Bayesian nonparametric models for ranked data

François Caron
INRIA
IMB - University of Bordeaux
Talence, France
Francois.Caron@inria.fr

Yee Whye Teh
Department of Statistics
University of Oxford
Oxford, United Kingdom
y.w.teh@stats.ox.ac.uk

Abstract

We develop 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 the prior specified by a gamma process. We derive a posterior characterization and a simple and effective Gibbs sampler for posterior simulation. We develop a time-varying extension of our model, and apply it to the New York Times lists of weekly bestselling books.

1 Introduction

Data in the form of partial rankings, i.e. in terms of an ordered list of the top-m items, arise in many contexts. For example, in this paper we consider datasets consisting of the top 20 bestselling books as published each week by the New York Times. The Plackett-Luce model [1, 2] is a popular model for modeling such partial rankings of a finite collection of M items. It has found many applications, including choice modeling [3], sport ranking [4], and voting [5]. [6, Chap. 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 appeared, 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 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 new ones. In this paper we propose such a Bayesian nonparametric Plackett-Luce model. Our approach is built upon recent work on Bayesian inference for the (finite) Plackett-Luce model and its extensions [7, 8, 9]. Our 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 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):

$$P(X_{\rho_1}, \ldots, X_{\rho_m}) = \prod_{i=1}^{m} \frac{w_{\rho_i}}{\left(\sum_{k=1}^{\infty} w_k\right) - \left(\sum_{j=1}^{i-1} w_{\rho_j}\right)}. \tag{2}$$

To formalize the framework, a natural representation to encapsulate the pool of items along with their ratings is using an atomic measure:

$$G = \sum_{k=1}^{\infty} w_k \delta_{X_k} \tag{3}$$

1

Using this representation, note that the top item X_{ρ_1} in our list is simply a draw from the probability measure obtained by normalizing 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 normalizing. Described this way, it is clear that the Plackett-Luce model is basically a partial size-biased permutation of the atoms in G [10], and the existing machinery of random measures and exchangeable random partitions [11] can be brought to bear on our problem.

In particular, in Section 2 we will use a gamma process as the prior over the atomic measure G. This is a completely random measure [12] with gamma marginals, such that the corresponding normalized probability measure is a Dirichlet process. We will show that with the introduction of a suitable set of auxiliary variables, we can characterize the posterior law of G given observations of top-m lists distributed according to (2). A simple Gibbs sampler can then be derived to simulate from the posterior distribution. In Section 3 we develop a time-varying extension of our model and derive a simple and effective Gibbs sampler for posterior simulation. In Section 4 we apply our time-varying Bayesian nonparametric Plackett- Luce model to the aforementioned New York Times bestsellers datasets, and conclude in Section 5.

2   A Bayesian nonparametric model for partial ranking

We start this section by briefly describing a Bayesian approach to inference in finite Plackett-Luce models [9], and taking the infinite limit to arrive at the nonparametric model. This will give good intuitions for how the model operates, before we rederive the same nonparametric model more formally using gamma processes. Throughout this paper we will suppose that our data consists of L partial rankings, with $\rho_\ell = (\rho_{\ell 1}, \ldots, \rho_{\ell m})$ for $\ell \in [L]$. For notational simplicity we assume that all the partial rankings are length m.

2.1   Finite Plackett-Luce model with gamma prior

Suppose we have M choice items, with item $k \in [M]$ having a positive desirability parameter w_k. A partial ranking $\rho_\ell = (\rho_{\ell 1}, \ldots, \rho_{\ell m})$ can be constructed generatively by picking the ith item $\rho_{\ell i}$ at the ith stage for $i = 1, \ldots, m$, with probability proportional to $w_{\rho_{\ell i}}$ as in (1). An alternative Thurstonian interpretation, which will be important in the following, is as follows: For each item k let $z_{\ell k} \sim \mathrm{Exp}(w_k)$ be exponentially distributed with rate 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 ith smallest value among $(z_{\ell k})_{k=1}^M$). The resulting probability of ρ_ℓ can then be shown to still be (1). In this interpretation $(z_{\ell k})$ can be understood as latent variables, and the EM algorithm can be applied to derive an algorithm to find a ML parameter setting for $(w_k)_{k=1}^M$ given multiple partial rankings. Unfortunately the posterior distribution of $(z_{\ell k})$ given ρ_ℓ is difficult to compute directly, so we instead consider an alternative parameterization: Let $Z_{\ell i} = z_{\rho_{\ell i}} - z_{\rho_{\ell\, i-1}}$ be the waiting time for the ith item to arrive after the $i-1$th item (with $z_{\rho_{\ell 0}}$ defined to be 0). Then it can be shown that the joint probability is:

