BAYESIAN MIXTURE OF PLACKETT-LUCE MODELS FOR PARTIALLY RANKED DATA

CRISTINA MOLLICA AND LUCA TARDELLA

Abstract. The elicitation of an ordinal judgment on multiple alternatives is often required in many psychological and behavioral experiments to investigate preference/choice orientation of a specific population. The Plackett-Luce model is one of the most popular and frequently applied parametric distributions to analyze partial top rankings of a finite set of items. The present work introduces a Bayesian finite mixture of Plackett-Luce models, that extends a device recently proposed in the literature in order to account for unobserved sample heterogeneity of ranked data. We describe an efficient way to incorporate the latent group structure in the data augmentation approach and the derivation of existing maximum likelihood procedures as special instances of the proposed Bayesian method. Inference can be conducted with the combination of the Expectation-Maximization algorithm for maximum a posteriori estimation and the Gibbs sampling iterative procedure. Our contribution also focuses on the practical impact of the identifiability issues that prevent from a straightforward implementation of the mixture modeling. We extensively investigate the label-switching phenomenon and describe the implementation of several relabeling algorithms to obtain meaningful posterior estimates. We additionally discuss Bayesian model checking and diagnostic tools for assessing parametric ranking distributions. Specifically, we perform posterior predictive check of goodness-of-fit by means of a discrepancy measure based on paired comparisons. The utility of the novel Bayesian Plackett-Luce mixture is illustrated with several applications to both simulated ranked data and real preference data.

1. Introduction

Choice behavior is a theme of great interest in several research areas, such as social and psychological sciences, but its investigation usually involves variables which cannot be directly observed and measured in an objective and precise manner. For this reason the evidence in choice experiments is often collected in ordinal form, that is, in terms of ranking data. More specifically, ranked data arise in those studies where a sample of \(N\) people is presented a finite set of \(K\) alternatives, called...
items, and is asked to rank them according to a certain criterion, typically personal preferences or attitudes. Thus, a generic ranking is the result of a comparative judgment on the competing alternatives expressed in the form of order relation. Interest in ranked data analysis is motivated, for example, by marketing and political surveys, but also by psychological and behavioral studies consisting, for instance, in the ordering of words/topics according to the perceived association with a reference subject.

Ranked data analysis has been addressed from numerous perspectives, as revealed by a wide and consolidated literature (Marden, 1995). Of course, a significant role is played by the parametric modeling of ranking data, which sometimes is inspired by possible patterns underlying the (random) mechanism of formation of individual preference. Nowadays there is a large number of parametric ranking distributions and they can be broadly grouped in four main classes: models based on order statistics, pairwise comparisons, distance-based and stagewise models. Despite this large availability of options often none of them is able to embody the appropriate flexibility to represent the heterogeneous nature of real data and it is natural to extend their flexibility through a mixture approach. Our work focuses on the finite mixture approach using the Plackett-Luce as parametric component within a Bayesian inferential framework. It parallels the frequentist approach in Gormley and Murphy (2006). On the other hand it differs from the Bayesian mixture approaches entertained in Gormley and Murphy (2009), and Caron et al. (2014) for the following reasons: the former deals with a soft grade of membership, whereas the more recent nonparametric extensions to an infinite number of items in the latter are based on an infinite mixture derived by means of a stick-breaking construction of the weights and a group structure similar to those derived through Dirichlet process mixtures. Moreover, in the present work we devote more attention to alternative criteria for the determination of the appropriate number of components and to suitable diagnostic tools to detect possible deficiencies of the parametric class in capturing the underlying dependence structure and hence an appropriate recognition of the parsimonious group structure.

The outline of the article is the following. In Section 2, after providing some basic definitions, we review the Bayesian literature concerning ranking models and describe the Plackett-Luce model. Section 3 presents the novel Bayesian mixture of Plackett-Luce models and details the related inferential procedures, whereas the label-switching issue and the relabeling algorithms are described in Section 4. We define the novel discrepancy variable for the Bayesian model assessment in Section 5 and illustrate applications of the proposed methods to both simulated and real ranking data in Section 6. The paper ends with concluding remarks and suggestions for future developments provided in Section 7.

2. Statistical models for ranked data

2.1. Preliminaries. Formally, a full (or complete) ranking \( \pi : I \rightarrow R \) is a bijective mapping of a finite set \( I = \{1, \ldots, K\} \) of labeled items into a set of ranks \( R = \{1, \ldots, K\} \), resulting from the attribution of a rank to each item. Hence, \( K \) is the total number of items to be ranked and the result of the mapping can be represented in terms of an ordered \( K \)-tuple \( \pi = (\pi(1), \ldots, \pi(K)) \), where the generic entry \( \pi(i) \) indicates the rank assigned to the \( i \)-th item. According to the underlying convention in the ranking theory, if \( \pi(i) < \pi(i') \) then item \( i \) is said to be ranked
higher than/preferred to item \(i'\). In the psychological literature ranked sequences are also referred to as ordinal ipsative data because of the constant sum of their components (Chan and Bentler [1998]).

Ranking data admit an alternative representation in terms of orderings. Specifically, the full (or complete) ordering \(\pi^{-1} : R \rightarrow I\) is simply the inverse function of the ranking \(\pi\), yielding a vector \(\pi^{-1} = (\pi^{-1}(1), \ldots, \pi^{-1}(K))\) whose generic component \(\pi^{-1}(j)\) denotes the item ranked in the \(j\)-th position. Before performing a ranking analysis, it is important to know exactly the format of the data and employ the suitable one, in order to avoid misleading results. The observed \(K\)-tuple, in fact, can be read in two different ways. As example for \(K = 4\), the vector (3,1,4,2) can be regarded as 

\[
\pi = (3, 1, 4, 2) \neq \pi^{-1} = (3, 1, 4, 2),
\]

where in the former notation \(\pi\) the first position is occupied by Item 2, whereas in the latter interpretation \(\pi^{-1}\) the top position is occupied by Item 3.

In many real applications, it often happens that the ranking elicitation is not completely carried out. A typical situation is when the ranker specifies only her best \(t < K\) preferences and leaves the remaining positions undefined. In this case the generic observation consists in the so-called top-\(t\) partial ordering of the form 

\[
\pi^{-1} = (\pi^{-1}(1), \ldots, \pi^{-1}(t))
\]

implying, with a slightly abuse of notation, that \(\pi(i) > t\) for all \(i \in I \setminus \{\pi^{-1}(1), \ldots, \pi^{-1}(t)\}\). Top-\(t\) rankings are just one type in the wide variety of incomplete data that can emerge from a ranking experiment. See Lebanon and Mao (2008) for further examples. In the past it was a usual practice to discard incomplete rankings from the analysis, with the consequent sample size reduction and loss of inferential accuracy. In the last decades this tendency has been reduced by the increasing number of ranking model extensions aimed at the simultaneous handling of mixed-type (full+partial) ranked data (Critchlow, 1985; Busse et al., 2007; Lebanon and Mao, 2008).

2.2. Probability models for random rankings. The space where complete observations take values is usually denoted with \(S_K\). It represents the finite discrete set containing all \(K!\) permutations of the first \(K\) positive integers, also known as symmetric group. Following the reviews in Critchlow et al. (1991); Marden (1995), the main approaches developed in the literature to conceive non-uniform parametric models on \(S_K\) can be classified in four main categories:

- order statistics models, where the fundamental approach is given by the Thurstone model (TM);
- models based on paired comparisons;
- distance-based models;
- stagewise models.

We concentrate on the last model class characterized by the decomposition of the ranking process into \(K - 1\) independent stages, corresponding to sequential selections from the item set \(I\). Specifically, our interest is on a well-established parametric distribution, the Plackett-Luce model (PL), widely described in Section 2.4 after a brief review of the Bayesian methods for the ranking modeling.

2.3. Bayesian modeling for ranked data. Bayesian ranking models have been mainly developed in the last decades, in conjunction with the fundamental advances.
achieved in the computational techniques that assisted analytically intractable posterior inference. After a pioneering attempt in the work of Fligner and Verducci (1990) to define a conjugate prior class for the distance-based model parameters, subsequent contributions appeared in the literature entail the posterior inference for the ordered statistics models. Yao and Böckenholt (1999) discuss the difficulties related to the estimation of the TM parameters, due in particular to the likelihood evaluation. They show that the GS can provide an efficient answer to high-dimensional integration thanks to the model completion with the latent item scores. Yu (2000) enlarges further such an approach accounting for the introduction of individual covariates in the TM through a linear regression framework. In this regard, Johnson and Kuhn (2013) provide JAGS code (Plummer et al., 2003) for a user-friendly implementation of the Bayesian TM. The GS is also applied for the distance-based models in Gupta and Damien (2002). Gupta and Damien (2005) also introduce a very general class of priors, named Binary Tree, on the group of permutations $S_K$. Stagewise ranking models are considered in Guiver and Snelson (2009), who derive a deterministic approach to perform Bayesian inference on the basic PL. Their method is based on the Power Expectation-Propagation algorithm, which simplifies the treatment of the annoying denominator of the PL likelihood by taking its reciprocal. Caron and Doucet (2012) show that an efficient Bayesian PL estimation can be achieved also by means of a suitable data augmentation approach. We recall their method in the next section, since it represents the starting point of our proposal.

