Bayesian nonparametric Plackett-Luce models for the analysis of clustered ranked data

Francois Caron, Yee Whye Teh, Thomas Brendan Murphy

Bayesian nonparametric Plackett-Luce models for the analysis of clustered ranked data

François Caron, Yee Whye Teh, Thomas Brendan Murphy

RESEARCH
REPORT
N° 8143
November 22, 2012
Project-Team ALEA

ISSN 0249-6399 ISRN INRIA/RR--8143--FR+ENG

Bayesian nonparametric Plackett-Luce models for the analysis of clustered ranked data

François Caron[*], Yee Whye Teh[†], Thomas Brendan Murphy[‡]

Project-Team ALEA

Research Report n° 8143 — November 22, 2012 — 28 pages

Abstract: In this paper we propose a Bayesian nonparametric model for clustering partial ranking data. We start by developing a Bayesian nonparametric extension of the popular Plackett-Luce choice model that can handle an infinite number of choice items. Our framework is based on the theory of random atomic measures, with prior specified by a completely random measure. We characterise the posterior distribution given data, and derive a simple and effective Gibbs sampler for posterior simulation. We then develop a Dirichlet process mixture extension of our model and apply it to clustering the preferences for university programmes of Irish secondary school graduates.

Key-words: choice models, generalized Bradley-Terry model, Plackett-Luce model, completely random measure, Mixture model, Dirichlet process, Markov Chain Monte Carlo

[*] INRIA, Institut de Mathématiques de Bordeaux, University of Bordeaux, Talence, France
[†] Department of Statistics, University of Oxford, Oxford, United Kingdom
[‡] School of Mathematical Sciences, University College Dublin, Dublin, Ireland

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

351, Cours de la Libération
Bâtiment A 29
33405 Talence Cedex

Modèles de Plackett-Luce bayésiens non paramétriques pour l'analyse de regroupement de données de rangs

Résumé : Dans cet article nous proposons un modèle bayésien non paramétrique pour la classification non supervisée de données de rangs partiels. On s'intéresse dans un premier temps à développer une extension bayésienne non paramétrique du modèle de Plackett-Luce pouvant traiter un nombre potentiellement infini d'éléments. Notre cadre se base sur la théorie des mesures complètement aléatoires, avec comme a priori une mesure complètement aléatoire. Nous dérivons une caractérisation de la loi a posteriori et un échantillonneur de Gibbs simple pour approcher la loi a posteriori. Nous développons ensuite une extension de notre modèle utilisant des processus de Dirichlet à mélange, et l'appliquons à la classification non supervisée des préférences pour les programmes universitaires de diplômés irlandais du secondaire.

Mots-clés : Modèles de choix, modèle de Bradley-Terry généralisé, modèle de Plackett-Luce, mesure complètement aléatoire, méthodes de Monte Carlo par chaîne de Markov, modèle de mélange, processus de Dirichlet

1 Introduction

In this paper we consider partial ranking data consisting of ordered lists of the top-m items among a set of objects. Data in the form of partial rankings arise in many contexts. For example, in this paper we shall consider data pertaining to the top ten preferences of Irish secondary school graduates for university programmes. The Plackett-Luce model (Luce, 1959; Plackett, 1975) is a popular model for modeling such partial rankings of a finite collection of M items. It has found many applications, including choice modeling (Luce, 1977; Chapman and Staelin, 1982), sport ranking (Hunter, 2004), and voting (Gormley and Murphy, 2008). Diaconis (1988, Chapter 9) provides detailed discussions on the statistical foundations of this model.

In the Plackett-Luce model, each item $k \in [M] = \{1, \dots, M\}$ is assigned a positive rating parameter w_k, which represents the desirability or rating of a product in the case of choice modeling, or the skill of a player in sport rankings. The Plackett-Luce model assumes the following generative story for a top-m list $\rho = (\rho_1, \dots, \rho_m)$ of items $\rho_i \in [M]$: At each stage $i = 1, \dots, m$, an item is chosen to be the ith item in the list from among the items that have not yet been chosen, with the probability that ρ_i is selected being proportional to its desirability w_{ρ_i}. The overall probability of a given partial ranking ρ is then

$$P(\rho) = \prod_{i=1}^{m} \frac{w_{\rho_i}}{\left(\sum_{k=1}^{M} w_k\right) - \left(\sum_{j=1}^{i-1} w_{\rho_j}\right)}, \tag{1}$$

with the denominator in (1) being the sum over all items not yet selected at stage i.

In many situations the collection of available items can be very large and/or potentially unknown. In this case a nonparametric approach can be sensible, where the pool of items is assumed to be infinite and the model allows for the possibility of items not observed in previous top-m lists to appear in future ones. A naïve approach, building upon recent work on Bayesian inference for the (finite) Plackett-Luce model and its extensions (Gormley and Murphy, 2009; Guiver and Snelson, 2009; Caron and Doucet, 2012), is to first derive a Markov chain Monte Carlo sampler for the finite model, then to "take the infinite limit" of the sampler, where the number of available items becomes infinite, but such that all unobserved items are grouped together for computational tractability.

Such an approach, outlined in Section 2, is reminiscent of a number of previous approaches deriving the (Gibbs sampler for the) Dirichlet process mixture model as the infinite limit of (a Gibbs sampler for) finite mixture models (Neal, 1992; Rasmussen, 2000; Ishwaran and Zarepour, 2002). Although intuitively appealing, this is not a satisfying approach since it is not clear what the underlying nonparametric model actually is, as it is actually the algorithm whose infinite limit was taken. It also does not directly lead to more general and flexible nonparametric models with no obvious finite counterpart, nor does it lead to alternative perspectives and characterisations of the same model, or resultant alternative inference algorithms. Orbanz (2010) further investigates the approach of constructing nonparametric Bayesian models from finite dimensional parametric Bayesian models.

Caron and Teh (2012) recently proposed a Bayesian nonparametric Plackett-Luce model based on a natural representation of items along with their ratings as an atomic measure. Specifically, the model assumes the existence of an infinite pool of items $\{X_k\}_{k=1}^{\infty}$, each with its own rating parameter, $\{w_k\}_{k=1}^{\infty}$. The atomic measure then consists of an atom located at each X_k with a mass of w_k:

$$G = \sum_{k=1}^{\infty} w_k \delta_{X_k}. \tag{2}$$

The probability of a top-m list of items, say $(X_{\rho_1}, \dots, X_{\rho_m})$, is then a direct extension of the finite case (1):

$$P(X_{\rho_1}, \dots, X_{\rho_m} | G) = \prod_{i=1}^{m} \frac{w_{\rho_i}}{\left(\sum_{k=1}^{\infty} w_k\right) - \left(\sum_{j=1}^{i-1} w_{\rho_j}\right)}. \tag{3}$$

Using this representation, note that the top item X_{ρ_1} in the list is simply a draw from the probability measure obtained by normalising G, while subsequent items in the top-m list are draws from probability measures obtained by first removing from G the atoms corresponding to previously picked items and normalising. Described this way, it is clear that the Plackett-Luce model is none other than a partial size-biased permutation of the atoms in G (Patil and Taillie, 1977), and the existing machinery of random measures and exchangeable random partitions (Pitman, 2006) can be brought to bear on our problem.

For example, we may use a variety of existing stochastic processes to specify a prior over the atomic measure G. Caron and Teh (2012) considered the case, described in Section 3, where G is a gamma process. This is a completely random measure (Kingman, 1967) with gamma marginals, such that the corresponding normalised probability measure is a Dirichlet process (Ferguson, 1973). They showed that with the introduction of a suitable set of auxiliary variables, it is possible to characterise the posterior law of G given observations of top-m lists distributed according to (3). A simple Gibbs sampler can then be derived to simulate from the posterior distribution which corresponds to the infinite limit of the Gibbs sampler for finite models.

In Section 4, we show that this construction can be extended from gamma processes to general completely random measures, and we discuss extensions of the Gibbs sampler to this more general case. In particular, we show that a simple Gibbs sampler can still be derived for the generalised gamma class of completely random measures.

In Section 5 we describe a Dirichlet process mixture model (Ferguson, 1973; Lo, 1984) for heterogeneous partial ranking data, where each mixture component is a gamma process nonparametric Plackett-Luce model. As we will see, in this model it is important to allow the same atoms to appear across the different random measures of the mixture components, otherwise the model becomes degenerate with all observed items that ever appeared together in some partial ranking being assigned to the same mixture component. To allow for this, we use a tree-structured extension of the time varying model of Caron and Teh (2012). In Section 6 we apply this mixture model to the Irish university preferences data mentioned earlier, showing that the model is able to recover clusters of students with similar and interpretable preferences.

Finally, we conclude in Section 7 with a discussion of the important contributions of this paper and proposals for future work.

2 A Bayesian nonparametric model for partial ranking

We start this section with a review of a Bayesian approach to inference in finite Plackett-Luce models (Gormley and Murphy, 2009; Guiver and Snelson, 2009; Caron and Doucet, 2012), and taking the infinite limit to arrive at a nonparametric model. This will give good intuitions for how the model operates, before we rederive the same nonparametric model more formally in the next section using gamma processes.

Recall that we have M choice items indexed by $[M] = \{1, \dots, M\}$, with item $k \in [M]$ having a positive desirability parameter w_k. We will suppose that our data consists of L partial rankings of the M choice items, with the ℓth ranking being denoted $\rho_\ell = (\rho_{\ell 1}, \dots, \rho_{\ell m})$, for $\ell = 1, \dots, L$, where each $\rho_{\ell i} \in [M]$. For notational simplicity we assume that all the partial rankings are of length m.

2.1 Finite Plackett-Luce model with gamma prior

As noted in the introduction, the Plackett-Luce model constructs a partial ranking $\rho_\ell = (\rho_{\ell 1}, \dots, \rho_{\ell m})$ iteratively. At the ith stage, with $i = 1, 2, \dots, m$, we pick $\rho_{\ell i}$ as the ith item from among those not yet picked with probability proportional to $w_{\rho_{\ell i}}$. The probability of the partial ranking ρ_ℓ is then as given in (1). An alternative Thurstonian interpretation, which will be important in the following, is as follows:

For each item k let $z_{\ell k}$ be exponentially distributed with rate w_k:

$$z_{\ell k} \sim \mathrm{Exp}(w_k)$$

Thinking of $z_{\ell k}$ as the arrival time of item k in a race, let $\rho_{\ell i}$ be the index of the ith item to arrive (the index of the ith smallest value among $(z_{\ell k})_{k=1}^M$). The resulting probability of the first m items to arrive being ρ_ℓ can be shown to be the probability (1) from before. In this interpretation $(z_{\ell k})$ can be understood as latent variables, and the EM algorithm (Dempster et al., 1977) can be applied to derive an algorithm to find a ML setting for the parameters $(w_k)_{k=1}^M$ given multiple partial rankings. Unfortunately the posterior distribution of $(z_{\ell k})_{k=1}^M$ given ρ_ℓ is difficult to compute, so we can instead consider an alternative parameterisation: Let $Z_{\ell i}$ be the waiting time for the ith item to arrive after the $i-1$th item. That is,

$$Z_{\ell i} = z_{\rho_{\ell i}} - z_{\rho_{\ell\, i-1}}$$

with $z_{\rho_{\ell 0}}$ defined to be 0. Then it is easily seen that the joint probability of the observed partial rankings, along with the alternative latent variables $(Z_{\ell i})$, is:

$$P((\rho_\ell)_{\ell=1}^L, ((Z_{\ell i})_{i=1}^m)_{\ell=1}^L | (w_k)_{k=1}^M) = \prod_{\ell=1}^L \prod_{i=1}^m w_{\rho_{\ell i}} \exp\left(-Z_{\ell i}\left(\sum_{k=1}^M w_k - \sum_{j=1}^{i-1} w_{\rho_{\ell j}}\right)\right) \tag{4}$$

In particular, the posterior of $(Z_{\ell i})_{i=1}^m$ is simply factorised, with

$$Z_{\ell i} | (\rho_\ell)_{\ell=1}^L, (w_k)_{k=1}^M \sim \mathrm{Exp}\left(\sum_{k=1}^M w_k - \sum_{j=1}^{i-1} w_{\rho_{\ell j}}\right)$$

being exponentially distributed. The M step of the EM algorithm can be easily derived as well. The resulting algorithm was first proposed by Hunter (2004) as an instance of the MM (majorisation-maximisation) algorithm (Lange et al., 2000) and its re-interpretation as an EM algorithm was recently given by Caron and Doucet (2012).