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

Note that the posterior of $(Z_{\ell i})_{i=1}^m$ is simply factorized with $Z_{\ell i} | \rho, w \sim \mathrm{Exp}(\sum_{k=1}^M w_k - \sum_{j=1}^{i-1} w_{\rho_{\ell j}})$, and the ML parameter setting can be easily derived as well. Taking a further step, we note that a factorized gamma prior over (w_k) is conjugate to (4), say $w_k \sim \mathrm{Gamma}(\frac{\alpha}{M}, \tau)$ with hyperparameters $\alpha, \tau > 0$. Now Bayesian inference can be carried out either with a VB EM algorithm, or a Gibbs sampler. In this paper we shall consider only Gibbs sampling algorithms. In this case the parameter updates are of the form

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

where n_k is the number of occurrences of item k among the observed partial rankings, and $\delta_{\ell i k} = 0$ if there is a $j < i$ with $\rho_{\ell j} = k$ and 1 otherwise. These terms arise by regrouping those in the exponential in (4).

A nonparametric Plackett-Luce model can now be easily derived by taking the limit as the number of choice items $M \to \infty$. For those items k that have appeared among the observed partial rankings,

2

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 infinitely many previously unobserved items, and show that

$$w_*|(\rho_\ell), (Z_{\ell i}), (w_k)_{k:n_k>0} \sim \text{Gamma}\left(\alpha, \tau + \sum_{\ell=1}^{L} \sum_{i=1}^{m} Z_{\ell i}\right) \qquad (6)$$

The Gibbs sampler thus alternates between updating $(Z_{\ell i})$, and updating the ratings 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 consistent manner. While intuitive, this derivation is ad hoc in the sense that it arises as the infinite limit of the Gibbs sampler for finite 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 subsection to be a gamma process.

2.2 A Bayesian nonparametric Plackett-Luce model

Let $\mathbb{X}$ be a measurable space of choice items. A gamma process is a completely random measure over $\mathbb{X}$ with gamma marginals. Specifically, it is a random atomic measure of the form (3), 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 $\{w_k\}$, it can be shown that such a random measure can be constructed as follows: 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 intensity $\lambda(w) = \alpha w^{-1} e^{-w\tau}$ where $\alpha > 0$ is the concentration parameter and $\tau > 0$ the inverse scale. We write this as $G \sim \Gamma(\alpha, \tau, H)$. Under this parametrization, we have that $G(A) \sim \text{Gamma}(\alpha H(A), \tau)$.

Each atom X_k is 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 $(X_{\rho_{\ell 1}} \ldots X_{\rho_{\ell m}})$ can be generated as the first m items to arrive in a race. In particular, for each atom X_k let $z_{\ell k} \sim \text{Exp}(w_k)$ be the time of arrival of X_k and $X_{\rho_{\ell i}}$ the ith item to arrive. The first m items to arrive $(X_{\rho_{\ell 1}} \ldots X_{\rho_{\ell m}})$ then constitutes our top-m list, with probability as given in (2). Again reparametrizing using inter-arrival durations, let $Z_{\ell i} = z_{\rho_{\ell i}} - z_{\rho_{\ell i-1}}$ for $i = 1, 2, \ldots$ (with $z_{\rho_0} = 0$). Then the joint probability is:

$$P((X_{\rho_{\ell i}})_{i=1}^{m}, (Z_{\ell i})_{i=1}^{m}|G) = P((z_{\rho_{\ell 1}} \ldots z_{\rho_{\ell m}}), \text{and } z_{\ell k} > z_{\rho_{\ell m}} \text{ for all } k \notin \{\rho_{\ell 1}, \ldots, \rho_{\ell m}\}) \quad (7)$$

$$= \left(\prod_{i=1}^{m} w_{\rho_{\ell i}} e^{-w_{\rho_{\ell i}} z_{\rho_{\ell i}}}\right) \left(\prod_{k \notin \{\rho_{\ell i}\}_{i=1}^{m}} e^{-w_k z_{\rho_{\ell m}}}\right) = \prod_{i=1}^{m} w_{\rho_{\ell i}} \exp\left(-Z_{\ell i}\left(\sum_{k=1}^{\infty} w_k - \sum_{j=1}^{i-1} w_{\rho_{\ell j}}\right)\right)$$