Model-based clustering in the ranking literature has been addressed within the Bayesian framework. Gormley and Murphy (2009) estimate a grade of membership model for ranked data through a Metropolis-within-Gibbs algorithm in order to derive a soft (not mutually exclusive) classification of the sample units. As a further attempt to handle sample heterogeneity, one can also mention the mixture of distance-based models in ? and Meilă and Chen (2010), where the group structure is essentially induced by a Dirichlet process placed in the prior specification. The resulting model is referred to as Dirichlet process mixture model (DPMM) and the authors describe in great detail an efficient GS scheme to conduct inference on it. See Ali et al. (2010) for a real application of the DPMM to the CAO data set of the college applicant degree choices. Nonparametric Bayesian clustering via DPMM has been recently proposed also by Caron et al. (2012) and Caron et al. (2014). In this series of works the generative model for partial rankings is the extended version of the PL model for infinite rankings, originally presented in Caron and Teh (2012).

2.4. The Plackett-Luce model. A ranking can be elicited through a series of sequential comparisons in which a single item is preferred to all the remaining alternatives and, after being selected, is removed from the next comparisons. This is the basic construction underlying the PL, introduced by Luce (1959) and Plackett (1975), parametrized by the so-called support parameters $p = (p_1, \ldots, p_K)$. These represent positive constants associated to each item such that a higher value of the support parameter $p_i$ implies a greater probability for the corresponding item $i$ to be preferred at each selection stage. Let $\pi^{-1} = \{\pi^{-1}_s\}_{s=1}^N$ be a random sample consisting of $N$ partial top orderings of the form $\pi^{-1}_s = (\pi^{-1}_s(1), \ldots, \pi^{-1}_s(n_s))$, where $n_s$ specifies the number of top positions expressed by unit $s$. A full ordering corresponds to both cases $n_s = K - 1$ and $n_s = K$ since, once $K - 1$ items have been ranked, the last position is automatically determined. Under the PL the
contribution to the likelihood of a single partial observation is given by

\begin{equation}
\mathbf{P}_{PL}(\pi^{-1}|p) = \prod_{t=1}^{n_s} \frac{p_{\pi^{-1}(t)}}{\sum_{\nu=t}^{n_s} p_{\pi^{-1}(\nu)}}.
\end{equation}

A first aspect to be emphasized in (2.1) is the property of internal consistency, allowing a straightforward treatment of the incomplete sequences. The probability of a generic partial ordering, computed through marginalization over missing positions, is equal to the probability of completely ordering only the items actually ranked (see Hunter (2004) for a formal proof). The PL property of internal consistency leads to a unit-dependent normalization of the support parameters at each stage, in the sense that it is taken w.r.t. the items exactly allocated in every specific ordinal sequence. One of the main inferential issues related to formulation (2.1) concerns the presence of the annoying normalization term \(\sum_{\nu=t}^{n_s} p_{\pi^{-1}(\nu)}\), that does not permit the direct maximization of the likelihood. In the ML framework Hunter (2004) overcomes this difficulty applying the Minorization-Maximization (MM) algorithm, an iterative method relying on the replacement of the original PL log-likelihood with a minorizing surrogate objective function. In the Bayesian perspective, instead, a related efficient solution is proposed by Caron and Doucet (2012). They realize a data augmentation step with the introduction of latent quantitative variables \(y = (y_{st})\) for \(s = 1, \ldots, N\) and \(t = 1, \ldots, n_s\), with conditional distribution given by

\begin{equation}
f(y|\pi^{-1}, p) = \prod_{s=1}^{N} \prod_{t=1}^{n_s} f_{Exp}(y_{st} | \sum_{\nu=t}^{n_s} p_{\pi^{-1}(\nu)}) ,
\end{equation}

where \(f_{Exp}(\cdot; \lambda)\) denotes the Negative Exponential density function parametrized by the rate parameter \(\lambda\). The parametric assumption (2.2) entails remarkable simplifications for the implementation of both the posterior optimization and the GS algorithm. The success of the Bayesian device introduced by Caron and Doucet (2012) is due to the combination of (2.2) with a conjugate prior specification. This latter aspect moves from the Thurstonian interpretation of (2.1), that is, the TM reduces to the PL when the Gumbel distribution is employed as distribution of the latent scores. The Gumbel distribution admits the Gamma density function as conjugate prior and this allows Caron and Doucet (2012) to derive a simple GS scheme for the approximation of the posterior distribution.

3. Bayesian mixture of Plackett-Luce models

To our knowledge Bayesian inference of a finite PL mixture for partially ranked data has not been previously developed in the literature, although a wide variety of research contexts requires a model-based analysis accounting for the presence of differential patterns in the sample. Bayesian PL estimation appeared so far in the literature is either limited to the homogeneous case (see Guiver and Snelson (2009) and Caron and Doucet (2012)) or, in the mixture context, accounts for an infinite number of items through a nonparametric method (Caron et al. 2012, 2014). In the next sections we detail the novel Bayesian PL mixture for partial top rankings.
3.1. **Model and prior specification.** Let \( \pi^{-1} \) be a random sample drawn from a PL mixture with \( G \) components, in symbols

\[
\pi_{1}^{-1}, \ldots, \pi_{N}^{-1} \mid \mathbf{p}, \omega \overset{i.i.d.}{\sim} \sum_{g=1}^{G} \omega_{g} P_{PL}(\pi_{s}^{-1} \mid \mathbf{p}_{g}),
\]

where \( \mathbf{p}_{g} \) is the support parameter vector specific of the \( g \)-th mixture component and \( \omega_{g} \) is the corresponding weight. In order to suitably generalize the data augmentation approach in Caron and Doucet (2012) within the finite mixture framework, we need to introduce an additional latent feature of each generic sample unit \( s \), represented by the unknown group membership labels

\[
\mathbf{z}_{s} = (z_{s1}, \ldots, z_{sG}) \mid \omega \overset{i.i.d.}{\sim} \text{Mult}(1, \omega = (\omega_{1}, \ldots, \omega_{G}))
\]

whose generic marginal is a Bernoulli r.v.

\[
z_{sg} = \begin{cases} 
1 & \text{if unit } s \text{ belong to } g \text{-th mixture component}, \\
0 & \text{otherwise},
\end{cases}
\]

governed by the mixture weight \( \omega_{g} \). We propose to include the unobserved group structure in the data augmentation strategy as follows

\[
f(y \mid \pi^{-1}, \mathbf{z}, \mathbf{p}, \omega) = \prod_{s=1}^{N} \prod_{t=1}^{n_{s}} f_{\text{Exp}} \left( y_{st} \mid \prod_{g=1}^{G} \left( \sum_{\nu=1}^{n_{s}} p_{g} \pi_{s}^{-1}(\nu) \right) z_{sg} \right).
\]

This implies that the latent group membership determine the cluster-specific support parameters acting on the underlying quantitative mechanism. Once the model governing observed and latent variables is specified, a fully Bayesian approach requires the elicitation of the joint prior distribution for the unknown parameters. Postulating prior independence between \( \mathbf{p} \) and \( \omega \), that is \( f_{0}(\mathbf{p}, \omega) = f_{0}(\mathbf{p})f_{0}(\omega) \), similarly to the homogeneous population context, we choose prior distributions with a convenient conjugate structure. For the support parameters, in fact, we extend the initial distribution in Caron and Doucet (2012), defining

\[
p_{g i} \overset{i.i.d.}{\sim} \text{Ga}(c_{gi}, d_{g})
\]

for \( g = 1, \ldots, G \) and \( i = 1, \ldots, K \), where the Gamma r.v.’s are indexed by the shape and the rate parameters. For the mixture weights, taking values in the \((G - 1)\)-dimensional simplex, we make the standard prior assumption

\[
\omega \sim \text{Dir}(\alpha_{1}, \ldots, \alpha_{G}),
\]

because of the conjugacy with the Multinomial likelihood assumed for the latent group labels \( \mathbf{z} \).

