbol{Y} = \text{rank}(\boldsymbol{u})$ if and only if $Y_i = \sum_{h=1}^{n} \mathbb{1}(u_h \geq u_i)$, $i = 1, ..., n$.

A different way to build a likelihood for the data $\boldsymbol{R}_1, ..., \boldsymbol{R}_N$ is to transform these full rankings into pair comparisons, in the spirit of DPP. The most famous models in this class are the Mallows-Bradley-Terry (MBT) and the $\phi-$ and $\rho-$models, both developed by Mallows (1957).

The MBT assumes that the probability that an item A_i is preferred to an item A_k in each comparison, $P(A_i \prec A_k)$, has the Bradley-Terry form (Bradley & Terry, 1952), $P(A_i \prec A_k) = \frac{\mu_i}{\mu_i + \mu_k}$, where $\mu_i > 0$, $i = 1, ..., n$, and $\sum_{i=1}^{n} \mu_i = 1$. That is to say, it only depends on the item-specific score parameters of the two items under comparison (larger values of μ_i correspond to more preferred items). The MBT model likelihood term for one assessor is then,

$$P(\boldsymbol{R}|\boldsymbol{\mu}) = c(\boldsymbol{\mu}) \prod_{i=1}^{n-1} \mu_i^{n-R_i}, \tag{3}$$

where $c(\boldsymbol{\mu})$ is the normalizing factor. If the N users express independent preferences, then the maximum likelihood estimate of $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$ is obtained by maximizing $\prod_{j=1}^{N} P(\boldsymbol{R}_j|\boldsymbol{\mu})$. Finally, the consensus ranking is obtained by ordering the vector of estimated scores using $\text{rank}(\boldsymbol{\mu})$.

The Plackett-Luce (PL) model (Luce, 1959; Plackett, 1975) is a multistage ranking model. The PL model assumes that, given a vector of score parameters $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$, an ordering $\boldsymbol{X}$ arises through the following process: the top ranked item is A_i with probability $\frac{\mu_i}{\sum_{k=1}^{n} \mu_k}, i = 1, ..., n$; then the second to the top item is chosen among the remaining items, excluding the item chosen as

first, say A_m, with renormalized multinomial probabilities $\frac{\mu_i}{\sum_{k \neq m} \mu_k}, i \neq m$; the process continues by excluding items already selected and renormalizing the multionomial probabilities, until a full ordering is obtained. The probability of an ordering $\boldsymbol{X} = (X_1, ..., X_n)$ is

$$P(\boldsymbol{X}|\boldsymbol{\mu}) = \prod_{i=1}^{n-1} \frac{\mu_i}{\sum_{j=i}^{n} \mu_j} \cdot \tag{4}$$

In view of the one-to-one correspondence between ordering and ranking vectors, the probabilities of eq. (4) can be converted to those for rankings $\boldsymbol{R} = \boldsymbol{X}^{-1}$, by performing a simple manipulation. Inferring the parameters of the PL model is typically done by maximum likelihood, using a Majorize-Minimization algorithm (see e.g. Hunter, 2004).

Among the Bayesian approaches, Caron & Doucet (2012) proposed an efficient data augmentation scheme, which, combined with the introduction of a conjugate prior specification, allows to perform a Gibbs sampling for the PL parameters. Recently, Mollica & Tardella (2016b) developed the `PLMIX` R package, which provides basic functions to make Bayesian inference on the PL model parameters. Both Caron & Doucet (2012) and Mollica & Tardella (2016a) specify independent gamma priors for each element of the score parameter vector $\boldsymbol{\mu}$, that is, they assume the prior $\pi(\boldsymbol{\mu}) = \prod_{i=1}^{n} \text{Ga}(\mu_i; a, b)$.

A different way of building statistical models for ranking data is to design appropriate distributions over the space of permutations $\mathcal{P}_n$, avoiding the use of a real valued score. A class of such models is the distance-based family of distributions for rankings, referred to as the Mallows model (MM) (Mallows, 1957).

The MM, formalized in its general form by Diaconis (1988), assumes that the probability of an observed ranking decays as the distance between the ranking and the consensus increases. The probability of a ranking $\boldsymbol{R} = (R_1, ..., R_n) \in \mathcal{P}_n$ is

$$P(\boldsymbol{R} \,|\, \alpha, \boldsymbol{\rho}) = \frac{1}{Z_n(\alpha)} \exp\left[-\frac{\alpha}{n} d(\boldsymbol{R}, \boldsymbol{\rho})\right], \tag{5}$$

where $\boldsymbol{\rho} \in \mathcal{P}_n$ is the location parameter (representing the shared true consensus ranking), $\alpha > 0$ is the scale parameter (describing the concentration around the shared consensus), $d(\cdot, \cdot)$ is a right-invariant distance function, which is unaffected by any relabelling of the items (Diaconis, 1988), between two $n-$dimensional permutations. $Z_n(\alpha)$ is the partition function.

Depending on the choice of the distance, several models arise: the $\phi-$model, originally proposed in Mallows (1957), corresponds to the MM with the Kendall distance $d_K(\boldsymbol{R}, \boldsymbol{\rho}) = \sum_{i<k} \mathbb{1}\left[(\rho_i - \rho_k)(R_i - R_k) < 0\right], \quad 1 \leq i < k \leq n$, and the $\rho-$model (Mallows, 1957) uses the Spearman's (l_2) distance $d_S(\boldsymbol{R}, \boldsymbol{\rho}) = \sum_{i=1}^{n} (\rho_i - R_i)^2$. Other relevant distance functions are the footrule (l_1) distance $d_F(\boldsymbol{R}, \boldsymbol{\rho}) = \sum_{i=1}^{n} |\rho_i - R_i|$, the Hamming distance $d_H(\boldsymbol{R}, \boldsymbol{\rho}) = n - \sum_{i=1}^{n} \mathbb{1}_{\rho_i}(R_i)$, the Cayley distance $d_C(\boldsymbol{R}, \boldsymbol{\rho})$, which measures the minimum number of transpositions to convert $\boldsymbol{R}$ into $\boldsymbol{\rho}$, and the Ulam distance $d_U(\boldsymbol{R}, \boldsymbol{\rho})$, which is the number of deletion-insertion operations to convert $\boldsymbol{R}$ into $\boldsymbol{\rho}$, see Marden (1995) for details. The Kendall distance, which measures the number of adjacent transpositions required to convert $\boldsymbol{R}$ into $\boldsymbol{\rho}$ or, equivalently, the number of discordant pairs in $\boldsymbol{R}$ and $\boldsymbol{\rho}$, is by far the most popular one in the

literature on the MM, mainly for computational reasons. The partition function $Z_n(\alpha)$ exists in closed form only for some choices of right-invariant distances, in particular for the Kendall distance (Lu & Boutilier, 2014; Meilă & Chen, 2010), the Hamming distance (Irurozki et al., 2014) and the Cayley distance (Irurozki et al., 2018).

Inferencing on $\boldsymbol{\rho}$ in the MM by maximum likelihood is challenging. Assuming N observed rankings, $\{\boldsymbol{R}_1, ..., \boldsymbol{R}_N\}$, the maximum likelihood estimator for $\boldsymbol{\rho}$ is obtained as $\boldsymbol{\rho} = \text{argmin}_{\boldsymbol{\rho}\in\mathcal{P}_n} \sum_{j=1}^{N} d(\boldsymbol{R}_j, \boldsymbol{\rho})$. The search space for $\boldsymbol{\rho}$ is $\mathcal{P}_n$, and computation is infeasible except for certain distances.

Vitelli et al. (2018) introduced a Bayesian approach for the Mallows model, which we will refer to as the Bayesian Mallows Model, or BMM. In the BMM, a uniform prior distribution is usually chosen for the consensus parameter $\boldsymbol{\rho}$, $\pi(\boldsymbol{\rho}) = n!^{-1}$. For the scale parameter α, an exponential prior may be employed, $\pi(\alpha) = \lambda_\alpha e^{\lambda_\alpha \alpha} 1_{[0,\infty)}(\alpha)$. The hyper-parameter λ_α is determined by tuning or by subjective reasonings, see Vitelli et al. (2018). The joint posterior distribution of $\boldsymbol{\rho}$ and α is given (up to a constant of proportionality) by

$$\pi(\boldsymbol{\rho}, \alpha | \boldsymbol{R}^{(N)}) \propto \frac{\pi(\boldsymbol{\rho})\pi(\alpha)}{[Z_n(\alpha)]^N} \exp\left[-\frac{\alpha}{n} \sum_{j=1}^{N} d(\boldsymbol{R}_j, \boldsymbol{\rho}) \right] . \tag{6}$$

Posterior samples of $\boldsymbol{\rho}$ and α can be obtained by applying a version of the Metropolis-Hastings algorithm.

The Mallows models have parameters varying in the discrete space of all $n!$ permutations of the integers $1, ...n$, which is more structured than the continuous parameters in the MBT and PL models. We will see in some of our experiments that this makes a difference. On the other hand, the continuous parametrization is computationally advantageous.