Marginalizing out $(Z_{\ell i})_{i=1}^{m}$ gives the probability of $(X_{\rho_{\ell i}})_{i=1}^{m}$ in (2). Further, conditional on ρ_ℓ it is seen that the inter-arrival durations $Z_{\ell 1} \ldots Z_{\ell m}$ are mutually independent and exponentially distributed:

$$Z_{\ell i}|(X_{\rho_{\ell i}})_{i=1}^{m}, G \sim \text{Exp}\left(\sum_{k=1}^{\infty} w_k - \sum_{j=1}^{i-1} w_{\rho_{\ell j}}\right) \qquad (8)$$

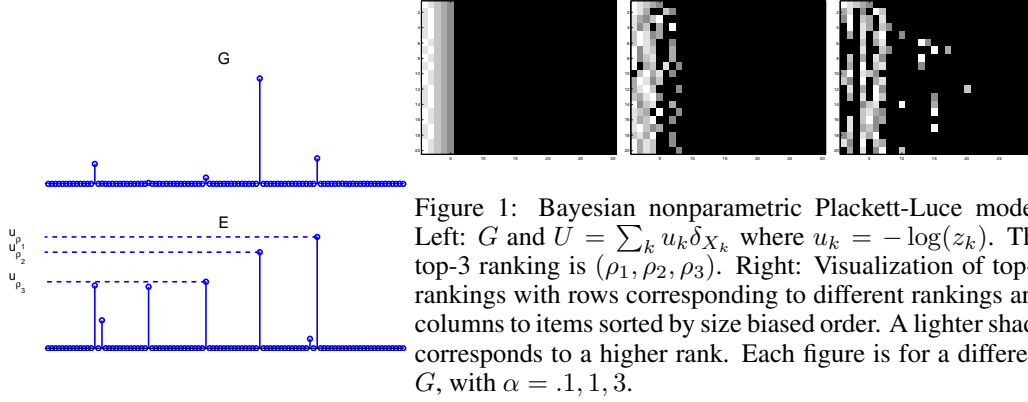
The above construction is depicted on Figure 1(left). We visualize on right some top-m lists generated from the model, with $\tau = 1$ and different values of α.

2.3 Posterior characterization

Consider a number L of partial rankings, with the ℓth list denoted $Y_\ell = (Y_{\ell 1} \ldots Y_{\ell m_\ell})$, for $\ell \in [L]$. While previously our top-m list $(X_{\rho_1} \ldots X_{\rho_m})$ consists of an ordered list of the atoms in G. Here G is unobserved and $(Y_{\ell 1} \ldots Y_{\ell m_\ell})$ is simply a list of observed choice items, which is why they were not expressed as an ordered list of atoms in G. The task here is then to characterize the posterior law of G under a gamma process prior and supposing that the observed partial rankings were drawn iid from the nonparametric Plackett-Luce model given G. Re-expressing the conditional distribution (2) of Y_ℓ given G, we have:

$$P(Y_\ell|G) = \prod_{i=1}^{m_\ell} \frac{G(\{Y_{\ell i}\})}{G(\mathbb{X} \backslash \{Y_{\ell 1} \ldots Y_{\ell\, i-1}\})} \qquad (9)$$

3

Figure 1: Bayesian nonparametric Plackett-Luce model. Left: G and $U = \sum_k u_k \delta_{X_k}$ where $u_k = -\log(z_k)$. The top-3 ranking is (ρ_1, ρ_2, ρ_3). Right: Visualization 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. Each figure is for a different G, with $\alpha = .1, 1, 3$.

