Biostatistics (2018) **19**, 3, *pp.* 342–358
doi:10.1093/biostatistics/kxx042
Advance Access publication on September 6, 2017

Dependent generalized Dirichlet process priors for the analysis of acute lymphoblastic leukemia

WILLIAM BARCELLA*, MARIA DE IORIO

Department of Statistical Science, University College London, 1-19 Torrington Place, London WC1E 7HB, UK

william.barcella.13@ucl.ac.uk

STEFANO FAVARO

Department of Economics and Statistics, University of Torino, Corso Unione Sovietica 218/bis, Torino 10134, Italy

GARY L. ROSNER

Oncology Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, 550 N. Broadway, Suite 1103, Baltimore, MD 21205,USA

Summary

We propose a novel Bayesian nonparametric process prior for modeling a collection of random discrete distributions. This process is defined by including a suitable Beta regression framework within a generalized Dirichlet process to induce dependence among the discrete random distributions. This strategy allows for covariate dependent clustering of the observations. Some advantages of the proposed approach include wide applicability, ease of interpretation, and availability of efficient MCMC algorithms. The motivation for this work is the study of the impact of asparginage metabolism on lipid levels in a group of pediatric patients treated for acute lymphoblastic leukemia.

Keywords: Bayesian nonparametrics; Beta regression; Dependent random probability measures; Generalized Dirichlet process; Stick-breaking processes.

1. Introduction

Very often real-world applications involve observational data that are collected in groups or clusters that can be characterized, for example, by spatial or temporal coordinates, as samples from the same experimental unit, or more generally by shared levels of covariates. While the groupings may be known at the time of data collection, some clustering may be unobserved. While covariates may allow consolidation of observations into subgroups, unobserved factors may also lead to latent clusters. Learning about these latent clusters in the presence of measured covariates may provide insight into underlying mechanisms. An example, the one that motivated this research, is the analysis of longitudinally collected measurements with the goal of learning about temporal relationships and how patient characteristics may affect these associations.

*To whom correspondence should be addressed.

This study sought to learn about the relationship between levels of an anti-cancer drug (asparaginase), administered corticosteroids (dexamethasone), and alterations of circulating lipids. Asparaginase is used to treat children with acute lymphoblastic leukemia (ALL). Osteonecrosis, a condition that leads to bone cell death and pain, occasionally occurs as a result of anti-ALL therapy. Complex relationships between asparaginase and dexamethasone on albumin and triglycerides may affect a patient's risk of osteonecrosis. We wanted to model trajectories of triglyceride levels and their relationships with asparaginase activity and albumin to improve our understanding of the relationship and the ability to predict these trajectories.

In settings with spatially or temporally grouped measurements, a common strategy is to model the data by introducing random effects to account for the correlation of the observations within each group. The main consequence of this approach is that the parameters shared by all clusters are robustly estimated. Generalized linear mixed models are an example in the regression framework. Common distributions for random effects are e.g. normal distributions or Student-t distributions, but these may be too restrictive in some circumstances. A variety of solutions have been presented as more flexible alternatives. Among these proposals, nonparametric techniques, such as infinite mixture models, are gaining popularity. The most general proposals for random effects' distributions assume an infinite mixture model for groups of observations and introduce dependence among the parameters of the mixture models (i.e., the weights and/or the locations). Each infinite mixture model is a convolution of a parametric density kernel with a discrete random probability measure that has (*a priori*) an infinite number of locations and weights. Thus, the problem of inducing dependence among the infinite mixture models can be rewritten in terms of the dependence among the discrete random probability measures indexed by the different groups or clusters of observations.

A seminal contribution in this field is the extension of the Dirichlet process (DP, Ferguson, 1973) called the dependent DP (DDP, MacEachern, 1999; MacEachern, 2000). The DDP is constructed in such a way that each group of observations is distributed as a DP. The random effects' distributions thereby become DP mixture models (DPM, Lo (1984)) and the observations are implicitly clustered by belonging to different mixture components of the sampling model. Dependence among the different DP probability measures is induced by specifying convenient stochastic process priors indexed by the groups of observations, leading to group-specific weights and locations. One can specify such models by enriching the structure of the stick-breaking representation of the DP presented by Sethuraman (1994).

Recently, a large number of extensions of the DDP have been introduced to incorporate covariates. These contributions may be classified in two groups. One group is a variation of the DDP in which only the locations of the discrete random measures are indexed by the covariate space (De Iorio *and others*, 2004, 2009; Gelfand *and others*, 2005, among the others). Most of the proposals that belong to this group preserve the DP marginals, as happens for the DDP. The second group of variations of the DDP has the weights indexed by the covariate space but maintains common cluster locations across covariate levels. Some examples include the works of Griffin and Steel (2006), Rodriguez and Dunson (2011), Ren *and others* (2011), Dunson and Park (2008), and Karabatsos *and others* (2012).

Alternative solutions introduce dependence within the more general construction of the discrete random measures based on Poisson random measures (Kingman, 1967). See for instance the works of Müller *and others* (2004), Griffin and Leisen (2017), and Lijoi *and others* (2014).

In this article, we propose a novel approach that generalizes the DDP by MacEachern (2000) by assuming that the discrete random measure associated with each group of observations is distributed according to a generalized Dirichlet process prior (GDP, Hjort (2000)). The GDP employs a richer parameterization compared to the usual DP and, for this reason, allows more flexibility. The dependence among the different random measures is induced by specifying a convenient prior for the weights of the measures, while assuming the locations to be the same across all groups of observations (although alternatives with also covariate dependent locations can be easily specified). We call the resulting process the dependent GDP (DGDP). The DGDP has a better control of the implicit partition of the observations defined by the

different mixture components compared to the DDP's case in terms of the distributions of number and size of the clusters. The law of the partition induced by samples from a GDP can be derived analytically allowing for a better interpretation of that quantity and an increased number of computational strategies compared to other processes where this cannot be derived. Furthermore, including the covariates within the weights of the process, the DGDP leads to improved predictive power compared to processes including covariate information only within the locations (Cruz-Marcelo *and others*, 2013).