Both distance and score based methods can be extended to data that originates from the comparison of items in pairs. For a complete review of these models, we refer to Cattelan (2012). The two traditional probabilistic models for pairwise comparison data is the Thurstone (1927) and the Bradley & Terry (1952) models. Based on these, many extensions were derived in the past, mostly in the econometric and psychometric literatures. The models for pairwise comparison data are designed for real pair preferences (RPP), but can obviously be applied to derived pair preferences (DPP).

We here present the Bradley-Terry (BT) model, first proposed in Zermelo (1929), for which the probability of a pair comparison can be expressed in the form

$$P(A_i \prec A_k | \boldsymbol{u}) = \frac{1}{1 + e^{-(u_i - u_k)}} = \frac{\mu_i}{\mu_i + \mu_k} = P(A_i \prec A_k | \boldsymbol{\mu}), \tag{7}$$

where the transformation $\mu_i = e^{u_i}$, $i = 1, ...n$, shows the relation to the MBT model. The key assumption of the BT model is that all pairwise probabilities are conditionally independent given the vector of scores $\boldsymbol{u}$ (or $\boldsymbol{\mu}$), and that they depend only on the relative sizes of the corresponding score parameters.

Suppose we have collected a number of pairwise comparisons among n items. Let $\boldsymbol{W}$ be the $n \times n$ data matrix with entries w_{ik} being the number of times item A_i is preferred to item A_k among all assessors. Under the BT model, the

likelihood of $\boldsymbol{\mu}$ is

$$P(\boldsymbol{W}|\boldsymbol{\mu}) = \prod_{1\leq i\neq k\leq n} \left(\frac{\mu_i}{\mu_i + \mu_k}\right)^{w_{ik}}. \tag{8}$$

The BT model has been generalized in several directions (see e.g. Davidson, 1970; Agresti, 1996; Wu et al., 2015). Maximum likelihood estimation is typically performed through iterative algorithms and Majorized - Minimization algorithms (Hunter, 2004). In the Bayesian framework, the scheme of Caron & Doucet (2012), was specifically developed for the BT model.

The BT model has a major drawback: inference fails when the data is very sparse, and in particular when the strong connection condition (Ford, 1957) is not met. This condition guarantees the existence and uniqueness of the maximum likelihood estimates of the BT parameters. It holds for a data set if for any partition of the n items into two sets, some items in the second set have been compared to some items in the first set at least once by some assessor, see Yan (2016). This might not be the case in situations where only few comparisons have been made.

The BMM has also been extended to pair comparison data by means of data augmentation, see Vitelli et al. (2018). The underlying principle is to augment, for every assessor, her comparison data into a latent full ranking of all n items, which is compatible with the pair preferences. These augmented individual latent rankings $\tilde{\boldsymbol{R}}_j$, for all assessors j, resemble the shared latent consensus $\boldsymbol{\rho}$ via (6). This construction assumes that the pair comparisons expressed by each assessor are transitive, that is, compatible with some latent $\boldsymbol{R}_j$. For example preferring item A_1 to item A_2 and this to item A_3 but then prefering A_3 to item A_1 would not be coherent. Crispino et al. (2018) extend the BMM to non-transitive pair comparsons, by adding a further component of the model which captures such mistakes (or incoherences) made by the individual user.

3.2 Methods for preference learning, inference on the individual rankings

Apart from performing rank aggregation or consensus learning, another main objective of preference learning is learning individuals' preferences, when not readily available from the data. This situation is common when the data is in the form of partial rankings (that is, when assessors only rank a subset of all items, typically their most preferred), or pairwise preferences (when each assessor is repeatedly asked to choose the preferred item in a number of pairs). In such cases, the full individual ranking is unknown, and the interest is to infer it. Personalized recommendations can hence be made based on the inferred complete individual ranking: the recommended items are the ones on top.

The most popular approaches so far for recommender systems is the well-known collaborative filtering (CF) method (Koren et al., 2009). Unlike the model-based approaches mentioned in the previous sections, CF does not produce an aggregation of group preferences, but is a method that only infers on personal preferences. CF is grounded on matrix factorization technique. Suppose we have an $N \times n$ sparse matrix $\mathbf{R}$, which contains some explicit ratings given by N users to n items, as well as some missing values. Let κ be te set of (user, item) pairs, which are given (not missing). Typically $|\kappa| << nN$.

CF's goal is to obtain two factor matrices - a user matrix $\mathbf{U} \in \mathbb{R}^{N \times L}$, in which each user is associated with a vector $\boldsymbol{u}_j$, and an item matrix $\mathbf{V} \in \mathbb{R}^{n \times L}$, where each item is associated with an item vector $\boldsymbol{v}_i$. Both $\boldsymbol{u}_j$,$\boldsymbol{v_i} \in \mathbb{R}^L$, $j = 1, ..., N$, $i = 1, ..., n$. L is much smaller than n and N. The product $\mathbf{U}^T\mathbf{V}$ should approximate the non-missing entires κ of the matrix $\mathbf{R}$, and otherwise filling out all missing entries. This matrix can then be used to make personalized recommendations. The latent factors $\mathbf{U}$ and $\mathbf{V}$ can be learned by optimizing the function

$$\min_{\mathbf{U},\mathbf{V}} \sum_{(j,i)\in\kappa} (r_{ji} - \boldsymbol{u}_i{}^T \boldsymbol{v}_j)^2 + \gamma(||\boldsymbol{u}_i||^2 + ||\boldsymbol{v}_j||^2). \tag{9}$$

The first term in the objective function describes the squared error between the data matrix $\mathbf{R}$ and the estimated product matrix. The second term is a ridge-like penalization term to avoid overfitting. Popular optimization techniques employed here include the Alternating Least Square and Stochastic Gradient Descent algorithms. The regularization parameter γ and choice of L are often selected by cross-validation. The dimension-reduction nature of the technique reveals its close relationship to singular value decomposition and principal component analysis.

Rating is a form of explicit expression of preference. Explicit feedback has the advantage of easy interpretation - it can reflect both positive and negative preference. However, there are also limitations. First, explicit feedbacks can be biased, some users tend to give high ratings to most items while others are stricter. Second, explicit feedbacks are more difficult to collect. In order to tackle these challenges, implicit feedback, such as clicking data, can be utilized. Clicking data arises naturally when users are interacting with an interface, therefore, is easy to collect and exists in abundance. The biggest challenge in handling click data is the interpretation of negative feedback and missing values.

Hu et al. (2008) introduced a CF model that specifically handles implicit datasets. In this model, the user-item interaction is represented as r_{ji}. For example, r_{ji} can be the number of times a user has clicked on an item, or a measure of the extent to which a user has interacted with an item. In the case of movie streaming, r_{ji} can say whether the user had seen the whole movie or only a part. Based on r_{ji}, two sets of variables are defined. The first is binary and represents user preference x_{ij}=$\mathbb{1}_{\{r_{ij}>0\}}$. When $x_{ij} = 0$, it indicates that user j has not expressed preference on item i, and should not be regarded as a negative feedback. A second set of variables, called "confidence" variables c_{ji}, describe the strength of the preference. Different implicit feedbacks give the recommender various levels of confidence: for example, if a user finishes watching a movie, the level of positive preference is stronger than if the movie was only seen in part. In general, the level of confidence c_{ji} has a positive correlation with r_{ji}, and c_{ji} often is modeled as a linear function on r_{ji}, namely as $c_{ji} = 1 + \alpha r_{ji}$.

As the CF model for explicit feedback, the goal of this model for implicit data is to obtain the latent user- and item-factors, by optimizing the following function:

$$\min_{\mathbf{U},\mathbf{V}} \sum_{(j,i)} c_{ji}(x_{ji} - \boldsymbol{u}_i{}^T \boldsymbol{v}_j)^2 + \gamma(||\boldsymbol{u}_i||^2 + ||\boldsymbol{v}_j||^2), \tag{10}$$

and by cross-validating the dimension L, γ and α.

Once $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are computed, then the ranking of all items for user j is obtained by ordering $\boldsymbol{u}_i{}^T\boldsymbol{v}_j$ for $i = 1, ..., n$, and this can be used to perform individual recomendation.

Among the probabilistic approaches for individual preference learning, we here mention two examples: the Bayesian Mallows model (Vitelli et al., 2018), and the Hierarchical Bradley Terry, HBT, model (Crispino & Frigessi, 2018, draft).

