Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programmes

Francois Caron, Yee Whye Teh, Thomas Brendan Murphy

Submitted to the Annals of Applied Statistics

BAYESIAN NONPARAMETRIC PLACKETT-LUCE MODELS FOR THE ANALYSIS OF PREFERENCES FOR COLLEGE DEGREE PROGRAMMES

BY FRANÇOIS CARON, YEE WHYE TEH AND THOMAS BRENDAN MURPHY

University of Oxford and University College Dublin

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 investigate the clustering of preferences for college degree programmes amongst Irish secondary school graduates. The existence of clusters of applicants who have similar preferences for degree programmes is established and we determine that subject matter and geographical location of the third level institution characterise these clusters.

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 who are applying to undergraduate degree programmes offered in Irish third level institutions. The third level institutions consist of universities, institutes of technologies and private colleges. This application is described in detail in Section 2.

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 $k \in [M] = \{1, \ldots, M\}$ 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 $\rho = (\rho_1, \ldots, \rho_m)$ of items $\rho_i \in [M]$: At each stage $i = 1, \ldots, m$, 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 3, 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 (2009) 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 $(X_{\rho_1}, \ldots, X_{\rho_m})$, is then a direct extension of the finite case (1):

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

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; Lijoi and Prünster, 2010) 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 4, where G is a gamma process. This is a completely random measure (Kingman, 1967; Lijoi and Prünster, 2010) 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 Appendix, we show that the 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 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 shown in Section 2, such a model is relevant for capturing heterogeneity in preferences for college degree programmes. 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 college degree programme preferences data, 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. Irish College Degree Programmes. Applications to college degree programmes in Ireland are handled by a centralised applications system called the College Application Office (CAO) (www.cao.ie); a degree programme involves studying a specific subject (broad or focussed) in a particular third level institution. The CAO handle applications for 35 different third level institutions including universities, institutes of technologies and private colleges. In autumn of each year, a list of all degree programmes for the subsequent year is made available to applicants. Quite often new degree programmes are added to the list of potential choices after the initial list has been published, thus meaning that the potential list of degree programme choices is evolving and not always completely known. Applications are completed early in the year in which the students plan to enter their college degree programme. The list of available degree programmes changes from year-to-year but has been generally growing in size year-on-year. Many degree programmes have a specific subject area, for example Mathematics, History or Computer Science, but others are more general, for example Science, Commerce or Arts. In the year 2000, which we are examining herein, there were 533 degree programmes available to be selected by the applicants. When students apply for degree programmes they rank up to ten degree programmes, in order of preference, from the list of all degree programmes that are being offered. Two examples of such applications for two different applicants are shown in Table 1.

Places in these degree programmes are allocated on the basis of the applicants' performance in the Irish Leaving Certificate examination. Students typically take between seven and nine subjects in the Leaving Certificate examination. Points between zero and one hundred are awarded for each applicant's best six subjects in the Leaving Certificate examination and the points are totalled to give an overall points score. The allocation of applicants to most degree programmes is solely on the basis of the applicant's points score and applicants with a high points score are more likely to get their high preference choices. The minimum points score of all applicants accepted into a degree programme is publicly available and is called the points requirement. It is worth mentioning that even though degree programmes may have required Leaving Certificate subjects and grades as part of the minimum entry requirements, the subjects used in the applicant's points score calculation can be any six Leaving Certificate subjects.

The college applications system in Ireland is much debated in the educational sector and it receives much attention in the Irish media. The debate has two main parts: one part of the debate is whether the current system of allocating points to students on the basis of a single Leaving Certificate examination is a fair method, especially when the points can be gained from any Leaving Certificate subjects; the other part of the debate explores the choice behaviour of the applicants and whether students are choosing degree programmes in a coherent manner. We focus on the applicant's choices which are core to the second part of the debate.

Many people feel that students don't necessarily pick degree programmes on the basis of the courses offered but that they choose on other grounds, like the perceived prestige of the degree programme. However, other factors like geographical location of the third level institution may also have an impact on the applicant's choice behaviour. The two example applications in Table 1 illustrate that a number of factors influence applicants choices. The first applicant has selected degree programmes in medicine and other health sciences, so their choices appear to be largely

based on the course material. However, the second application includes a wide variety of different degree programmes; the applicant's first choice degree programme leads to a career in Primary Teaching whereas the other degree programmes are in different areas. However, the institutions that have been chosen are geographically close (within 100 km).

In the year 1997, the Department of Education and Science, commissioned a review of the Irish college applications system. A report (Hyland, 1999) reviewed the current system and made some recommendations concerning the future of the system. In addition, four research reports were published, one of which (Tuohy, 1998) examined the applicant's choices. Tuohy (1998) used a number of exploratory data analysis techniques to investigate the degree programmes selected, but without reference to the preference ordering, and he found that subject matter was an important factor in applicant choices. More recently, Gormley and Murphy (2006) used a finite mixture of Plackett-Luce models to find clusters of applications with similar choice profiles. They fitted their model using maximum likelihood and chose the number of mixture components using the Bayesian Information Criterion (BIC). Their results also indicated that subject matter and geographical location were strong determinants of student choices. However, the model fitting paradigm used in their analysis could not find small clusters of applicants because of the manner that BIC penalises each additional mixture component. Further, McNicholas (2007) used association rule mining to further explore college applicant choices, but he restricted his attention to degree programme choice combinations that were selected by at least 0.5% of the applicants; thus, that analysis emphasised only high frequency choice behaviour.

O'Connell, Clancy and McCoy (2006) conducted a survey of new college entrants (as opposed to applicants) in 2004 and found that the choice of college where they commenced their degree programme was influenced primarily by reputation and geographical location of the third level institution, and that the choice of degree programme was influenced by intrinsic interest in the subject matter and to a lesser extent future career prospects. Whilst that study only looks at students who entered college and the degree programme that they ultimately studied, it provides a further insight into the factors that influence choice of degree programme.

We investigate the complete degree programme choice data for the year 2000 cohort of applications to the College Application Office; these data correspond to top-10 rankings of college degree programmes for 53757 applicants. The model proposed herein has a number of appealing properties because it can account for choosing from the large number of degree programmes on offer, it allows for small differences in preference between degree programmes, it facilitates discovering large and small clusters of applicants with similar preferences and the fitting in the Bayesian paradigm facilitates a deep exploration of the clustering and co-clustering of applicants.

3. An extension of the Plackett-Luce model to countably infinite choice sets. 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 $[M] = \{1, \ldots, M\}$, 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 $\rho_\ell = (\rho_{\ell 1}, \ldots, \rho_{\ell m})$, for $\ell = 1, \ldots, L$, where each $\rho_{\ell i} \in [M]$. For notational simplicity we assume that all the partial rankings are of length m.