3.2. **MAP estimation.** In the presence of the latent variables \( y \) and \( \mathbf{z} \), we can construct an EM algorithm in order to optimize the posterior distribution and learn the posterior mode (MAP estimate). The complete-data likelihood turns out to be

\[
L_{c}(\mathbf{p}, \omega, y, \mathbf{z}) = P(y \mid \pi^{-1}, \mathbf{z} \mid \mathbf{p}, \omega) = f(y \mid \pi^{-1}, \mathbf{z}, \mathbf{p}, \omega) P(\pi^{-1}, \mathbf{z} \mid \mathbf{p}, \omega).
\]

equal to the product of the full-conditional (3.2) times the standard complete-data likelihood of a mixture model specification without the vector \( y \). With simple
computations detailed in the Appendix A both factors in (3.3) can be rearranged in order to explicit a multinomial form in the $z$ as follows

$$f(y \mid \pi^{-1}, z, p, \omega) = \prod_{s=1}^{N} \prod_{g=1}^{G} \left( \prod_{i=1}^{n_s} \sum_{\nu=t}^{n_s} p_{g \pi_s^{-1}(\nu)} \right) e^{-\sum_{i=1}^{n_s} y_{st} \sum_{\nu=t}^{n_s} p_{g \pi_s^{-1}(\nu)}}$$

and

$$P(\pi^{-1}, z \mid p, \omega) = \prod_{s=1}^{N} \prod_{g=1}^{G} \left( \omega_{g} \prod_{i=1}^{n_s} \frac{p_{g \pi_s^{-1}(i)}}{\sum_{\nu=t}^{n_s} p_{g \pi_s^{-1}(\nu)}} \right)^{z_{sg}}.$$ 

Hence,

$$L_c(p, \omega, y, z) = \prod_{s=1}^{N} \prod_{g=1}^{G} \left( \omega_{g} \prod_{i=1}^{K} \frac{p_{g \pi_s^{-1}(i)} u_{s i} e^{-p_{g \pi_s^{-1}(t)} \sum_{\nu=t}^{n_s} \delta_{st \nu}}}}{\sum_{i=1}^{n_s} p_{g \pi_s^{-1}(t)}} \right)^{z_{sg}},$$

where

$$u_{s i} = \begin{cases} 1 & \text{if } i \in \{ \pi_s^{-1}(1), \ldots, \pi_s^{-1}(n_s) \}, \\ 0 & \text{otherwise}, \end{cases}$$

and

$$\delta_{st i} = \begin{cases} 1 & \text{if } i \in \{ \pi_1^{-1}(t), \ldots, \pi_s^{-1}(n_s) \}, \\ 0 & \text{otherwise}. \end{cases}$$

Indicating with $l_c(p, \omega, y, z)$ the complete-data log-likelihood, the implementation of the EM algorithm requires the iterative maximization of the following objective function

$$Q((p, \omega), (p^*, \omega^*)) = \mathbb{E}_{y, z \mid \pi^{-1}, p^*, \omega^*}[l_c(p, \omega, y, z)] + \log f_0(p, \omega),$$

where the expectation is computed w.r.t. the joint distribution of all the latent variables, given by

$$(3.4) \quad P(y, z \mid \pi^{-1}, p, \omega) = f(y \mid \pi^{-1}, z, p, \omega) P(z \mid \pi^{-1}, p, \omega).$$

Computing first the expectation w.r.t. $y$, one has

$$\sum_{s=1}^{N} \sum_{g=1}^{G} z_{sg} \left( \log \omega_{g} + \sum_{i=1}^{K} \left( u_{s i} \log p_{g i} - p_{g i} \sum_{t=1}^{n_s} \delta_{st i} \sum_{\nu=t}^{n_s} p_{g \pi_s^{-1}(\nu)} \right) \right).$$

For the expectation w.r.t. $z$ it is convenient to rewrite the last fractional term as

$$\prod_{g=1}^{G} \left( \sum_{\nu=t}^{n_s} p_{g \pi_s^{-1}(\nu)} \right)^{z_{sg}} \prod_{g=1}^{G} \frac{\delta_{st i}}{\sum_{\nu=t}^{n_s} p_{g \pi_s^{-1}(\nu)}} = \sum_{g=1}^{G} z_{sg} \frac{\delta_{st i}}{\sum_{\nu=t}^{n_s} p_{g \pi_s^{-1}(\nu)}}$$

and, noting that $P(z \mid \pi^{-1}, p, \omega)$ in the right-hand side of (3.4) is proportional to $P(\pi^{-1}, z \mid p, \omega)$, the E-step returns

$$Q((p, \omega), (p^*, \omega^*)) = \sum_{s=1}^{N} \sum_{g=1}^{G} z_{sg} \left( \log \omega_{g} + \sum_{i=1}^{K} \left( u_{s i} \log p_{g i} - p_{g i} \sum_{t=1}^{n_s} \delta_{st i} \sum_{\nu=t}^{n_s} p_{g \pi_s^{-1}(\nu)} \right) \right) + \sum_{g=1}^{G} (\alpha_{g} - 1) \log \omega_{g} + \sum_{g=1}^{G} \sum_{i=1}^{K} (c_{gi} - 1) \log p_{g i} - d_{g p_{gi}}.$$
where
\[ \hat{z}_{sg} = \frac{\omega_g^* P_{PL}(\pi_s^{-1}|p_g^*)}{\sum_{g'=}^{G} \omega_{g'}^* P_{PL}(\pi_s^{-1}|p_{g'}^*)}. \]
We have used the notation \( \hat{z}_{sg} \) for the conditional membership probability to stress the difference with the binary group label \( z_{sg} \). Differentiating the objective function \( Q \) w.r.t. each \( p_{gi} \) and equating to zero, the estimation formula for the support parameters resulting from the M-step turns out to be
\[ p_{gi} = \frac{c_{gi} - 1 + \hat{\gamma}_{gi}}{d_g + \sum_{s=1}^{N} \sum_{t=1}^{n_s} \delta_{sti} \sum_{\nu=t}^{\nu_s} p_{g\pi_s^{-1}(\nu)}}. \]
for \( g = 1, \ldots, G \) and \( i = 1, \ldots, K \), where
\[ \hat{\gamma}_{gi} = \sum_{s=1}^{N} \hat{z}_{sg} u_{si}. \]
The optimization of \( Q((p, \omega), (p^*, \omega^*)) \) w.r.t. \( \omega \), subject to the canonical constraint \( \sum_{g=1}^{G} \omega_g = 1 \), leads to the following updated estimates of the mixture weights
\[ \omega_g = \frac{\alpha_g - 1 + \sum_{s=1}^{N} \hat{z}_{sg}}{\sum_{g'=1}^{G} \alpha_{g'} - G + N}, \quad g = 1, \ldots, G. \]
In conclusion the EM reduces to the following iterative procedure:

**Initialization**: set starting values \( p^{(0)}, \omega^{(0)} \) for the parameters to be estimated;

**Computation**: at iteration \( l + 1 \), compute until convergence
- for \( s = 1, \ldots, N \) and \( g = 1, \ldots, G \)
\[ \hat{z}_{sg}^{(l+1)} = \frac{\omega_g^{(l)} P_{PL}(\pi_s^{-1}|p_g^{(l)})}{\sum_{g'=1}^{G} \omega_{g'}^{(l)} P_{PL}(\pi_s^{-1}|p_{g'}^{(l)})}. \]
- for \( g = 1, \ldots, G \)
\[ \omega_g^{(l+1)} = \frac{\alpha_g - 1 + \sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)}}{\sum_{g'=1}^{G} \alpha_{g'} - G + N}, \]
- for \( g = 1, \ldots, G \) and \( i = 1, \ldots, K \)
\[ p_{gi}^{(l+1)} = \frac{c_{gi} - 1 + \hat{\gamma}_{gi}^{(l+1)}}{d_g + \sum_{s=1}^{N} \sum_{t=1}^{n_s} \delta_{sti} \sum_{\nu=t}^{\nu_s} p_{g\pi_s^{-1}(\nu)}}. \]
Notice that when \( G = 1 \) this three-step procedure collapses into the single updating formula obtained by Caron and Doucet (2012). Moreover, similarly to their method, also in our mixture approach we can recover the MLE as special case of the uninformative Bayesian analysis with flat priors. In our scheme these can be obtained fixing \( c_{gi} = 1, d_g = 0 \) and \( \alpha_g = 1 \). Such a configuration for the hyperparameter reduces the proposed MAP procedure to the EMM algorithm described in Gormley and Murphy (2006). The EMM is a hybrid Expectation-Maximization algorithm, given by the combination of the standard EM with the MM scheme,
successfully applied by [Gormley and Murphy (2006)] to infer the PL mixture in the ML framework.