The model-based approach to individual preference learning recently proposed by Vitelli et al. (2018) extends the BMM to handle the problem of individual preference learning. In the top-k rankings case the model is specified as follows. Suppose each assessor j has ranked the subset of items $\mathcal{A}_j \subseteq \{A_1, A_2, \ldots, A_n\}$, giving them top ranks from 1 to $k_j = |\mathcal{A}_j|$. The data is then in the form of partial rankings, $\boldsymbol{R}_1, ..., \boldsymbol{R}_N$. The idea is to estimate the full latent rankings of each assessor, denoted by $\tilde{\boldsymbol{R}}_1, ..., \tilde{\boldsymbol{R}}_N$, where for each $\tilde{\boldsymbol{R}}_j$, the top-k_j ranks are fixed and identical to the data. The idea of the model is to define these augmented latent vectors $\tilde{\boldsymbol{R}}_1, ..., \tilde{\boldsymbol{R}}_N$, assuming they are i.i.d. given the Mallows model parameters $\boldsymbol{\rho}$ and α. The likelihood is then

$$P\left(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N | \alpha, \boldsymbol{\rho}\right) = \sum_{\tilde{\boldsymbol{R}}_1 \in \mathcal{S}_1} \cdots \sum_{\tilde{\boldsymbol{R}}_N \in \mathcal{S}_N} P(\tilde{\mathbf{R}}_1, \ldots, \tilde{\boldsymbol{R}}_N | \alpha, \boldsymbol{\rho}), \tag{11}$$

where $\mathcal{S}_j = \{\tilde{\boldsymbol{R}}_j \in \mathcal{P}_n : \tilde{R}_{ij} = R_{ij}, \forall i \leq n_j\},\ j = 1, \ldots, N$, is the set of rankings compatible with the partial information in the data $\boldsymbol{R}_j$. Inference in this model can be performed through an MCMC scheme with a data augmentation (Tanner & Wong, 1987) step to handle the estimation of the individual random rankings. The choice of prior distributions is the same as described in **Section 3.1**. Personalized recommendations can hence be made based on the inferred individual rankings $\tilde{\boldsymbol{R}}_j$.

In the case of transitive pairwise comparisons, the BMM is specified as follows. Suppose each assessor j has performed a number of comparisons $\mathcal{B}_j = \{(A_i \prec_j A_k)\}_{i \neq j \in \{1,..n\}}$, where the symbol $\prec_j$ indicates the pairwise preference ordering of assessor j. The data is then a collection of sets, $\mathcal{B}_1, ..., \mathcal{B}_N$. Similarly to the top-$k$ case, the idea is to estimate the full latent rankings of each assessor, where now the items seen by each assessor are, in general, not fixed to a given rank. Hence, each augmented latent vector $\tilde{\boldsymbol{R}}_j$ is required to agree, in the sense of eq. (1), with the partial information given in $\mathcal{B}_j$. The likelihood is

$$P\left(\mathcal{B}_1, \ldots, \mathcal{B}_N | \alpha, \boldsymbol{\rho}\right) = \sum_{\tilde{\boldsymbol{R}}_1 \in \mathcal{T}_1} \cdots \sum_{\tilde{\boldsymbol{R}}_N \in \mathcal{T}_N} P(\tilde{\mathbf{R}}_1, \ldots, \tilde{\boldsymbol{R}}_N | \alpha, \boldsymbol{\rho}), \tag{12}$$

where, for each j, $\mathcal{T}_j$ is the set of rankings compatible with the pairwise preferences in $\mathcal{B}_j$. This model was extended to non-transitive pairwise comparisons by Crispino et al. (2018).

Note that the BMM can also be applied to implicit data. In this case, the implicit data is first transformed into pairwise comparisons such that each each clicked item is preferred to each of the non-clicked items, and inference can hence be made using the pairwise comparison version of the BMM.

Another model-based approach to personalized preference learning is a recently proposed extension to the BT model, named Hierarchical Bradley Terry, HBT (Crispino & Frigessi, 2018, draft). As in the BT model, suppose that the

preferences expressed by an assessor are pair comparisons as in (7), but now allowing the score parameters to vary from one assessor to another. In other words, denoting the parameter vector of assessor j by $\boldsymbol{\mu}_j = (\mu_{j1}, ..., \mu_{jn})$, the probability that assessor j prefers item A_i to item A_k has the BT form

$$P(A_i \prec_j A_k | \mu_{ji}, \mu_{jk}) = \frac{\mu_{ji}}{\mu_{ji} + \mu_{jk}} \cdot \tag{13}$$

The pairwise comparisons for assessor j depend only on the relative sizes of the individual score parameters of the items being compared. We denote by $\boldsymbol{W}_j$ the $n \times n$ data matrix, where each element w_{jik} is the number of times item A_i was preferred to A_k by assessor j. Assuming that the individual pairwise comparisons of assessor j are conditionally independent given $\boldsymbol{\mu}_j$, the likelihood of $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_N$ is

$$P(\boldsymbol{W}_1, ..., \boldsymbol{W}_N | \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_N) = \prod_{j=1}^{N} \prod_{1 \leq i \neq k \leq n} \left(\frac{\mu_{ji}}{\mu_{ji} + \mu_{jk}} \right)^{w_{jik}} \cdot \tag{14}$$

The assessor specific parameters $\boldsymbol{\mu}_j$ are assumed a priori independent and distributed according to a gamma distribution with a common mean vector $\boldsymbol{\mu}$; for $\boldsymbol{\mu}$ an inverse gamma prior is assumed. Inference on this model's parameters is performed with an extension of the Gibbs sampler of Caron & Doucet (2012).

Personalized recommendation in the HBT model is then performed by means of the estimated individual latent score parameters $\boldsymbol{\mu}_j$. The individual full ranking of the items, which corresponds to the $\tilde{\boldsymbol{R}}_j$ rankings of eq. (11), can be obtained by the *rank* transformation rank($\boldsymbol{\mu}_j$).

Comparing the three approaches, HBT, BMM and CF, we observe that each has a latent individual vector, namely $\boldsymbol{\mu}_j$, $\tilde{\boldsymbol{R}}_j$ and the vector $(\boldsymbol{u}_1{}^T \boldsymbol{v}_j, ..., \boldsymbol{u}_n^T \boldsymbol{v}_j)$, respectively. The first and third are vectors with continuous real components; the second is a permutation of n integers. The product form of the parametrization of CF is much more sparse compared to the one of HBT, because $L << n, N$. This gives computational advantages, but it also results in less flexibility.

4 Preference learning in practice

In this section, we illustrate and assess the performance of the methods mentioned in **Section 3** with a simple dataset. We purchased a bag of $n = 20$ potatoes (items), which were placed on a table, randomly labeled by letters A,B,C,D..., T. $N = 12$ assessors were asked to visually inspect the potatoes, and to rank them according to their weight, with rank 1 assigned to the heaviest and 20 to the lightest potato. There was no communication between assessors. After the experiment, each potato was weighed, thus obtaining the true ranking of the potatoes, denoted by $\boldsymbol{\rho}_{true}$. This dataset therefore contains $N = 12$ full rankings of $n = 20$ items.The data will be referred to as $\boldsymbol{R}^N = \{\boldsymbol{R}_1, ..., \boldsymbol{R}_N\}$, where each vector $\boldsymbol{R}_j \in \mathcal{P}_n$ is the full ranking provided by assessor j. The purpose is not to compare the precision of the different methods, for which purpose a systematic study would be needed, but to illustrate the methods via a simple real data set. We aim at showing how the methods can be used, which choices must be made in order to perform the analyses, what results can be obtained and how they can be visualised. There is an interesting literature on how to compare methods that estimate rankings (see Gunawardana & Shani, 2015).

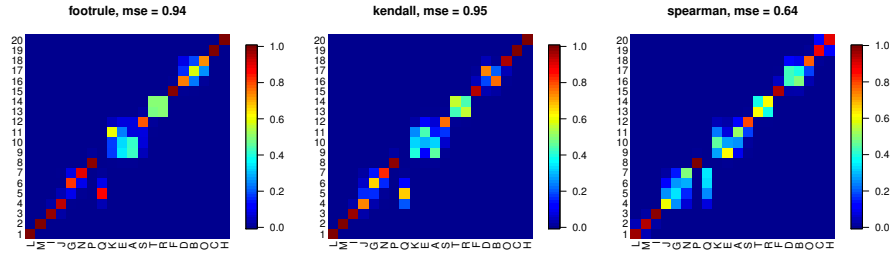

Figure 1: Consensus estimation of the potato experiment with full ranking data produced by the Bayesian Mallows Model. Posterior probability of each item to be ranked as the k-th heaviest potato (y-axis) versus the true ranking of the item (x-axis). Three distances. The posterior expected mse is reported on top of each panel.

4.1 Estimating the consensus ranking

4.1.1 Inference from full rankings

In this section, we illustrate the performance of the BMM and the PL model as rank aggregation methods, when a full ranking of all items is given by each assessor. In the case of the BMM we use three different distance measures: the footrule, Spearman and Kendall distances. The exponential prior hyper-parameter λ_α was chosen to be $\lambda_\alpha = 1/10$ for the footrule and Kendall distances; and $\lambda_\alpha = n/20$ for the Spearman distance, following the recommendations in Vitelli et al. (2018). As $n = 20$, the partition function is computed exactly. We used the authors' R package (soon available on CRAN). For the analysis of the PL model, we exploit the `PLMIX` R package of Mollica & Tardella (2016b), setting the number of mixture components to 1, and with the default choices for the prior hyper-parameters in the gamma priors.