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

$$Z_{\ell i}|Y_\ell, G \sim \mathrm{Exp}(G(\mathbb{X} \backslash \{Y_{\ell 1}, \ldots, Y_{\ell\, i-1}\})) \tag{10}$$

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

$$P((Y_\ell, Z_\ell)_{\ell=1}^L | G) = \prod_{\ell=1}^L \prod_{i=1}^{m_\ell} G(\{Y_{\ell i}\}) \exp(-Z_{\ell i} G(\mathbb{X} \backslash \{Y_{\ell 1}, \ldots, Y_{\ell\, i-1}\})) \tag{11}$$

Note that under the generative process described in Section 2.2, 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

$$\delta_{\ell i k} = \begin{cases} 0 & \text{if } \exists j < i \text{ with } Y_{\ell j} = X_k^*; \\ 1 & \text{otherwise.} \end{cases} \tag{12}$$

i.e. $\delta_{\ell i k}$ is the indicator of the occurence that item X_k^* does not appear at a rank lower than i in the ℓth list. Then the joint probability under the nonparametric Plackett-Luce model is:

$$P((Y_\ell, Z_\ell)_{\ell=1}^L | G) = \prod_{k=1}^K G(\{X_k^*\})^{n_k} \times \prod_{\ell=1}^L \prod_{i=1}^{m_\ell} \exp(-Z_{\ell i} G(\mathbb{X} \backslash \{Y_{\ell 1}, \ldots, Y_{\ell\, i-1}\}))$$

$$= \exp\left(-G(\mathbb{X}) \sum_{\ell i} Z_{\ell i}\right) \prod_{k=1}^K G(\{X_k^*\})^{n_k} \exp\left(-G(\{X_k^*\}) \sum_{\ell i} (\delta_{\ell i k} - 1) Z_{\ell i}\right) \tag{13}$$

Taking expectation of (13) with respect to G using the Palm formula gives:

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

$$P((Y_\ell, Z_\ell)_{\ell=1}^L) = e^{-\psi(\sum_{\ell i} Z_{\ell i})} \prod_{k=1}^K h(X_k^*) \kappa\left(n_k, \sum_{\ell i} \delta_{\ell i k} Z_{\ell i}\right) \tag{14}$$

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

$$\psi(z) = -\log \mathbb{E}\left[e^{-z G(\mathbb{X})}\right] = \int_{\mathbb{R}+} \lambda(w)(1 - e^{-zw})dw = \alpha \log\left(1 + \frac{z}{\tau}\right) \tag{15}$$

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

$$\kappa(n, z) = \int_{\mathbb{R}+} \lambda(w) w^n e^{-zw} dw = \frac{\alpha}{(z + \tau)^n} \Gamma(n) \tag{16}$$

Details are given in the supplementary material. Another application of the Palm formula now allows us to derive a posterior characterisation of G:

4

Theorem 2 *Given the observations and associated auxiliary 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,*

$$G|(Y_\ell, Z_\ell)_{\ell=1}^L = G^* + \sum_{k=1}^K w_k^* \delta_{X_k^*} \tag{17}$$

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

$$G^*|(X_\ell, Z_\ell)_{\ell=1}^L \sim \Gamma(\alpha, \tau^*, h) \qquad\qquad \tau^* = \tau + \sum_{\ell i} Z_{\ell i} \tag{18}$$

while the masses have distributions,

$$w_k^*|(Y_\ell, Z_\ell)_{\ell=1}^L \sim \mathrm{Gamma}\left(n_k, \tau + \sum_{\ell i} \delta_{\ell i k} Z_{\ell i}\right) \tag{19}$$

2.4 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^*) and the auxiliary variables $(Z_{\ell i})$. The update for $Z_{\ell i}$ is given by (10), while those for the masses are given in Theorem 2:

Gibbs update for $Z_{\ell i}$: $\qquad Z_{\ell i}|\mathrm{rest} \sim \mathrm{Exp}\left(w_*^* + \sum_k \delta_{\ell i k} w_k^*\right) \tag{20}$

Gibbs update for w_k^*: $\qquad w_k^*|\mathrm{rest} \sim \mathrm{Gamma}\left(n_k, \tau + \sum_{\ell i} \delta_{\ell i k} Z_{\ell i}\right) \tag{21}$

Gibbs update for w_*^*: $\qquad w_*^*|\mathrm{rest} \sim \mathrm{Gamma}\left(\alpha, \tau + \sum_{\ell i} Z_{\ell i}\right) \tag{22}$

Note that the auxiliary 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 $\mathrm{Gamma}(a, b)$ prior is placed on it, its conditional distribution is still gamma:

Gibbs update for α: $\qquad \alpha|\mathrm{rest} \sim \mathrm{Gamma}\left(a + K, b + \log\left(1 + \frac{\sum_{\ell i} Z_{\ell i}}{\tau}\right)\right) \tag{23}$

Note that this update was derived with w_*^* marginalized out, so after an update to α it is necessary to immediately update w_*^* via (22) before proceeding to update other variables.