3.1. *Finite Plackett-Luce model with gamma prior.* As noted in the introduction, the Plackett-Luce model constructs a partial ranking $\rho_\ell = (\rho_{\ell 1}, \ldots, \rho_{\ell m})$ iteratively. At the ith stage, with $i = 1, 2, \ldots, m$, we pick $\rho_{\ell i}$ as the ith item from among those not yet picked with probability

TABLE 1
Two samples from the CAO preference data. Each rank observation is an ordered list of up to ten degree programmes.

Rank	CAO code	College	Degree Programme
1	DN002	University College Dublin	Medicine
2	GY501	NUI - Galway	Medicine
3	CK701	University College Cork	Medicine
4	DN006	University College Dublin	Physiotherapy
5	TR053	Trinity College Dublin	Physiotherapy
6	DN004	University College Dublin	Radiotherapy
7	TR007	Trinity College Dublin	Clinical Speech
8	FT223	Dublin IT	Human Nutrition
9	TR084	Trinity College Dublin	Social Work
10	DN007	University College Dublin	Social Science

Rank	CAO code	College	Degree Programme
1	MI005	Mary Immaculate Limerick	Education - Primary Teaching
2	CK301	University College Cork	Law
3	CK105	University College Cork	European Studies
4	CK107	University College Cork	Language - French
5	CK101	University College Cork	Arts

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:

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

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, Laird and Rubin, 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

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

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, Hunter and Yang, 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 $w_k \sim \text{Gamma}(\frac{\alpha}{M}, \tau)$ with hyperparameters $\alpha, \tau > 0$. Now Bayesian inference can be carried out, for example, using 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):

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

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.

3.2. *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.

4. A Bayesian nonparametric Plackett-Luce model based on the gamma process. Let $\mathbb{X}$ be a measurable space of choice items. In the case of college applications, the space $\mathbb{X}$ is the space of all possible Irish programme courses. 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, Chapter 9): 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

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

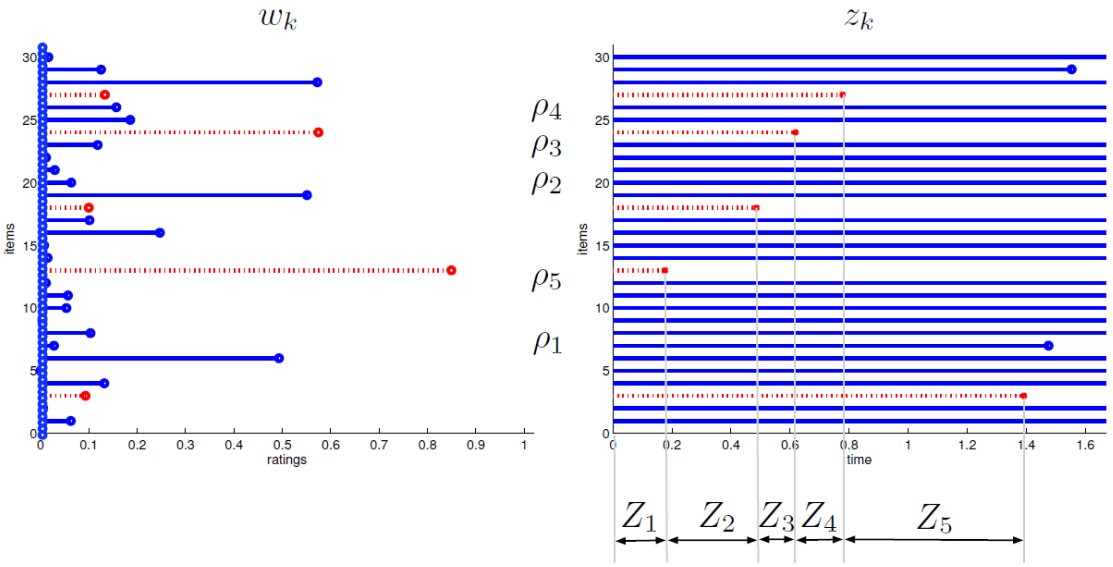


FIG 1. *Bayesian nonparametric Placket-Luce model. Left: an instantiation of the atomic measure G encapsulating both the items and their ratings. Right: Arrival times z_k and latent variables $Z_k = z_{\rho_k} - z_{\rho_{k-1}}$. The top 5 items are $(\rho_1, \rho_2, \ldots, \rho_5)$.*

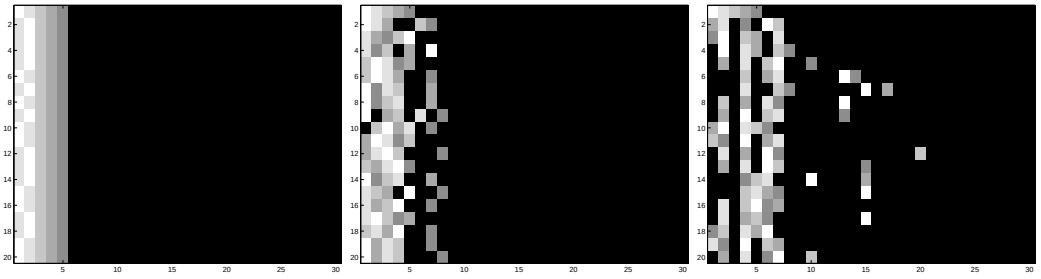
where $\alpha > 0$ is the concentration parameter and $\tau > 0$ the inverse scale. We write this as $G \sim \Gamma(\alpha, \tau, H)$. Under this parametrisation, we have that $G(A) \sim \text{Gamma}(\alpha H(A), \tau)$. $\lambda(w)h(x)$ is known as the Lévy intensity of the homogeneous CRM G. The jump part $\lambda(w)$ of the Lévy intensity verifies the necessary condition

$$\int_0^\infty (1 - \exp(-w))\lambda(w)dw < \infty \tag{6}$$

and plays a significant role in characterising the properties of the gamma process.

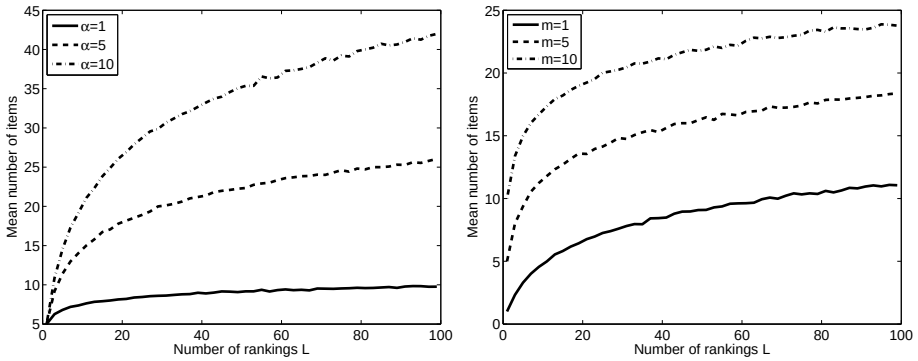
We shall interpret each atom X_k as a choice item, with its mass $w_k > 0$ corresponding to the desirability parameter. The Thurstonian view described in the finite model can be easily extended to the nonparametric one, where a partial ranking $(X_{\rho_1}, \ldots, X_{\rho_m})$ can be generated as the 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 $(X_{\rho_1}, \ldots, X_{\rho_m})$ 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 $i = 1, 2, \ldots$ (with $z_{\rho_0} = 0$). The joint probability of an observed partial ranking of length m along with the m