We use the DGDP in order to consider potential (latent) groupings of patients over and above groupings based on measured covariates including less restrictive prior distributions, while preserving computational simplicity.

The article is organized as follows. Section 2 reviews the main properties of the GDP and presents some new results. In Section 3, we introduce the DGDP and we present two possible MCMC algorithms for posterior inference in Section 4. The analysis of the ALL data with the DGDP is in Section 5. We conclude with a discussion in Section 6. Proofs and details of two MCMC algorithms for posterior inference are deferred to supplementary material available at *Biostatistics* online.

2. GENERALIZED DIRICHLET PROCESS

2.1. *Definition*

Let us consider a measurable space $(\Theta, \mathcal{A})$ and an associated probability measure $G \in \mathcal{G}$. We say that G is distributed according to a GDP (Ishwaran and James, 2001; Hjort, 2000) with parameters $\boldsymbol{\phi} = \{\phi_h\}_{h=1}^{\infty}$ (with each element belonging to $\mathbb{R}^+$), $\boldsymbol{\mu} = \{\mu_h\}_{h=1}^{\infty}$ (with each element belonging to $(0, 1)$), and center measure G_0 (a non-atomic probability measure on Θ) if it admits the following stick-breaking representation:

$$G = \sum_{h=1}^{\infty} W_h \delta_{\theta_h}, \tag{2.1}$$

where $\{\theta_h\}_{h=1}^{\infty} \overset{iid}{\sim} G_0$ and $\{W_h\}_{h=1}^{\infty}$ are constructed via the stick-breaking procedure. The procedure involves a sequence of random variables $\{V_h\}_{h=1}^{\infty}$ taking values on $(0, 1)$. Common practice is to assume these are Beta distributed, which we do here. We specify the Beta density function as

$$\mathbb{P}[V_h \in \mathrm{d}v_h \mid \phi_h, \mu_h] = \frac{\Gamma(\phi_h)}{\Gamma(\phi_h\mu_h)\Gamma(\phi_h(1-\mu_h))} v_h^{\phi_h\mu_h - 1}(1 - v_h)^{\phi_h(1-\mu_h)-1}\mathrm{d}v_h,$$

where $\mu_h = \mathbb{E}[V_h]$ and $\phi_h = \mu_h(1 - \mu_h)/\mathbb{V}[V_h] - 1$, with $\mathbb{E}[\cdot]$ and $\mathbb{V}[\cdot]$ denoting the expectation and variance operators, respectively. Thus, we assume $\{V_h\}_{h=1}^{\infty} \overset{ind}{\sim} \mathrm{Beta}(v_h \mid \phi_h\mu_h, \phi_h(1 - \mu_h))$, independent from $\{\theta_h\}_{h=1}^{\infty}$, and we specify the infinite sequence of weights setting $W_1 = V_1$ and obtaining the other weights as

$$W_r = V_r \prod_{l=1}^{r-1}(1 - V_l), \quad r = 2, 3, \ldots. \tag{2.2}$$

The resulting measure, G, is a proper random distribution function. Indeed it can be easily verified that $\sum_{h=1}^{\infty} \mathbb{E}[\log(1 - V_h)] = -\infty$, which is a necessary and sufficient condition for $\sum_{h=1}^{\infty} W_h = 1$. (See Ishwaran and James (2001) for a detailed proof). We write $G \sim \mathrm{GDP}(\boldsymbol{\phi}, \boldsymbol{\mu}, G_0)$.

A more parsimonious formulation of the GDP, described by Hjort (2000), assumes $\{\phi_h\}_{h=1}^{\infty} = \phi$ and $\{\mu_h\}_{h=1}^{\infty} = \mu$; we denote it as GDP$(\phi, \mu, G_0)$. In Section S.1 of supplementary material available at *Biostatistics* online, we include a plot with the cumulative distribution functions of realizations from a GDP.

As the name suggests, the GDP generalizes the well-known Dirichlet Process (DP, Ferguson (1973)), which can be specified by a GDP with $\{\phi_h = \mu_h^{-1}\}_{h=1}^{\infty}$ and $\{\mu_h\}_{h=1}^{\infty} = \mu$.

2.2. *Moments*

If $G \sim \text{GDP}(\boldsymbol{\phi}, \boldsymbol{\mu}, G_0)$, we have that

$$\mathbb{E}[G(A)] = G_0(A).$$

The variance of $G(A)$ is given by

$$\mathbb{V}[G(A)] = (1 - G_0(A))G_0(A)\mathbb{E}\left[\sum_{h=1}^{\infty} W_h^2\right]. \tag{2.3}$$

The expectation in the last equation cannot be computed explicitly, unless we consider the constant-parameter case GDP(ϕ, μ, G_0). With constant parameters, Hjort (2000) showed that

$$\mathbb{E}\left[\sum_{h=1}^{\infty} W_h^2\right] = \frac{\mathbb{E}[V^2]}{2\mathbb{E}[V] - \mathbb{E}[V^2]},$$

where V is a Beta random variable with parameters $(\phi\mu, \phi(1-\mu))$. Thus, $\mathbb{E}[V] = \mu$ and $\mathbb{E}[V^2] = \mu(1-\mu)/(\phi+1) + \mu^2$.

Hjort (2000) discussed also the computation of higher order moments of $G(A)$ to demonstrate the extra flexibility gained by the richer parameterization employed by the GDP, compared to the DP.