Taking a further step, we note that the joint probability (4) is conjugate to a factorised gamma prior over the parameters, say $w_k \sim \mathrm{Gamma}(\frac{\alpha}{M}, \tau)$ with hyperparameters $\alpha, \tau > 0$. Now Bayesian inference can be carried out, for example, using with a variational Bayesian EM algorithm, or a Gibbs sampler. In this paper we shall consider only Gibbs sampling algorithms. By regrouping the terms in the exponential in (4), the parameter updates are derived to be (Caron and Doucet, 2012):

$$w_k | \rho, (Z_{\ell i}), (w_{k'})_{k' \neq k} \sim \mathrm{Gamma}\left(\frac{\alpha}{M} + n_k, \tau + \sum_{\ell=1}^L \sum_{i=1}^m \delta_{\ell i k} Z_{\ell i}\right) \tag{5}$$

where n_k is the number of occurrences of item k among the observed partial rankings, and

$$\delta_{\ell i k} = \begin{cases} 0 & \text{if there is a } j < i \text{ with } \rho_{\ell j} = k, \\ 1 & \text{otherwise.} \end{cases}$$

Note that the definitions of n_k and $\delta_{\ell i k}$ slightly differ from those in (Hunter, 2004) and (Caron and Doucet, 2012). In these articles, the authors consider full m-rankings of subsets of $[M]$ whereas we consider here partial top-m rankings of all M items.

2.2 Taking the infinite limit

A Gibbs sampler for a nonparametric Plackett-Luce model can now be easily derived by taking the limit as the number of choice items $M \to \infty$. If item k has appeared among the observed partial rankings, the limiting conditional distribution (5) is well defined since $n_k > 0$. For items that did not appear in the observations, (5) becomes degenerate at 0. Instead we can define $w_* = \sum_{k:n_k=0} w_k$ to be the total desirability among all the infinitely many unobserved items. Making use of the fact that sums of independent gammas with the same scale parameter is a gamma with shape parameter given by the sum of the shape parameters,

$$w_*|\rho, (Z_{\ell i}), (w_k)_{k:n_k>0} \sim \text{Gamma}\left(\alpha, \tau + \sum_{\ell=1}^{L}\sum_{i=1}^{m} Z_{\ell i}\right)$$

The resulting Gibbs sampler alternates between updating the latent variables $(Z_{\ell i})$, and updating the desirabilities of the observed items $(w_k)_{k:n_k>0}$ and of the unobserved ones w_*.

This nonparametric model allows us to estimate the probability of seeing new items appearing in future partial rankings in a coherent manner. While intuitive, the derivation is ad hoc in the sense that it arises as the infinite limit of the Gibbs sampler for finite Plackett-Luce models, and is unsatisfying as it did not directly capture the structure of the underlying infinite dimensional object, which we will show in the next section to be a gamma process.

3 A Bayesian nonparametric Plackett-Luce model

Let $\mathbb{X}$ be a measurable space of choice items. A gamma process is a completely random measure over $\mathbb{X}$ with gamma marginals. Specifically, it is a random atomic measure of the form (2), such that for each measurable subset A, the (random) mass $G(A)$ is gamma distributed. Assuming that G has no fixed atoms (that is, for each element $x \in \mathbb{X}$ we have $G(\{x\}) = 0$ with probability one) and that the atom locations $\{X_k\}$ are independent of their masses $\{w_k\}$ (that is, the gamma process is homogeneous), it can be shown that such a random measure can be constructed as follows (Kingman, 1967): each X_k is iid according to a base distribution H (which we assume is non-atomic with density $h(x)$), while the set of masses $\{w_k\}$ is distributed according to a Poisson process over $\mathbb{R}^+$ with mean intensity

$$\lambda(w) = \alpha w^{-1} e^{-w\tau}$$

where $\alpha > 0$ is the concentration parameter and $\tau > 0$ the inverse scale. We write this as $G \sim \Gamma(\alpha, \tau, H)$. Under this parametrisation, we have that $G(A) \sim \text{Gamma}(\alpha H(A), \tau)$. The intensity $\lambda(w)$ is known as the Lévy intensity and plays a significant role in characterising the properties of the gamma process.

We shall interpret each atom X_k as a choice item, with its mass $w_k > 0$ corresponding to the desirability parameter. The Thurstonian view described in the finite model can be easily extended to the nonparametric one, where a partial ranking $(X_{\rho_1}, \dots, X_{\rho_m})$ can be generated as the first m items to arrive in a race. In particular, for each atom X_k let $z_k \sim \text{Exp}(w_k)$ be the time of arrival of X_k and X_{ρ_i} the ith item to arrive. The first m items to arrive $(X_{\rho_1}, \dots, X_{\rho_m})$ then constitutes our partial ranking, with probability as given in (3). This construction is depicted on Figure 1. The top row of Figure 2 visualises some top-5 rankings generated from the model, with $\tau = 1$ and different values of α. Figure 3 shows the mean number of items appearing in L top-m rankings. For $m = 1$, one recovers the well-known result on the number of clusters for a Dirichlet process model.

Again reparametrising using inter-arrival durations, let $Z_i = z_{\rho_i} - z_{\rho_{i-1}}$ for $i = 1, 2, \dots$ (with $z_{\rho_0} = 0$). The joint probability of an observed partial ranking of length m along with the m associated latent variables can be derived to be:

$$\begin{aligned}&P((X_{\rho_1},\dots,X_{\rho_m}),(Z_1,\dots,Z_m)|G)\\ =&P((z_{\rho_1},\dots,z_{\rho_m}),\text{and } z_k > z_{\rho_m} \text{ for all } k\notin\{\rho_1,\dots,\rho_m\})\\ =&\left(\prod_{i=1}^m w_{\rho_i}\exp(-w_{\rho_i}z_{\rho_i})\right)\left(\prod_{k\notin\{\rho_1,\dots,\rho_m\}}\exp(-w_k z_{\rho_m})\right)\\ =&\prod_{i=1}^m w_{\rho_i}\exp\left(-Z_i\left(\sum_{k=1}^\infty w_k-\sum_{j=1}^{i-1}w_{\rho_j}\right)\right)\end{aligned} \tag{6}$$

Marginalising out $(Z_1,\dots,Z_m)$ gives the probability of $(X_{\rho_1},\dots,X_{\rho_m})$ as in (3). Further, conditional on $\rho=(\rho_i)_{i=1}^m$ it is seen that the inter-arrival durations $Z_1\dots Z_m$ are mutually independent, with

$$Z_i|(X_{\rho_1},\dots,X_{\rho_m}),G\sim\mathrm{Exp}\left(\sum_{k=1}^\infty w_k-\sum_{j=1}^{i-1}w_{\rho_j}\right)$$

In the next section we shall characterise the posterior distribution over G given observed partial rankings and their associated latent variables. We end this subsection with two observations. Firstly, note that the Lévy intensity $\lambda(w)$ of the gamma process satisfies the following two properties:

$$\int_0^\infty \lambda(w)dw=\infty, \qquad\qquad \int_0^\infty(1-e^{-w})\lambda(w)dw<\infty \tag{7}$$

The first property is equivalent (via Campbell's Theorem) to the fact that there are an infinite number of atoms in G with probability one. In other words that we are dealing with a nonparametric model with an infinite number of choice items. The second is equivalent to the fact that G has finite total mass with probability one, so that it is a well-defined operation to pick an item with probability proportional to its rating parameter, as in the generative story for the Plackett-Luce model.

The second observation is with regard to a subtle but important difference between the atomic measure approach described in this section and the finite Plackett-Luce model of the previous section. In particular, here we specified the choice items X_k as locations in a space $\mathbb{X}$ with a prior given by the base distribution H, while in the finite Plackett-Luce model we simply index the M choice items using $1,\dots,M$. One may wonder if it is possible to simply index the infinitely many choice items using the natural numbers, and dispense with the atom locations $\{X_k\}$ altogether. This turns out to be impossible, if we were to make the following reasonable assumptions: That item desirabilities are a priori mutually independent, that they are positive with probability one, and that item desirabilities do not depend on the index of their corresponding items. With these assumptions, along with an infinite number of choice items, it is easy to see that the sum of all item desirabilities will be infinite with probability one, so that the Plackett-Luce generative model becomes ill-defined. Using the atomic measure approach, it is possible to satisfy all assumptions while making sure the Plackett-Luce generative model is well-defined.

3.1 Posterior characterisation

In this section we develop a characterisation of the posterior law of G under a gamma process prior and given Plackett-Luce observations consisting of L partial rankings. We shall denote the ℓth partial ranking as $Y_\ell=(Y_{\ell 1},\dots,Y_{\ell m})$, where each $Y_{\ell i}\in\mathbb{X}$. Note that previously our partial rankings $(X_{\rho_1},\dots,X_{\rho_m})$ were denoted as ordered lists of the atoms in G. Since G is unobserved here, this is no longer possible, so we instead simply use a list of observed choice items $(Y_{\ell 1},\dots,Y_{\ell m})$. Re-expressing the conditional

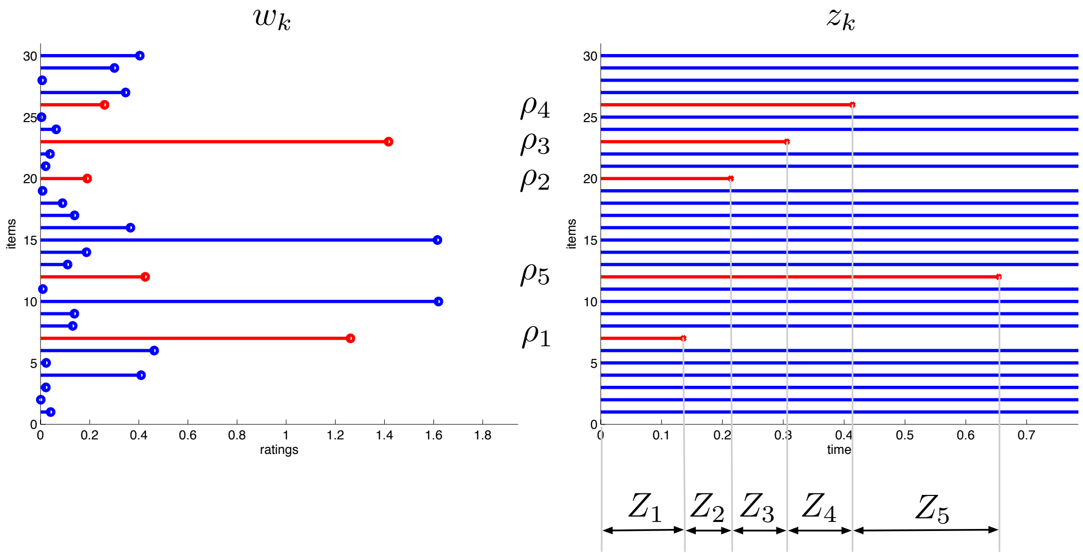


Figure 1: Bayesian nonparametric Plackett-Luce model. Left: an instantiation of the atomic measure G encapsulating both the items and their ratings. Right: Arrival times z_k and latent variables $Z_k = z_{\rho_k} - z_{\rho_{k-1}}$. The top 5 items are $(\rho_1, \rho_2, \dots, \rho_5)$.