The heatplots in **Figure 1** represent the marginal posterior distribution of $\boldsymbol{\rho}$ obtained using the BMM. The x-axis represents the 20 potatoes, re-arranged according to their true weights in descending order. The y-axis lists the ranks and each column represents the posterior distribution of the corresponding potato, i.e. $P(\rho_i = k \mid \boldsymbol{R}^N)$, for each potato i. These plots do not compare the estimated consensus rankings with the unknown group consensus ranking, but with the ranking based on the actual weights, i.e., $\boldsymbol{\rho}_{true}$. Comparing the estimated consensus with the known $\boldsymbol{\rho}_{true}$ allows us to evaluate how the assessors performed as a group.

To assess the accuracy of the inferred group consensus $\boldsymbol{\rho}$, incorporating the uncertainty associated with it, we use the posterior expected mean squared error (`mse`), defined as

$$\texttt{mse} = \mathbb{E}\left\{ \frac{1}{n} ||\boldsymbol{\rho} - \boldsymbol{\rho}_{true}||_2^2 \,\middle|\, \boldsymbol{R}_1, \ldots \boldsymbol{R}_N \right\} = \frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{n} (r - \rho_{true,i})^2 P(\rho_i = r \mid \boldsymbol{R}_1, \ldots \boldsymbol{R}_N).$$

From **Figure 1**, it can be observed that the BMM is able to estimate a group consensus $\boldsymbol{\rho}$ that resembles $\boldsymbol{\rho}_{true}$ well. Consistently in all three distances, BMM produces a consensus $\boldsymbol{\rho}$ with high accuracy for items L, M and I, the

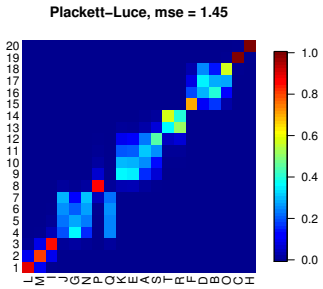

Figure 2: Consensus estimation of the potato experiment produced by the Plackett-Luce model

top-weighed potatoes, as well as items C and H, the lightest potatoes. However, the consensus rankings of the potatoes in the middle-range, i.e. items K,E and A, are not estimated as accurately and certainly. This can be explained by the fact that the true weights of these potatoes are differed only by a few grams, and the assessors were not in agreement with each other regarding their rankings. We also notice that potato Q was ranked higher than its actual weight implied.

The choice of distances also affects the ranking aggregation performance of BMM. Shown in **Figure 1**, the footrule and Kendall distances produce very similar posterior distributions. This is consistent with Diaconis & Graham (1977), where it is proved that $\forall \boldsymbol{r} \in \mathcal{P}_n,\ d_K(\boldsymbol{r}, \mathbf{1}) \leq d_F(\boldsymbol{r}, \mathbf{1}) \leq 2d_K(\boldsymbol{r}, \mathbf{1})$, where $\mathbf{1} = (1, ..., n)$. The Spearman distance produces a flatter marginal posterior distribution for each item. Overall, the Spearman distance has a slight edge in terms of `mse` in this case, and this is largely attributed to item Q, for which both the footrule and Kendall distances produce wrong estimations with a high degree of certainty.

Next we consider the same data and estimation problem, but apply the PL model. Since this method produces estimates of $\boldsymbol{\mu}$, we applied the rank transformation to obtain a ranking. In the case of the PL model, the marginal posterior distribution produced for each item is flatter and less certain compared to the BMM, as shown in **Figure 2**. Only for items C and H, the two bottom-ranked items, the PL model produces estimations with high certainty. All four estimations of the consensus indicate that the twelve assessors did a rather good job in evaluating the weight of our potatoes. Three medium-weight potatoes were difficult to rank: the estimated consensus obtained by the PL model and the BMM with the Kendall distance is rather uniform among a group of ranks. The BMM with the footrule and Spearman distances take a clearer decision. One potato (Q) was missplaced by most assessors: the BMM with the footrule and Kendall distances estimate its consensus rank most likely as fifth, which implies that the next potato (G) has to be sixth. For the PL model, the pattern is more uniform, because an estimate of μ_Q does not force particularly μ_G. We here see the difference between a continuous real-valued parametrization and one with permutations of n integers. We also see that the choice of the distance in the BMM matters.

4.1.2 Inference from top-k rankings

In this section, we illustrate the rank aggregation performance of the BMM and the PL model when each assessor j only ranks her top k_j items, $k_j < n$, exploiting a modified version of the potato dataset. We sample k_j from a Poisson distribution truncated in the interval $[1, 17]$, and with the expected number of ranked items set to 7. After the k_j-th item, further items are considered as missing values. Ten experiments are conducted using 10 different datasets. The parameters of both BMM and PL are chosen as in **Section 4.1.1**, and only the BMM with the footrule distance is discussed in this section.

The two left panels in **Figure 3** show the heatplots corresponding to the first experiment run, summarizing the marginal posterior distribution for each item in the consensus $\boldsymbol{\rho}$, using the BMM (left) and the PL model (middle). Both models, compared to the other items, produce clearer consensus estimations for the top-ranked items, i.e., items L, M and I, as there is sufficient and non-noisy information about these items. Both models also correctly produce uniform distributions for the 6 bottom-ranked items, since in this experiment run, no assessor included these 6 items in her top-k_j ranked items. As a consequence there is no preference information regarding these items in the data (see also Vitelli et al., 2018).

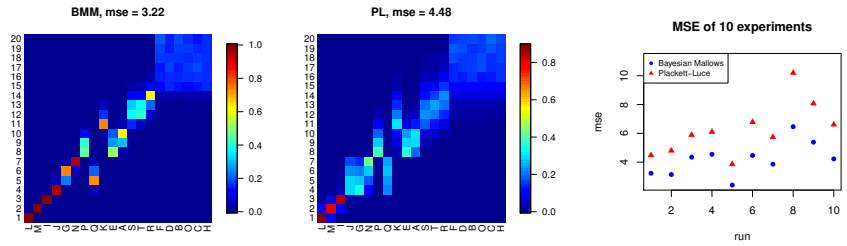

Figure 3: Left: Consensus estimation of the top-k potato experiment run number 1, using the BMM with footrule distance. Middle: Consensus estimation of the top-k potato experiment using the PL, run number 1. Right: scatter plot of the `mse` of the 10 experiment runs

Consistently with the description in **Section 4.1.1**, the PL model produces a consensus estimation with more uncertainty compared to the BMM, and the BMM has a tendency to produce a more certain estimation on some items. Compared to the actual weight of the potato, BMM is making more mistakes for the middle-ranked potatoes. However, overall, as shown in the right panel of **Figure 3**, the BMM produces consensus estimation with lower `mse` (with respect to the true weight) compared to the PL model, consistently throughout all 10 runs of the experiment.

4.1.3 Inference form pairwise comparisons

In this section, we illustrate the rank aggregation performance of the BMM and the BT models when data is in the form of pairwise comparisons (RPP). We exploit the MCMC algorithm of Vitelli et al. (2018) for making inference on the BMM parameters, and the Gibbs sampling of Caron & Doucet (2012) for

making inference on the BT parameter.

We assume that each assessor j provides a fraction $n_j \leq n(n-1)/2$ of the possible pairwise preferences, using the potato dataset. Here, $n_j = \bar{n}$, $\forall\, j$, and $\bar{n}$ is chosen deterministically, as we wish to control the number of pair comparisons made by each assessor. However the pairs that are actually compared are chosen uniformly at random among all pairs, without replacement. For each assessor j, we sample randomly $\bar{n} \leq 190$ different pairs of items $\{A_i, A_k\}$ (i.e., the potatoes) from the collection of all possible pairs. We then assign the preference between the tested pair of potatoes according to the individual ranking $\boldsymbol{R}_j$, available in the full potato data. We are left with the data $\boldsymbol{D}^N = \{\boldsymbol{D}_1, ..., \boldsymbol{D}_N\}$, where each $\boldsymbol{D}_j$, $j = 1, ..., N$, is a collection of pairwise preferences in the form $(A_i \prec_j A_k)$, where, as in equation (1), $(A_i \prec_j A_k) \iff R_{ji} < R_{jk},\ i, k = 1, ..., n, \quad i \neq k$. The notation $(A_i \prec_j A_k)$ means that item A_i is preferred to item A_k by assessor j.

We here investigate how the accuracy of the BMM and the BT is affected by data incompleteness. As such, we create 10 nested datasets, each containing a different number of pairs: from 10% of all available pairs, to 100% of available pairs (DPP). The datasets are nested in the sense that when we increase the fraction of pairs given to each assessor, we only add new pairs, hence ensuring that any changes in prediction accuracy is solely due to the increase of data availability.