2.3. *Distributional sampling properties*

We now derive some properties of the GDP. Consider G as in (2.1). Because G is discrete, a sample $(\theta_1, \ldots, \theta_n)$ from G induces a random partition of the set $\{1, \ldots, n\}$ into $K_n = k$ blocks with frequencies $(N_1, \ldots, N_{K_n}) = (n_1, \ldots, n_k)$. We denote by $p(n_1, \ldots, n_k)$ the probability of any particular partition of $\{1, \ldots, n\}$, with k blocks and block-specific frequencies $(n_1, \ldots, n_k)$. In Definition 4 of Pitman (1995) this is referred to as the partially exchangeable partition probability function. Under the GDP with constant parameters, an application of Corollary 7 of Pitman (1995) leads to an explicit expression for $p(n_1, \ldots, n_k)$, i.e.

$$\begin{aligned} &p(n_1, \ldots, n_k) \\ &= \mathbb{E}\left[\left(\prod_{h=1}^{k} W_h^{n_h - 1}\right)\prod_{h=1}^{k-1}\left(1 - \sum_{i=1}^{h} W_i\right)\right] \\ &= \frac{(\phi(1-\mu))^{k-1}}{(\phi)_{(n-1)}}\prod_{i=1}^{k-1}\frac{(\phi(1-\mu)+1)_{(\sum_{j=i+1}^{k} n_j - 1)}}{(\phi)_{(\sum_{j=i+1}^{k} n_j - 1)}}\prod_{i=1}^{k}(\phi\mu)_{(n_i-1)}, \end{aligned} \tag{2.4}$$

where $(a)_{(b)} = a(a+1)\cdots(a+b-1)$ is the rising factorial number. We note that if we set $\phi = \mu^{-1}$, then the first product over i in (2.4) cancels, leading to a symmetric distribution with respect to the frequencies n_i's. In other term if $\phi = \mu^{-1}$ then (2.4) becomes an exchangeable partition probability function. This is the celebrate Ewens partition probability function (Ewens, 1972) induced by a sample drawn from a Dirichlet process (Blackwell and MacQueen, 1973).

According to the theory of partially exchangeable random partitions developed in Pitman (1995), (2.4) characterizes the predictive probabilities of the GDP with constant parameters. See Proposition 10 in Pitman (1995). In particular, consider a sample of size n from a GDP(ϕ, μ, G_0) and assume that it induces a partition of $\{1,\ldots,n\}$ into $K_n = k$ blocks, labeled by $\theta_1^*,\ldots,\theta_{K_n}^*$, with corresponding frequencies $(N_1,\ldots,N_{K_n}) = (n_1,\ldots,n_k)$. Then

$$\mathbb{P}[X_{n+1} \notin \{\theta_1^*,\ldots,\theta_{K_n}^*\}] = \frac{\phi(1-\mu)}{\phi+n-1}\prod_{i=1}^{k-1}\frac{\phi(1-\mu)+\sum_{j=i+1}^{k} n_j}{\phi+\sum_{j=i+1}^{k} n_j-1} \tag{2.5}$$

and

$$\mathbb{P}[X_{n+1} = \theta_r^*] = \frac{\phi\mu+n_r-1}{\phi+n-1}\prod_{i=1}^{r}\frac{\phi(1-\mu)+\sum_{j=i+1}^{k} n_j}{\phi+\sum_{j=i+1}^{k} n_j-1} \tag{2.6}$$

for any $r = 1,\ldots,k$. Unfortunately, due to the cumbersome dependency on k and the frequencies n_i's, the predictive probabilities (2.5) and (2.6) neither allow to obtain moments of the distribution of K_n or moments of the distribution of the number of blocks with certain frequencies.

We now determine the asymptotic behavior of K_n as n grows. Using results in Karlin (1967), one can show that

$$\frac{K_n}{\log(n)} \to \frac{1}{\psi^{(0)}(\phi)-\psi^{(0)}(\phi(1-\mu))} \tag{2.7}$$

almost surely, as $n \to +\infty$. In (2.7), $\psi^{(0)}(x)$ denotes the polygamma function, i.e., the first derivative of the logarithm of the Gamma function with respect to x. Details of the derivation of this result are in Section S.2 of supplementary material available at *Biostatistics* online. If $\phi = \mu^{-1}$, then the large n asymptotic result in (2.7) reduces to the well-known large n asymptotic behavior of K_n under the assumption of the Dirichlet process. Indeed, $\psi^{(0)}(1/\mu)-\psi^{(0)}(1/\mu-1) = (1/\mu-1)^{-1}$ and, hence, $K_n/\log(n) \to (1/\mu-1)$ almost surely, as $n \to +\infty$.

The richer parameterization of the GDP allows controlling simultaneously different important features of the partition (see Rodriguez and Dunson (2014)). For instance, fixing the $\mathbb{E}(K_n)$, the parameters of GDP can control quantities such as the cardinality of the largest clusters, the average cluster size for different values or the number of clusters with cardinality equal one. This is in contrast with what happens using the DP, where the precision parameter governs all this quantities at once. Similarly, the additional flexibility can be appreciated also by looking at the distribution of K_n under the GDP and the DP, after matching the first moment as in Figure 1.

2.4. *Truncated GDP*

We next consider a modified version of (2.1) that includes a finite number H of atoms. We write:

$$G_H = \sum_{h=1}^{H} W_h \delta_{\theta_h}. \tag{2.8}$$

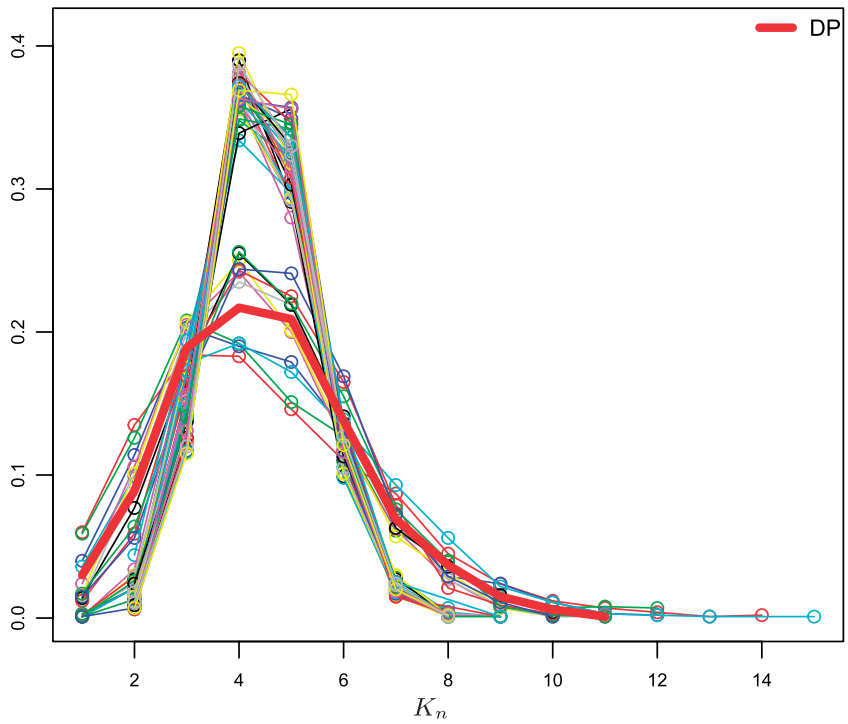

Fig. 1. Distributions of the number of clusters K_n under DP (thick line) and GDP (thin lines), all having $\mathbb{E}(K_n) \approx 4$. The different distributions under the GDP correspond to different combinations of the parameters μ and ϕ.