distribution (3) of Y_ℓ given G, we have:

$$P(Y_\ell|G) = \prod_{i=1}^{m} \frac{G(\{Y_{\ell i}\})}{G(\mathbb{X}\backslash\{Y_{\ell 1}, \dots, Y_{\ell\, i-1}\})}$$

In addition, for each ℓ, we will also introduce a set of auxiliary variables $Z_\ell = (Z_{\ell 1}, \dots, Z_{\ell m})$ (the inter-arrival times) that are conditionally mutually independent given G and Y_ℓ, with:

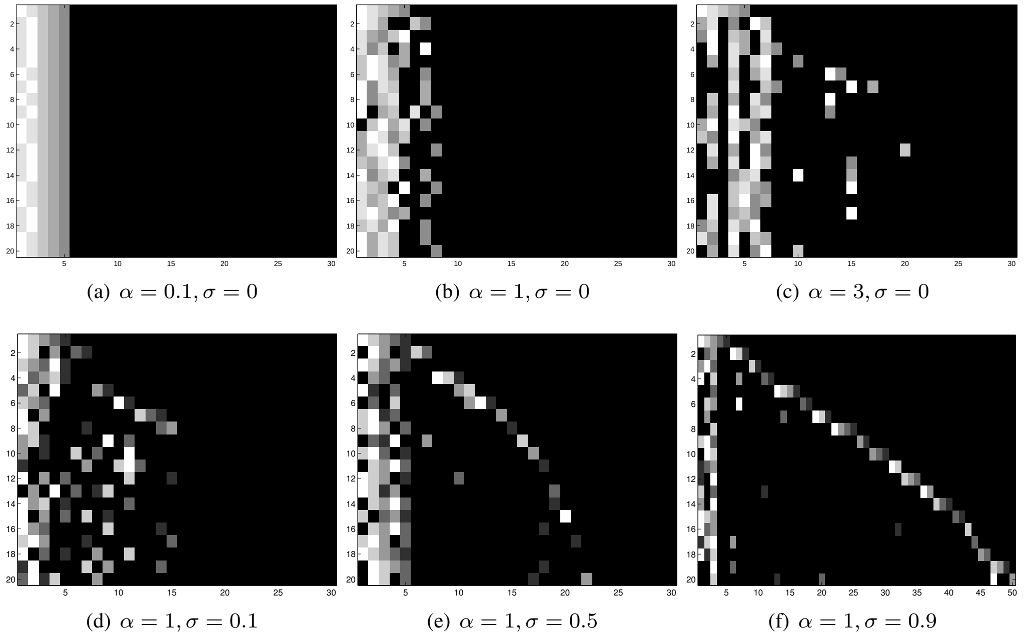
$$Z_{\ell i}|Y_\ell, G \sim \text{Exp}(G(\mathbb{X}\backslash\{Y_{\ell 1}, \dots, Y_{\ell i-1}\})) \tag{8}$$

The joint probability of the item lists and auxiliary variables is then (c.f. (6)):

$$P((Y_\ell, Z_\ell)_{\ell=1}^L|G) = \prod_{\ell=1}^{L}\prod_{i=1}^{m} G(\{Y_{\ell i}\}) \exp(-Z_{\ell i} G(\mathbb{X}\backslash\{Y_{\ell 1}, \dots, Y_{\ell\, i-1}\}))$$

Note that under the generative process described in Section 3, there is positive probability that an item appearing in a list Y_ℓ appears in another list $Y_{\ell'}$ with $\ell' \neq \ell$. Denote the unique items among all L lists by $X_1^*, \dots, X_K^*$, and for each $k = 1, \dots, K$ let n_k be the number of occurrences of X_k^* among the item lists. Finally define occurrence indicators

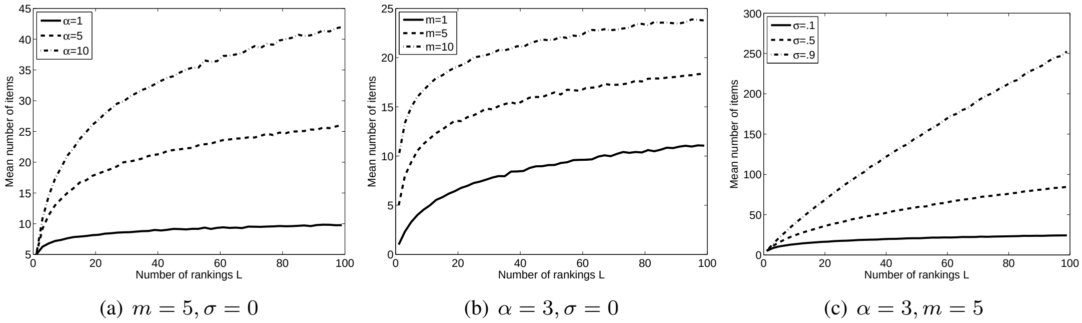
$$\delta_{\ell ik} = \begin{cases} 0 & \text{if } \exists j < i \text{ with } Y_{\ell j} = X_k^*; \\ 1 & \text{otherwise.} \end{cases} \tag{9}$$



(a) $\alpha = 0.1, \sigma = 0$ (b) $\alpha = 1, \sigma = 0$ (c) $\alpha = 3, \sigma = 0$

(d) $\alpha = 1, \sigma = 0.1$ (e) $\alpha = 1, \sigma = 0.5$ (f) $\alpha = 1, \sigma = 0.9$

Figure 2: Visualisation of top-5 rankings with rows corresponding to different rankings and columns to items sorted by size biased order. A lighter shade corresponds to a higher rank. Results are shown for a generalised gamma process with $\lambda(w) = \frac{\alpha}{\Gamma(1-\sigma)} w^{-\sigma-1} \exp(-\tau w)$ with $\tau = 1$ and different values of α and σ.



(a) $m = 5, \sigma = 0$ (b) $\alpha = 3, \sigma = 0$ (c) $\alpha = 3, m = 5$

Figure 3: Mean number of items appearing in L top-m rankings for a generalised gamma process with $\lambda(w) = \frac{\alpha}{\Gamma(1-\sigma)} w^{-\sigma-1} \exp(-\tau w)$ with $\tau = 1$ and different values of α, m and σ.

Then the joint probability under the nonparametric Plackett-Luce model is:

$$\begin{aligned} & P((Y_\ell, Z_\ell)_{\ell=1}^L | G) \\ & = \prod_{k=1}^K G(\{X_k^*\})^{n_k} \times \prod_{\ell=1}^L \prod_{i=1}^m \exp(-Z_{\ell i} G(\mathbb{X} \backslash \{Y_{\ell 1}, \dots, Y_{\ell\, i-1}\})) \\ & = \exp\left(-G(\mathbb{X}) \sum_{\ell i} Z_{\ell i}\right) \prod_{k=1}^K G(\{X_k^*\})^{n_k} \exp\left(-G(\{X_k^*\}) \sum_{\ell i} (\delta_{\ell i k} - 1) Z_{\ell i}\right) \end{aligned} \tag{10}$$

Taking expectation of (10) with respect to G gives:

Theorem 1 *The marginal probability of the L partial rankings and latent variables is:*

$$P((Y_\ell, Z_\ell)_{\ell=1}^L) = e^{-\psi(\sum_{\ell i} Z_{\ell i})} \prod_{k=1}^K h(X_k^*) \kappa\left(n_k, \sum_{\ell i} \delta_{\ell i k} Z_{\ell i}\right) \tag{11}$$

where $\psi(z)$ is the Laplace transform of $\lambda(w)$,

$$\psi(z) = -\log \mathbb{E}\left[e^{-zG(\mathbb{X})}\right] = \int_0^\infty (1 - e^{-zw}) \lambda(w) dw = \alpha \log\left(1 + \frac{z}{\tau}\right)$$

and $\kappa(n, z)$ is the nth moment of the exponentially tilted Lévy intensity $\lambda(w)e^{-zw}$:

$$\kappa(n, z) = \int_0^\infty w^n e^{-zw} \lambda(w) dw = \frac{\alpha}{(z + \tau)^n} \Gamma(n)$$

The proof, using the Poisson process characterisation of completely random measures and the Palm formula, is given in the appendix.

Another application of the Palm formula now allows us to derive a posterior characterisation of G.

Theorem 2 *Given the observations and associated latent variables $(Y_\ell, Z_\ell)_{\ell=1}^L$, the posterior law of G is also a gamma process, but with atoms with both fixed and random locations. Specifically,*

$$G | (Y_\ell, Z_\ell)_{\ell=1}^L = G^* + \sum_{k=1}^K w_k^* \delta_{X_k^*} \tag{12}$$

where G^* and $w_1^*, \dots, w_K^*$ are mutually independent. The law of G^* is still a gamma process,

$$G^* | (X_\ell, Z_\ell)_{\ell=1}^L \sim \Gamma(\alpha, \tau^\star, H), \qquad \tau^* = \tau + \sum_{\ell i} Z_{\ell i}$$

while the masses have distributions,

$$w_k^* | (Y_\ell, Z_\ell)_{\ell=1}^L \sim \text{Gamma}\left(n_k, \tau + \sum_{\ell i} \delta_{\ell i k} Z_{\ell i}\right)$$

Proof. Let $f : \mathbb{X} \to \mathbb{R}$ be measurable with respect to H. Then the characteristic functional of the posterior G is given by:

$$\mathbb{E}[e^{-\int f(x) G(dx)} | (Y_\ell, Z_\ell)_{\ell=1}^L] = \frac{\mathbb{E}[e^{-\int f(x) G(dx)} P((Y_\ell, Z_\ell)_{\ell=1}^L | G)]}{\mathbb{E}[P((Y_\ell, Z_\ell)_{\ell=1}^L | G)]} \tag{13}$$

The denominator is as given in Theorem 1, while the numerator is obtained using the same Palm formula technique as Theorem 1, with the inclusion of the term $e^{-\int f(x) G(dx)}$. Some algebra then shows that the resulting characteristic functional of the posterior G coincides with that of (12). The proof details are given in the appendix. ■

3.2 Gibbs sampling

Given the results of the previous section, a simple Gibbs sampler can now be derived, where all the conditionals are of known analytic form. In particular, we will integrate out all of G^* except for its total mass $w^*_* = G^*(\mathbb{X})$. This leaves the latent variables to consist of the masses w^*_*, $(w^*_k)_{k=1}^K$ and the latent variables $((Z_{\ell i})_{i=1}^m)_{\ell=1}^L$. The update for $Z_{\ell i}$ is given by (8), while those for the masses are given in Theorem 2:

$$\begin{aligned}
&\text{Gibbs update for } Z_{\ell i}: && Z_{\ell i}|\text{rest} \sim \text{Exp}\left(w^*_* + \textstyle\sum_k \delta_{\ell ik} w^*_k\right) \\
&\text{Gibbs update for } w^*_k: && w^*_k|\text{rest} \sim \text{Gamma}\left(n_k, \tau + \textstyle\sum_{\ell i} \delta_{\ell ik} Z_{\ell i}\right) \\
&\text{Gibbs update for } w^*_*: && w^*_*|\text{rest} \sim \text{Gamma}\left(\alpha, \tau + \textstyle\sum_{\ell i} Z_{\ell i}\right) && (14)
\end{aligned}$$

Note that the latent variables are conditionally independent given the masses and vice versa. Hyperparameters of the gamma process can be simply derived from the joint distribution in Theorem 1. Since the marginal probability of the partial rankings is invariant to rescaling of the masses, it is sufficient to keep τ fixed at 1. As for α, if a $\text{Gamma}(a, b)$ prior is placed on it, its conditional distribution is still gamma:

$$\text{Gibbs update for } \alpha: \qquad \alpha|\text{rest} \sim \text{Gamma}\left(a + K, b + \log\left(1 + \tfrac{\sum_{\ell i} Z_{\ell i}}{\tau}\right)\right)$$

Note that this update was derived with w^*_* marginalised out, so after an update to α it is necessary to immediately update w^*_* via (14) before proceeding to update other variables.