As in the previous section, we conducted 10 experiments for each scenario to take care of the randomness involved in pairs selection. The BMM parameters are chosen as in **Section 4.1.1**, while for the BT we follow the specification of Caron & Doucet (2012). **Figure 4** shows the 95% confidence interval for the expected posterior mse for the different scenarios of data incompleteness, and for the two methods: BMM (black dots), and BT (red dots). The posterior sample of BT parameters $\boldsymbol{\mu}$ was converted into a sample of rankings simply by applying the *rank* transformation, similarly to what we did with the PL parameters earlier.

It can be observed that the `mse` of both methods dramatically decreases when we increase the fraction of available pairs from 0.1 to 0.2. When 20% of pairwise comparisons are collected from each assessor, the BMM can already estimate the consensus ranking with reasonable accuracy, with an `mse` below 2. The BMM does not gain much prediction accuracy after we give each assessor more than 50% of all available pairs. The BT is considerably less accurate than the BMM when fewer pairs are given to the assessors (less than 30%), in terms of `mse`, but the two methods perform similarly as the number of pairs increases.

4.2 Individual preference learning

Here we consider the situation where assessors have expressed incomplete preferences and we are interested in *completing* the understanding of their individual prteferences.

4.2.1 Inference from top-k rankings

Consider the situation where the assessors rank only their top preferred items. Often it is of interest to recommend to each assessor her "next" items, in order of preference. For instance, given that an assessor j has ranked her top k_j items,

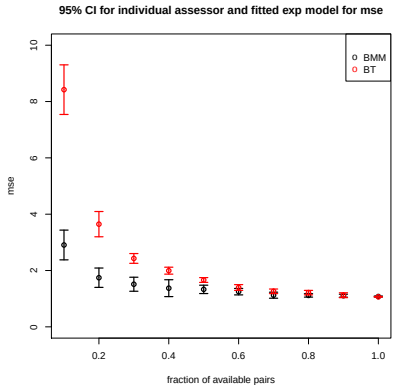

Figure 4: Pairwise experiment with different fractions of pairs compared by the 12 assessors, from 10% to 100% of the total number of possible pairs (190). Expected postrior `mse` of the consensus with 95% confidence interval based on ten independent runs. Red dots: BT; Black dots: BMM. x-axis: fractions of pairs; y-axis: `mse`.

Table 1: Next-$m = 1, 2$ items estimation accuracy based on top-k potato data, exp. correct refers to assignment recommendations at random

scenario	mean correct	95% CI	exp. correct	improvement
Next 1 item	0.450	0.37 - 0.53	0.077	485% ±106%
Next 2 items	1.325	1.230 - 1.421	0.333	298%± 29%

it may be interesting to estimate which items are most likely to be ranked by the same assessor j as the $(k_j + 1)$-th and $(k_j + 2)$-th, which we call the "next two" items. Among the methods mentioned in **Section 3**, BMM produces complete augmented ranking vectors $\tilde{\boldsymbol{R}}_j$ for each individual j. We demonstrate how BMM makes inference on the "next m" (with $m = 1$ or 2) individual preferences, using the top-k potato datasets (with 10 repeated experiments) explained in **Section 4.1.2**, where the expected number of ranked items per assessor is 7. Notice that the potatoes are difficult to order in that range. The "next m" items recommended to each assessor j are those associated to the highest posterior probability of being her most preferred, among the items that assessor j did not rank. Such posterior probabilities are computed on the basis of the individual augmented latent ranking $\tilde{R}_j$ as follows: for each assessor j, and each item i, we compute $P_{ij} = P(k_j + 1 \leq \tilde{R}_{ij} \leq k_j + m|\boldsymbol{R}_1, ..., \boldsymbol{R}_N)$. Then, for each assessor j, the m items associated with the highest posterior probabilities, P_{ij}, are recommended to the assessor. Obviously, the items that are already ranked by an assessor are not considered for recommendations. After estimating the next m items for each assessor, we can compare the prediction with the original full potato dataset.

Table 1 presents the average number of correctly predicted next 1 and 2 items for the 12 assessors, with the 95% confidence intervals over 10 experiments. We compare the performance with the reference of choosing the recommended items at random among the non-ranked items. The expected value of this

Table 2: Summary of the posterior probabilities of being next-m items, based on top-k potato data, $m = 1, 2$.

P_{ij}	min	25%	median	***mean***	75%	max
Next 1 item	0.172	0.418	0.489	**0.484**	0.559	0.642
Next 2 items	0.210	0.576	0.673	**0.651**	0.755	0.837

number is shown in the second to the last column and is calculated as

$$\mathbb{E}[\text{number of correct random guesses}] = \sum_{i=1}^{m} i \cdot \frac{\binom{m}{i}\binom{n-\lambda-m}{m-1}}{\binom{n-\lambda}{m}},$$

where λ is the parameter of the truncated Poisson. BMM is able to correctly predict roughly 45% of the next 1 item, and 66% out of the next 2 items. The BMM makes a very large improvement for both next 1 and next 2 items compared to naive guess.

The Bayesian approach gives an estimate of the uncertainty of the prediction of the "next" items, namely the posterior probability P_{ij} for the item i to be among the "next m" items for assessor j. **Table 2** shows the summary of these probabilities across all assessors and all recommended items. The mean of the estimated P_{ij} for $m = 1$ is 0.484. For some assessors, the probabilities are so low that the item should probably not be recommended. On the other hand, for $m = 1$, for more than 25% of the assessors the probability is above 0.5. All probabilities are larger for $m = 2$, because the item can be either of the "next two": the median is 0.673. Comparing the estimated mean of P_{ij}, which is 0.484 for $m = 1$, with the average number of correctly predicted next 1 item shown in **Table 1** (45%), and similarly for $m = 2$, we see that the BMM predictions are well calibrated, and the produced uncertainty is reliable. The results are rather good, given the diffiuclty of the potato data in the middle range.

4.2.2 Inference from pairwise comparisons

When the assessors perform pairwise comparisons, it is relevant to predict individual preferences that are not in the data, usually the "top" preferred items for an assessor. We here use the same datasets used in **Section 4.1.3** to illustrate the performance of two methods: the BMM and the HBT. Both methods produce full augmented vectors, the latent ranking $\tilde{\boldsymbol{R}}_j$ (BMM) and the individual vector of score parameters $\boldsymbol{\mu}_j$ (HBT), one for each assessor j, which are then used to recommend items to the assessors.

After estimating the $\tilde{\boldsymbol{R}}_j$ and $\boldsymbol{\mu}_j$, we can refer to the original $\mathbf{R}^N$ data matrix, to check whether they are correctly estimated. **Figure 5** shows the average posterior expected `mse` of the individual rankings of all assessors against the proportion of compared pairs in the data, for the two methods: BMM (black dots), and HBT (red dots). Since we know the true ranking of the assessors, in this case the comparison of this with the estimate is meaningful. It is clear that in terms of `mse` the BMM outperforms the HBT. As we mentioned before, this may be due to the fact that the BMM augments the partial information with full rankings in the space of n-dimensional permutations. For this reason, the number of possible rankings that are compatible with the partial information in the data is limited. The HBT on the other hand, augments the partial

information in the data with the real-valued vectors $\boldsymbol{\mu}_j$, $j = 1, ..., N$, which are much more variable.

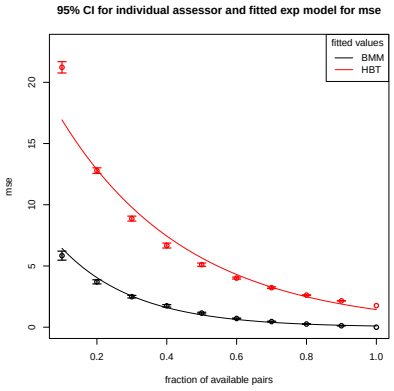

Figure 5: Average `mse` and 95% confidence interval for the pairwise experiment with different fractions of pairs available in the data, 10 independent repetitions. Red dots: HBT; Black dots: BMM. The lines are fitted from an exponential model.

A similar effect can be observed from **Figure 6**, which presents the heatplots of the posterior distribution for each item plotted against the true ranking for assessor 1, when assessor 1 is given 20%, 60% and 100% of all available pairs (run 1). The three lower plots are obtained with the BMM, the three upper ones with the HBT. Notice the special case where all pairwise comparisons are given to the assessors (right panels): the BMM heatplot (bottom-right) does not show any variability. This is due to the fact that there is only one ranking vector $\tilde{\boldsymbol{R}}_j$ which is consistent with the given $n(n-1)/2$ possible pairs. On the other hand, the plot obtained with HBT in the same situation (top-right) is very accurate, but still shows some variability around the correct ranking. We clearly see that BMM has more peaked marginal posteriors, and that they often peak around the diagonal, i.e. the correct ranking.