3.3. **Gibbs sampling.** In order to draw a sample from the joint posterior distribution and learn about the uncertainty associated to the final estimates, we detail the implementation of a GS procedure. The conjugate prior configuration described in Section 3.1, combined with the complete likelihood $L_c(p, \omega, \gamma, \mathbf{z})$, leads to a sampling scheme with parametric distributions which are simple to draw from. In particular, the full-conditionals of the latent component labels are easily derived noting that

$$P(\mathbf{z} | \pi^{-1}, p, \omega) \propto L_c(p, \omega, \gamma, \mathbf{z}),$$

implying the following multinomial structure

$$P(\mathbf{z} | \pi^{-1}, p, \omega) \propto \prod_{g=1}^{G} \prod_{i=1}^{K} \omega_g \prod_{s=1}^{n_s} \prod_{t=1}^{n_s} \delta_{st} y_{gt} z_{sg}.$$

The full-conditionals of the support parameters are are still members of the Gamma family with hyperparameters suitably updated as follows

$$P(p_{gi} | \pi^{-1}, \gamma, \mathbf{z}, p_{[-gi]}, \omega) \propto f_0(p_{gi}) L_c(p, \omega, \gamma, \mathbf{z}) = p_{gi}^{\gamma_{gi} + \gamma_{gi} - 1} e^{-p_{gi} (d_g + \sum_{s=1}^{N} z_{sg} \sum_{t=1}^{n_s} \delta_{st} y_{gt})},$$

where $\gamma_{gi} = \sum_{s=1}^{N} z_{sg} u_{si}$ is the number of units assigned to cluster $g$ who have ranked item $i$ and $p_{[-gi]}$ denotes the matrix $p$ of the support parameters without the $(g \times i)$-th component. Also the full-condional of the mixture weights has the same form of the corresponding prior family, given by

$$P(\omega | \pi^{-1}, \gamma, \mathbf{z}, \mathbf{p}) \propto f_0(\omega) L_c(p, \omega, \gamma, \mathbf{z}) \propto f_0(\omega) P(\mathbf{z} | \omega) = \prod_{g=1}^{G} \omega_g^{\alpha_g + \sum_{s=1}^{N} z_{sg} - 1}.$$

Finally, notice that the full-condional of the $y$'s is given by construction in the assumption (3.2). In conclusion, the GS algorithm to approximate the joint posterior distribution $P(\mathbf{z}, \gamma, \mathbf{p}, \omega | \pi^{-1})$ consists in performing iteratively the following steps

**Initialization:** set the total number $T$ of iterations and the starting values $\mathbf{y}^{(0)}, \mathbf{z}^{(0)}, p^{(0)}$ and $\omega^{(0)}$ (note that we need starting values only for $\mathbf{z}$ and $p$);

**Sampling:** at iteration $(l + 1) \leq T$, sample

- the mixture weights

  $$\omega_{s(l+1)} | \mathbf{z}^{(l)} \sim \text{Dir}\left(\alpha_1 + \sum_{s=1}^{N} z_{s1}^{(l)}, \ldots, \alpha_G + \sum_{s=1}^{N} z_{sG}^{(l)}\right),$$

  - for $s = 1, \ldots, N$ and $t = 1, \ldots, n_s$

- $y_{st}^{(l+1)} | \pi^{-1}, \mathbf{z}^{(l)}, \mathbf{p}^{(l)} \sim \text{Exp}\left(\prod_{g=1}^{G} \left(\sum_{s=1}^{N} p_{g(l)}^{(l)} \pi_s^{-1} (v) \right) z_{sg}^{(l)}\right)$

  - for $g = 1, \ldots, G$ and $i = 1, \ldots, K$

- $p_{gi}^{(l+1)} | \pi^{-1}, \mathbf{y}^{(l+1)}, \mathbf{z}^{(l+1)} \sim \text{Ga}\left(c_{gi} + \gamma_{gi}, d_g + \sum_{s=1}^{N} z_{sg}^{(l)} \sum_{t=1}^{n_s} \delta_{st} y_{gt}^{(l+1)}\right)$,
- for \( s = 1, \ldots, N \)
\[
Z_s^{(l+1)} | \pi_s^{(-1)}, \sum_s^{(l+1)}, \varphi^{(l+1)}_s \sim \text{Mult} \left( 1, \left( m_{s1}^{(l+1)}, \ldots, m_{sG}^{(l+1)} \right) \right),
\]
where
\[
m_{sg}^{(l+1)} = \frac{\omega_g^{(l+1)} \prod_{k=1}^{K} (p_{gi}^{(l+1)})^{u_{gs} - p_{gi}^{(l+1)} \sum_{t=1}^{N_s} \delta_{st}^{(l+1)}} \sum_{g'=1}^{G} \omega_{g'}^{(l+1)} \prod_{k=1}^{K} (p_{g'i}^{(l+1)})^{u_{gs} - p_{g'i}^{(l+1)} \sum_{t=1}^{N_s} \delta_{st}^{(l+1)}}}{\sum_{g'=1}^{G} \omega_{g'}^{(l+1)} \prod_{k=1}^{K} (p_{g'i}^{(l+1)})^{u_{gs} - p_{g'i}^{(l+1)} \sum_{t=1}^{N_s} \delta_{st}^{(l+1)}}}.
\]

Notice that the EM and the GS can be conveniently combined employing the MAP solution as good initialization of the chain in the MCMC simulation. Once convergence has been achieved, the posterior marginal distributions and related summaries can be approximated by their empirical counterparts computed on a suitably large posterior sample.

### 3.4. Bayesian model comparison.

In the estimation procedures previously described the number \( G \) of groups is fixed \textit{a priori} and we can then perform a separate inference on PL mixtures with a different number of components. Thus, a method for discriminating among the competing models is needed. Due to the computational difficulties related to the marginal likelihood-based methods, such as the Bayes factor, for the selection of the number of groups we rely on two alternative Bayesian criteria, i.e., the Deviance Information Criterion (DIC) and the Bayesian Predictive Information Criterion (BPIC). The DIC was originally introduced in the fundamental work by Spiegelhalter et al. (2002) and is a very popular penalized measure of fitting for Bayesian model comparison. Moreover, it is very easily approximated when inference is carried out with an MCMC sampling scheme. Denoting with \( D(\theta) = -2 \log L(\theta) \) the deviance, the DIC has the following expression
\[
\text{DIC} = \bar{D} + p_D.
\]
The former term \( \bar{D} = \mathbb{E}[D(\theta)|x] \) is the posterior expected deviance, whereas \( p_D \) is the \textit{effective number of parameters} acting as penalty term. The penalty \( p_D \) is computed as the difference \( \bar{D} - D(\hat{\theta}) \), where \( D(\hat{\theta}) \) is the deviance evaluated at the single point estimate \( \hat{\theta} \). In the literature an alternative method to assess model complexity was proposed by Gelman et al. (2004), consisting in setting \( p_D \) equal to half the posterior variance of the deviance. Both computations of the effective number of parameters are justified by the asymptotic posterior distribution of the deviance, given by
\[
D(\theta)|x \sim D(\hat{\theta}_{ML}) + \chi^2_{q^*},
\]
where \( D(\hat{\theta}_{ML}) \) is the ordinary frequentist measure of the goodness-of-fit and \( q^* \) is the true number of parameters. We will exploit the above asymptotic results and explore, in the large sample analyses presented in Section 6.1 and 6.3, both DIC formulations. We will denote the criterion respectively with DIC1 when the penalty is computed as \( p_D = \bar{D} - D(\hat{\theta}_{MAP}) \), and with DIC2 when based on \( p_D = \text{VAR}[D(\theta)|x]/2 \).

One aspect often debated on the DIC is its tendency to overfitting due to the double usage of the observed data. So, we additionally considered the BPIC suggested by Ando (2007), which penalizes the fitting measure \( \bar{D} \) with \( 2p_D \).

The optimal model is identified with the one that minimizes one of the former criteria.
4. Label-switching

When one adopts an MCMC simulation to derive approximate Bayesian inference of a mixture model, an annoying identifiability issue can affect the posterior sample, the so-called label-switching phenomenon (LS).