3 Dynamic Bayesian nonparametric ranking models

In this section we develop an extension of the Bayesian nonparametric Plackett-Luce model to model time-varying rankings, where the rating parameters of items may change smoothly over time and reflected in a changing series of rankings. Given a series of times indexed by $t = 1, 2, \ldots$, we may model the rankings at time t using a gamma process distributed random measure G_t as in Section 2.2, with Markov dependence among the sequence of measures (G_t) enabling dependence among the rankings over time.

3.1 Pitt-Walker dependence model

We will construct a dependent sequence (G_t) which marginally follow a gamma process $\Gamma(\alpha, \tau, H)$ using the construction of [13]. Suppose $G_t \sim \Gamma(\alpha, \tau, H)$. Since G_t is atomic, we can write it in the form:

$$G_t = \sum_{k=1}^\infty w_{tk} \delta_{X_{tk}} \tag{24}$$

Define a random measure C_t with conditional law:

$$C_t|G_t = \sum_{k=1}^\infty c_{tk} \delta_{X_{tk}} \qquad\qquad c_{tk}|G_t \sim \mathrm{Poisson}(\phi_t w_{tk}) \tag{25}$$

where $\phi_t > 0$ is a dependence parameter. Using the same method as in Section 2.3, we can show:

5

Proposition 3 *Suppose the law of G_t is $\Gamma(\alpha, \tau, H)$. The conditional law of G_t given C_t is then:*

$$G_t = G_t^* + \sum_{k=1}^{\infty} w_{tk}^* \delta_{X_{tk}} \tag{26}$$

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

$$G_t^*|C_t \sim \Gamma(\alpha, \tau + \phi_t, H) \qquad\qquad w_{tk}^*|C_t \sim \text{Gamma}(c_{tk}, \tau + \phi_t) \tag{27}$$

The idea of [13] is to define the conditional law of G_{t+1} given G_t and C_t to coincide with the conditional law of G_t given C_t as in Proposition 3. In other words, define

$$G_{t+1} = G_{t+1}^* + \sum_{k=1}^{\infty} w_{t+1,k} \delta_{X_{tk}} \tag{28}$$

where $G_{t+1}^* \sim \Gamma(\alpha, \tau + \phi_t, H)$ and $w_{t+1,k} \sim \text{Gamma}(c_{tk}, \tau + \phi_t)$ are mutually independent. If the prior law of G_t is $\Gamma(\alpha, \tau, H)$, the marginal law of G_{t+1} will be $\Gamma(\alpha, \tau, H)$ as well when both G_t and C_t are marginalized out, thus maintaining a form of stationarity. Further, although we have described the process in order of increasing t, the joint law of G_t, C_t, G_{t+1} can equivalently be described in the reverse order with the same conditional laws as above. Note that if $c_{tk} = 0$, the conditional distribution of $w_{t+1,k}$ will be degenerate at 0. Hence G_{t+1} has an atom at X_{tk} if and only if C_t has an atom at X_{tk}, that is, if $c_{tk} > 0$. In addition, it also has atoms (those in G_{t+1}^*) where C_t does not (nor does G_t). Finally, the parameter ϕ_t can be interpreted as controlling the strength of dependence between G_{t+1} and G_t. Indeed it can be shown that

$$\mathbb{E}[G_{t+1}|G_t] = \frac{\phi_t}{\phi_t + \tau} G_t + \frac{\tau}{\phi_t + \tau} H. \tag{29}$$