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

FIG 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 gamma process with* $\lambda(w) = \alpha w^{-1} \exp(-\tau w)$ *with* $\tau = 1$ *and different values of* α*. The parameter* α *tunes the variability in the partial rankings. The larger* α *the higher the variability. As the probability of partial rankings 3 is invariant to rescaling of the weights, the scaling parameter* τ *has no effect on the partial rankings.*



(a) $m = 5$ (b) $\alpha = 3$

FIG 3. *Mean number of items appearing in* L *top-*m *rankings for a gamma process with* $\lambda(w) = \alpha w^{-1} \exp(-\tau w)$ *with* $\tau = 1$ *and different values of* α *and* m*.*

associated latent variables can be derived to be:

$$
\begin{aligned}
& P((X_{\rho_1},\dots,X_{\rho_m}),(Z_1,\dots,Z_m)|G) \\
=& P((z_{\rho_1},\dots,z_{\rho_m}), \text{and } z_k > z_{\rho_m} \text{ for all } k \notin \{\rho_1,\dots,\rho_m\}) \\
=& \Bigg(\prod_{i=1}^{m} w_{\rho_i}\exp(-w_{\rho_i}z_{\rho_i})\Bigg)\Bigg(\prod_{k\notin\{\rho_1,\dots,\rho_m\}}\exp(-w_k z_{\rho_m})\Bigg) \\
=& \prod_{i=1}^{m} w_{\rho_i}\exp\Bigg(-Z_i\Bigg(\sum_{k=1}^{\infty} w_k - \sum_{j=1}^{i-1} w_{\rho_j}\Bigg)\Bigg)
\end{aligned} \tag{7}
$$

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

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

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 jump part $\lambda(w)$ of the Lévy intensity of the gamma process satisfies the following property:

$$
\int_0^{\infty} \lambda(w)dw = \infty, \tag{8}
$$

This 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. It is also a necessary and sufficient condition for the homogeneous CRM G to have finite and strictly positive total mass $0 < G(\mathbb{X}) < \infty$ (Regazzini, Lijoi and Prünster, 2003). It therefore ensures that the generative Plackett-Luce probability (3) is well defined.

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. Note that the atoms locations X_k are just used for modelling purposes. When considering inference, they are assumed to be known and need not to be defined explicitly so that to make inference on the item desirabilities.

4.1. *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. Posterior characterisation for our model is a variation of posterior characterisation for

normalised random measures in density estimation (Prünster, 2002; James, 2002; James, Lijoi and Prünster, 2009; Lijoi and Prünster, 2010). We shall denote the ℓth partial ranking as $Y_\ell = (Y_{\ell 1}, \ldots, Y_{\ell m})$, where each $Y_{\ell i} \in \mathbb{X}$. Note that previously our partial rankings $(X_{\rho_1}, \ldots, X_{\rho_m})$ 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 $(Y_{\ell 1}, \ldots, Y_{\ell m})$. Re-expressing the conditional distribution (3) of Y_ℓ given G, we have:

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

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

$$Z_{\ell i} | Y_\ell, G \sim \mathrm{Exp}(G(\mathbb{X} \backslash \{Y_{\ell 1}, \ldots, Y_{\ell i-1}\})) \tag{9}$$

The joint probability of the item lists and auxiliary variables is then (c.f. (7)):

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

Note that under the generative process described in Section 4, 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 $X_1^*, \ldots, X_K^*$, and for each $k = 1, \ldots, K$ let n_k be the number of occurrences of X_k^* among the item lists. Finally define occurrence indicators

$$\delta_{\ell i k} = \begin{cases} 0 & \text{if } \exists j < i \text{ with } Y_{\ell j} = X_k^*; \\ 1 & \text{otherwise.} \end{cases} \tag{10}$$

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

$$\begin{aligned} & P((Y_\ell, Z_\ell)_{\ell=1}^L | G) \\ &= \prod_{k=1}^{K} G(\{X_k^*\})^{n_k} \times \prod_{\ell=1}^{L} \prod_{i=1}^{m} \exp(-Z_{\ell i} G(\mathbb{X} \backslash \{Y_{\ell 1}, \ldots, Y_{\ell\, i-1}\})) \\ &= \exp\left(-G(\mathbb{X}) \sum_{\ell i} Z_{\ell i}\right) \prod_{k=1}^{K} G(\{X_k^*\})^{n_k} \exp\left(-G(\{X_k^*\}) \sum_{\ell i} (\delta_{\ell i k} - 1) Z_{\ell i}\right) \end{aligned} \tag{11}$$

Taking expectation of (11) with respect to G gives:

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

$$P((Y_\ell, Z_\ell)_{\ell=1}^L) = e^{-\psi(\sum_{\ell i} Z_{\ell i})} \prod_{k=1}^{K} h(X_k^*) \kappa\left(n_k, \sum_{\ell i} \delta_{\ell i k} Z_{\ell i}\right) \tag{12}$$

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 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 (James, Lijoi and Prünster, 2009), is given in the appendix.

Another application of the Palm formula (James, Lijoi and Prünster, 2009) now allows us to derive a posterior characterisation of G. The posterior CRM can be decomposed as the sum of a CRM with fixed atoms and a CRM whose jump part of the Lévy intensity is updated to $\lambda^*(w)$ in a conjugate fashion, similar to deriving a conjugate posterior for a parametric distribution.

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,*

$$G|(Y_\ell, Z_\ell)_{\ell=1}^L = G^* + \sum_{k=1}^K w_k^* \delta_{X_k^*} \tag{13}$$

where G^* and $w_1^*, \ldots, w_K^*$ 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:

$$\mathbb{E}[e^{-\int f(x)G(dx)}|(Y_\ell, Z_\ell)_{\ell=1}^L] = \frac{\mathbb{E}[e^{-\int f(x)G(dx)} P((Y_\ell, Z_\ell)_{\ell=1}^L|G)]}{\mathbb{E}[P((Y_\ell, Z_\ell)_{\ell=1}^L|G)]} \tag{14}$$

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 (13). The proof details are given in the appendix. ■

4.2. *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 (9), while those for the masses are given in Theorem 2:

$$\begin{aligned}
&\text{Gibbs update for } Z_{\ell i}: && Z_{\ell i}|\text{rest} \sim \text{Exp}\left(w_*^* + \textstyle\sum_k \delta_{\ell i k} w_k^*\right) \\
&\text{Gibbs update for } w_k^*: && w_k^*|\text{rest} \sim \text{Gamma}\left(n_k, \tau + \textstyle\sum_{\ell i} \delta_{\ell i k} Z_{\ell i}\right) \\
&\text{Gibbs update for } w_*^*: && w_*^*|\text{rest} \sim \text{Gamma}\left(\alpha, \tau + \textstyle\sum_{\ell i} Z_{\ell i}\right)
\end{aligned} \tag{15}$$

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 (15) before proceeding to update other variables.