4.1. Formal definition. The LS represents a specific form of unidentifiability concerning mixture models, that reflects the arbitrary attribution of the indices \( \{1, \ldots, G\} \) to denote the mixture components. The application of a permutation \( \xi \in S_G \) of the \( G \) indices to a given parameter point, which corresponds to a relabeling of the latent classes, does not modify the resulting sampling distribution. Formally, let \( \theta = (\omega_1, \eta_1, \ldots, \omega_G, \eta_G) \) be the parameter vector indexing a generic \( G \)-component mixture model \( f_{\text{mix}}(\cdot|\theta) \), where \( \eta_g \) collects the parameters of the \( g \)-th group and \( \omega_g \) is the corresponding weight. The LS translates into the following invariance condition on the sampling distribution

\[
f_{\text{mix}}(\cdot|\theta) = \sum_{g=1}^{G} \omega_g f(\cdot|\eta_g) = \sum_{g=1}^{G} \omega_{\xi(g)} f(\cdot|\eta_{\xi(g)}) = f_{\text{mix}}(\cdot|\xi(\theta)),
\]

implying

\[
L(\theta) = L(\xi(\theta))
\]

for all \( \xi \in S_G \). Equality (4.1) explains the presence of \( G! \) symmetric modes in the likelihood function, independently on the sample realization. If the same permutation invariance is fulfilled by the prior distribution, that is, no discrimination is made \textit{a priori} among the mixture components, then the symmetry transfers also in the posterior distribution. In this situation, the marginal posterior distribution of each parameter is the same for all the mixture components. Hence, the posterior summaries coincide, nullifying the standard practice to use them as point estimates.

4.2. Relabeling algorithms. As better clarified shortly with an example on synthetic data, the LS induces important difficulties in the application of sampling-based methods for the Bayesian analysis of mixture models. Different strategies have been proposed in the statistical literature to solve the LS issue in the MCMC analysis. Following the review by Jasra et al. (2005), they can be summarized in three classes: (i) introduction of artificial identifiability constraints, (ii) relabeling algorithms (RA) and (iii) employment of label-invariant loss functions. We focus on the RA, whose basic idea is the ex-post relabeling of the raw MCMC simulations in order to make them lie in a unique posterior mode among the \( G! \) possible modal regions. If the MCMC sample after the burn-in phase has size \( L \), the RA determine a total of \( L \) permutations for the rearrangement of each single parameter drawing from the posterior distribution. For example, the Pivotal Reordering algorithm proposed by Marin et al. (2005) switches the elements of each simulated value \( \theta^{(l)} \) according to a permutation \( \xi_l \) so that a certain distance from a target mode is minimized. The target mode plays the role of pivot and can be easily identified with the MAP solution. The class of RA includes also the clustering-oriented method proposed by Stephens (2000), the Probabilistic RA (Sperrin et al., 2010), the Equivalence Classes Representatives technique (Papastamoulis and Iliopoulos, 2010) and the Data RA recently introduced by Rodríguez and Walker (2014). With the only exception of the Data RA, they are all implemented in a recently released
R package, called \texttt{label.switching} \cite{Papastamoulis2014}, that we exploited in our Bayesian PL mixture applications.

5. Bayesian model assessment

Once the model has been selected, a comprehensive inferential analysis should also contemplate the adequacy of the estimated model in describing the observed data \cite{Gelman1996}. Despite its practical relevance, the fitting performance is often neglected by practitioners, especially in frequentist analysis of ranking data. In the frequentist domain, in fact, this issue is typically addressed with the computation of the \textit{p-value} associated to a goodness-of-fit statistic, such as the likelihood ratio or the Pearson test. However, in sparse data situations, where the sample size \( N \) is small w.r.t. the cardinality \( K! \) of \( S_K \), serious problems arise with this approach. The chi-square sampling distribution of the above statistics under the posited model \( H \), for example, does not longer apply. Cohen and Mallows \cite{Cohen1983} suggest to overcome this difficulty comparing observed and expected frequencies for relevant partitions of the ranking space. Within the Bayesian paradigm Yao and Böckenholt \cite{Yao1999}, Yu \cite{Yu2000} and Gormley and Murphy \cite{Gormley2009} follow this suggestion to generalize the classical test statistic into a parameter-dependent quantity, referred to in the literature as \textit{discrepancy variable} \cite{Gelman1996, Meng1994}.

In order to assess the adequacy of the Bayesian PL mixture we concentrate on a popular summary statistic in ranked data analysis, i.e. the paired comparison matrix \( \tau \). When the sample is entirely composed of complete orderings, the generic entry of \( \tau \) is

\begin{equation}
\tau_{ii'} = \sum_{s=1}^{N} I_{\pi_s(i) < \pi_s(i')} \quad 1 \leq i < i' \leq K, \tag{5.1}
\end{equation}

corresponding to the total number of wins of item \( i \) over item \( i' \). Since ties are not contemplated in this setting, the sum \( T_{ii'} = \tau_{ii'} + \tau_{i'i} \) indicating the total number of pairwise comparisons between item \( i \) and \( i' \) is equal to the sample size \( N \). In the case of partial orderings, definition (5.1) becomes

\[ \tau_{ii'} = \sum_{s=1}^{N} (1 - (1 - u_{si})(1 - u_{si'})) I_{\pi_s(i) < \pi_s(i')} = \sum_{s=1}^{N} (u_{si} + u_{si'} - u_{si}u_{si'}) I_{\pi_s(i) < \pi_s(i')} \]

As described in Yao and Böckenholt \cite{Yao1999}, one could employ the sample quantities \( \tau_{ii'} \) for defining a chi-square discrepancy as follows

\[ X^2(\pi^{-1}; \theta) = \sum_{i < i'} \frac{(\tau_{ii'} - \tau^*_{ii'}(\theta))^2}{\tau^*_{ii'}(\theta)} \]

which compares the observed pairwise preferences with the ones expected under the posited model \( H \). Assuming a PL mixture as the data generating mechanism implies

\[ \tau^*_{ii'}(\theta) = T_{ii'} \frac{p_i}{p_i + p_{i'}}, \]

where the overall support to item \( i \) is obtained as \( p_i = \sum_{g=1}^{G} \omega_g p_{gi} \). The reader can refer to Cohen and Mallows \cite{Cohen1983} for the expression of \( \tau^*_{ii'}(\theta) \) under different
model specifications. The posterior predictive p-value for checking the model fitness is then obtained as
\begin{equation}
  p_B = P(X^2(\hat{\pi}_{\text{rep}}^{-1}; \theta) \geq X^2(\hat{\pi}_{\text{obs}}^{-1}; \theta)|\hat{\pi}_{\text{obs}}^{-1}, H),
\end{equation}
where the randomness is induced by the joint posterior distribution of \((\hat{\pi}_{\text{rep}}^{-1}, \theta)\) given \(H\). As emphasized by Gelman et al. (1996), once a sample of size \(L\) from the posterior distribution is available, the computation of (5.2) does not require substantial additional efforts, but the application of the following two-step procedure for all the parameter draws: one first simulates a new data set \(\pi_{\text{rep}}^{-1}(l)\) from the sampling distribution induced by \(\theta(l)\) and subsequently compares \(X^2(\pi_{\text{rep}}^{-1}(l); \theta(l))\) with \(X^2(\pi_{\text{obs}}^{-1}; \theta(l))\). The posterior predictive p-value is finally estimated as the empirical relative frequency of the times that \(X^2(\pi_{\text{rep}}^{-1}(l); \theta(l))\) exceeds \(X^2(\pi_{\text{obs}}^{-1}; \theta(l))\).

6. Illustrative applications

Although the Bayesian approaches described in Section 3.2 and 3.3 allows to convey specific (subjective) prior knowledge on the parameters, in the following analyses we will rely upon weakly or noninformative priors and make a direct comparison with the MLE.

6.1. Simulation study. We have first evaluated the effectiveness of the estimation algorithms applying them to synthetic data, paying special attention to the behavior of the GS regarding the exploration of the parameter space and the occurrence of LS. For this purpose we simulated a sample of \(N = 300\) complete orderings of length \(K = 4\) from a 2-component PL mixture under three different population scenarios. True parameter values are reported in the upper-panel of Table 1 and describe bimodal populations with decreasing Kendall distances \(d_K\) between the two modes, specifically equal to \(d_K = 6\) in Scenario 1, \(d_K = 3\) in Scenario 2 and \(d_K = 1\) in Scenario 3. Adopting uninformative prior densities with hyperparameters equal to \(c_g = 1\), \(d_g = .001\) and \(\alpha_g = 1\), we recorded for each simulated data set the MAP estimate and posterior summaries obtained with the help of the GS initialized with a random start. For the third scenario we considered also the initialization of the GS with the MAP estimate. We run the GS for a total of 22000 iterations discarding the first 2000 drawings as burn-in period. The LS problem did not take place at all in the GS application to the first two scenarios, because the sampler explored only one of the two symmetric modes. This aspect confirms a well-known feature of the GS (see Celeux et al. (2000)), which is unable to suitably move over the parameter space and capture in full the multimodal profile of the posterior distribution. For the case of artificial multimodality induced by the LS this behavior is paradoxically convenient since it allows straightforward inference. As remarked by Celeux et al. (2000), in fact, the visit of different symmetric modal regions is unnecessary from an inferential point of view, because the artificial peaks equivalently characterize the sampling distribution. Thus, the absence of LS for Scenario 1 and 2 allows us to directly derive meaningful GS estimates which, as indicated in Table 1 are indeed very close to the MAP results and in turn to the true parameter values, especially for the case of maximum distance (Scenario 1) between the two mixture components.