Another measure of dependence can be gleaned by examining the "lifetime" of an atom. Suppose X is an atom in G_1 with mass $w > 0$. The probability that X is an atom in C_2 with positive mass is $1 - \exp(-\phi_1 w)$, in which case it has positive mass in G_2 as well. Conversely, once it is not an atom, it will never be an atom in the future since the base distribution H is non-atomic. The lifetime of the atom is then the smallest t such that it is no longer an atom. We can show by induction that: (details in supplementary material)

Proposition 4 *The probability that an atom X in G_1 with mass $w > 0$ is dead at time t is given by*

$$P(G_t(\{X\}) = 0|w) = \exp(-y_{t|1} w)$$

where $y_{t|1}$ can be obtained by the recurrence $y_{t|t-1} = \phi_{t-1}$ and $y_{t|s-1} = \frac{y_{t|s} \phi_{s-1}}{\phi_{s-1} + \tau + y_{t|s}}$.

3.2 Posterior characterization and Gibbs sampling

Assume for simplicity that at each time step $t = 1, \ldots, T$ we observe one top-m list $Y_t = (Y_{t1}, \ldots, Y_{tm})$ (it trivially extends to multiple partial rankings of differing sizes). We extend the results of the previous section in characterizing the posterior and developing a Gibbs sampler for the dynamical model.

Since each observed item at time t has to be an atom in its corresponding random measure G_t, and atoms in G_t can propagate to neighboring random measures via the Pitt-Walker dependence model, we conclude that the set of all observed items (through all times) has to include all fixed atoms in the posterior of G_t. Thus let $X^* = (X_k^*)$, $k = 1, \ldots, K$ be the set of unique items observed in $Y_1, \ldots, Y_T$, let $n_{tk} \in \{0, 1\}$ be the number of times the item X_k^* appears at time t, and let ρ_t be defined as $Y_t = (X_{\rho_1}^*, \ldots, X_{\rho_m}^*)$. We write the masses of the fixed atoms as $w_{tk} = G_t(\{X_k^*\})$, while the total mass of all other random atoms is denoted $w_{t*} = G_t(\mathbb{X}\backslash X^*)$. Note that w_{tk} has to be positive on a random contiguous interval of time that includes all observations of X_k^*—it's lifetime—but is zero outside of the interval. We also write $c_{tk} = C_t(\{X_k^*\})$ and $c_{t*} = C_t(\mathbb{X}\backslash X^*)$. As before, we introduce, for $t = 1, \ldots, T$ and $i = 1, \ldots, m$, latent variables

$$Z_{ti} \sim \text{Exp}\left(w_{t*} + \sum_{k=1}^{K} w_{tk} - \sum_{j=1}^{i-1} w_{t\rho_j}\right) \tag{30}$$

6

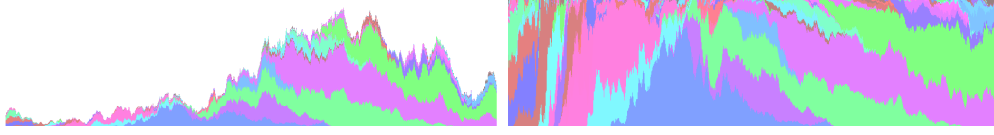



Figure 2: Sample path drawn from the Dawson-Watanabe superprocess. Each colour represents an atom, with height being its (varying) mass. Left shows (G_t) and right $(G_t/G_t(\mathbb{X}))$, a Fleming-Viot process.

Each iteration of the Gibbs sampler then proceeds as follows (details in supplementary material). The latent variables (Z_{ti}) are updated as above. Conditioned on the latent variables (Z_{ti}), (c_{tk}) and (c_{t*}), we update the masses (w_{tk}), which are independent and gamma distributed since all likelihoods are of gamma form. Note that the total masses $(G_t(\mathbb{X}))$ are not likelihood identifiable, so we introduce an extra step to improve mixing by sampling them from the prior (integrating out (c_{tk}), (c_{t*})), scaling all masses along with it. Directly after this step we update (c_{tk}), (c_{t*}). We update α along with the random masses (w_{t*}) and (c_{t*}) efficiently using a forward-backward recursion. Finally, the dependence parameters (ϕ_t) are updated.