In Section C in Appendix, we show that the 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.

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 degree 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:

$$\begin{aligned} \pi &\sim \mathrm{GEM}(\gamma) \\ c_\ell|\pi &\sim \mathrm{Discrete}(\pi) \quad \text{for } \ell = 1, \dots, L, \\ Y_\ell|c_\ell, G_{c_\ell} &\sim \mathrm{PL}(G_{c_\ell}) \end{aligned} \tag{16}$$

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 4. 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:

$$U_j|G_0 = \sum_{k=1}^{\infty} u_{jk}\delta_{X_k}$$
$$u_{jk}|G_0 \sim \text{Poisson}(\phi w_{0k}) \tag{17}$$

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 4.1, 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 (17). 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,

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

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

$$G_j = G_j^* + \sum_{k=1}^{\infty} w_{jk}^*\delta_{X_k} \tag{18}$$

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 $\Gamma(\alpha, \tau, H)$, it can be seen that G_j will marginally follow the same law $\Gamma(\alpha, \tau, H)$ as well. More compactly, we can write the dependence model as:

$$U_j|G_0 \sim \text{Poisson}(\phi G_0)$$
$$G_j|U_j \sim \Gamma\left(\alpha + U_j(\mathbb{X}), \tau + \phi, \frac{\alpha H + U_j}{\alpha + U_j(\mathbb{X})}\right) \tag{19}$$

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. Larger ϕ may also favour a larger number of clusters as similar partial rankings are more likely to be clustered in different groups.

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 (16), with a DP mixture with each component specified by a random atomic measure, is reminiscent of the nested DP of Rodríguez, Dunson and Gelfand (2008) as well, though our model has an additional hierarchical structure allowing the sharing of atoms among different component measures. In this respect, it also shares similarities with the hierarchical Dirichlet process model of Müller, Quintana and Rosner (2004).

We focused here on a DP mixture for its simplicity, with a single parameter γ tuning the clustering structure. The model can be generalised to more flexible random measures, such as Pitman-Yor processes (Pitman, 1995) or normalised random measures (Regazzini, Lijoi and Prünster, 2003; Lijoi, Mena and Prünster, 2007).

5.2. *Posterior characterisation and Gibbs sampling.* Assume for simplicity we have observed L top-m partial ranking $Y_\ell = (Y_{\ell 1}, \ldots, Y_{\ell m})$ (the following will trivially extend to partial rankings of differing sizes). We extend the results of Section 4 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 $Y_1, \ldots, Y_L$. 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 $Y_\ell = (X^*_{\rho_{\ell 1}}, \ldots, X^*_{\rho_{\ell m}})$, and $\delta_{\ell ik}$ be occurrence indicators similar to (10).

As in Section 4, 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 $\ell = 1, \ldots, L$ and $i = 1, \ldots, m$:

$$Z_{\ell i} | Y_\ell, c_\ell, G_{c_\ell} \sim \mathrm{Exp}\left(w_{c_\ell *} + \sum_{k=1}^{K} \delta_{\ell ik} w_{c_\ell k} \right) \tag{20}$$

The overall graphical model is described in Figure 4.

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*

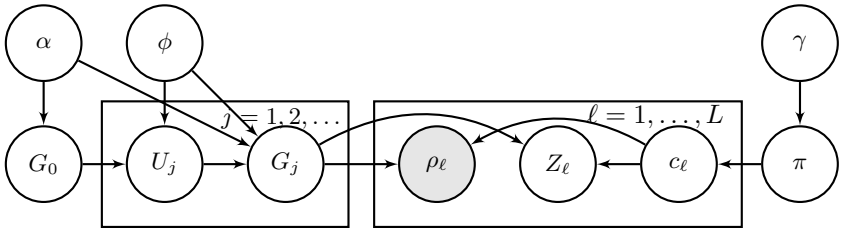


FIG 4. *Graphical model of the Dirichlet process mixture of nonparametric Plackett-Luce components. The variables at the top are hyperparameters, (ρ_ℓ) are the observed partial rankings, while the other variables are unobserved variables.*

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,

$$G_j^*|(Y_\ell),(Z_{\ell i}),(u_{jk}),(u_{j*}),(c_\ell) \sim \Gamma\left(\alpha + u_{j*}, \tau + \phi + \sum_{\ell|c_\ell=j}\sum_{i=1}^{m} Z_{\ell i}, H\right), \tag{21}$$

while the masses have distributions,

$$w_{jk}|(Y_\ell),(Z_{\ell i}),(u_{jk}),(u_{j*}),,(c_\ell) \sim \text{Gamma}\left(n_{jk} + u_{jk}, \tau + \phi + \sum_{\ell|c_\ell=j}\sum_{i=1}^{m} \delta_{\ell ik} Z_{\ell i}\right) \tag{22}$$

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:

$$p(u_{jk}|w_{0k},w_{jk}) \propto f_{\text{Gamma}}(w_{jk};u_{jk},\tau+\phi) f_{\text{Poisson}}(u_{jk};\phi w_{0k}) \tag{23}$$

$$p(u_{j*}|w_{0*},w_{j*}) \propto f_{\text{Gamma}}(w_{j*};\alpha+u_{j*},\tau+\phi) f_{\text{Poisson}}(u_{j*};\phi w_{0*}) \tag{24}$$

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):

$$\begin{aligned} p(w_{jk}|w_{0k}) =& \exp(-\phi w_{0k}) 1_{w_{jk},0} \\ &+ \mathcal{I}_{-1}\left(2\sqrt{w_{jk}\phi w_{0k}(\tau+\phi)}\right)\left(\frac{\phi(\tau+\phi)w_{0k}}{w_{jk}}\right)^{1/2} \exp\left(-\phi(w_{jk}+w_{0k}) - \tau w_{jk}\right) \end{aligned} \tag{25}$$

(26)

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

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 (23) and (24) 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:

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

   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 (23) and (24).
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 $k = 1, \ldots, K$, the masses are conditionally independent with distributions:

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

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

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

4. The latent variables $(Z_{\ell i})$ are updated as in (20).
5. Conditioned on $(Z_{\ell i})$, (u_{jk}) and (u_{j*}), the masses (w_{jk}) are updated via (22), while the total mass of the unobserved atoms is $w_{j*} \sim \text{Gamma}(\alpha_j^*, \tau_j^*)$ from (21).
6. The mixture weights π and the allocation variables c_ℓ are updated using a slice sampler for mixture models (Walker, 2007; Kalli, Griffin and Walker, 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 (25) and (26) 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. The computational cost scales as $O(K \times J \times m \times L)$ where J is the average number of clusters. However, it is possible to parallelise over the different items in the algorithm to obtain an algorithm that scales as $O(J \times m \times L)$.