In the GS analysis of the third scenario the effects of the LS are evident. They manifest with step-like configurations of the traceplots of the parameters indicating
Table 1. Population scenarios considered in the simulation study and corresponding inferential results from the MAP estimate and the GS procedure initialized with random starting values. Posterior means were computed on both the raw and the relabeled MCMC samples. For Scenario 3 the GS was initialized also with the MAP solution, whose results are highlighted in grey.

| Scenario 1          | Scenario 2          | Scenario 3          |
|---------------------|---------------------|---------------------|
| True value          | True value          | True value          |
| 1.700 (1, 2, 3, 4)  | 1.700 (1, 2, 3, 4)  | 1.700 (1, 2, 3, 4)  |
| 2.300 (4, 3, 2, 1)  | 2.300 (4, 3, 2, 1)  | 2.300 (4, 3, 2, 1)  |
| MAP                 | MAP                 | MAP                 |
| 1.702 (1, 2, 3, 4)  | 1.702 (1, 2, 3, 4)  | 1.702 (1, 2, 3, 4)  |
| 2.298 (4, 3, 2, 1)  | 2.298 (4, 3, 2, 1)  | 2.298 (4, 3, 2, 1)  |
| Raw GS              | Raw GS              | Raw GS              |
| 1.699 (1, 2, 3, 4)  | 1.699 (1, 2, 3, 4)  | 1.699 (1, 2, 3, 4)  |
| 2.301 (4, 3, 2, 1)  | 2.301 (4, 3, 2, 1)  | 2.301 (4, 3, 2, 1)  |
| Relabeled GS        | Relabeled GS        | Relabeled GS        |
| 1.699 (1, 2, 3, 4)  | 1.699 (1, 2, 3, 4)  | 1.699 (1, 2, 3, 4)  |
| 2.301 (4, 3, 2, 1)  | 2.301 (4, 3, 2, 1)  | 2.301 (4, 3, 2, 1)  |

Figure 1. Traceplots and posterior marginal densities for the weights of the 2-component PL mixture resulting from the raw MCMC sample for the third population scenario.

transitions of the sampler from one mode to another (see Figures 1 and 2). In particular, the traceplots point out that almost half of the chain is affected by LS, leading to marginal posterior densities which pretty much overlap, as shown in Figure 1 and 2 and yielding very similar empirical marginal means (Table 1). The LS still occurs when the MCMC chain is initialized with the MAP estimation, but at a reduced extent. The dependence on the initialization reveals the instability of the LS occurrence for different runs of the MCMC algorithm.
Consequently, we assess how alternative relabeling methods act in addressing the LS issue in the third scenario. The raw MCMC output has been post-processed using the functions implemented in the R package `label.switching`. The check of the traceplots pointed out a good performance of all methods in removing the artificial multimodality. All strategies consistently rearranged about 49% of the drawings and returned very similar results in terms of adjusted estimates. Posterior means derived specifically from the application of the algorithm proposed by Stephens (2000) are shown on the bottom-right of Table 1. Analogous considerations are valid for the results relative to the MCMC chain initialized with the MAP but, in this case, only 21% of the simulations has been relabeled (see results in Table 1 highlighted in grey). As consequence of the adjacency of the two mixture components in the third scenario, we note that the closeness of the adjusted

**Figure 2.** Traceplots for the support parameters of the 2-component PL mixture resulting from the raw MCMC sample for the third population scenario. Solid lines indicate the true parameter values.
Figure 3. Posterior marginal densities for the support parameters of the 2-component PL mixture resulting from the raw MCMC sample for the third population scenario.

GS estimates to the true values is slightly reduced than the one observed in both Scenario 1 and 2, although the actual order of the support parameters within each group is fully recovered (Table 1). As expected, when the two mixture components considerably overlap it is more difficult to reconstruct the actual group membership of the sample units, with possible negative effects on the final estimates. In this regard, the performance of the GS turns out to be better than the MAP procedure, which indeed corresponds to the MLE because of the flat priors, since the latter completely fails in inferring the minor mixture component. The posterior mode exhibits the tendency to privilege the homogenous model, i.e., to allocate the sample units into the major mode, returning very poor estimates for the minor subpopulation. This aspect strongly affects also the criteria for Bayesian model comparison. Table 2 shows that in the third scenario both the DIC and BPIC are minimized by the homogeneous model, suggesting a mild evidence of heterogeneity induced by the strong similarity of the two generating mixture components. However, we
verified that with a double sample size \((N = 600)\) in Scenario 3 the 2-component PL mixture is fully recovered by the selection criteria (results not shown). For Scenario 1 and 2, instead, all the criteria recognize the true number of groups.

### 6.2. The HPQ data

In this section we illustrate the Bayesian PL mixture model with an application to the real data from the Hamburger Preparation Quiz (HPQ), carried out as part of the Menu Census Survey organized by the Market Research Corporation of America during the period March 1996−February 1997. The HPQ data set collects \(N = 594\) complete rankings of \(K = 5\) hamburger cooking methods ordered by the respondents according to their taste preferences. Hamburger preparation types, labeled from 1 to 5, are respectively: rare, medium-rare, medium, medium-well and well-done. The same data set has been previously analyzed by Gormley and Murphy (2010), who evaluated the possible information contribution provided by the inclusion of socio-demographic covariates in the mixture setting, and by Bao and Meil˘ a (2008), where sample heterogeneity was investigated by means of a nonparametric approach. In our Bayesian analysis we first set weakly/noninformative hyperparameters values for the prior specification \((c_{gi} = 1, d_g = .001\) and \(\alpha_g = 1\)) and implemented the EM algorithm to obtain the MAP estimates for the PL mixtures, with a number of components varying from \(G = 1\) up to \(G = 7\). We subsequently employed the MAP solution to initialize the GS procedures. A total of 15000 iterations, after a 5000 burn-in period, has been considered adequate subsequently to a positive graphical inspection of the algorithm.
Table 4. MAP estimates and posterior means of the relabeled MCMC samples for the best Bayesian PL mixture fitted to the HPQ data. Ordered sequences indicate the component-specific modal profiles, whereas posterior standard deviations are shown in parentheses.

| Estimation | $g$ | $\hat{\omega}_g$ | $\hat{p}_{g1}$ | $\hat{p}_{g2}$ | $\hat{p}_{g3}$ | $\hat{p}_{g4}$ | $\hat{p}_{g5}$ |
|------------|-----|------------------|----------------|----------------|----------------|----------------|----------------|
| MAP        | 1   | .412             | (.00)          | (.00)          | (.00)          | (.255)         | (.736)         |
|            | 2   | .238             | (.006)         | (.262)         | (.594)         | (.132)         | (.006)         |
|            | 3   | .013             | (.000)         | (.988)         | (.000)         | (.012)         | (.000)         |
|            | 4   | .177             | (.000)         | (.002)         | (.378)         | (.613)         | (.007)         |
|            | 5   | .160             | (.316)         | (.454)         | (.220)         | (.009)         | (.001)         |
| Relabeled GS | 1 | .389 (.02) | (.000 (<.01)) | (.000 (<.01)) | (.007 (<.01)) | (.245 (.03)) | (.747 (.03)) |
|            | 2 | .234 (.03) | (.008 (<.01)) | (.262 (.04)) | (.594 (.04)) | (.129 (.03)) | (.007 (<.01)) |
|            | 3 | .025 (<.01) | (.136 (.06)) | (.384 (.13)) | (.028 (.03)) | (.351 (.10)) | (.102 (.06)) |
|            | 4 | .197 (.03) | (.001 (<.01)) | (.010 (<.01)) | (.356 (.05)) | (.596 (.05)) | (.037 (.02)) |
|            | 5 | .155 (.02) | (.319 (.05)) | (.450 (.04)) | (.219 (.04)) | (.012 (<.01)) | (.001 (<.01)) |

convergence, supported by the good mixing of the traceplots and the weak sample autocorrelation. The marginal traceplots revealed the presence of LS in the posterior MCMC samples. Similarly to the simulation study we applied alternative relabeling strategies to the GS outputs.

DIC and PBIC values for all the fitted mixture models are shown in Table 3 and consistently select $G = 5$ components. The evidence of sample heterogeneity is reinforced also by the posterior predictive check conducted with the novel discrepancy measure introduced in Section 5 whose results are shown in the last column of Table 3. The null posterior predictive $p$-value $p_B$ associated to the basic PL indicates the clear reject of the homogeneity assumption. Notice also that $p_B$ increases substantially up to $G = 4$ and then it remains essentially unchanged around 48%, revealing a very good fit of the final 5-component PL mixture.