3.3 Continuous time formulation using superprocesses

The dynamic model described in the previous section is formulated for discrete time data. When the time interval between ranking observations is not constant, it is desirable to work with dynamic models evolving over continuous-time instead, with the underlying random measures (G_t) defined over all $t \in \mathbb{R}$, but with observations at a discrete set of times $t_1 < t_2 < \cdots$. Here we propose a continuous-time model based on the Dawson-Watanabe superprocess [14, 15] (see also [16, 17, 18, 19]). This is a diffusion on the space of measures with the gamma process $\Gamma(\alpha, \tau, H)$ as its equilibrium distribution. It is defined by a generator

$$\mathcal{L} = \xi \left(\int G(dX) \frac{\partial^2}{\partial G(X)^2} + \alpha \int H(dX) \frac{\partial}{\partial G(X)} - \tau \int G(dX) \frac{\partial}{\partial G(X)} \right)$$

with ξ parametrizing the rate of evolution. Figure 2 gives a sample path, where we see that it is continuous but non-differentiable. For efficient inference, it is desirable to be able to integrate out all G_t's except those $G_{t_1}, G_{t_2}, \ldots$ at observation times. An advantage to using the Dawson-Watanabe superprocess is that, the conditional distribution of G_{t_s} given $G_{t_{s-1}}$ is remarkably simple [20]. In particular it is simply given by the discrete-time process of the previous section with dependence parameter $\phi_{t_s|t_{s-1}} = \frac{\tau}{e^{\tau \xi (t_s - t_{s-1})} - 1}$. Thus the inference algorithm developed previously is directly applicable to the continuous-time model too.

4 Experiments

We apply the discrete-time dynamic Plackett-Luce model to the New York Times bestsellers data. These consist of the weekly top-20 best-sellers list from June 2008 to April 2012 in various categories. We consider here the categories paperback nonfiction (PN) and hardcover fiction (HF), for which respectively 249 and 916 books appear at least once in the top-20 lists over the 200 weeks. We consider that the correlation parameter $\phi_t = \phi$ is constant over time, and assign flat improper priors $p(\alpha) \propto 1/\alpha$ and $p(\phi) \propto 1/\phi$. In order to take into account the publication date of a book, we do not consider books in the likelihood before their first appearance in a list. We run the Gibbs sampler with 10000 burn-in iterations followed by 10000 samples. Mean normalized weights for the more popular books in both categories are shown in Figure 3.

The model is able to estimate the weights associated to each book that appeared at least once, as well as the total weight associated to all other books, i.e. the probability that a new book enters at the first rank in the list, represented by the black curve. Moreover, the Bayesian approach enables us to have a measure of the uncertainty on the weights. The hardcover fiction category is characterized by rapid changes in successive lists, compared to the paperback nonfiction. This is quantified by the estimated value of the parameter ϕ, which are respectively 85 ± 20 and 140 ± 40 for PN and HF. The estimated values of the shape parameter α are 7 ± 1.5 and 2 ± 1 respectively.