Let us finally look at the estimates of the assessors' top-3 items using the two methods, shown in **Figure 7**. For run 1, when the proportion of pairs assessed by each individual assessor is 50%, **Figure 7** was obtained as follows: in the separate column on the left, we display the posterior probability $P(\rho_{A_i} \leq 3 \mid \boldsymbol{D}^N)$ that a given potato $A_i, i = 1, ..., 20$, is among the top-3 in the estimated consensus $\boldsymbol{\rho}$. In the other columns we show, for each potato A_i, the individual posterior probabilities $\tilde{P}_{ij} = P(\tilde{R}_{j,A_i} \leq 3 \mid \boldsymbol{D}^N)$ of being among the top-3 for assessor j, $j = 1, ..., 12$. The left panel is obtained with HBT, the right one with BMM. These probabilities are used to assign the top-3 potatoes to each assessor individually. We see again how more marked BMM's posterior probabilities are, though individual posterior rankings are allowed to depart from the consensus.

Figure 8 shows the prediction outcome for the top three items using the BMM and HBT, averaged among all assessors and runs, plotted against the fraction of pairs evaluated by each assessor. For both BMM and HBT, we observe an expected increase of prediction accuracy as the number of assessed pairs increases. BMM performs well in predicting the top-3 items for each

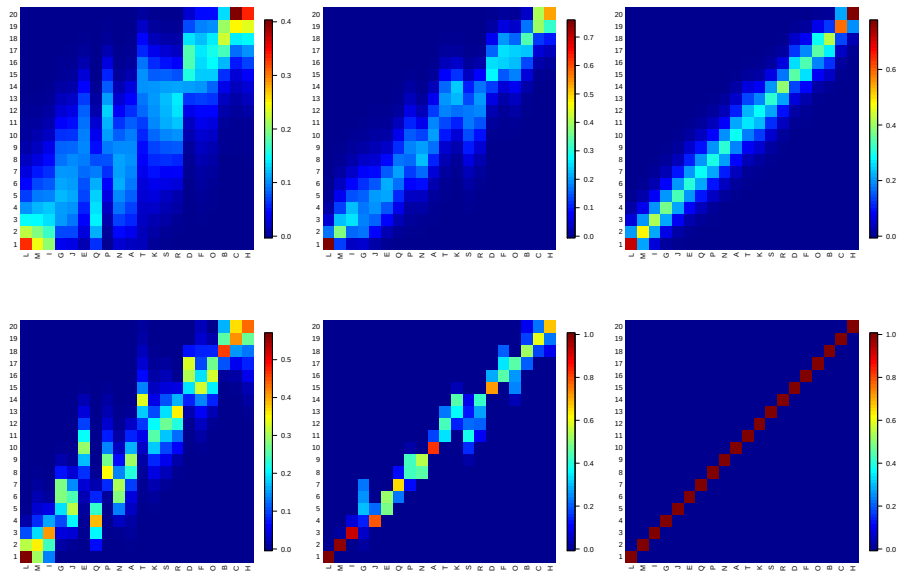

Figure 6: Heatplots of the posterior distribution for each item for assessor 1, when assessor 1 is given (left panels) 20%, (middle panels) 60% and (right panels) 100% of all available pairs (in run 1). The three lower plots are obtained with the BMM, the three upper ones with the HBT. Note that the ranges of the color scales in the different panels are different.

assessor: already when only 10% of the pairs are compared, on average 2.5 items are correct, and the average number of correct items increases then slowly. One could conclude that in this data set it is probably not worth to extend the learning task very much. For HBT the situation is different. The same curve starts around 1.7 and reaches 2.5 items approximately when 50% of the pairs are compared in the data. Here, asking the assessors to express more preferences appears to be important, result in line with the strong connection condition introduced in **Section 3.1**. We can see that the BMM has an edge in terms of accuracy for the potato data, we believe also due to the discrete parametrization in $\mathcal{P}_n$.

4.2.3 Top-k rankings with implicit data

The most common and abundant type of preference data is in an implicit form. In order to use clicking data to make inference on individual preferences, an underlying assumption is needed, namely that the items that are clicked by assessor j are preferred, compared to the non-cliked ones. This assumption naturally forms pairwise comparisons. Denote the set of items that are clicked by assessor j as $\mathcal{C}_j$, and the non-clicked set as $\mathcal{C}_j^c$. Then $A_i \prec_j A_k$ if $A_i \in \mathcal{C}_j$ and $A_k \in \mathcal{C}_j^c$. We can therefore use also in this case the same methods as in the case of pair comparison data. The BMM and the CF methods for implicit datasets (Hu et al., 2008) will be considered. For each assessor j, the aim is to infer on the next $m = 2$ items (contained in $\mathcal{C}_j^c$), which are ranked next to the

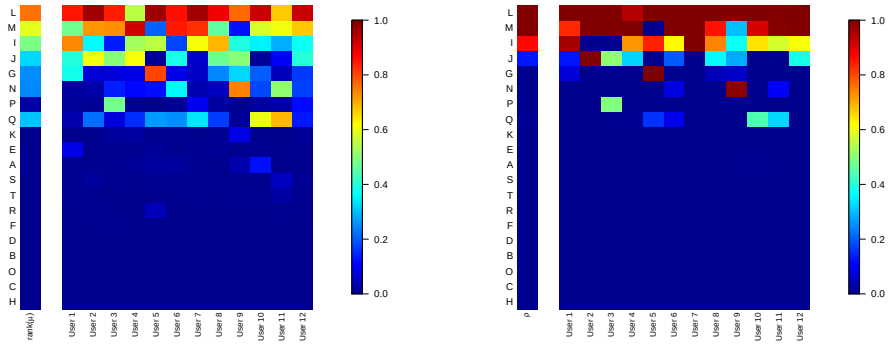

Figure 7: Posterior probability, for each potato, of being ranked among the top-3 in $\boldsymbol{\rho}$ (column 1), and in $\tilde{\boldsymbol{R}}_j, j = 1, ..., 12$ (following columns). Left: results with HBT; Right: results with BMM.

ones in $\mathcal{C}_j$.

We continue with the potato dataset, which we convert into the implicit form as follows. For each assessor j, we sample the number k_j of items that she has clicked, using $k_j \sim$Truncated Poisson$_{[1,n]}(\lambda)$. We let λ vary to increase the amount of data. Then we look to the potato data and define the implicit interaction between assessor j and item i as

$$I_{ij} = \begin{cases} 1 & \text{if } R_{ij} \leq k_j \\ 0 & \text{otherwise.} \end{cases}$$

The BMM first converts the implicit data into all implied pairwise comparisons, and then uses the pairwise comparison version of the method, as in **Section 4.2.2**, to obtain the augmented ranking vectors $\tilde{\boldsymbol{R}}_j$ for each assessor. The CF method follows the matrix factorization approach, and handles the implicit dataset more directly. Note that in this experiment, r_{ji} and x_{ji} as described above in the CF model take the same value, since the implicit dataset is in a binary form. We consider data with $\lambda = 3, 4, ...12$, and record the average number of correctly predicted next 2 items across all assessors. Ten independent implicit datasets are generated for each λ to produce 95% confidence intervals.

Figure 9 shows the prediction outcome using the BMM and CF. Both models perform well in predicting the next $m = 2$ items for each individual assessor: on average more than 50% of the items are correctly predicted. The BMM appears to have a slight edge in terms of accuracy, with narrower 95% confidence interval, because of its better performance on the more difficult of the ten datasets. Quite surprisingly, for both BMM and CF, we do not observe an increasing prediction accuracy as λ increases. This can be explained by the strong consensus shared by the assessors in the potato experiment, as described in **Section 4.1.1**. Except from the middle-ranked items, most assessors tend to agree on the rankings of the top and bottom-ranked items. To make inference on each assessor's next $m = 2$ items, the models borrow information about the non-clicked items, $A_j \in \mathcal{C}_j^c$. In this example however, as λ increases, due to the

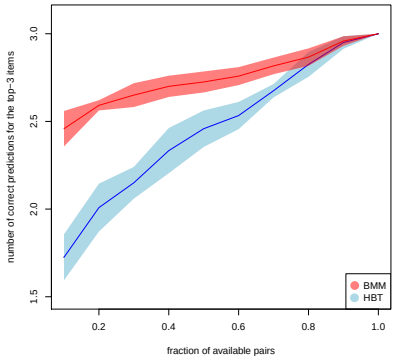

Figure 8: Average number of correctly predicted top-3 potatoes across all assessors and runs, and 95% confidence interval (y-axis) plotted against fraction of pairs assessed (x-axis) in the potato data set. Blue: prediction outcome using the HBT. Red: prediction outcome using the BMM approach.

strong consensus shared between assessors, this information does not increase much.

5 Non-homogeneous assessors

In real situations, assessors often do not all share the same preference on a set of items. The methods which are built around a common consensus can be extended to handle this more complex situation. We demonstrate an example of inference from top-k rankings when the assessors in the sample do not share one single consensus, but multiple group consensuses exist. Clustering the assessors can be handled by many of the methods presented in **Section 3**. In this section, we illustrate the BMM, extended to finite mixtures in Vitelli et al (2017), and the PL model, which was generalized to infinite non-parametric mixture by Caron et al (2014), and to parametric mixtures by Mollica and Tardella (2016). In this section, we also use a larger simulated data set, in order to discuss briefly scaling issues, for large N and n.