As in the infinite-dimensional case, the sequence of locations is an *i.i.d.* sample from G_0. The weights are constructed with the same stick-breaking procedure presented above, with the exception of the last weight, W_H, which is set to the value that makes the weights sum to 1. We denote this truncated process $\text{GDP}_H(\boldsymbol{\phi}, \boldsymbol{\mu}, G_0)$.

Truncated versions of the DP and other random probability measures have been employed in the literature, because they allow simplified computation when used as prior mixing distributions. Obviously, the use of a truncated process introduces an approximation error. The most common way to control this error was proposed by Ishwaran and James (2001) (Theorem 1) and has been adapted for many other processes. This consists in setting an upper bound to L_1 distance between marginal densities obtained under the original and the truncated version of the process. An adaptation to GDP of this result is in Rodriguez and Dunson (2014).

In mixture models, it is common to truncate the mixing measure to a specific level for computational purposes. This is particularly true when the mixing measure is not distributed as a DP for which simple and efficient Gibbs samplers are available. When the mixing measure is a GDP (consequently, also a DP), however, the joint distribution of the truncated sequence of weights, namely $\boldsymbol{W} = (W_1, \ldots, W_H)$, has a known distribution. This distribution is the generalized Dirichlet distribution (Connor and Mosimann, 1969; Ishwaran and Zarepour, 2000), which is conjugate with the multinomial distribution. This leads to simple calculations in the case GDP_H with constant parameters when we want to sample from the posterior of ϕ and μ, encouraging the use of parsimonious models. Discussion of this point continues in Section S.4 of supplementary material available at *Biostatistics* online.

In Section S.2 of supplementary material available at *Biostatistics* online, we discuss the result obtained when the truncation of the GDP is random, extending a similar result of DP introduced in Muliere and Tardella (1998). This can give useful insights about the number of components to include in (2.8) to approximate a sample from GDP matching pre-specified approximation levels.

3. DEPENDENT GDP

Recalling the definition of the GDP in (2.1), a realization from a GDP is an almost surely discrete probability measure. While the discreteness of G may seem unappealing, the use of such objects as random prior distributions is common in Bayesian nonparametrics, such as when dealing with density estimation. The most famous example is the Dirichlet process mixture (DPM, Lo, 1984), which results from convolving a density kernel parameterized by some quantity with a random prior distribution that is distributed according to a DP. One may adopt an equivalent strategy using a GDP. The resulting model is represented by the following hierarchy:

$$\begin{aligned} y_1, \ldots, y_n \mid G &\overset{iid}{\sim} \int f(y_i \mid \theta) \mathrm{d}G(\theta) \\ G \mid \boldsymbol{\phi}, \boldsymbol{\mu}, G_0 &\sim \mathrm{GDP}(\boldsymbol{\phi}, \boldsymbol{\mu}, G_0), \end{aligned} \tag{3.1}$$

where the quantities $\boldsymbol{\phi} = \{\phi_h\}_{h=1}^{\infty}$ and $\boldsymbol{\mu} = \{\mu_h\}_{h=1}^{\infty}$ require the specification of suitable hyperprior distributions. According to the hierarchical formulation (3.1), the resulting sampling model is equivalent to an infinite mixture model with weights constructed as in (2.2).

Using a similar argument to the one presented in MacEachern (1999) and MacEachern (2000), the model in (3.1) can be enriched when covariates are available, assuming the observations are generated by a collection of infinite mixture models indexed by the covariate space and sharing hyperparameters. We achieve this result by modifying the GDP in such a way that the sequence $\{W_h\}_{h=1}^{\infty}$ is a function of the covariates. Given the parameterization of the Beta distribution that we used as the prior for the sequence $\{V_h\}_{h=1}^{\infty}$, we can express the expectations of the latter quantities as functions of the covariates. We call the resulting process the DGDP. More specifically, for a generic point $x \in \mathcal{X}$, where $\mathcal{X}$ is the covariate space, a sample from DGDP is

$$G_x = \sum_{h=1}^{\infty} W_{h,x} \delta_{\theta_h},$$

where $\{\theta_h\}_{h=1}^{\infty} \overset{iid}{\sim} G_0$, $W_{1,x} = V_{1,x}$ and

$$W_{r,x} = V_{r,x} \prod_{l=1}^{r-1} (1 - V_{l,x}), \quad r = 2, 3, \ldots$$

Each $V_{h,x}$ is independently distributed following a $\mathrm{Beta}(v_{h,x} \mid \phi_h \mu_h(x), \phi_h(1 - \mu_h(x)))$, where $\mu_h(\cdot)$ is a random mean function mapping into the set $(0, 1)$. Using the DGPD, the hierarchical model in (3.1) can be rewritten as

$$\begin{aligned} y_1, \ldots, y_n \mid G_{x_1}, \ldots, G_{x_n} &\overset{ind}{\sim} \int f(y_i \mid \theta) \mathrm{d}G_{x_i}(\theta) \\ G_{x_1}, \ldots, G_{x_n} \mid \boldsymbol{\phi}, \boldsymbol{\mu}(\cdot), G_0 &\overset{ind}{\sim} \mathrm{DGDP}(\boldsymbol{\phi}, \boldsymbol{\mu}(x_i), G_0), \end{aligned} \tag{3.2}$$

where $\boldsymbol{\phi} = \{\phi_h\}_{h=1}^{\infty}$ and $\boldsymbol{\mu}(\cdot) = \{\mu_h(\cdot)\}_{h=1}^{\infty}$. In case $\mathcal{X}$ is a dense set, each y_i is associated with an individual random measure, i.e. G_{x_i}. If $\mathcal{X}$ is not dense, then there may be ties in the vector $(x_1, \ldots, x_n)$, which leads to ties in the corresponding random measures $(G_{x_1}, \ldots, G_{x_n})$, i.e. groups of observations having the same covariates share the same random measure. Furthermore, it is trivial to generalize the DGDP to the case with non-common location parameters, which can be obtained substituting G_0 with a stochastic process indexed by $x \in \mathcal{X}$.