7

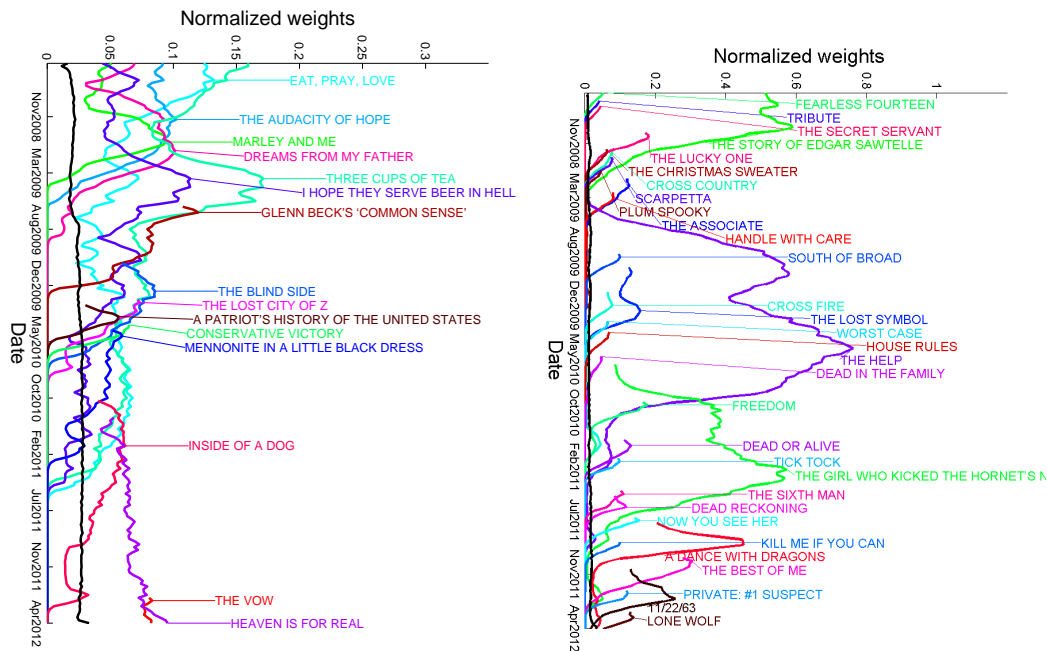



Figure 3: Mean normalized weights for paperback nonfiction (left) and hardcover fiction (right). The black lines represent the weight associated to all the books that have not appear in the top-20 lists.

5   Discussion

We have proposed a Bayesian nonparametric Plackett-Luce model for ranked data. Our approach is based on the theory of atomic 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 characterized the posterior distribution, and derived a simple MCMC sampling algorithm for posterior simulation. Our approach can be see as a multi-stage generalization of posterior inference in normalized random measures [21, 22, 23], and can be easily extended from gamma processes to general completely random measures.

We also proposed dynamical extensions of our model for both discrete and continuous time data, and applied it to modeling the bestsellers' lists on the New York Times. Our dynamic extension may be useful for modeling time varying densities or clusterings as well. In our experiments we found that our model is insufficient to capture the empirical observation that bestsellers often start off high on the lists and tail off afterwards, since our model has continuous sample paths. We adjusted for this by simply not including books in the model prior to their publication date. It may be possible to model this better using models with discontinuous sample paths, for example, the Orstein-Uhlenbeck approach of [24] where the process evolves via a series of discrete jump events instead of continuously.

Acknowledgements

YWT thanks the Gatsby Charitable Foundation for generous funding.

8

References

[1] R.D. Luce. *Individual choice behavior: A theoretical analysis*. Wiley, 1959.

[2] R. Plackett. The analysis of permutations. *Applied Statistics*, 24:193–202, 1975.

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

[4] D.R. Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32:384–406, 2004.

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

[6] P. Diaconis. *Group representations in probability and statistics, IMS Lecture Notes*, volume 11. Institute of Mathematical Statistics, 1988.

[7] I.C. Gormley and T.B. Murphy. A grade of membership model for rank data. *Bayesian Analysis*, 4:265–296, 2009.

[8] J. Guiver and E. Snelson. Bayesian inference for Plackett-Luce ranking models. In *International Conference on Machine Learning*, 2009.

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

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

[11] J. Pitman. *Combinatorial stochastic processes. Ecole d'été de Probabilités de Saint-Flour XXXII - 2002*, volume 1875 of *Lecture Notes in Mathematics*. Springer, 2006.

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

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

[14] S. Watanabe. A limit theorem of branching processes and continuous state branching processes. *Journal of Mathematics of Kyoto University*, 8:141–167, 1968.

[15] D. A. Dawson. Stochastic evolution equations and related measure processes. *Journal of Multivariate Analysis*, 5:1–52, 1975.

[16] S.N. Ethier and RC Griffiths. The transition function of a measure-valued branching diffusion with immigration. *Stochastic Processes. A Festschrift in Honour of Gopinath Kallianpur (S. Cambanis, J. Ghosh, RL Karandikar and PK Sen, eds.)*, 71:79, 1993.

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

[18] S. Feng. *Poisson-Dirichlet Distribution and Related Topics*. Springer, 2010.

[19] J.C. Cox, J.E. Ingersoll Jr, and S.A. Ross. A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society*, pages 385–407, 1985.

[20] S. N. Ethier and R. C. Griffiths. The transition function of a measure-valued branching diffusion with immigration. *Stochastic Processes*, 1993.

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

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

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

[24] J. E. Griffin. The Ornstein-Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference. *Journal of Statistical Planning and Inference*, 141:3648–3664, 2011.

9