We sampled $N = 2000$ assessors from the Mallows distribution using the simulator described in Vitelli et al. (2018), each ranking $n = 100$ items. We assume that the assessors belong to $C = 3$ distinct clusters, characterised by three true cluster consensuses: $\{n, n-1, ...1\}$, $\{1, 2, ...n\}$,and $\{n/2, n/2-1, ...1, n, n-1, ..., n/2+1\}$. The central cluster is twice as large as the other two, as we use proportions $\{0.25, 0.5, 0.25\}$, respectively. The three clusters share the same $\alpha = 3$. Similar to **Section 4.1.2**, the dataset is then converted into top-k_j rankings, where $k_j \sim$ Truncated Poisson$_{[1,97]}(20)$, thus on average, 20% of the items' rankings are reported.

The prior and distance choices for the BMM are the same as in **Section 4.1.2** . The normalizing constant Z_n is approximated by the IPFP method of Mukherjee (2016), and the MCMC is run for 10^5 iterations.

First we specify the correct number of clusters, $C = 3$, and let both the PL

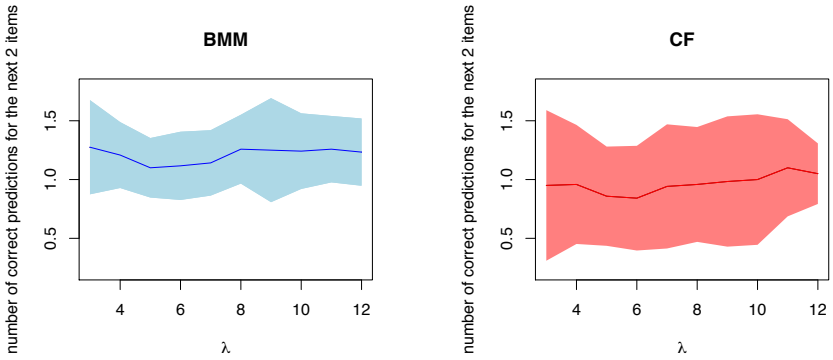

Figure 9: Average number of correctly predicted next 2 items across all assessors and 95% confidence interval (y-axis) plotted against average number of items clicked (x-axis). Left: prediction outcome using the BMM. Right: prediction outcome using the CF approach.

model and the BMM estimate the group consensus rankings. The heatplots of the estimated consensuses for the three clusters and the two methods are shown in **Figure 10**. The hundred items are not ordered in the x-axis according to the true consensus. Careful inspection of the heatplots reveals that the top ranked items are estimated with higher accuracy, while for the other items an almost flat posterior distribution appears. This is as expected, as each assessor provides on average only the 20 top items of the 100 items, and there is little information regarding the other items. Both methods deliver estimated consensuses which resembles to a reasonable extent the true consensuses of each of the three clusters. The `mse` (reported on top of each panel) is significantly smaller for cluster 2 because this cluster has more assessors. The two methods perform rather similarly in terms of `mse`.

The approach to estimating the clusters when applying the PL model is different from when using the BMM. The PL model is a mixture model with the form $p(\boldsymbol{R}_j|\mu_1, ..., \mu_K) = \sum_{k=1}^{K} w_k\, p(\boldsymbol{R}_j|\mu_k)$, where the weights $\{w_1, ...,w_C\}$ are parameters to be estimated. The BMM, on the other hand, estimates the auxiliary variables, $z_j \in \{1, ..., C\}$ for each assessor j, which indicates which cluster the observation $\boldsymbol{R}_j$ belongs to. We refer to the auxiliary variables as cluster assignments. **Table 3** shows the cluster weights estimated by the PL model and BMM. In the case of the PL model, the posterior means of the weight parameters are shown; for the BMM, the mode of the posterior marginal density of the z_j's is used to determine cluster assignments. Both methods perform well. The left panel of **Figure 11** shows for each assessor the posterior probabilities of being assigned to each of the three clusters, for BMM. It is clear that the cluster assignment is more certain for users that belong to cluster 1. This is due to the fact that cluster 1 is more distinct from the other 2 clusters - the footrule distances between $\boldsymbol{\rho}_1$ and the other two clusters are much larger than the footrule distance between $\boldsymbol{\rho}_2$ and $\boldsymbol{\rho}_3$.

In this experiment, the number of clusters $C = 3$ was known. As this is usually is not the case, the number of clusters needs be estimated. While classical

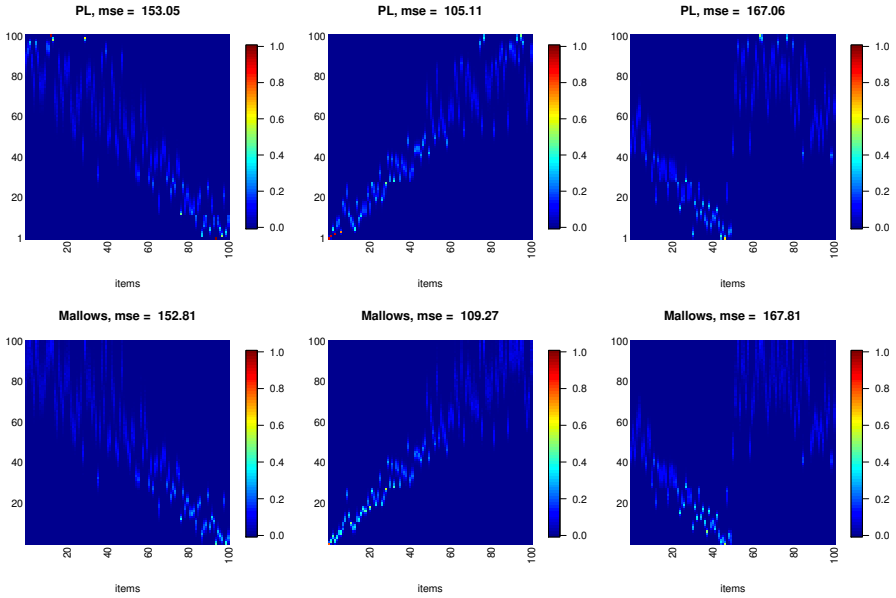

Figure 10: Heatplots of the posterior distributions of 3 clusters group consensuses. Top: the PL model; Bottom: BMM

Table 3: cluster weights estimation of PL and BMM

method	cluster 1	cluster 2	cluster 3
PL	0.250	0.495	0.255
BMM	0.252	0.492	0.257

methods to estimate the number of components in a mixture are useful, and some will lead to a Reversible jump MCMC, here we simply run the BMM with an increasing number of clusters and use a within cluster distance to determine the most likely number of clusters. In the right panel of **Figure 11** we plot the within cluster footrule distance, $\sum_{c=1}^{C}\sum_{j:z_j=c} d(\tilde{\boldsymbol{R}}_j, \boldsymbol{\rho}_c)$ for $C = 2, 3, ..., 7$ clusters. The boxplots indicate correctly that $C = 3$ should be preferred.

The computing time for this simulation using the BMM is dependent on C. When using one core of the Intel Xeon e-8890 processor, running at 2.5 GHz, the computation with $C = 2$ took approximately 19.95 minutes and approximately an addition of 4.3 minutes to compute for each extra cluster. The time complexity is linear in terms of N. The increase of n does not greatly affect computing time per iteration, but a longer chain is needed to reach convergence.

6 Discussion and further developments

In this paper we have reviewed the most important model-based approaches to estimate a consensus ranking of items or the individual rankings of the assessors. Model-based here stands for a statistical model which represents the stochastic process generating the data. In some cases, the decisions made by

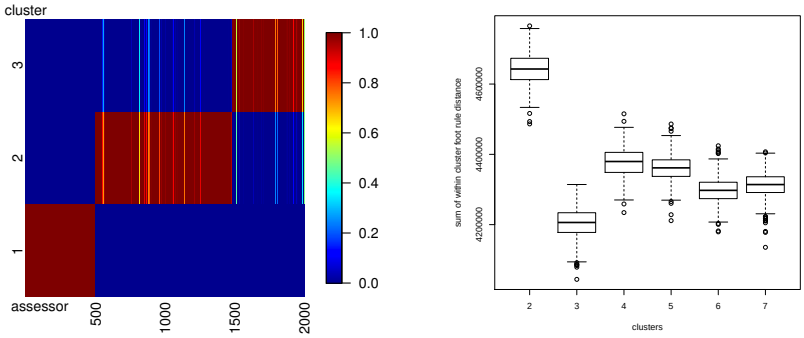

Figure 11: Left: Posterior distribution of cluster assignment for each assessor using the BMM when the number of clusters is $C = 3$. Right: Boxplots of the sum of within cluster footrule distances for different choices of the number of clusters, using the BMM.

the assessors are assumed to be sequential, from the most preferred downwards; in other cases pair comparisons are thought to be performed; or a complete ranking is assumed to exist so that expressions of preference correspond to this ranking. The key component is a latent vector of scores, utilities or ranks associated with the items, shared by a group of assessors, representing the consensus preference; or a collection of such latent vectors, one per assessor, representing her personal preference. These latent vectors are assumed to be continuous, real and positive, or a permutation of the n integers. We have seen how these choices can be justified, also on the basis of the form of the available data. Both likelihood based and Bayesian approaches are fruitful. Priors on the latent vectors are usually non-informative or uniform, but informative forms can be used when appropriate. In addition to point estimates, which synthesize posterior distributions or likelihoods, it is important to quantify the uncertainty of the estimated consensus or individual rankings.