4 Generalisation to completely random measures

The posterior characterisation we have developed along with the Gibbs sampler can be easily extended to completely random measures (CRM) (Kingman, 1967). To keep the exposition simple, we shall consider homogeneous CRMs without fixed atoms. These can be described, as for the gamma process before, with atom locations $\{X_k\}$ iid according to a non-atomic base distribution H, and with atom masses $\{w_k\}$ being distributed according to a Poisson process over $\mathbb{R}^+$ with a general Lévy measure $\lambda(w)$ which satisfies the constraints (7) leading to a normalisable measure G with infinitely many atoms. We will write $G \sim \text{CRM}(\lambda, H)$ if G follows the law of a homogeneous CRM with Lévy intensity $\lambda(w)$ and base distribution H.

Both Theorems 1 and 2 generalise naturally to homogeneous CRMs. In fact the statements and the proofs in the appendix still hold with the more general Lévy intensity, along with its Laplace transform $\psi(z)$ and moment function $\kappa(n, z)$:

Theorem 1' *The marginal probability of the L partial rankings and latent variables is:*

$$P((Y_\ell, Z_\ell)_{\ell=1}^L) = e^{-\psi(\sum_{\ell i} Z_{\ell i})} \prod_{k=1}^K h(X^*_k) \kappa\left(n_k, \sum_{\ell i} \delta_{\ell ik} Z_{\ell i}\right)$$

where $\psi(z)$ is the Laplace transform of $\lambda(w)$,

$$\psi(z) = -\log \mathbb{E}\left[e^{-zG(\mathbb{X})}\right] = \int_0^\infty (1 - e^{-zw}) \lambda(w) dw$$

and $\kappa(n, z)$ is the nth moment of the exponentially tilted Lévy intensity $\lambda(w)e^{-zw}$:

$$\kappa(n, z) = \int_0^\infty w^n e^{-zw} \lambda(w) dw$$

Theorem 2' *Given the observations and associated latent variables $(Y_\ell, Z_\ell)_{\ell=1}^L$, the posterior law of G is also a homogeneous CRM, but with atoms with both fixed and random locations. Specifically,*

$$G|(Y_\ell, Z_\ell)_{\ell=1}^L = G^* + \sum_{k=1}^K w_k^* \delta_{X_k^*}$$

where G^* and $w_1^*, \dots, w_K^*$ are mutually independent. The law of G^* is a homogeneous CRM with an exponentially tilted Lévy intensity:

$$G^*|(X_\ell, Z_\ell)_{\ell=1}^L \sim \mathrm{CRM}(\lambda^\star, H) \qquad\qquad \lambda^*(w) = \lambda(w) e^{-w\sum_{\ell i} Z_{\ell i}}$$

while the masses have densities:

$$P(w_k^*|(Y_\ell, Z_\ell)_{\ell=1}^L) = \frac{(w_k^*)^{n_k} e^{-w_k^* \sum_{\ell i} Z_{\ell i}} \lambda(w_k^*)}{\kappa(n_k, \sum_{\ell i} Z_{\ell i})}.$$

Examples of CRMs that have been explored in the literature for Bayesian nonparametric modelling include the stable process (Kingman, 1975), the inverse Gaussian process (Lijoi et al., 2005), the generalised gamma process (Brix, 1999), and the beta process (Hjort, 1990). The generalised gamma process forms the largest known simple and tractable family of CRMs, with the gamma, stable and inverse Gaussian processes included as subfamilies. It has a Lévy intensity of the form:

$$\lambda(w) = \frac{\alpha}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-\tau w}$$

where the concentration parameter is $\alpha > 0$, the inverse scale is $\tau \geq 0$, and the index is $0 \leq \sigma < 1$. The gamma process is recovered when $\sigma = 0$, the stable when $\tau = 0$, and the inverse Gaussian when $\sigma = 1/2$. The Laplace transform and the moment function of the generalised gamma process are:

$$\psi(z) = \frac{\alpha}{\sigma}((\tau + z)^\sigma - \tau^\sigma) \qquad\qquad \kappa(n, z) = \frac{\alpha}{(\tau + z)^{n-\sigma}} \frac{\Gamma(n-\sigma)}{\Gamma(1-\sigma)}.$$

The Gibbs sampler developed for the gamma process can be generalised to homogeneous CRMs as well. Recall that given the observed partial rankings, the parameters consist of the ratings $(w_k^*)_{k=1}^K$ of the observed items and the total ratings w_*^* of the unobserved ones, while the latent variables are $(Z_{\ell i})$. A corollary of Theorems 1' and 2' which will prove useful is the joint probability of these along with the observed partial rankings:

$$P((Y_{\ell i}, Z_{\ell i}), (w_k^*), w_*^*) = e^{-w_*^*(\sum_{\ell i} Z_{\ell i}))} f(w_*^*) \prod_{k=1}^K h(X_k^*)(w_k^*)^{n_k} e^{-w_k^*(\sum_{\ell i} \delta_{\ell i k} Z_{\ell i})} \lambda(w_k^*) \tag{15}$$

where $f(w)$ is the density (assumed to exist) of the total mass w_*^* under a CRM with the *prior* Lévy intensity $\lambda(w)$. Note that integrating out the parameters $(w_k^*), w_*^*$ from (15) gives the marginal probability in Theorem 1'. From the joint probability (15), the Gibbs sampler can now be derived:

Gibbs update for $Z_{\ell i}$: $\quad Z_{\ell i}|\text{rest} \sim \text{Exp}\left(w_*^* + \sum_k \delta_{\ell i k} w_k^*\right)$

Gibbs update for w_k^*: $\quad P(w_k^*|\text{rest}) \propto (w_k^*)^{n_k} e^{-w_k^* \sum_{\ell i} Z_{\ell i}} \lambda(w_k^*)$

Gibbs update for w_*^*: $\quad P(w_*^*|\text{rest}) \propto e^{-w_*^*(\sum_{\ell i} Z_{\ell i}))} f(w_*^*)$

To be concrete, consider the updates for a generalised gamma process. The conditional distribution for w_k^* can be seen to be $\mathrm{Gamma}(n_k - \sigma, \tau + \sum_{\ell i} Z_{\ell i})$, while the conditional distribution for w_*^* can be seen to be an exponentially tilted stable distribution. This is not a standard distribution (nor does it have known analytic forms for its density), but can be effectively sampled using recent techniques (Devroye, 2009). Another approach is to marginalise out w_*^* first:

$$P((Y_{\ell i}, Z_{\ell i}), (w_k^*)) = e^{-\psi(\sum_{\ell i} Z_{\ell i})} \prod_{k=1}^{K} h(X_k^*)(w_k^*)^{n_k} e^{-w_k^*(\sum_{\ell i} \delta_{\ell ik} Z_{\ell i})} \lambda(w_k^*)$$

The MCMC algorithm then consists of sampling the ratings (w_k^*) and auxiliary variables $(Z_{\ell i})$. Marginalising out w_*^* introduces additional dependencies among the latent variables $Z_{\ell i}$. Fortunately, since the Laplace transform for a generalised gamma process is of simple form, it is possible to update the latent variables $(Z_{\ell i})$ using a variety of standard techniques, including Metropolis-Hastings, Hamiltonian Monte Carlo, or adaptive rejection sampling. For these techniques to work well we suggest reparametrising each $Z_{\ell i}$ using its logarithm $\log Z_{\ell i}$ instead.

5 Mixtures of Nonparametric Plackett-Luce Components

In this section we propose a mixture model for heterogeneous ranking data consisting of nonparametric Plackett-Luce components. Using the same data augmentation scheme, we show that an efficient Gibbs sampler can be derived, and apply the model to a dataset of preferences for Irish university programmes by high school graduates.

5.1 Statistical model

Assume that we have a set of L rankings (Y_ℓ) for $\ell \in [L]$ of top-m preferred items, and our objective is to partition these rankings into clusters of similar preferences. We consider the following Dirichlet process (DP) mixture model:

$$\begin{aligned} \pi &\sim \mathrm{GEM}(\gamma) \\ c_\ell | \pi &\sim \mathrm{Discrete}(\pi) \quad \text{for } \ell = 1, \dots, L, \\ Y_\ell | c_\ell, G_{c_\ell} &\sim \mathrm{PL}(G_{c_\ell}) \end{aligned}$$

where $\mathrm{GEM}(\gamma)$ denotes the Griffiths-Engen-McCloskey (GEM) distribution (Pitman, 2006) with concentration parameter γ (also known as the stick-breaking construction) and $\mathrm{PL}(G)$ denotes the nonparametric Plackett-Luce model parameterised by the atomic measure G described in Section 3. The jth cluster in the mixture model is parameterised by an atomic measure G_j and has mixing proportion π_j.

To complete the model, we have to specify the prior on the component atomic measures G_j. An obvious choice would be to use independent draws from a gamma process $\Gamma(\alpha, \tau, H)$ for each G_j. This unfortunately does not work. The reason is because if H is smooth then different atomic measures will never share the same atoms. On the other hand, notice that all items appearing in some observed partial ranking has to come from the same Plackett-Luce model, thus has to appear as atoms in the corresponding atomic measure. Putting these two observations together, the result is that any observed pair of partial rankings that share a common item will have to be assigned to the same component, and the mixture model will degenerate to using a few very larger components only. In consequence the model will not capture the fine-scale preference structure that may be present in the partial rankings. This is a similar problem that motivated the hierarchical DP (Teh et al., 2006), and the solution there as in here is to allow different atomic measures to share the same set of atoms, but to allow different atom masses.

Our solution, which is different from Teh et al. (2006), is to make use of the Pitt-Walker (Pitt and Walker, 2005) dependence model for gamma processes. Consider a tree-structured model where there is a single root G_0 and each component atomic measure G_j is a leaf which connects directly to G_0. The Pitt-Walker model allows us to construct the dependence structure between the root G_0 and the leaves (G_j) such that each G_j marginally follows a gamma process $\Gamma(\alpha, \tau, H)$. At the root, G_0 is first given a gamma process prior:

$$G_0 \sim \Gamma(\alpha, \tau, H)$$

Since G_0 is atomic, we can write it in the form:

$$G_0 = \sum_{k=1}^{\infty} w_{0k}\delta_{X_k}$$

Now for each j, define a random measure U_j with conditional law:

$$\begin{aligned} U_j|G_0 &= \sum_{k=1}^{\infty} u_{jk}\delta_{X_k} \\ u_{jk}|G_0 &\sim \text{Poisson}(\phi w_{0k}) \end{aligned} \tag{16}$$

where $\phi > 0$ is a parameter which, as we shall see, governs the strength of dependence between G_0 and each G_j. Note that since G_0 has finite total mass, U_j consists only of a finite number of atoms with positive masses; the other atoms all have masses equal to zero. Using the same Palm formula method as Section 3.1, we can show the following proposition:

Proposition 3 *Suppose the prior law of G_0 is $\Gamma(\alpha, \tau, H)$ and U_j has conditional law given by (16). The posterior law of G_0 given U_j is then:*

$$G_0 = G_0^* + \sum_{k=1}^{\infty} w_{0k}^*\delta_{X_k}$$

where G_0^* and $(w_{0k}^*)_{k=1}^{\infty}$ are all mutually independent. The law of G_0^* is given by a gamma process while the masses are conditionally gamma,

$$\begin{aligned} G_0^*|U_j &\sim \Gamma(\alpha, \tau + \phi, H) \\ w_{0k}^*|U_j &\sim \text{Gamma}(u_{jk}, \tau + \phi) \end{aligned}$$

Note that if $u_{jk} = 0$, we define w_{0k}^* to be degenerate at 0, thus the posterior of G_0 consists of a finite number of atoms in common with U_j, along with an infinite number of atoms (those in G_0^*) not in common. The total mass of G_0^* has distribution $\text{Gamma}(\alpha, \tau + \phi)$.