Corresponding parameter estimates obtained from the MAP procedure and from the posterior means of the relabeled MCMC samples are shown in Table 4. Support parameter estimates are also represented via mosaic plots in Figure 4. It is interesting to compare the optimal Bayesian PL mixture with the ML inference performed by Gormley and Murphy (2010) which, in the present uninformative prior setting, is very similar to our MAP estimates. The four main clusters recognized by our Bayesian model essentially agree with those pointed out by Gormley and Murphy (2010), both for the estimated size and for the group-specific preference patterns. Nevertheless, some differences can be highlighted. The final Bayesian model selected by the DIC and BPIC turns out to be more parsimonious because it identifies a single component (the first one in Table 4) to represent those sample units who strongly prefer the well-done cooking type (Item 5) and, as second best choice, the medium-done hamburger (Item 4). Gormley and Murphy (2010), instead, identify two components with this preference profile in their analysis. Another difference concerns the group with smaller size, which is labeled as the third cluster in Table 4. In Gormley and Murphy (2010) this component is characterized by the exclusive preference for the medium-rare hamburger (Item 2) and a substantial indifference towards the remaining alternatives. The GS estimates, instead, describe a more assorted liking profile with larger support to both Item 2 and 4. The last choice,
Figure 4. Mosaic plots for the support parameter estimates of the best Bayesian PL mixture fitted to the HPQ data. Bar widths are proportional to the estimated group weights.

Table 5. DIC, BPIC and posterior predictive $p-$values for the Bayesian PL mixtures with varying number of components fitted to the CARCONF data. Optimal values of the criteria are indicated in bold.

| $G$ | $\text{DIC}_1$ | $\text{DIC}_2$ | $\text{BPIC}_1$ | $\text{BPIC}_2$ | $\text{BIC}$ | $\rho_B$ |
|-----|-----------------|-----------------|-----------------|-----------------|--------------|----------|
| 1   | 4969.16         | 4969.13         | 4974.21         | 4974.15         | 4989.44      | 0.051    |
| 2   | **4950.14**     | **4951.48**     | **4961.68**     | **4964.35**     | 4993.90      | 0.117    |
| 3   | 4957.38         | 4953.85         | 4979.75         | 4972.69         | 5015.92      | 0.161    |
| 4   | 4966.50         | 4953.72         | 4999.44         | 4973.87         | 5040.36      | 0.177    |

with very low estimated support equal to .028, is the medium cooking (Item 3). Compared to the other clusters, this one exhibits a peculiar pattern since most of the support is not placed on contiguous/similar levels of hamburger doneness.

6.3. The CARCONF data. The second application concerns a marketing study described in [Dabic and Hatzinger (2009)](Dabic and Hatzinger (2009)), aimed at investigating customer preferences towards different car features. The car configurator (CARCONF) data set consists of $N = 435$ top orderings and is available in the R package [prefmod](Hatzinger and Dittrich, 2012). Customers were asked to construct their car using an online configurator system. The respondents were presented a set of $K = 6$ car modules to carry out their personal arrangement, namely exterior and interior design, technical equipment, brand, price and producing country. The survey did not require a complete ranking elicitation, therefore the sample is composed of partial top orderings with varying lengths. The distribution of the number of missing...
bottom positions is illustrated in Figure 5 and reveals that for the most part the preference profiles are complete, with 365 (84%) of the sample units who provided a full ranking of all the car modules. Among the partial observations 43 (10%) subjects left only two bottom positions undefined. A single unit, instead, provided only her most liked car feature. The vector (42,17,0,29,62) lists the number of missing responses for each item and points out that all respondents assigned a rank to the brand (Item 3), whereas the producing country (Item 5) is the one whose exact position is more frequently lacking (62 occurrences corresponding to 89% of the total number of incomplete responses). Item 5 is also associated with the lowest mean rank, as indicated by the fifth entry of the average rank vector $\bar{\pi} = (3.56, 2.88, 3.17, 3.11, 4.49, 3.20)$. The graphical inspection of the marginal rank distributions for each item, reported in the form of c.d.f. in Figure 5, provides additional details on the overall preferences. We can note that the c.d.f. for Item 5 is remarkably stochastically dominated by the other ones, reinforcing the suspect of a minor global interest in the car production country. An important aspect to be highlighted is also the presence of intersections among the c.d.f.’s, that can be interpreted as an empirical violation to the assumption of an underlying homogeneous PL, under which the rank distributions are instead expected to be marginally stochastically ordered (Marden, 1995). The observed behavior of the rank distributions could be explained with the existence of differential preference patterns in the sample. Rather than with a basic PL, these could be adequately captured by an underlying group structure with unknown number of groups.

We performed an uninformative Bayesian PL mixture analysis to the CARCONF data in a similar fashion as described for the HPQ data set, fitting up to $G =$
Table 6. Posterior means of the relabeled MCMC samples for the best Bayesian 2-component PL mixture fitted to the CARCONF data. Ordered sequences indicate the component-specific modal profiles, whereas posterior standard deviations are shown in parentheses.

| Estimation | \( \omega_g \) | \( \sigma_g^{-1} \) | \( \hat{p}_{g1} \) | \( \hat{p}_{g2} \) | \( \hat{p}_{g3} \) | \( \hat{p}_{g4} \) | \( \hat{p}_{g5} \) | \( \hat{p}_{g6} \) |
|------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Relabeled GS 1 | .702 (.11) | (2,6,4,3,1,5) | .085 (.02) | .258 (.02) | .165 (.02) | .195 (.02) | .080 (<.01) | .217 (.02) |
| 2 | .298 (.11) | (1,4,3,2,6,5) | .448 (.12) | .130 (.04) | .134 (.04) | .145 (.03) | .045 (.02) | .098 (.03) |

Figure 6. Mosaic plots for the support parameter estimates of the best Bayesian PL mixture fitted to the CARCONF data. Bar widths are proportional to the estimated group weights.

4 components. Bayesian selection criteria are shown in Table 5, together with the BIC in order to make a comparison with the frequentist solution. The BIC does not recognize the existence of an underlying group structure despite the large sample size, whereas DIC and BPIC values agree in selecting the 2-component PL mixture as the optimal solution. Corresponding parameter estimates, after a suitable adjustment for the LS phenomenon, are shown in Table 6 and represented via a mosaic plot in Figure 6. Our analysis suggests the presence of a major cluster, whose weight is estimated as 0.702. This cluster is comprised of those customers who are mainly interested in the aesthetic car features, such as exterior and interior design. The minor group, instead, is characterized by a great attention in the economic aspect, expressed by the price. Both groups share a minor interest in the production country. These results seem to better agree with the typical preference patterns observed in the car market than the MLE solution. Bayesian predictive p-values are reported in the last column of Table 5 and point out evidence in favor of heterogeneity over the basic PL, that provides a remarkable poor description of the CARCONF data.
7. Concluding remarks and future work

We detailed a Bayesian finite PL mixture and described efficient algorithms to conduct approximate inference. The proposal contemplates a completion method of the sample space with the latent group structure and allows model-based classification of partially top orderings. Our approach can be seen as a direct extension of the basic PL introduced by Caron and Doucet (2012) into the finite mixture context, aimed at identifying and characterizing possible groups of rankers with similar preference/attitudes. At the same time, our contribution can be interpreted as the generalization within the Bayesian domain of the PL mixture employed in Gormley and Murphy (2006), whose frequentist approach can be recovered as a by-product of the uninformative analysis. An important advantage in this perspective lies in the possibility to address estimation uncertainty without relying on large sample approximations and additional computational burden.

Despite our proposal is based on a relatively standard data augmentation approach and estimation procedures, mixture model applications can indeed turn out to be less straightforward in practice, mainly due to identifiability issues that can prevent from a direct posterior inference. The LS phenomenon, in fact, may hamper posterior summaries based on the crude MCMC samples. For this reason, in order to verify the actual utility of the Bayesian PL mixture, much effort of the present work has been devoted to the discussion of these critical aspects and to the attempt to find an adequate solution in several applications to both simulated and real ranking data.

We additionally made use of a diagnostic device to evaluate the fitting of our proposal by means of posterior predictive check. The adopted chi square discrepancy measure is a parameter-dependent quantity that compares the observed pairwise preferences with those expected in future replications of the data under the inferred models. This method allows to compare the realized test statistic with the entire postulated parametric family averaging w.r.t. the posterior distribution, rather than with the single best fitting solution as required by the classical $p$-value.