A key aspect of the construction above is the infinite sequence of random functions $\{\mu_h(\cdot)\}_{h=1}^{\infty}$, which incorporates the dependence of the random measures on the covariates and the association between random measures indexed by different covariate values in $\mathcal{X}$. One way to evaluate the dependence between random distributions is by considering a measurable set $A \in \mathcal{A}$, two locations $x, x' \in \mathcal{X}$, and the covariance $\mathbb{C}[G_x(A), G_{x'}(A)]$. Considering location-specific mean functions and precisions, the covariance is equal to

$$\mathbb{C}[G_x(A), G_{x'}(A)] = (1 - G_0(A))G_0(A)\mathbb{E}\left[\sum_{h=1}^{\infty} W_{h,x} W_{h,x'}\right],$$

which converts to $\mathbb{V}[G_x(A)]$ when $x = x'$ (compare to (2.3)). Assuming a constant mean function and precision across locations simplifies the calculations, as was the case with such an assumption for the moments of the GDP. In particular, considering $\{\mu_h(\cdot)\}_{h=1}^{\infty} = \mu(\cdot)$ and $\{\phi_h\}_{h=1}^{\infty} = \phi$ allows one to write

$$\mathbb{E}\left[\sum_{h=1}^{\infty} W_{h,x} W_{h,x'}\right] = \frac{\mathbb{E}[V_x V_{x'}]}{\mathbb{E}[V_x] + \mathbb{E}[V_{x'}] - \mathbb{E}[V_x V_{x'}]},$$

where V_r is a Beta random variable with parameters $(\phi\mu(r), \phi(1 - \mu(r)))$.

Hatjispyros *and others* (2016) argue that another convenient way to learn about similarities among to dependent random measures is to look at the distance between the infinite mixture models induced by two random measures indexed at two different locations in the covariate space. We apply this to two random measures distributed according to a DGDP. In particular, defining $f_x(y) = \int f(y \mid \theta)\mathrm{d}G_x(\theta)$ and $f_{x'}(y) = \int f(y \mid \theta)\mathrm{d}G_{x'}(\theta)$ to be two mixture sampling models indexed at $x, x' \in \mathcal{X}$, respectively, with $G_\cdot \sim \mathrm{DGDP}(\boldsymbol{\phi}, \boldsymbol{\mu}(\cdot))$, the expected L_2-distance (denoted $\|\cdot\|_2$) between $f_x(y)$ and $f_{x'}(y)$ is given by

$$\mathbb{E}[\|f_x(y) - f_{x'}(y)\|_2] = (a - b)\mathbb{E}\left[\sum_{h=1}^{\infty}(W_{h,x} - W_{h,x'})^2\right], \tag{3.3}$$

where $a = \mathbb{E}\left[\int f(y \mid \theta_h)^2\mathrm{d}y\right]$ and $b = \mathbb{E}\left[\int f(y \mid \theta_h) f(y \mid \theta_j)\mathrm{d}y\right]$. The latter equation shows that using covariate-dependent weights allows one to set mixture models to be arbitrarily close, despite the fact that the mixture models share common locations. This could be an argument in favor of using a stochastic process with only the weights indexed by the covariates.

Using the same approach, we employed for calculating the moments of the GDP and assuming $\{\phi_h\}_{h=1}^{\infty} = \phi$ and $\{\mu_h(\cdot)\}_{h=1}^{\infty} = \mu(\cdot)$, we can write,

$$\mathbb{E}\left[\sum_{h=1}^{\infty}(W_{h,x} - W_{h,x'})^2\right] = \frac{\mathbb{E}[V_x^2]}{2\mathbb{E}[V_x] - \mathbb{E}[V_x^2]} + \frac{\mathbb{E}[V_{x'}^2]}{2\mathbb{E}[V_{x'}] - \mathbb{E}[V_{x'}^2]} - \frac{2\mathbb{E}[V_x V_{x'}]}{\mathbb{E}[V_x] + \mathbb{E}[V_{x'}] - \mathbb{E}[V_x V_{x'}]}. \tag{3.4}$$

We can derive expressions for the latter expectations for different choices of $\mu(\cdot)$ and ϕ. We can gain some insight into the distance measure represented by the previous equation by assuming