6. Application: Irish College Degree Programmes. We now consider the application of the proposed model to study the choices made by the 53757 degree programme applicants to the College Application Office (CAO) in the year 2000.

6.1. *Model Set-Up & Implementation Details .* The following flat priors are used for the hyperparameters

$$p(\alpha) \propto 1/\alpha \qquad p(\phi) \propto 1/\phi \qquad p(\gamma) \propto 1/\gamma$$

We run the Gibbs sampler with $N = 20000$ iterations. In order to obtain a point estimate of the partition from the posterior distribution, we use the approach proposed by Dahl (2006). Let $c^{(i)}, i = 1, \ldots, N$ be the Monte Carlo samples. The point estimate $\widehat{c}$ is obtained by

$$\widehat{c} = \underset{c^{(i)} \in \{c^{(1)}, \ldots, c^{(N)}\}}{\arg\min} \sum_k \sum_\ell (\delta_{c_k^{(i)} c_\ell^{(i)}} - \zeta_{k\ell})^2$$

TABLE 2
Description of the different clusters. The size of the clusters, the entropy and a cluster description are provided.

Cluster	Size	Entropy	Description	Cluster	Size	Entropy	Description
1	3325	0.72	Social Science/Tourism	14	1918	0.71	Engineering
2	3214	0.71	Science	15	1835	0.48	Teaching/Arts
3	3183	0.64	Business/Commerce	16	1835	0.68	Art/Music - Dublin
4	2994	0.58	Arts	17	1740	0.71	Engineering - Dublin
5	2910	0.63	Business/Marketing - Dublin	18	1701	0.55	Medicine
6	2879	0.68	Construction	19	1675	0.70	Arts/Religion/Theology
7	2803	0.66	CS - outside Dublin	20	1631	0.76	Arts/History - Dublin
8	2225	0.67	CS - Dublin	21	1627	0.66	Galway
9	2303	0.67	Arts/Social - outside Dublin	22	1392	0.70	Limerick
10	2263	0.63	Business/Finance - Dublin	23	1273	0.65	Law
11	2198	0.65	Arts/Psychology - Dublin	24	1269	0.72	Business - Dublin
12	2086	0.63	Cork	25	1225	0.79	Arts/Bus./Language - Dublin
13	2029	0.64	Comm./Journalism - Dublin	26	47	0.96	Mixed

where the coclustering matrix ζ is obtained with

$$\zeta_{k\ell} = \frac{1}{N}\sum_{i=1}^{N} \delta_{c_k^{(i)} c_\ell^{(i)}}$$

and $\delta_{k\ell} = 1$ if $k = \ell$, 0 otherwise. Given this partition $\widehat{c}$, we run a Gibbs sampler with 2000 iterations to obtain the posterior mean Plackett-Luce parameters for each cluster. Clusters are then reordered by decreasing size. Table 2 shows the sizes of the 26 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.).

6.2. *Results.* 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 2). For example, clusters 6, 18 and 23 are characterised as construction, medicine and law, respectively. Besides the type of degree, geographical location is a strong determinant of degree programme choice. Clusters 12, 21 and 22 are respectively concerned with applications to college degree programmes in Cork, Galway and Limerick. There is a lot of heterogeneity in the subject area of the college degree programmes for these clusters, as can be seen for example for the Cork cluster 12 in Table 5. 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 3 and 4, which correspond to computer science respectively outside and inside Dublin.

As mentioned in Section 2, there is a common perception in the Irish society and media that students pick degree programme based on prestige rather than subject area. Another perception is that the points requirement for a degree programme is a measure of prestige; in fact the points requirement is determined by a number of factors including the number of available places, the number of applicants who list the degree programme in their top-10 preferences and the quality of the applicants who apply for the degree programme. Such a selection-by-prestige phenomenon should be evidenced by a cluster of students picking degree programmes in medicine and law, both of which have very high points requirements, but no such cluster was found. In fact, medicine and law applicants are clustered separately into clusters 18 and 23, respectively. Therefore, the clustering suggests that students are primarily picking degree programmes on the basis of subject area