The idea, inspired by Pitt and Walker (2005), is to define the conditional law of G_j given G_0 and U_j to be independent of G_0 and to coincide with the conditional law of G_0 given U_j as in Proposition 3. In other words, define

$$G_j = G_j^* + \sum_{k=1}^{\infty} w_{jk}^*\delta_{X_k} \tag{17}$$

where $G_j^* \sim \Gamma(\alpha, \tau + \phi, H)$ and $w_{jk}^* \sim \text{Gamma}(u_{jk}, \tau + \phi)$ are mutually independent. Note that if $u_{jk} = 0$, the conditional distribution of w_{jk}^* will be degenerate at 0. Hence G_j has an atom at X_k if and only if U_j has an atom at X_k, that is, if $u_{jk} > 0$. In addition, it also has an infinite number of atoms (those in G_j^*) which are in neither U_j nor G_0.

Since the conditional laws of G_j and G_0 given U_j coincide, and G_0 has prior $\mathrm{Gamma}(\alpha, \tau, H)$, it can be seen that G_j will marginally follow the same law $\mathrm{Gamma}(\alpha, \tau, H)$ as well. More compactly, we can write the dependence model as:

$$\begin{aligned} U_j|G_0 &\sim \mathrm{Poisson}(\phi G_0) \\ G_j|U_j &\sim \Gamma\left(\alpha + U_j(\mathbb{X}), \tau + \phi, \frac{\alpha H + U_j}{\alpha + U_j(\mathbb{X})}\right) \end{aligned}$$

As a final observation, the parameter ϕ can be interpreted as controlling the strength of dependence between G_0 and each G_j. Indeed it can be shown that

$$\mathbb{E}[G_j|G_0] = \frac{\phi}{\phi + \tau} G_0 + \frac{\tau}{\phi + \tau} H$$

so that larger ϕ corresponds to each G_j being more similar to G_0.

Our construction to inducing sharing of atoms has a number of qualitative differences from that of the hierarchical DP (Teh et al., 2006). Firstly, the marginal law of each G_j is known: it is marginally a gamma process. For the hierarchical DP the marginal laws of the individual random measures are not of simple analytical forms. Since normalising a gamma process gives a DP, our construction can be used as an alternative method to induce sharing of atoms across multiple random measures, each of which still has marginal DP law. Secondly, in our construction only a finite number of atoms will be shared across random measures (though the number shared can be controlled by the dependence parameter ϕ), while in the hierarchical DP all infinitely many atoms are shared. In Caron and Teh (2012) we used the Pitt-Walker construction for a different purpose: we constructed a dynamical nonparametric Plackett-Luce model, where at each time t, G_t is a gamma process, with the Pitt-Walker construction used to define a Markov dependence structure for the sequence of random measures (G_t). The structure of our model, with a DP mixture with each component specified by a random atomic measure, is reminiscent of the nested DP of Rodríguez et al. (2008) as well, though our model has an additional hierarchical structure allowing the sharing of atoms among different component measures.

5.2 Posterior characterisation and Gibbs sampling

Assume for simplicity we have observed L top-m partial ranking $Y_\ell = (Y_{\ell 1}, \dots, Y_{\ell m})$ (the following will trivially extend to partial rankings of differing sizes). We extend the results of Section 3 in characterising the posterior and developing a Gibbs sampler for the mixture model.

Let $X^* = (X_k^*)_{k=1}^K$ be the set of unique items observed among $Y_1, \dots, Y_L$. For each cluster index j, let n_{jk} be the number of occurrences of item X_k^* among the set of item lists Y_ℓ in cluster j, that is, where $c_\ell = j$. Let $\rho_\ell = (\rho_{\ell i})_{i=1}^m$ be defined such as $Y_\ell = (X_{\rho_{\ell 1}}^*, \dots, X_{\rho_{\ell m}}^*)$, and $\delta_{\ell ik}$ be occurrence indicators similar to (9).

As in Section 3, the observed items X^* will contain the set of fixed atoms in the posterior law of the atomic measures $G_0, (G_j)$. We write the masses of the fixed atoms as $w_{0k} = G_0(\{X_k^*\})$, $w_{jk} = G_j(\{X_k^*\})$, while the total masses of all other random atoms are denoted $w_{0*} = G_0(\mathbb{X}\backslash X^*)$ and $w_{j*} = G_j(\mathbb{X}\backslash X^*)$. We also write $u_{jk} = U_j(\{X_k^*\})$ and $u_{j*} = U_j(\mathbb{X}\backslash X^*)$. As before, we will introduce latent variables for each $\ell = 1, \dots, L$ and $i = 1, \dots, m$:

$$Z_{\ell i}|Y_\ell, c_\ell, G_{c_\ell} \sim \mathrm{Exp}\left(w_{c_\ell *} + \sum_{k=1}^K \delta_{\ell ik} w_{c_\ell k}\right) \tag{18}$$

The overall graphical model is described in Figure 4.

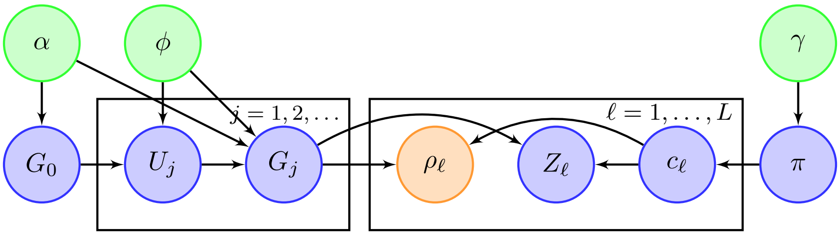


Figure 4: Graphical model of the Dirichlet process mixture of nonparametric Plackett-Luce components. The variables at the top in green are hyperparameters, (ρ_ℓ) (in orange) are the observed partial rankings, while the other variables (in blue) are unobserved variables.

Proposition 4 *Given the partial rankings (Y_ℓ) and associated latent variables $(Z_{\ell i})$, (u_{jk}), (u_{j*}), and cluster indicators (c_ℓ), the posterior law of G_j is a gamma process with atoms with both fixed and random locations. Specifically,*

$$G_j|(Y_\ell),(Z_{\ell i}),(u_{jk}),(u_{j*}),(c_\ell) = G_j^* + \sum_{k=1}^K w_{jk}\delta_{X_k^*}$$

where G_j^* and $w_{j1},\dots,w_{jK}$ are mutually independent. The law of G_j^* is a gamma process,

$$G_j^*|(Y_\ell),(Z_{\ell i}),(u_{jk}),(u_{j*}),(c_\ell) \sim \Gamma\left(\alpha + u_{j*}, \tau + \phi + \sum_{\ell|c_\ell=j}\sum_{i=1}^m Z_{\ell i}, H\right), \tag{19}$$

while the masses have distributions,

$$w_{jk}|(Y_\ell),(Z_{\ell i}),(u_{jk}),(u_{j*}),,(c_\ell) \sim \text{Gamma}\left(n_{jk} + u_{jk}, \tau + \phi + \sum_{\ell|c_\ell=j}\sum_{i=1}^m \delta_{\ell ik} Z_{\ell i}\right) \tag{20}$$

Note that if $n_{jk} + u_{jk} = 0$, then $w_{jk} = 0$ and G_j will not have a fixed atom at X_k^*. To complete the posterior characterisation, note that conditioned on G_0 and G_j the variables $u_{j1},\dots,u_{jK}$ and u_{j*} are independent, with u_{jk} dependent only on w_{0k} and w_{jk} and similarly for u_{j*}. The conditional probabilities are:

$$p(u_{jk}|w_{0k},w_{jk}) \propto f_{\text{Gamma}}(w_{jk}; u_{jk}, \tau+\phi) f_{\text{Poisson}}(u_{jk}; \phi w_{0k}) \tag{21}$$

$$p(u_{j*}|w_{0*},w_{j*}) \propto f_{\text{Gamma}}(w_{j*}; \alpha + u_{j*}, \tau+\phi) f_{\text{Poisson}}(u_{j*}; \phi w_{0*}) \tag{22}$$

where f_{Gamma} is the density of a Gamma distribution and f_{Poisson} is the probability mass function for a Poisson distribution. The normalising constants are available in closed form (Mena and Walker, 2009):

$$\begin{aligned} p(w_{jk}|w_{0k}) =& \exp(-\phi w_{0k}) 1_{w_{jk},0} \\ & + \mathcal{I}_{-1}\left(2\sqrt{w_{jk}\phi w_{0k}(\tau+\phi)}\right)\left(\frac{\phi(\tau+\phi)w_{0k}}{w_{jk}}\right)^{1/2} \exp\left(-\phi(w_{jk}+w_{0k}) - \tau w_{jk}\right) \end{aligned} \tag{23}$$

$$p(w_{j*}|w_{0*}) = \mathcal{I}_{\alpha-1}\left(2\sqrt{w_{j*}\phi w_{0*}(\tau+\phi)}\right)(\tau+\phi)^{\frac{\alpha+1}{2}}\left(\frac{w_{j*}}{\phi w_{0*}}\right)^{\frac{\alpha-1}{2}} \exp(-\phi(w_{j*}+w_{0*}) - \tau w_{j*}) \tag{24}$$

where $1_{a,b} = 1$ if $a = b$, 0 otherwise, and $\mathcal{I}$ is the modified Bessel function of the first kind. It is therefore possible to sample exactly from the discrete distributions (21) and (22) using standard retrospective sampling for discrete distributions, see for example Papaspiliopoulos and Roberts (2008). Alternatively, we describe in the appendix a Metropolis-Hastings procedure that worked well in the applications.

Armed with the posterior characterisation, a Gibbs sampler can now be derived. Each iteration of the Gibbs sampler proceeds in the following order (details are in appendix):

1. First note that the total masses $G_j(\mathbb{X})$ are not likelihood identifiable, so we introduce a step to improve mixing. We simply sample them from the prior:

$$\begin{aligned} G_0(\mathbb{X}) &\sim \mathrm{Gamma}(\alpha, \tau) \\ U_j(\mathbb{X})|G_0(\mathbb{X}) &\sim \mathrm{Poisson}(\phi G_0(\mathbb{X})) \\ G_j(\mathbb{X})|U_j(\mathbb{X}) &\sim \mathrm{Gamma}(\alpha + U_j(\mathbb{X}), \tau + \phi) \end{aligned}$$

   The individual atom masses (w_{jk}, w_{j*}) are scaled along with the update to the total masses. Then the Poisson masses (u_{jk}), (u_{j*}) are updated using (21) and (22).

2. The concentration parameter α and the masses w_{0*}, (w_{j*}) and (u_{j*}) associated with other unobserved items are updated efficiently using a forward-backward recursion detailed in the appendix.