There are several themes that we have not discussed in this paper. The potato data set allows to illustrate the use of different methods for different types of data, but is not useful for discussing scaling in the number of assessors or items. MCMC algorithms suffer particularly when N and n are large and the data is very incomplete. The amount of expressed preferences by individual assessors of course matters, starting from completely new assessors (the so called "cold start" situation). Bandit models (Li et al., 2010) are very useful for learning preferences rapidly, by dynamically suggesting items to each assessor, in order to test, and in this way, learn their preferences. In many practical situations, additional information is available on the assessors and the items. Such features or covariates can be very useful for recognizing similarities across assessors and items. Recommender systems using covariates related to the interaction between user and item are called context-aware, see Adomavicius & Tuzhilin (2015). Extracting features that give good descriptions of items or users (for example books or movies, or individual CV's) is difficult and deep learning has been proposed (Karatzoglou & Hidasi, 2017). Preferences change

in time and models can account for time dependence: in Asfaw et al. (2017) changes are assumed to be smooth in time. Crispino et al. (2018) consider pair comparison data, where assessors are not consistent with themselves, and non-transitivities appear in the data. Sometimes, preference expressions of various types have been collected from users, for example ratings, rankings and pair comparisons. Mallows models can in principle be extended in this direction, because of the underlying latent structure, conditioned on which likelihood terms can be assumed independent. Regarding Mallows models, it is difficult to recommend one distance with respect to another on the basis of the actual application. Finally, a well studied, difficult question is concerned with the evaluation of preference learning algorithms, see for example Gunawardana & Shani (2015). This includes also running designed experiments on online systems.

In a world where data and digital information is massively available, users like us would be overloaded, incapable of sorting and finding the needed pieces of knowledge, if not assisted by algorithms. These algorithms have to understand our goals, intentions and preferences: the underlying inferential task is evident. Algorithms must learn from our behavior, and therefore, traces of our actions, roles, conducts and expressions need to be collected. Who should have access to such data, who should own it, how to consent (or not) to its use, and use for what purpose: these are all fundamental issues that question privacy, freedom, democracy, and trust. Mallows, Plackett-Luce, Bradley-Terry, collaborative filtering and all the other preference learning methods contribute to making the information based society possible, but they do not guarantee a fair, legal, good, social use of private data. Science and scientists must engage and keep alert.

Acknowledgement

Special thanks to Elja Arjas for important discussions, for having conceived the potato experiment and for having bought the potatoes. Øystein Sørensen's MCMC codes have been used in this paper.

References

Adomavicius G, Tuzhilin A. 2015. Context-aware recommender systems. In *Recommender systems handbook*. Springer, 191–226

Agarwal DK, Chen BC. 2016. Statistical methods for recommender systems. Cambridge University Press

Aggarwal CC. 2016. Recommender systems. Springer

Agresti A. 1996. Categorical data analysis. New York: John Wiley & Sons

Alvo M, Yu PLH. 2014. Statistical methods for ranking data. Frontiers in Probability and the Statistical Sciences. New York, NY, USA: Springer

Asfaw D, Vitelli V, Sørensen Ø, Arjas E, Frigessi A. 2017. Time-varying rankings with the Bayesian Mallows model. *Stat* 6:14–30

Bradley RA, Terry ME. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39:324–345

Caron F, Doucet A. 2012. Efficient Bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics* 21:174–196

Cattelan M. 2012. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science* :412–433

Crispino M, Arjas E, Vitelli V, Barrett N, Frigessi A. 2018. A Bayesian Mallows approach to non-transitive pair comparison data: how human are sounds? *Forthcoming in the Annals of Applied Statistics*

Crispino M, Frigessi A. 2018. The Hierarchical Bradley-Terry model. *draft*

Davidson RR. 1970. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association* 65:317–328

Desarkar MS, Sarkar S, Mitra P. 2016. Preference relations based unsupervised rank aggregation for metasearch. *Expert Systems with Applications* 49:86–98

Diaconis P. 1988. Group representations in probability and statistics, vol. 11 of *Lecture Notes - Monograph Series*. Hayward, CA, USA: Institute of Mathematical Statistics

Diaconis P, Graham RL. 1977. Spearman's Footrule as a Measure of Disarray. *Journal of the Royal Statistical Society B* 39:262–268

Fang H, Wang Y, Jin Q, Ma J. 2017. RankwithTA: A robust and accurate peer grading mechanism for MOOCs, In *Teaching, Assessment, and Learning for Engineering (TALE), 2017 IEEE 6th International Conference on*. IEEE

Ford LR. 1957. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly* 64:28–33

Fürnkranz J, Hüllermeier E. 2010. Preference learning: An introduction. Springer

Glickman ME. 1999. Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics* :377–394

Gunawardana A, Shani G. 2015. Evaluating recommender systems. In *Recommender Systems Handbook*. Springer, 265–308

Hu Y, Koren Y, Volinsky C. 2008. Collaborative filtering for implicit feedback datasets, In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08. Washington, DC, USA: IEEE Computer Society

Hunter DR. 2004. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics* 32:384–406

Irurozki E, Calvo B, Lozano A. 2014. Sampling and learning the Mallows and generalized Mallows models under the Hamming distance. *Bernoulli (submitted)*

Irurozki E, Calvo B, Lozano JA. 2018. Sampling and learning mallows and generalized mallows models under the cayley distance. *Methodology and Computing in Applied Probability* 20:1–35

Jacques J, Biernacki C. 2014. Model-based clustering for multivariate partial ranking data. *Journal of StatistiAli2012cal Planning and Inference* 149:201–217

Karatzoglou A, Hidasi B. 2017. Deep learning for recommender systems, In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM

Kemeny JG, Snell JL. 1962. Mathematical models in the social sciences. Blaisdell Publishing Company

Koren Y, Bell R, Volinsky C. 2009. Matrix factorization techniques for recommender systems. *Computer* 42

Lansdowne ZF, Woodward BS. 1996. Applying the borda ranking method. *Air Force Journal of Logistics* 20:27–29

Li L, Chu W, Langford J, Schapire RE. 2010. A contextual-bandit approach to personalized news article recommendation, In *Proceedings of the 19th international conference on World wide web*. ACM

Lu T, Boutilier C. 2014. Effective sampling and learning for Mallows models with pairwise-preference data. *Journal of Machine Learning Research* 15:3783–3829

Luce RD. 1959. Individual choice behavior: A theoretical analysis. New York, NY, USA: Wiley

Mallows CL. 1957. Non-null ranking models. I. *Biometrika* 44:114–130

Marden JI. 1995. Analyzing and Modeling Rank Data, vol. 64 of *Monographs on Statistics and Applied Probability*. Cambridge, MA, USA: Chapman & Hall

Meilă M, 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)*. Corvallis, OR, USA: AUAI Press

Mollica C, Tardella L. 2016a. Bayesian Plackett-Luce mixture models for partially ranked data. *Psychometrika* (published on line)

Mollica C, Tardella L. 2016b. PLMIX: An R package for modeling and clustering partially ranked data. *arXiv preprint arXiv:1612.08141*

Mukherjee S. 2016. Estimation in exponential families on permutations. *The Annals of Statistics* 44:853–875

Plackett RL. 1975. The Analysis of Permutations. *Journal of the Royal Statistical Society C* 24:193–202

Raman K, Joachims T. 2015. Bayesian ordinal peer grading, In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. ACM

Ricci F, Rokach L, Shapira B. 2015. Recommender systems: introduction and challenges. In *Recommender systems handbook*. Springer, 1–34

Tanner M, Wong W. 1987. The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* 82:528–550

Thurstone LL. 1927. A law of comparative judgment. *Psychological review* 34:273

Vitelli V, Sørensen Ø, Crispino M, Frigessi A, Arjas E. 2018. Probabilistic preference learning with the Mallows rank model. *Journal of Machine Learning Research* 18:1–49

Wu R, Xu J, Srikant R, Massoulié L, Lelarge M, Hajek B. 2015. Clustering and inference from pairwise comparisons, In *ACM SIGMETRICS Performance Evaluation Review*, vol. 43/1. ACM

Yan T. 2016. Ranking in the generalized Bradley–Terry models when the strong connection condition fails. *Communications in Statistics-Theory and Methods* 45:340–353

Zermelo E. 1929. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* 29:436–460