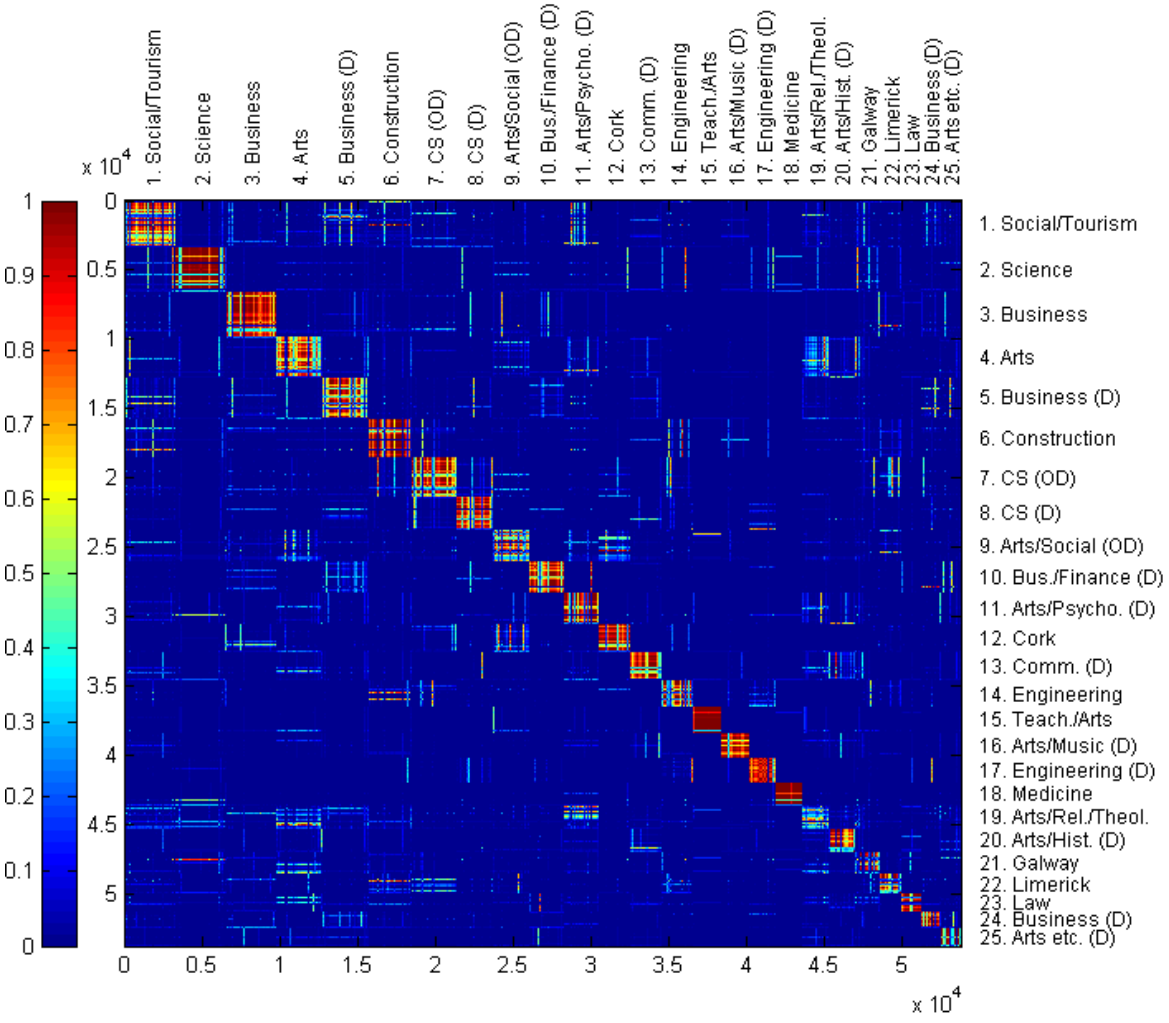


FIG 5. *Coclustering matrix of the 53757 college applicants for the CAO data. The posterior probability that two applicants belong to the same cluster is indicated by a color between blue (0) and red (1). Applicants are arranged by their cluster membership, and the clusters are ordered by size. The clusters are described in Table 2.*

TABLE 3
Cluster 7: Computer Science - outside Dublin

Rank	Aver. Norm. Weight	College	Degree Programme
1	0.081	Cork IT	Computer Applications
2	0.075	Limerick IT	Software Development
3	0.072	University of Limerick	Computer Systems
4	0.064	Waterford IT	Applied Computing
5	0.061	Cork IT	Software Dev & Comp Net
6	0.046	IT Carlow	Computer Networking
7	0.038	Athlone IT	Computer and Software Engineering
8	0.036	University College Cork	Computer Science
9	0.033	Dublin City University	Computer Applications
10	0.033	University of Limerick	Information Technology

TABLE 4
Cluster 8: Computer Science - Dublin

Rank	Aver. Norm. Weight	College	Degree Programme
1	0.141	Dublin City University	Computer Applications
2	0.054	University College Dublin	Computer Science
3	0.049	NUI - Maynooth	Computer Science
4	0.043	Dublin IT	Computer Science
5	0.040	National College of Ireland	Software Systems
6	0.038	Dublin IT	Business Info. Systems Dev.
7	0.036	Trinity College Dublin	Computer Science
8	0.035	Dublin IT	Applied Sciences/Computing
9	0.030	Trinity College Dublin	Information & Comm. Tech.
10	0.029	University College Dublin	B.A. (Computer Science)

TABLE 5
Cluster 12: Cork

Rank	Aver. Norm. Weight	College	Degree Programme
1	0.105	University College Cork	Arts
2	0.072	University College Cork	Computer Science
3	0.072	University College Cork	Commerce
4	0.067	University College Cork	Business Information Systems
5	0.057	Cork IT	Computer Applications
6	0.049	Cork IT	Software Dev & Comp Net
7	0.035	University College Cork	Finance
8	0.031	University College Cork	Law
9	0.031	University College Cork	Accounting
10	0.026	University College Cork	Biological and Chemical Sciences

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 normalised entropy, which takes its values between 0 and 1, and defined for each cluster j by

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

where $\widehat{w}_{jk}$ are the averaged normalised weights of item k in cluster j obtained from the second MCMC run; the normalised entropy values for each cluster are reported in Table 2. 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 normalised entropy, where 56% of the students in that cluster are likely to take one of the three most popular degree programmes of that cluster (Drumcondra, Froebel or Marina) as their first choice; these degree programmes are the main primary teacher education degree programmes in Dublin 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 degree programmes, and only 23% of the students are likely to take one of the three most popular degree programmes as their first choice.

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 4 and 19 which are both in the arts. Cluster 4 is characterised by arts degrees which do not require the applicants to select their major in advance, whereas cluster 19 is characterised by arts degrees where the student needs to specify their major in advance. It is worth observing that the clusters are fairly well separated, and very few

clusters exhibit the phenomenon of sharing applicants, which is further evidence that the applicants are only selecting degree programmes of a particular type (as described by the cluster names in Table 2).

Marginal Posterior distributions of the hyperparameters α, γ and ϕ are respectively in the ranges $[3, 8]$, $[2, 5]$ and $[100, 200]$. Correlation parameter ϕ is rather high. This is due to the fact that some degree programmes, such as Arts in University College Dublin or Cork often appear in the top-ten list of applicants, whatever their main subject matter is. Parameter γ is associated to the number of clusters, which is around 35. Parameter α relates to the variability of the weights within clusters (and thus to the entropy of the clusters).

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 (Regazzini, Lijoi and Prünster, 2003; James, Lijoi and Prünster, 2009; Griffin and Walker, 2011; Favaro and Teh, 2013).

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 college degree 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, Orbanz and Buhmann, 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 generalised Mallows models can be difficult.

The mixture model established the existence of clusters of applicants with similar degree programme preferences and characterises these clusters and their coherence in terms of choices. The results support the previous hypotheses that subject matter and geographical location are the primary drivers of degree programme choice (Gormley and Murphy, 2006; McNicholas, 2007). These factors are important because they reflect the intrinsic interest in the subject matter of the degree programmes and the economic and practical aspects of choosing a third level institution for study. The geographical location influence is further supported by results on acceptances to degree programmes (O'Connell, Clancy and McCoy, 2006) and studies on how students fund their education which found that 45% of Irish university students live in their family home (Clancy and Kehoe, 1999) and thus attend an institution that is geographically close by.

An interesting extension of the proposed model would be to consider inhomogeneous completely random measures, where the preferences would depend on a set of covariates (e.g. location).

Acknowledgements. The authors thank Igor Prünster for very helpful feedback on an earlier version of this work. F.C. acknowledges the support of the European Commission under the Marie

Curie Intra-European Fellowship Programme[1].

APPENDIX A: PROOF OF THEOREM 1

The marginal probability (12) is obtained by taking the expectation of (11) with respect to G. Note however that (11) 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.

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

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,

$$
=\mathbb{E}\left[e^{-\int wN(dw,dx)\sum_{\ell i} Z_{\ell i}} \prod_{k=1}^{K}\sum_{j=1}^{\infty} w_j^{n_k} \mathbb{1}(X_j = X_k^*) e^{-w_j \sum_{\ell i}(\delta_{\ell i k}-1)Z_{\ell i}}\right] \tag{27}
$$

We now recall the Palm formula (see e.g. Bertoin (2006, Lemma 2.3)).