A possible future development to add further flexibility and improve the ranking modeling could be the introduction of extra information provided by individual or item-specific covariates. As revealed by previous applications in the MLE framework (Gormley and Murphy, 2008, 2010), explanatory variables may fruitfully contribute to characterizing choice patterns and supporting decisions for better satisfying specific preference profiles or segments.

APPENDIX A. COMPLETE-DATA LIKELIHOOD FOR THE PL MIXTURE

Given the special binary characterization of the $z$’s, it is often possible to simplify analytic steps exploiting the following equivalence

$$
\prod_{g=1}^{G} \left( \sum_{\nu = k}^{n_x} p_{g \pi^{-1}_{x}(\nu)} \right) \overset{z_{xg}}{=} \sum_{g=1}^{G} \sum_{\nu = k}^{n_x} p_{g \pi^{-1}_{x}(\nu)}
$$

As indicated in (3.3), one has

$$
L_c(p, \omega, y, z) = P(y, \bar{\pi}^{-1}, z | p, \omega) = f(y | \bar{\pi}^{-1}, z, p, \omega) P(\bar{\pi}^{-1}, z | p, \omega)
$$
where, respectively,
\[ f(y | \pi^{-1}, z_p, \omega) = f(y | \pi^{-1}, z, p) \]
\[ = \prod_{s=1}^{N} \prod_{t=1}^{n_s} \prod_{g=1}^{G} \left( \sum_{\nu_t=1}^{n_s} p_{g, \pi^{-1}_s(\nu)} \right)^{z_{sg}} e^{-\sum_{t=1}^{n_s} y_{st} \sum_{\nu_t=1}^{n_s} p_{g, \pi^{-1}_s(\nu)}} \]
\[ = \prod_{s=1}^{N} \prod_{t=1}^{n_s} \prod_{g=1}^{G} \left( \sum_{\nu_t=1}^{n_s} p_{g, \pi^{-1}_s(\nu)} \right)^{z_{sg}} \]
\[ = \prod_{s=1}^{N} \prod_{g=1}^{G} \left( \sum_{t=1}^{n_s} \sum_{\nu_t=1}^{n_s} p_{g, \pi^{-1}_s(\nu)} \right)^{z_{sg}} \]
\[ \]
References

Alnur Ali, Thomas Brendan Murphy, Marina Meilä, and Harr Chen. Preferences in college applications—a nonparametric Bayesian analysis of top-10 rankings, 2010. URL http://www.cs.umass.edu/~wallach/workshops/nips2010css/papers/ali.pdf. NIPS Workshop on Computational Social Science and the Wisdom of Crowds, 2010 (slides).

Tomohiro Ando. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94(2):443–458, 2007.

Le Bao and Marina Meilä. Clustering permutations by Exponential Blurring Mean-Shift algorithm. Technical Report 524, Dept of Statistics, University of Washington, 2008.

Ludwig M. Busse, Peter Orbanz, and Joachim M. Buhmann. Cluster analysis of heterogeneous rank data. In Zoubin Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning – ICML 2007*, pages 113–120. Omnipress, 2007.

François Caron and Arnaud Doucet. Efficient Bayesian inference for generalized Bradley-Terry models. *J. Comput. Graph. Statist.*, 21(1):174–196, 2012. ISSN 1061-8600. doi: 10.1080/10618600.2012.638220.

François Caron and Yee W. Teh. Bayesian nonparametric models for ranked data. In *Neural Information Processing Systems – NIPS 2012*, pages 1529–1537, 2012.

François Caron, Yee Whye Teh, and Thomas Brendan Murphy. Bayesian nonparametric Plackett-Luce models for the analysis of clustered ranked data, 2012. ArXiv:1211.5037.

François Caron, Yee Whye Teh, and Thomas Brendan Murphy. Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics*, 8(2):1145–1181, 2014.

Gilles Celeux, Merrilee Hurn, and Christian P Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.

Wai Chan and Peter M Bentler. Covariance structure analysis of ordinal ipsative data. *Psychometrika*, 63(4):369–399, 1998.

Ayala Cohen and Colin L. Mallows. Assessing goodness of fit of ranking models to data. *J. Roy. Statist. Soc. Ser. D (The Statistician)*, 32(4):361–374, 1983. ISSN 0039-0526.

Douglas E. Critchlow. *Metric methods for analyzing partially ranked data*, volume 34 of *Lecture Notes in Statistics*. Springer-Verlag, 1985. ISBN 3-540-96288-3. doi: 10.1007/978-1-4612-1106-8.

Douglas E. Critchlow, Michael A. Fligner, and Joseph S. Verducci. Probability models on rankings. *J. Math. Psych.*, 35(3):294–318, 1991. ISSN 0022-2496. doi: 10.1016/0022-2496(91)90050-4.

M. Dabic and R. Hatzinger. Zielgruppenadäquate ablaufe in konfigurationssystemen ?? eine empirische studie im automobilmarkt. In R. Hatzinger, R. Dittrich, and T. Salzberger, editors, *Präferenzanalyse mit R: Anwendungen aus Marketing, Behavioural Finance und Human Resource Management*. Facultas, 2009.

Michael A. Fligner and Joseph Stephen Verducci. Posterior probabilities for a consensus ordering. *Psychometrika*, 55(1):53–63, 1990. ISSN 0033-3123. doi: 10.1007/BF02294743.
Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2004. ISBN 1-58488-388-X.

Isobel Claire Gormley and Thomas Brendan Murphy. Analysis of irish third-level college applications data. *Journal of the Royal Statistical Society: Series A*, 169(2):361–379, 2006. ISSN 0964-1998. doi: 10.1111/j.1467-985X.2006.00412.x.

Isobel Claire Gormley and Thomas Brendan Murphy. A mixture of experts model for rank data with applications in election studies. *Ann. Appl. Stat.*, 2(4):1452–1477, 2008. ISSN 1932-6157. doi: 10.1214/08-AOAS178.

John Guiver and Edward Snelson. Bayesian inference for Plackett-Luce ranking models. In Léon Bottou and Michael Littman, editors, *Proceedings of the 26th International Conference on Machine Learning – ICML 2009*, pages 377–384. Omnipress, 2009. ISBN 978-1-60558-516-1.

Jayanti Gupta and Paul Damien. Conjugacy class prior distributions on metric-based ranking models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):433–445, 2002. ISSN 1369-7412. doi: 10.1111/1467-9868.00343.

Jayanti Gupta and Paul Damien. Prior distributions on symmetric groups. *Psychometrika*, 70(1):169–177, 2005.

Reinhold Hatzinger and Regina Dittrich. prefmod: An r package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software*, 48(10):1–31, 5 2012. ISSN 1548-7660. URL [http://www.jstatsoft.org/v48/i10](http://www.jstatsoft.org/v48/i10).

David R. Hunter. MM algorithms for generalized Bradley-Terry models. *Ann. Statist.*, 32(1):384–406, 2004. ISSN 0090-5364. doi: 10.1214/aos/1079120141.

A Jasra, CC Holmes, and DA Stephens. Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.

Timothy R Johnson and Kristine M Kuhn. Bayesian Thurstonian models for ranking data using JAGS. *Behavior research methods*, 45(3):857–872, 2013.

Guy Lebanon and Yi Mao. Non-parametric modeling of partially ranked data. *J. Mach. Learn. Res.*, 9:2401–2429, 2008. ISSN 1532-4435.

R. D. Luce. *Individual choice behavior: A theoretical analysis*. John Wiley & Sons Inc., 1959.

John I. Marden. *Analyzing and modeling rank data*, volume 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1995. ISBN 0-412-99521-2.

Jean-Michel Marin, Kerrie Mengersen, and Christian P Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, 25:459–507, 2005.
Marina Meilà and Harr Chen. Dirichlet process mixtures of Generalized Mallows models. In Peter D. Grünwald and Peter Spirtes, editors, *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence – UAI 2010*, pages 358–367. AUAI Press, 2010.

Xiao-Li Meng. Posterior predictive p-values. *The Annals of Statistics*, pages 1142–1160, 1994.

Panagiotis Papastamoulis. *label.switching: Relabelling MCMC outputs of mixture models*, 2014. URL http://CRAN.R-project.org/package=label.switching R package version 1.3.

Panagiotis Papastamoulis and George Iliopoulos. An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics*, 19(2), 2010.

Robin L. Plackett. The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, 24(2):193–202, 1975. ISSN 0035-9254.

Martyn Plummer et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, pages 20–22, 2003.

Carlos E Rodríguez and Stephen G Walker. Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45, 2014.

Matthew Sperrin, Thomas Jaki, and Ernst Wit. Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20(3):357–366, 2010.

David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

Grace Yao and Ulf Böckenholt. Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, 52(1):79–92, 1999.

Philip L. H. Yu. Bayesian analysis of order-statistics models for ranking data. *Psychometrika*, 65(3):281–299, 2000.