3. The masses (w_{0k}) and w_{0*} of the atoms in G_0 are updated via an extension of Proposition 3. In particular, for each item $k = 1, \dots, K$, the masses are conditionally independent with distributions:

$$w_{0k}|u_{1:J,k}, \phi \sim \mathrm{Gamma}\left(\textstyle\sum_{j=1}^J u_{jk}, J\phi + \tau\right)$$

   while the total mass of the remaining atoms have conditional distribution:

$$w_{0*}|u_{1:J*}, \phi \sim \mathrm{Gamma}\left(\alpha + \textstyle\sum_{j=1}^J u_{j*}, J\phi + \tau\right)$$

4. The latent variables $(Z_{\ell i})$ are updated as in (18).

5. Conditioned on $(Z_{\ell i})$, (u_{jk}) and (u_{j*}), the masses (w_{jk}) are updated via (20), while the total mass of the unobserved atoms is $w_{j*} \sim \mathrm{Gamma}(\alpha_j^*, \tau_j^*)$ from (19).

6. The allocation variables c_ℓ are updated using a slice sampler for mixture models (Walker, 2007; Kalli et al., 2011).

7. Finally, the scale parameter γ of the Dirichlet process is updated using (West, 1992) and the dependence parameter ϕ is updated by a Metropolis-Hastings step using (23) and (24) with the latent (u_{jk}) and (u_{j*}) marginalised out.

The resulting algorithm is a valid partially collapsed Gibbs sampler (Van Dyk and Park, 2008). Note however that permutations of the above steps could result in an invalid sampler.

6 Irish University Programmes

Applications to third level degree programmes in Ireland are handled by a centralised applications system called the College Application Office (CAO). When students apply for degree programmes they rank up to ten courses, in order of preference, from a list of more than five hundred degree programmes. Places in these degree programmes are allocated on the basis of the applicants performance in the Irish

Table 1: Description of the different clusters. The size of the clusters, the entropy and a cluster description are provided.

No	Size	Ent.	Description	No	Size	Ent.	Description
1	3091	0.71	Social Science/Tourism	15	1832	0.47	Teaching/Arts
2	3081	0.70	Science	16	1819	0.68	Art/Music - Dublin
3	3026	0.58	Arts	17	1671	0.54	Medicine
4	2869	0.63	Bus./Marketing - Dublin	18	1658	0.71	Engineering
5	2756	0.67	Construction	19	1615	0.66	Galway
6	2598	0.63	Business/Commerce	20	1558	0.69	Arts/Religion/Theology
7	2549	0.65	CS - outside Dublin	21	1514	0.76	Arts/History - Dublin
8	2225	0.66	CS - Dublin	22	1508	0.68	Engineering - Dublin
9	2224	0.66	Arts/Social - ouside Dublin	23	1262	0.70	Limerick
10	2154	0.63	Business/Finance - Dublin	24	1249	0.79	Art/Bus./Language - Dublin
11	2089	0.65	Arts/Psychology - Dublin	25	1249	0.65	Law
12	2001	0.63	Comm./Journalism - Dublin	26	1215	0.72	Business - Dublin
13	1994	0.62	Cork	27	1075	0.74	Sciences/Maths - Dublin
14	1858	0.71	Bus./Tourism/Waterford				

Leaving Certificate examination; students with a high “points” score are more likely to get their high preference choices. We consider the application of the mixture of Plackett-Luce to the year 2000 cohort of applications to the College Application Office; these data correspond to top-10 rankings of college degree courses for 53757 applicants.

Flat priors are used for the hyperparameters and we run the Gibbs sampler with 20000 iterations. The partition from the last sample is taken as the point estimate. Given this partition, we run a Gibbs sampler with 2000 iterations so that to obtain the posterior mean Plackett-Luce parameters for each cluster. Clusters are then reordered by decreasing size. Table 1 shows the sizes of the 27 clusters which have a size larger than 10. In addition, a coclustering matrix was computed based on the first MCMC run which records for each pair of students the probability of them belonging to the same cluster. Figure 5 shows the coclustering matrix to summarise the clustering of the 53757 students, where students are rearranged by their cluster membership (members of the first cluster first, then members of the second cluster, etc.).

An examination of the Plackett-Luce parameter for each cluster reveals that the subject matter of the degree programme is a strong determinant of the clustering of students (Table 1). For example, clusters 5, 17 and 25 are characterised as construction, medicine and law, respectively. Besides the type of degree, geographical location is a strong determinant of course choice. Clusters 13, 19 and 23 are respectively concerned with applications to college degrees in Cork, Galway and Limerick. There is a lot of heterogeneity in the subject area of the college degrees for these clusters, as can be seen for example for the Cork cluster 13 in Table 4. A number of clusters are also defined by a combination of both subject area and location, for example, for clusters 7 and 8 in Tables 2 and 3, which correspond to computer science respectively outside and inside Dublin.

There is a common perception in the Irish society and media that students pick courses based on prestige rather than subject area. Such a phenomenon should be evidenced by a cluster of students picking courses in medicine, actuarial science and law, but no such cluster was found. In fact, medicine, law and actuarial science applicants are clustered separately into clusters 17, 25 and 27, respectively. Therefore, the clustering suggests that students are primarily picking courses on the basis of subject area and geographical considerations; this finding is in agreement with the results found in (Gormley and Murphy, 2006; McNicholas, 2007).

It is also of interest to look at the variability of the student choices within each cluster. This can be quantified by the normalized entropy, which takes its values between 0 and 1, and defined for each cluster

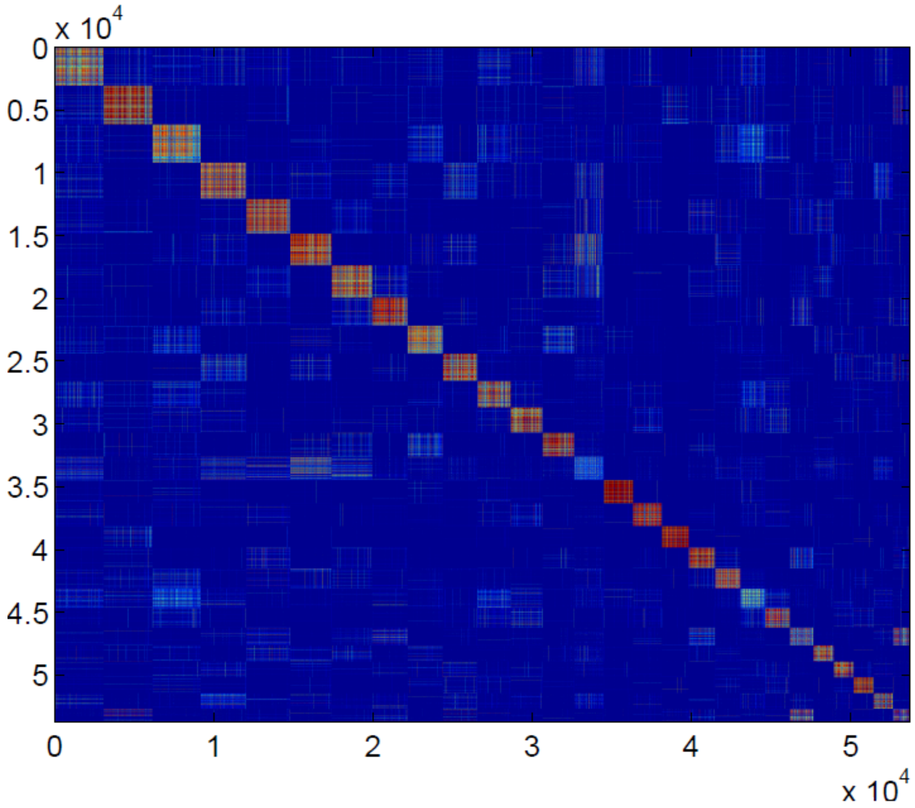


Figure 5: Coclustering for the CAO data. The clusters are arranged according to size and are described in Table 1.

Table 2: Cluster 7: Computer Science - outside Dublin

Rank	Aver. Norm. Weight	University	Degree
1	0.080	Cork IT	Computer Applications
2	0.079	University of Limerick	Computer Systems
3	0.078	Limerick IT	Software Development
4	0.062	Cork IT	Software Dev & Comp Net
5	0.059	Waterford IT	Applied Computing
6	0.041	IT Carlow	Computer Networking
7	0.044	University College Cork	Computer Science
8	0.038	Athlone IT	Computer and Software Engineering
9	0.036	University of Limerick	Information Technology
10	0.036	Dublin City University	Computer Applications

Table 3: Cluster 8: Computer Science - Dublin

Rank	Aver. Norm. Weight	University	Degree
1	0.144	Dublin City University	Computer Applications
2	0.057	University College Dublin	Computer Science
3	0.050	NUI - Maynooth	Computer Science
4	0.047	Dublin IT	Computer Science
5	0.041	National College of Ireland	Software Systems
6	0.040	Dublin IT	Business Info. Systems Dev.
7	0.038	Trinity College Dublin	Computer Science
8	0.036	Dublin IT	Applied Sciences/Computing
9	0.030	University College Dublin	B.A. (Computer Science)
10	0.030	Trinity College Dublin	Information & Comm. Tech.

Table 4: Cluster 13: Cork

Rank	Aver. Norm. Weight	University	Degree
1	0.103	University College Cork	Arts
2	0.074	University College Cork	Computer Science
3	0.073	University College Cork	Commerce
4	0.068	University College Cork	Business Information Systems
5	0.059	Cork IT	Computer Applications
6	0.052	Cork IT	Software Dev & Comp Net
7	0.036	University College Cork	Finance
8	0.032	University College Cork	Accounting
9	0.029	University College Cork	Law
10	0.024	University College Cork	Biological and Chemical Sciences

j by

$$\frac{-\sum_{k=1}^{K}(\widehat{\pi}_{jk}\log\widehat{\pi}_{jk})-\widehat{\pi}_{j*}\log\widehat{\pi}_{j*}}{\log(K+1)}$$

where $\widehat{\pi}_{jk}$ are the averaged normalized weights of item k in cluster j obtained from the second MCMC run; the normalized entropy values for each cluster are reported in Table 1. A low value indicates low variability in the choices within a cluster, whereas a large value indicates a lot of variability. Interestingly, cluster 15 has very low normalized entropy, where 56% of the students in that cluster are likely to take one of the three most popular courses of that cluster (Drumcondra, Froebel or Marina) as their first choice; these courses are the main teacher education courses in Ireland and thus many members of this cluster have a strong interest in teacher education as a degree choice. Further, there is much more variability in cluster 7, where students choices are spread across various computing degrees, and only 24% of the students are likely to take one of the three most popular courses as their first choice. The highest normalized entropy is for cluster 24, where only 11% of the students are likely to take one of the top three courses as their first choice and course preferences are approximately equally spread between the various courses in arts and courses involving business with a language subject that characterize this cluster.

The coclustering matrix reveals some interesting connections between clusters, which have not been explored in previous analyses of the CAO data. For example, the plot reveals that a number of applicants have high probability of belonging to clusters 3 and 20 which are both in the arts. Cluster 3 is characterised by arts degrees which do not require the applicants to select their major in advance, whereas cluster 20 is characterised by arts degrees where the student needs to specify their major in advance. However, there is some evidence of co-membership of the medical (cluster 18) and law (cluster 22) clusters and the law (cluster 22) and mathematical science (cluster 27) clusters which suggests that there may be some applicants who are choosing courses by prestige. This phenomenon is difficult to observe in the individual cluster parameters but becomes more apparent in the coclustering results.

7 Discussion

We have proposed a Bayesian nonparametric Plackett-Luce model for ranked data. Our approach is based on the theory of completely random measures, where we showed that the Plackett-Luce generative model corresponds exactly to a size-biased permutation of the atoms in the random measure. We characterised the posterior distribution, and derived a simple MCMC sampling algorithm for posterior simulation. Our approach can be seen as a multi-stage generalisation of posterior inference in normalised random measures (James et al., 2009; Griffin and Walker, 2011; Favaro and Teh, 2012).