$$\mu(x) = \frac{\exp(x\mu)}{1 + \exp(x\mu)},$$

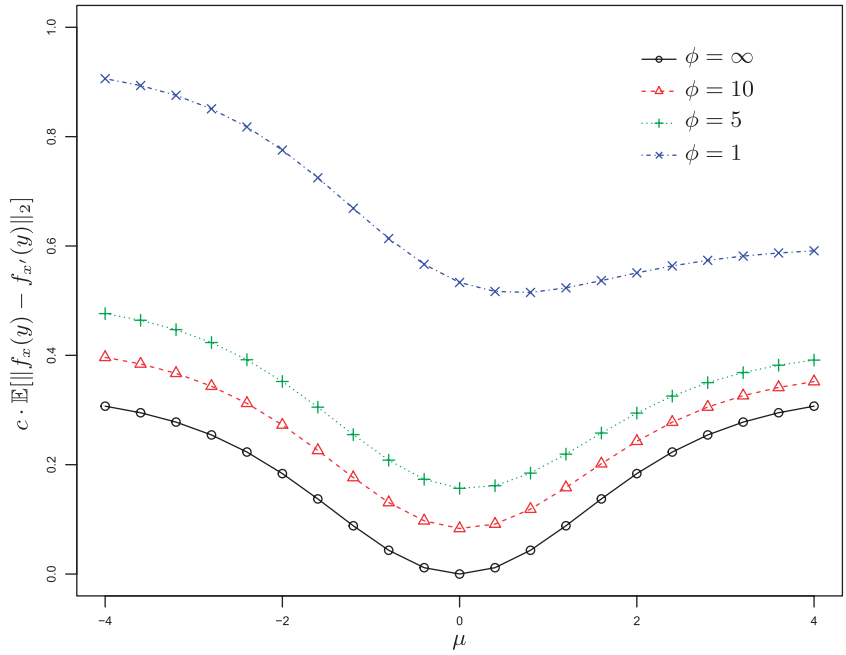

Fig. 2. Expected L_2-distance between two mixture models corresponding to $x = \{0, 1\}$ generated according to a DGDP (up to a constant dependent on the center measure c) for different values of ϕ and μ, the latter being the parameter of $\mu(x)$, a logistic regression without intercept.

which is the usual link for logistic regression. We consider two levels of $x = \{0, 1\}$ for simplicity and evaluate the expectation in (3.4) for different values of μ and ϕ. The results are shown in Figure 2.

We note that as ϕ tends to infinity, the DGDP defined as shown in this example becomes the logit stick-breaking prior introduced in Ren *and others* (2011). Similarly, assuming $\mu(\cdot)$ to be a probit regression and letting ϕ tend to infinity leads the DGDP to become the probit stick-breaking process introduced in Rodriguez and Dunson (2011).

4. POSTERIOR INFERENCE VIA MCMC

Posterior inference for DGDP mixture models can be performed adapting existing MCMC algorithms for DDP mixture models. Following the work in Ishwaran and James (2001), posterior distributions can be obtained using the blocked Gibbs sampler which provides an approximate inference based on the process in (2.8). An alternative approach which does not involve a deterministic truncation of the DGDP is the slice sampler described by Walker (2007). This is equivalent to the blocked Gibbs sampler, but involves auxiliary variables which produce a random truncation of the GDP. Details of the blocked Gibbs and the slice sampler are given in Section S.4 of supplementary material available at *Biostatistics* online. An alternative MCMC strategy to perform posterior inference is offered by the retrospective sampler of Papaspiliopoulos and Roberts (2008). Additionally, the result in equation (2.4) enables us to design a Gibbs sampler algorithm to draw from the distribution of the underlying partition of the observations when the DGDP has non-common locations across different levels of the covariates.

5. ALL AND DYSLIPIDEMIA

Childhood ALL is a cancer that affects the production of blood cells. The bone marrow produces an excess of lymphoblasts, which are immature white blood cells (WBC). Children affected by ALL are currently treated with combinations of chemotherapies, and the drug regimens include a class of steroids called glucocorticoids, such as dexamethasone. While this therapy has improved cure rates for patients, it is associated with a number of side effects. One adverse side effect is osteonecrosis, a disease that is associated with reduced blood flow to bones and joints, leading to bone cell death and possible fractures. The pathogenesis of osteonecrosis and its relationship with treatments for childhood ALL are described by Kawedia *and others* (2011). In particular, poor metabolism of the glucocorticoids included in the treatment of ALL may lead to this disease. The association between these steroids and the risk of osteonecrosis is thought to be through the glucocorticoid's effect on lipid levels. The effect leads to an increase in the size of lipocytes (fat cells) and subsequent marrow ischemia and apoptosis. These complications often result in bone necrosis, pain, and inability to use the joint.

Recent studies have shown that other drugs that are part of ALL therapy, such as asparaginase, may lead to osteonecrosis by a different mechanism than that of steroids. The objective of this analysis is to model the change of lipid measures over time (in particular triglycerides) during ALL therapy as a function of a biomarker of the pharmacological activity of asparaginase. We use albumin level as this biomarker, since higher asparaginase activity leads to lower albumin levels.

This study includes $n = 198$ ALL patients who have been classified by clinicians into two risk groups based on expected outcome. Children in the low-risk group (LR) have a better chance of cure than children in the standard-risk or high-risk groups. In the study data, the standard-risk patients are combined with the high-risk patients into the standard/high risk group (SHR). Factors at baseline that determine a patient's risk group are age (younger children tend to have better outcomes than older children), initial WBC count (very high counts require more intensive treatment), sex (females have a somewhat greater chance of cure than males), race (Caucasian children tend to have better outcomes), and subtype of the disease, to name a few. The data set includes 93 ALL patients in the SHR group and 105 patients in the LR group.

Because the LR group tends to have a better prognosis than the SHR group, the treatment regimens for the risk groups differ. The SHR group receives more intensive therapy than the LR group. The different treatment regimens include different doses and schedules of dexamethasone and asparaginase, the two drugs that are associated with risk of osteonecrosis. The analysis considers each patient's measurements of triglycerides (mg/dL) and albumin (g/dL) from blood samples at baseline ($t = 0$), week 7 ($t = 7$), week 8 ($t = 8$), and week 12 ($t = T = 12$) of treatment. Patients received both drugs at the start of weeks 7 and 8 but not at baseline or week 12.