Proposition 5 *Palm Formula. Let N be a Poisson process on S with mean measure ν. Let S_p denote the set of point measures on S, $f : S \to [0, +\infty[$ and $\mathcal{G} : S \times S_p \to [0, +\infty[$ be some measurable functional. Then we have the so-called Palm formula*

$$
\mathbb{E}\left[\int_S f(x)\mathcal{G}(x, N)N(dx)\right] = \int_S \mathbb{E}[\mathcal{G}(x, N + dx)]f(x)\nu(dx) \tag{28}
$$

where the expectation is with respect to N.

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

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

[1]The contents reflect only the authors views and not the views of the European Commission.

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

$$
\begin{aligned}
&= \mathbb{E}\left[e^{-G(\mathbb{X})\sum_{\ell i} Z_{\ell i}}\right] \prod_{k=1}^{K} h(X_k^*) \int (w_k^*)^{n_k} e^{-w_k^* \sum_{\ell i} \delta_{\ell i k} Z_{\ell i}} \lambda(w_k^*) dw_k^* \\
&= e^{-\psi(\sum_{\ell i} Z_{\ell i})} \prod_{k=1}^{K} h(X_k^*) \kappa\left(n_k, \sum_{\ell i} \delta_{\ell i k} Z_{\ell i}\right) \qquad (29)
\end{aligned}
$$

This completes the proof of Theorem 1.

APPENDIX B: PROOF OF THEOREM 2

The proof is essentially obtained by calculating the numerator and denominator of (14). 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),

$$
\begin{aligned}
&= \exp\left(-\int (1 - e^{-w(f(x)+\sum_{\ell i} Z_{\ell i})}) \lambda(w) h(x) dw dx\right) \\
&\times \prod_{k=1}^{K} h(X_k^*) \int (w_k^*)^{n_k} e^{-w_k^*(f(X_k^*)+\sum_{\ell i} \delta_{\ell i k} Z_{\ell i})} \lambda(w_k^*) dw_k^* \qquad (30)
\end{aligned}
$$

Dividing the numerator (30) by the denominator (29), the characteristic functional of the posterior G is:

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

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.

APPENDIX C: 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; Regazzini, Lijoi and Prünster, 2003; Lijoi and Prünster, 2010). 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 (8) leading to a normalisable measure G with infinitely many atoms. We will write $G \sim \mathrm{CRM}(\lambda, H)$ 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:*

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

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

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

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, Mena and Prünster, 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:

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

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:

$$P((Y_{\ell i}, Z_{\ell i}), (w_k^*), w_*^*) = e^{-w_*^*(\sum_{\ell i} Z_{\ell i}))} f(w_*^*) \prod_{k=1}^K h(X_k^*)(w_k^*)^{n_k} e^{-w_k^*(\sum_{\ell i} \delta_{\ell i k} Z_{\ell i})} \lambda(w_k^*) \tag{31}$$

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 (31) gives the marginal probability in Theorem 1'. From the joint probability (31), the Gibbs sampler can now be derived:

Gibbs update for $Z_{\ell i}$:	$Z_{\ell i}\|\text{rest} \sim \text{Exp}\left(w_*^* + \sum_k \delta_{\ell i k} w_k^*\right)$
Gibbs update for w_k^*:	$P(w_k^*\|\text{rest}) \propto (w_k^*)^{n_k} e^{-w_k^* \sum_{\ell i} Z_{\ell i}} \lambda(w_k^*)$
Gibbs update for w_*^*:	$P(w_*^*\|\text{rest}) \propto e^{-w_*^*(\sum_{\ell i} Z_{\ell i}))} f(w_*^*)$

To be concrete, consider the updates for a generalised gamma process. The conditional distribution for w_k^* can be seen to be $\text{Gamma}(n_k - \sigma, \tau + \sum_{\ell i} Z_{\ell i})$, 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:

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

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.

APPENDIX D: GIBBS SAMPLER FOR THE MIXTURE OF NONPARAMETRIC PLACKETT-LUCE COMPONENTS

Let J be the number of different values taken by c (number of clusters). Please note that the number of clusters is not set in advance and its value may change at each iteration. The Gibbs sampler proceeds with each of the following updates in turn:

1. a. Update $G_0(\mathbb{X})$ given α, then for $j = 1, \ldots, J$, update $G_j(\mathbb{X})$ given $(G_0(\mathbb{X}), \alpha, \phi, c)$
   b. For $j = 1, \ldots, J$, 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 $j = 1, \ldots, J$, update u_{j*} given $(Z, \phi, c, \alpha, w_{0*})$
   d. For $j = 1, \ldots, J$, update w_{j*} given $(Z, \alpha, u_{j*}, \phi, c)$

3. Update $(w_{0k}), w_{0*}$ given $(U_{1:J}, \alpha)$
4. For $\ell = 1, \ldots, L$, update Z_ℓ given $(w_{c_\ell}, w_{c_\ell *}, c_\ell)$
5. For $j = 1, \ldots, J$, update (w_j, w_{j*}) given $(Z, \alpha, u_j, u_{j*}, \phi, c)$
6. For $\ell = 1, \ldots, L$, update c_ℓ and the mixture weights π 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 $j = 1, \ldots, J$, 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 $j = 1, \ldots, J$

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

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

1.b) For $j = 1, \ldots, J$, 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 $j = 1, \ldots, J$ and $k = 1, \ldots, K$

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

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 $u^*_{jk} \sim$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 $u^*_{jk} \sim \text{Poisson}(\phi w_{0k})$

then set $w^*_{jk} = 0$ if $u^*_{jk} = 0$ otherwise sample $w^*_{jk} \sim \text{Gamma}(u_{jk}, \tau + \phi)$. 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 $w^*_{j*} \sim \text{Poisson}(\phi w_{0*})$ 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

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

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 $k = 1, \dots, K$, 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 $\ell = 1, \dots, L$, update Z_ℓ given $(w_{c_\ell}, w_{c_\ell *}, c_\ell)$
For $\ell = 1, \dots, L$ and $i = 1, \dots m$, 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 $j = 1, \dots, J$, update $(w_{jk}), w_{j*}$ given $(Z, \alpha, u_j, u_{j*}, \phi, c)$
For each cluster $j = 1, \dots, J$

- For each item $k = 1, \dots, K$, 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 $\ell = 1, \dots, L$, update c_ℓ and the weights π given $w_{1:J}, w_{1:J*}$
The allocation variables $(c_1, \dots, c_L)$ are updated using the slice sampling technique described in (Walker, 2007; Kalli, Griffin and Walker, 2011; Fall and Barat, 2012). It builds on the introduction of additional latent slice variables, and does not require to set any truncation. For completeness, we briefly recall here the details of the sampler. From Eq. (16), we have

$$f(Y_\ell | \pi, G) = \sum_{k=1}^{\infty} \pi_k PL(Y_\ell; G_k) \tag{32}$$

where the π_k admit the following stick-breaking representation

$$\pi_1 = v_1, \pi_k = v_k \prod_{j<k} (1 - v_j) \tag{33}$$

where the v_k are i.i.d. from $\text{Beta}(1, \gamma)$. For each observation Y_ℓ, slice sampling introduces latent variable ω_ℓ such that the joint distribution of Y_ℓ, ω_ℓ and c_ℓ is given by

$$f(Y_\ell, \omega_\ell, c_\ell | \pi, G) = 1(\omega_\ell < \pi_{c_\ell}) \, \text{PL}(Y_\ell; G_{c_\ell}) \tag{34}$$

For simplicity, assume that the c_ℓ take values in $\{1, 2, \dots, J\}$. Let μ_k be the number of allocation variables taking value $k \in \{1, \dots, J\}$. The sampler samples ω and v as a block given c, then c given v and ω.