We also developed a nonparametric mixture model consisting of nonparametric Plackett-Luce components to model heterogeneity in partial ranking data. In order to allow atoms to be shared across components, we made use of the Pitt-Walker construction, which was previously only used to define Markov dynamical models. Applying our model to a dataset of preferences for Irish university programmes, we find interesting clustering structure supporting the observation that students were choosing programmes mainly based on subject area and geographical considerations.

It is worthwhile comparing our mixture model to another nonparametric mixture model, DPM-GM, where each component is a generalised Mallows model (Busse et al., 2007; Meilă and Bao, 2008; Meilă and Chen, 2010). In the generalised Mallows model the component distributions are characterised by a (discrete) permutation parameter whereas in the Plackett-Luce model the component distributions are characterised by a continuous rating parameter. Thus the Plackett-Luce model offers greater modelling flexibility to capture the strength of preferences for each item. On the other hand, the scale parameters in the generalised Mallows model can accommodate varying precision in the ranking. Additionally, inference for the generalized Mallows models can be difficult.

A Proof of Theorem 1

The marginal probability (11) is obtained by taking the expectation of (10) with respect to G. Note however that (10) is a density, so to be totally precise here we need to work with the *probability* of infinitesimal neighborhoods around the observations instead, which introduces significant notational complexity. To keep the notation simple, we will work with densities, leaving it to the careful reader to verify that the calculations indeed carry over to the case of probabilities.

$$
\begin{aligned}
&P((Y_\ell, Z_\ell)_{\ell=1}^L) \\
=&\mathbb{E}\left[P((Y_\ell, Z_\ell)_{\ell=1}^L | G)\right] \\
=&\mathbb{E}\left[e^{-G(\mathbb{X})\sum_{\ell i} Z_{\ell i}} \prod_{k=1}^K G(\{X_k^*\})^{n_k} e^{-G(\{X_k^*\})\sum_{\ell i}(\delta_{\ell ik}-1)Z_{\ell i}}\right]
\end{aligned}
$$

The gamma prior on $G = \sum_{j=1}^\infty w_j \delta_{X_j}$ is equivalent to a Poisson process prior on $N = \sum_{j=1}^\infty \delta_{(w_j, X_j)}$ defined over the space $\mathbb{R}^+ \times \mathbb{X}$ with mean intensity $\lambda(w)h(x)$. Then,

$$
=\mathbb{E}\left[e^{-\int wN(dw,dx)\sum_{\ell i} Z_{\ell i}} \prod_{k=1}^K \sum_{j=1}^\infty w_j^{n_k} \mathbb{1}(X_j = X_k^*) e^{-w_j \sum_{\ell i}(\delta_{\ell ik}-1)Z_{\ell i}}\right]
$$

Applying the Palm formula for Poisson processes to pull the $k = 1$ term out of the expectation,

$$\begin{aligned}
=\int \mathbb{E}&\left[e^{-\int w(N+\delta_{w_1^*,x_1^*})(dw,dx)\sum_{\ell i} Z_{\ell i}} \prod_{k=2}^{K}\sum_{j=1}^{\infty} w_j^{n_k} \mathbb{1}(X_j = X_k^*) e^{-w_j \sum_{\ell i}(\delta_{\ell i k}-1)Z_{\ell i}}\right] \\
&\qquad \times (w_1^*)^{n_1} h(X_1^*) e^{-w_1^* \sum_{\ell i}(\delta_{\ell i 1}-1)Z_{\ell i}} \lambda(w_1^*)dw_1^* \\
=\mathbb{E}&\left[e^{-\int wN(dw,dx)\sum_{\ell i} Z_{\ell i}} \prod_{k=2}^{K}\sum_{j=1}^{\infty} w_j^{n_k} \mathbb{1}(X_j = X_k^*) e^{-w_j \sum_{\ell i}(\delta_{\ell i k}-1)Z_{\ell i}}\right] \\
&\qquad \times h(X_1^*) \int (w_1^*)^{n_1} e^{-w_1^* \sum_{\ell i}\delta_{\ell i 1}Z_{\ell i}} \lambda(w_1^*)dw_1^*
\end{aligned}$$

Now iteratively pull out terms $k = 2, \ldots, K$ using the same idea, and we get:

$$\begin{aligned}
=&\mathbb{E}\left[e^{-G(\mathbb{X})\sum_{\ell i} Z_{\ell i}}\right] \prod_{k=1}^{K} h(X_k^*) \int (w_k^*)^{n_k} e^{-w_k^* \sum_{\ell i}\delta_{\ell i k}Z_{\ell i}} \lambda(w_k^*)dw_k^* \\
=&e^{-\psi(\sum_{\ell i} Z_{\ell i})} \prod_{k=1}^{K} h(X_k^*) \kappa\left(n_k, \sum_{\ell i}\delta_{\ell i k}Z_{\ell i}\right) \qquad (25)
\end{aligned}$$

This completes the proof of Theorem 1.

B Proof of Theorem 2

The proof is essentially obtained by calculating the numerator and denominator of (13). The denominator is already given in Theorem 1. The numerator is obtained using the same technique with the inclusion of the term $e^{\int f(x)G(dx)}$, which gives:

$$\begin{aligned}
&\mathbb{E}\left[e^{-\int f(x)G(dx)} P((Y_\ell, Z_\ell)_{\ell=1}^L | G)\right] \\
=&\mathbb{E}\left[e^{-\int (f(x)+\sum_{\ell i} Z_{\ell i})G(dx)}\right] \prod_{k=1}^{K} h(X_k^*) \int (w_k^*)^{n_k} e^{-w_k^*(f(X_k^*)+\sum_{\ell i}\delta_{\ell i k}Z_{\ell i})} \lambda(w_k^*)dw_k^*
\end{aligned}$$

By the Lévy-Khintchine Theorem (using the fact that G has a Poisson process representation N),

$$\begin{aligned}
&= \exp\left(-\int (1 - e^{-w(f(x)+\sum_{\ell i} Z_{\ell i})})\lambda(w)h(x)dwdx\right) \\
&\times \prod_{k=1}^{K} h(X_k^*) \int (w_k^*)^{n_k} e^{-w_k^*(f(X_k^*)+\sum_{\ell i}\delta_{\ell i k}Z_{\ell i})} \lambda(w_k^*)dw_k^* \qquad (26)
\end{aligned}$$

Dividing the numerator (25) by the denominator (26), the characteristic functional of the posterior G is:

$$\begin{aligned}
&\mathbb{E}\left[e^{-\int f(x)G(dx)} | (Y_\ell, Z_\ell)_{\ell=1}^L\right] \\
&= \exp\left(-\int (1 - e^{-wf(x)})e^{-\sum_{\ell i} Z_{\ell i}}\lambda(w)h(x)dwdx\right) \\
&\times \prod_{k=1}^{K} h(X_k^*) \frac{\int e^{-f(X_k^*)}(w_k^*)^{n_k} e^{-w_k^* \sum_{\ell i}\delta_{\ell i k}Z_{\ell i}}\lambda(w_k^*)dw_k^*}{\int (w_k^*)^{n_k} e^{-w_k^* \sum_{\ell i}\delta_{\ell i k}Z_{\ell i}}\lambda(w_k^*)dw_k^*}
\end{aligned}$$

Since the characteristic functional is the product of $K+1$ terms, we see that the posterior G consists of $K+1$ independent components, one corresponding to the first term above (G^*), and the others corresponding to the K terms in the product over k. Substituting the Lévy measure $\lambda(w)$ for a gamma process, we note that the first term shows that G^* is a gamma process with updated inverse scale τ^*. The kth term in the product shows that the corresponding component is an atom located at X_k^* with density $(w_k^*)^{n_k} e^{-w_k^* \sum_{\ell i} \delta_{\ell i k} Z_{\ell i}} \lambda(w_k^*)$; this is the density of the gamma distribution over w_k^* in Theorem 2. This completes the proof.

C Gibbs sampler for the mixture of nonparametric Plackett-Luce components

Let J be the number of different values taken by c. The Gibbs sampler proceeds with each of the following updates in turn:

1. a. Update $G_0(\mathbb{X})$ given α, then for $j=1,\dots,J$, update $G_j(\mathbb{X})$ given $(G_0(\mathbb{X}),\alpha,\phi,c)$
   b. For $j=1,\dots,J$, update (u_j,u_{j*}) given $(w_0,w_{0*},w_j,w_{j*},\phi,\alpha,c)$
2. a. Update α given (Z,ϕ,c)
   b. Update w_{0*} given (Z,ϕ,c,α)
   c. For $j=1,\dots,J$, update u_{j*} given (Z,ϕ,c,α,w_{0*})
   d. For $j=1,\dots,J$, update w_{j*} given (Z,α,u_{j*},ϕ,c)
3. Update $(w_{0k}),w_{0*}$ given $(U_{1:J},\alpha)$
4. For $\ell=1,\dots,L$, update Z_ℓ given $(w_{c_\ell},w_{c_\ell *},c_\ell)$
5. For $j=1,\dots,J$, update (w_j,w_{j*}) given $(Z,\alpha,u_j,u_{j*},\phi,c)$
6. For $\ell=1,\dots,L$, update c_ℓ given $w_{1:J},w_{1:J*}$
7. Update γ given c
8. Update ϕ given $w_0,w_{0*},w_{1:J},w_{1:J*},\alpha,\phi$

The step are now fully described.

1.a) Update $G_0(\mathbb{X})$ given α, then for $j=1,\dots,J$, update $G_j(\mathbb{X})$ given $(G_0(\mathbb{X}),\alpha,\phi,c)$

We have

$$G_0(\mathbb{X})|\alpha \sim \text{Gamma}(\alpha,\tau)$$

and for $j=1,\dots,J$

$$G_j(\mathbb{X}) \sim \text{Gamma}(\alpha+M_j,\tau+\phi)$$

where $M_j \sim \text{Poisson}(\phi G_0(\mathbb{X}))$.