We denote the $\log_2$ transformation of the triglyceride level for the ith patient at time t by $y_{i,t}$. We assume the following model for the triglyceride trajectories, $\boldsymbol{y}_i = (y_{i,0}, \ldots, y_{i,T})^\intercal$,

$$\boldsymbol{y}_i \mid \boldsymbol{B}_i, \Omega_i \sim \mathrm{MN}_T \left(\begin{array}{c|c} y_0 & \boldsymbol{x}_{i,0}^\intercal \boldsymbol{\beta}_{i,0} \\ \vdots & \vdots \quad , \ \Omega_i \\ y_T & \boldsymbol{x}_{i,T}^\intercal \boldsymbol{\beta}_{i,T} \end{array} \right), \tag{5.1}$$

where $\mathrm{MN}_T(\cdot \mid \cdot, \cdot)$ denotes the $T-$dimensional Normal distribution, $\boldsymbol{B}_i = (\boldsymbol{\beta}_{i,0}, \ldots, \boldsymbol{\beta}_{i,T})$ is a matrix of coefficients, and $\boldsymbol{x}_{i,0}, \ldots, \boldsymbol{x}_{i,T}$ are time-dependent column vectors of covariates that include the measured albumin levels at different times, along with an intercept. Ω_i is the variance–covariance matrix, and we assume

$$\Omega_i = \sigma_i^2 H(\rho_i).$$

As in Quintana *and others* (2016), we specify the matrix $H(\rho_i)$ such that the covariance $\mathbb{C}(y_{i,t}, y_{i,s}) = \sigma_i^2 \rho_i^{|t-s|}$. This choice induces a correlation structure among the elements in $\boldsymbol{y}_i$ that is equivalent to one implied by an autoregressive model with time lag of one.

We account for possible heterogeneity between patients by assuming *a priori* that the trajectories come from a mixture of distributions. We also assume different but correlated mixing measures for patients belonging to the two risk groups (LR and SHR). This assumption allows us to control for information implied by being in a certain risk group, making more realistic the linear dependence of the triglyceride values on the albumin levels. The latter argument is similar to one described in Papageorgiou *and others* (2015). We formalize this assumption through the following hierarchical structure for the patient-specific parameters.

$$
\begin{aligned}
&(\boldsymbol{B}_i, \sigma_i^2, \rho_i) \mid G_{z_i} \sim G_{z_i} \\
&G_{z_i} \mid \phi, \mu(\cdot), G_0 \sim \mathrm{DGDP}(\phi, \mu(\boldsymbol{z}_i), G_0),
\end{aligned}
$$

where $\boldsymbol{z}_i = (1, z_i)$ and z_i is equal to 1 if the i-th patient belongs to the LR and 0 if SHR. We assume that the mean for the stick-breaking sequence is a logistic regression on z_i,

$$
\mu(\boldsymbol{z}_i) = \frac{\exp(\boldsymbol{z}_i \boldsymbol{\eta})}{1 + \exp(\boldsymbol{z}_i \boldsymbol{\eta})}.
$$

The regression parameters in the hypermean function are multivariate normal,

$$
\boldsymbol{\eta} \sim \mathrm{MN}_2(\boldsymbol{\eta} \mid \mathbf{0}_2, \sigma_\eta^2 I_2),
$$

where $\mathbf{0}_N$ is a N-dimensional vector of zeros, and I_N denotes the identity matrix of dimension $N \times N$. Finally, we specify a gamma hyperprior distribution for the precision of the DGDP

$$
\phi \sim \mathrm{Gamma}(\phi \mid a_\phi, b_\phi)
$$

and the following for the prior mean measure of the process,

$$
G_0 = \mathrm{U}(\sigma \mid a_\sigma, b_\sigma)\mathrm{U}(\rho \mid 0, 1) \prod_{t=0}^{T} \mathrm{MN}_T(\boldsymbol{\beta}_t \mid \mathbf{0}_2, \sigma_\beta^2 I_2).
$$

We fix σ_β^2 and σ_γ^2 to equal 100; set σ_μ^2, a_ϕ, and b_ϕ to 1; and let a_σ and b_σ equal 0 and 5, respectively. We run the blocked Gibbs sampler discussed in Section S.4 of supplementary material available at *Biostatistics* online with truncation level $H = 30$ and 50000 iterations after a burnin period of 30 000.

The expectations of the posterior predictive distributions for the triglycerides are depicted in Figure 3, showing different trajectories corresponding to different risk groups and different values of albumin at baseline and weeks 7, 8, and 12.

Overall, the SHR-specific trajectories are higher than those corresponding to the LR patients with the same albumin levels. The predicted triglyceride values at each time point, as a function of albumin, indicate a negative relationship between albumin and triglyceride levels for both risk groups. This relationship suggests that a reduction in the asparaginase activity, which is in turn related to an increase in albumin level, leads to a reduction in triglycerides in both risk groups, with a stronger effect among the SHR patients.

The largest difference in the values of the triglyceride trajectories is observed between the $t = 7$ and $t = 8$, when patients receive both the glucocorticoid and asparaginase. Figure 4 shows marginal density

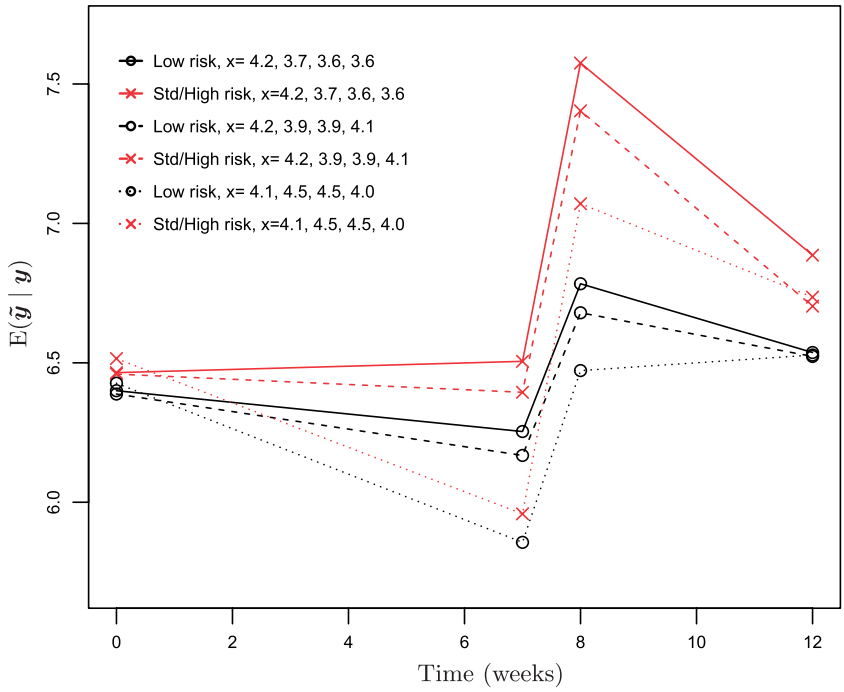

Fig. 3. Posterior predictive mean triglycerides for the two risk groups and different albumin values at baseline and weeks 7, 8, and 12.

estimates of the distributions of triglycerides at week 8, where the different type of lines and colors correspond to the legend in Figure 3.