1. (a) Sample $(\pi_1, \ldots, \pi_J, \pi_*) \sim \text{Dirichlet}(\mu_1, \ldots, \mu_J, \gamma)$
   (b) For $\ell = 1, \ldots, L$, sample $\omega_\ell \sim \text{Unif}([0, \pi_{c_\ell}])$
   (c) Set $k = J$. While $\sum_{j=1}^{k} \pi_k < (1 - \min(\omega_1, \ldots, \omega_L))$
   - Set $k = k + 1$
   - Sample $v_k \sim \text{Beta}(1, \gamma)$
   - Set $\pi_k = \pi_* v_k \prod_{j=J+1}^{k-1}(1 - v_j)$
   - Sample G_k given G_0 using Eq. (19)
2. For $\ell = 1, \ldots, L$, sample c_ℓ from

$$p(c_\ell = k) \propto 1(\pi_k > \omega_\ell)\, \text{PL}(Y_\ell; G_{c_\ell})$$

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

BERTOIN, J. (2006). *Random Fragmentation and coagulation processes*. Cambridge University Press.

BRIX, A. (1999). Generalized Gamma Measures and Shot-noise Cox Processes. *Advances in Applied Probability* **31** 929–953.

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

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

CARON, F. and TEH, Y. W. (2012). Bayesian nonparametric models for ranked data. In *Advances in Neural Information Processing Systems*, **25** 1529–1537.

CHAPMAN, R. and STAELIN, R. (1982). Exploiting Rank Ordered Choice Set Data within the Stochastic Utility Model. *Journal of Marketing Research* **19** 288–301.

CLANCY, P. and KEHOE, D. (1999). Financing Third-Level Students in Ireland. *European Journal of Education* **34** 43–57.

DAHL, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian inference for gene expression and proteomics* (K. Do, P. Muller and M. Vannucci, eds.) 201–218. Cambridge University Press.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39** 1–38.

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

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

FALL, M. D. and BARAT, E. (2012). Gibbs sampling methods for Pitman-Yor mixture models Technical Report, INRIA.

FAVARO, S. and TEH, Y. W. (2013). MCMC for normalized random measure mixture models. *Statistical Science* **28** 335-359.

FERGUSON, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics* **1** 209-230.

GORMLEY, I. C. and MURPHY, T. B. (2006). Analysis of Irish third-level college applications data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169** 361–379.

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

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

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

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

HJORT, N. L. (1990). Nonparametric Bayes Estimators Based on Beta Processes in Models for Life history Data. *Annals of Statistics* **18** 1259-1294.

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

HYLAND, A. (1999). *Commission on the Points System: Final Report and Recommendations*. *Commission on the Points System Reports*. The Stationery Office, Dublin, Ireland.

ISHWARAN, H. and ZAREPOUR, M. (2002). Exact and Approximate Sum-Representations for the Dirichlet Process. *Canadian Journal of Statistics* **30** 269–283.

JAMES, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *arXiv preprint math/0205093*.

JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics* **36** 76–97.

KALLI, M., GRIFFIN, J. E. and WALKER, S. G. (2011). Slice sampling mixture models. *Statistics and Computing* **21** 93–105.

KINGMAN, J. F. C. (1967). Completely Random Measures. *Pacific Journal of Mathematics* **21** 59–78.

KINGMAN, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society: Series B (Methodological)* **37** 1–22.

LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics* **9** 1-59.

LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2005). Hierarchical Mixture Modelling with Normalized Inverse-Gaussian Priors. *Journal of the American Statistical Association* **100** 1278–1291.

LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 715–740.

LIJOI, A. and PRÜNSTER, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (N. L. Hjort, P. M. C. Holmes and S. G. Walker, eds.) Cambridge University Press.

LO, A. Y. (1984). On a class of Bayesian Nonparametric estimates: I. Density Estimates. *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** 215–233.

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

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

MEILĂ, M. and CHEN, H. (2010). Dirichlet Process Mixtures of Generalized Mallows Models. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)* 358–367.

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

MÜLLER, P., QUINTANA, F. and ROSNER, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 735–749.

NEAL, R. M. (1992). Bayesian Mixture Modeling. In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis* **11** 197-211.

O'CONNELL, P. J., CLANCY, P. and MCCOY, S. (2006). *Who went to college in 2004? A national survey of new entrants to higher education*. The Higher Education Authority, Dublin, Ireland.

ORBANZ, P. (2009). Construction of Nonparametric Bayesian Models from Parametric Bayes Equations. In *Advances in Neural Information Processing Systems*, **22** 1392–1400.

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

PATIL, G. P. and TAILLIE, C. (1977). Diversity as a Concept and its Implications for Random Communities. *Bulletin of the International Statistical Institute* **47** 497–515.

PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102** 145–158.

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

PITT, M. K. and WALKER, S. G. (2005). Constructing stationary time series models using auxiliary variables with

applications. *Journal of the American Statistical Association* **100** 554–564.
PLACKETT, R. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **24** 193–202.
PRÜNSTER, I. (2002). Random probability measures derived from increasing additive processes and their application to Bayesian statistics PhD thesis, University of Pavia.
RASMUSSEN, C. E. (2000). The Infinite Gaussian Mixture Model. In *Advances in Neural Information Processing Systems* **12** 554-560.
REGAZZINI, E., LIJOI, A. and PRÜNSTER, I. (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics* **31** 560–585.
RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association* **103** 1131–1154.
TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* **101** 1566-1581.
TUOHY, D. (1998). *Demand for Third-Level Places*. *Commission on the Points System Research Papers* 1. The Stationery Office, Dublin, Ireland.
VAN DYK, D. A. and PARK, T. (2008). Partially collapsed Gibbs samplers. *Journal of the American Statistical Association* **103** 790–796.
WALKER, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation* **36** 45–54.
WEST, M. (1992). Hyperparameter estimation in Dirichlet process mixture models Technical Report No. 1992-03, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina, USA.

DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
OXFORD, UNITED KINGDOM
E-MAIL: francois.caron@stats.ox.ac.uk; y.w.teh@stats.ox.ac.uk

SCHOOL OF MATHEMATICAL SCIENCES
UNIVERSITY COLLEGE DUBLIN
DUBLIN, IRELAND
E-MAIL: brendan.murphy@ucd.ie