1.b) For $j=1,\dots,J$, update (u_j,u_{j*}) given $(w_0,w_{0*},w_j,w_{j*},\phi,\alpha,c)$

Consider first the sampling of u_j. We have, for $j=1,\dots,J$ and $k=1,\dots,K$

$$p(u_{jk}|w_{0k},w_{jk}) \propto p(u_{jk}|w_{0k})p(w_{jk}|u_{jk})$$

where

$$p(u_{jk}|w_{0k}) = f_{\text{Poisson}}(u_{jk}; \phi w_{0k})$$

and

$$p(w_{jk}|u_{jk}) = \begin{cases} \delta_0(w_{jk}) & \text{if } u_{jk} = 0 \\ f_{\text{Gamma}}(w_{jk}; u_{jk}, \tau + \phi) & \text{if } u_{jk} > 0 \end{cases}$$

Hence we can have the following MH update. If $w_{jk} > 0$, then we necessarily have $u_{jk} > 0$. We sample u^*_{jk} ∼zPoisson(ϕw_{0k}) where zPoisson(ϕw_{0k}) denotes the zero-truncated Poisson distribution and accept u^*_{jk} with probability

$$\min\left(1, \frac{f_{\text{Gamma}}(w_{jk}; u^*_{jk}, \tau + \phi)}{f_{\text{Gamma}}(w_{jk}; u_{jk}, \tau + \phi)}\right)$$

If $w_{jk} = 0$, we only have two possible moves: $u_{jk} = 0$ or $u_{jk} = 1$, given by the following probabilities

$$P(u_{jk} = 0|w_{jk} = 0, w_{0k}) = \frac{\exp(-\phi w_{0k})}{\exp(-\phi w_{0k}) + \phi w_{0k}\exp(-\phi w_{0k})(\tau + \phi)} = \frac{1}{1 + \phi w_{0k}(\tau + \phi)}$$

$$P(u_{jk} = 1|w_{jk} = 0, w_{0k}) = \frac{\phi w_{0k}\exp(-\phi w_{0k})(\tau + \phi)}{\exp(-\phi w_{0k}) + \phi w_{0k}\exp(-\phi w_{0k})(\tau + \phi)} = \frac{\phi w_{0k}(\tau + \phi)}{1 + \phi w_{0k}(\tau + \phi)}$$

Note that the above Markov chain is not irreducible, as the probability is zero to go from a state $(u_{jk} > 0, w_{jk} > 0)$ to a state $(u_{jk} = 0, w_{jk} = 0)$, even though the posterior probability of this event is non zero in the case item k does not appear in cluster j. We can add such moves by jointly sampling (u_{jk}, w_{jk}). For each k that does not appear in cluster j, sample $u^*_{jk} \sim$ Poisson(ϕw_{0k}) then set $w^*_{jk} = 0$ if $u^*_{jk} = 0$ otherwise sample $w^*_{jk} \sim$ Gamma$(u_{jk}, \tau + \phi)$. Accept (u^*_{jk}, w^*_{jk}) with probability

$$\min\left(1, \frac{\exp(-w^*_{jk}\sum_{\ell|c_\ell=j}\sum_{i=1}^m Z_{\ell i})}{\exp(-w_{jk}\sum_{\ell|c_\ell=j}\sum_{i=1}^m Z_{\ell i})}\right)$$

We now consider sampling of $u_{j*}, j = 1, \ldots, J$. We can use a MH step. Sample $w^*_{j*} \sim$ Poisson(ϕw_{0*}) and accept with probability

$$\min\left(1, \frac{f_{\text{Gamma}}(u_{j*}; \alpha + u^*_{j*}, \tau + \phi)}{f_{\text{Gamma}}(u_{j*}; \alpha + u^*_{j*}, \tau + \phi)}\right)$$

2.a) Update α given (Z, ϕ, c)

We can sample from the full conditional which is given by

$$\alpha|(Z, \gamma, \phi, c) \sim \text{Gamma}\left(a + K, b + y_0 + \log(1 + x_0)\right)$$

where

$$x_0 = \sum_{j=1}^J \frac{\phi \widetilde{Z}_j}{1 + \phi + \widetilde{Z}_j}$$

$$y_0 = -\sum_{j=1}^J \log\left(\frac{1 + \phi}{1 + \phi + \widetilde{Z}_j}\right)$$

with $\widetilde{Z}_j = \sum_{\ell|c_\ell=j}\sum_{i=1}^m Z_{\ell i}$.

2.b) Update w_{0*} given (Z, ϕ, c, α)
We can sample from the full conditional which is given by

$$w_{0*}|(Z, \phi, c, \alpha) \sim \text{Gamma}\,(\alpha, \tau + x_0)$$

where x_0 is defined above.

2.c) For $j = 1, \ldots, J$, update u_{j*} given $(Z, \phi, c, \alpha, w_{0*})$
We can sample from the full conditional which is given, for $j = 1, \ldots, J$ by

$$u_{j*}|(Z, \phi, c, \alpha, w_{0*}) \sim \text{Poisson}\left(\frac{1+\phi}{1+\phi+\widetilde{Z}_j}\phi w_{0*}\right)$$

where $\widetilde{Z}_j$ is defined above.

2.d) For $j = 1, \ldots, J$, update w_{j*} given $(Z, \alpha, u_{j*}, \phi, c)$
We can sample from the full conditional which is given, for $j = 1, \ldots, J$ by

$$w_{j*}|u_{j*}, Z, c, \alpha \sim \text{Gamma}\left(\alpha + u_{j*}, \tau + \phi + \widetilde{Z}_j\right)$$

where $\widetilde{Z}_j$ is defined above.

3) Update $(w_{0k}), w_{0*}$ given $(U_{1:J}, \alpha)$
For each item $k = 1, \ldots, K$, sample

$$w_{0k}|u_{1:J,k}, \phi \sim \text{Gamma}\left(\sum_{j=1}^{J} u_{jk}, J\phi + \tau\right)$$

Sample the remaining mass

$$w_{0*}|u_{1:J*}, \phi \sim \text{Gamma}\left(\alpha + \sum_{j=1}^{J} u_{j*}, J\phi + \tau\right)$$

4) For $\ell = 1, \ldots, L$, update Z_ℓ given $(w_{c_\ell}, w_{c_\ell *}, c_\ell)$
For $\ell = 1, \ldots, L$ and $i = 1, \ldots m$, sample

$$Z_{\ell i}|c, w, w_* \sim \text{Exp}\left(w_{c_\ell,*} + \sum_{k=1}^{K} \delta_{\ell ik} w_{c_\ell,k}\right)$$

5) For $j = 1, \ldots, J$, update $(w_{jk}), w_{j*}$ given $(Z, \alpha, u_j, u_{j*}, \phi, c)$
For each cluster $j = 1, \ldots, J$

- For each item $k = 1, \ldots, K$, sample

$$w_{jk}|u_{jk}, \{\rho_\ell | c_\ell = j\} \sim \text{Gamma}\left(n_{jk} + u_{jk}, \tau + \phi + \sum_{\ell | c_\ell = j}\left\{\sum_{i=1}^{m} \delta_{\ell ik} Z_{\ell i}\right\}\right)$$

  if $u_{jk} + n_{jk} > 0$, otherwise, set $w_{jk} = 0$.

- Sample the total mass

$$w_{j*}|u_{j*},\{\rho_\ell|c_\ell=j\}\sim\text{Gamma}\left(\alpha+u_{j*},\tau+\phi+\sum_{\ell|c_\ell=j}\sum_{i=1}^{m}Z_{\ell i}\right)$$

6) For $\ell=1,\ldots,L$, update c_ℓ given $w_{1:J},w_{1:J*}$

The allocation variables $(c_1,\ldots,c_L)$ are updated using the slice sampling technique described in (Walker, 2007; Kalli et al., 2011).

7) Update γ given c

The scale parameter γ of the Dirichlet process is updated using the data augmentation technique of West (1992).

8) Update ϕ given $w_0,w_{0*},w_{1:J},w_{1:J*},\alpha,\phi$

We sample ϕ using a MH step. Propose $\phi^*=\phi\exp(\sigma\varepsilon)$ where $\sigma>0$ and $\varepsilon\sim\mathcal{N}(0,1)$. And accept it with probability

$$\min\left(1,\frac{p(\phi^*)}{p(\phi)}\frac{\phi^*}{\phi}\prod_{j=1}^{J}\left[\frac{p(w_{j*}|\phi^*,w_{0*})}{p(w_{j*}|\phi,w_{0*})}\prod_{k=1}^{K}\frac{p(w_{jk}|\phi^*,w_{0k})}{p(w_{jk}|\phi,w_{0k})}\right]\right)$$

References

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability 31*(4), 929–953.

Busse, L., P. Orbanz, and J. Buhmann (2007). Cluster analysis of heterogeneous rank data. In *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 113–120. ACM.

Caron, F. and A. Doucet (2012). Efficient Bayesian inference for generalized Bradley-Terry models. *Journal of Computational and Graphical Statistics 21*(1), 174–196.

Caron, F. and Y. W. Teh (2012). Bayesian nonparametric models for ranked data. In *Neural Information Processing Systems (NIPS'2012)*, Lake Tahoe, USA.

Chapman, R. and R. Staelin (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research 19*(3), 288–301.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Transactions Modeling and Computer Simulation 19*(4), 18:1–18:20.

Diaconis, P. (1988). *Group representations in probability and statistics*, Volume 11 of *IMS Lecture Notes*. Institute of Mathematical Statistics.

Favaro, S. and Y. W. Teh (2012). MCMC for normalized random measure mixture models. Technical report, University of Turin.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics 1*(2), 209–230.

Gormley, I. C. and T. B. Murphy (2006). Analysis of Irish third-level college applications data. *Journal of the Royal Statistical Society Series A 169*(2), 361–379.

Gormley, I. C. and T. B. Murphy (2008). Exploring voting blocs with the Irish electorate: a mixture modeling approach. *Journal of the American Statistical Association 103*(483), 1014–1027.

Gormley, I. C. and T. B. Murphy (2009). A grade of membership model for rank data. *Bayesian Analysis 4*(2), 265–296.

Griffin, J. E. and S. G. Walker (2011). Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics 20*(1), 241–259.

Guiver, J. and E. Snelson (2009). Bayesian inference for Plackett-Luce ranking models. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*, New York, NY, USA, pp. 377–384. ACM.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics 18*(3), 1259–1294.

Hunter, D. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics 32*(1), 384–406.

Ishwaran, H. and M. Zarepour (2002). Exact and approximate sum-representations for the Dirichlet process. *Canadian Journal of Statistics 30*(2), 269–283.

James, L., A. Lijoi, and I. Prünster (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics 36*(1), 76–97.

Kalli, M., J. Griffin, and S. Walker (2011). Slice sampling mixture models. *Statistics and Computing 21*(1), 93–105.

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics 21*(1), 59–78.

Kingman, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society 37*(1), 1–22.

Lange, K., D. Hunter, and I. Yang (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics 9*, 1–59.

Lijoi, A., R. H. Mena, and I. Prünster (2005). Hierarchical mixture modelling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association 100*(472), 1278–1291.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density Estimates. *The Annals of Statistics 12*, 352–357.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* Wiley.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology 15*(3), 215–233.

McNicholas, P. (2007). Association rule analysis of CAO data. *Journal of the Statistical and Social Inquiry Society of Ireland 36*, 44–83.

Meilă, M. and L. Bao (2008). Estimation and clustering with infinite rankings. In *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence (UAI 2008)*, pp. 393–402.

Meilă, M. and H. Chen (2010). Dirichlet process mixtures of generalized Mallows models. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, pp. 358–367.

Mena, R. H. and S. G. Walker (2009). On a construction of Markov models in continuous time. *Metron-International Journal of Statistics 67*(3), 303–323.

Neal, R. M. (1992). Bayesian mixture modeling. In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, Volume 11, pp. 197–211.

Orbanz, P. (2010). Construction of nonparametric Bayesian models from parametric Bayes equations. In *Advances in Neural Information Processing Systems*.

Papaspiliopoulos, O. and G. Roberts (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika 95*(1), 169–186.

Patil, G. and C. Taillie (1977). Diversity as a concept and its implications for random communities. *Bulletin of the International Statistical Institute 47*, 497–515.

Pitman, J. (2006). *Combinatorial stochastic processes. Ecole d'été de Probabilités de Saint-Flour XXXII - 2002*, Volume 1875 of *Lecture Notes in Mathematics*. Springer.

Pitt, M. K. and S. G. Walker (2005). Constructing stationary time series models using auxiliary variables with applications. *Journal of the American Statistical Association 100*(470), 554–564.

Plackett, R. (1975). The analysis of permutations. *Applied Statistics 24*(2), 193–202.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*, Volume 12, pp. 554–560.

Rodríguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process. *Journal of the American Statistical Association 103*(483), 1131–1154.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*(476), 1566–1581.

Van Dyk, D. A. and T. Park (2008). Partially collapsed Gibbs samplers. *Journal of the American Statistical Association 103*(482), 790–796.

Walker, S. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation 36*(1), 45–54.

West, M. (1992). Hyperparameter estimation in Dirichlet process mixture models. Technical Report 1992-03, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina, USA.

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

351, Cours de la Libération
Bâtiment A 29
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399