Figure 4 shows the posterior predictive densities for week 8 triglycerides, which are mixtures with weights that vary across risk group. The curves corresponding to the SHR group assign high probability to a mixture component located around 9 $\log_2$(mg/dL). This component is centered at a relatively high value and leads to the differences seen in the expectations observed in Figure 3. The other apparent mixture component has a roughly equivalent location for both risk groups and is centered around 6.5 $\log_2$(mg/dL). This observation suggests that the risk group-specific differences in triglyceride values evident at week 8 are driven by a subset of the SHR patients, whereas the other SHR patients show similar triglyceride values as the LR risk group. An equivalent, although less evident, pattern can be seen in the marginal density distributions for the triglycerides at weeks $t = 7$ and 12.

In Figure 5, we show the posterior densities for the effects of albumin on triglycerides at each of the four time points under analysis (i.e., the time and group-specific regression coefficients).

While the relationship between albumin and triglycerides at baseline (top-left panel) seems similar for the risk groups, the densities diverge at $t = 7$. That is, after the start of treatment, a group of patients (mostly SHR patients) show a stronger negative relation between albumin and triglycerides, while a number of other patients (mostly belonging to the LR group) exhibit a weaker negative effect of albumin on triglycerides. This pattern also appears at week 8, although the albumin effect is less negative than at week 7 for the majority of LR patients. At week 12, the majority of the mass corresponding to the albumin effect on triglycerides among the LR patients is centered a little to the right of zero. The effect for the SHR group at $t = 12$, however, remains bimodal, with the left-hand component remaining strongly negative and the right-hand component looking much like the density corresponding to the majority of

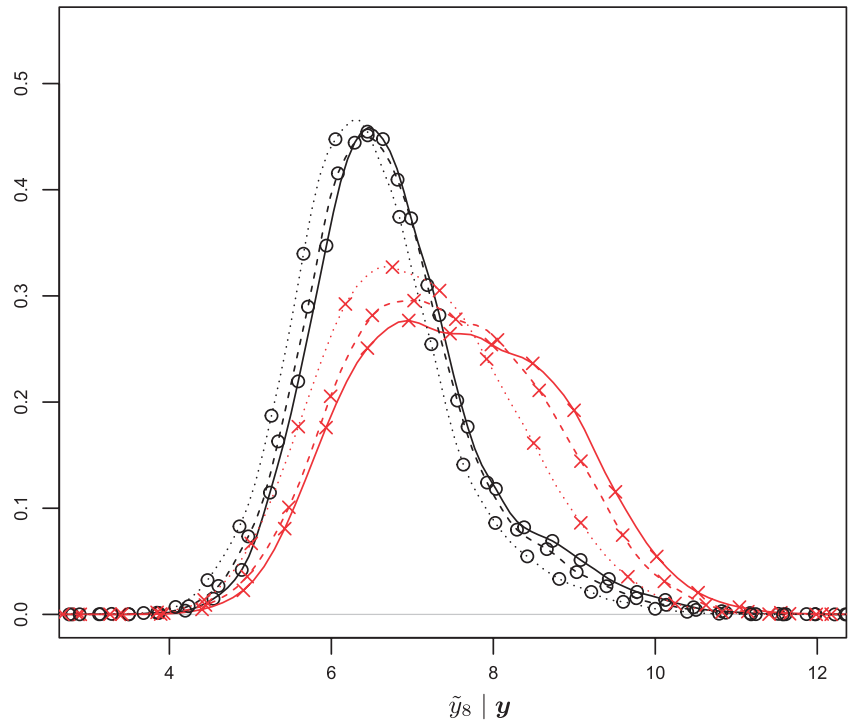

Fig. 4. Marginal posterior predictive densities of triglycerides at week 8. The different lines correspond to the legend in Figure 3.

the LR patients. These observations suggest that a subset of the SHR patients may be at higher risk of osteonecrosis, perhaps because of greater sensitivity to the drugs.

5.1. *Comparison with related methods*

We compare the performance of the DGDP mixture model described above for the data analysis in this section with those obtainable with related and more standard alternatives. The first competitor model is a parametric mixed-effect model, which is specified by a sampling model equivalent to the one in (5.1) where the individual effects, namely $\boldsymbol{B}_i$ and Ω_i, have been replaced with parameters shared by all patients. In addition, information regarding the risk groups is included via random effects within the model of the mean. Borrowing strength across groups is favored by a suitable hierarchical structure. The second competitor is a DDP mixture model which is specified as the DGDP mixture model above, except for the distribution of the mixing weights which follows marginally (for each value of $\boldsymbol{z}$) a DP with precision parameter equal $\alpha(\boldsymbol{z}) = \exp(\boldsymbol{z}\boldsymbol{\eta})$.

The competitor models are assessed using a pseudo Bayes factor (PSBF, Geisser and Eddy, 1979; Gelfand and Dey, 1994). When two models, M_l and M_r, are considered, the PSBF is defined as

$$\mathrm{PSBF}(M_l, M_r) = \frac{\prod_{i=1}^{n} p_{M_l}(\boldsymbol{y}_i \mid \boldsymbol{Y}_{-i})}{\prod_{i=1}^{n} p_{M_r}(\boldsymbol{y}_i \mid \boldsymbol{Y}_{-i})},$$

where $p_{M_l}(\boldsymbol{y}_i \mid \boldsymbol{Y}_{-i})$ and $p_{M_r}(\boldsymbol{y}_i \mid \boldsymbol{Y}_{-i})$ are posterior predictive densities of $\boldsymbol{y}_i$ under M_l and M_r, including the information of $\boldsymbol{Y}_{-i}$, the matrix containing all observations except for $\boldsymbol{y}_i$. All these posterior predictive

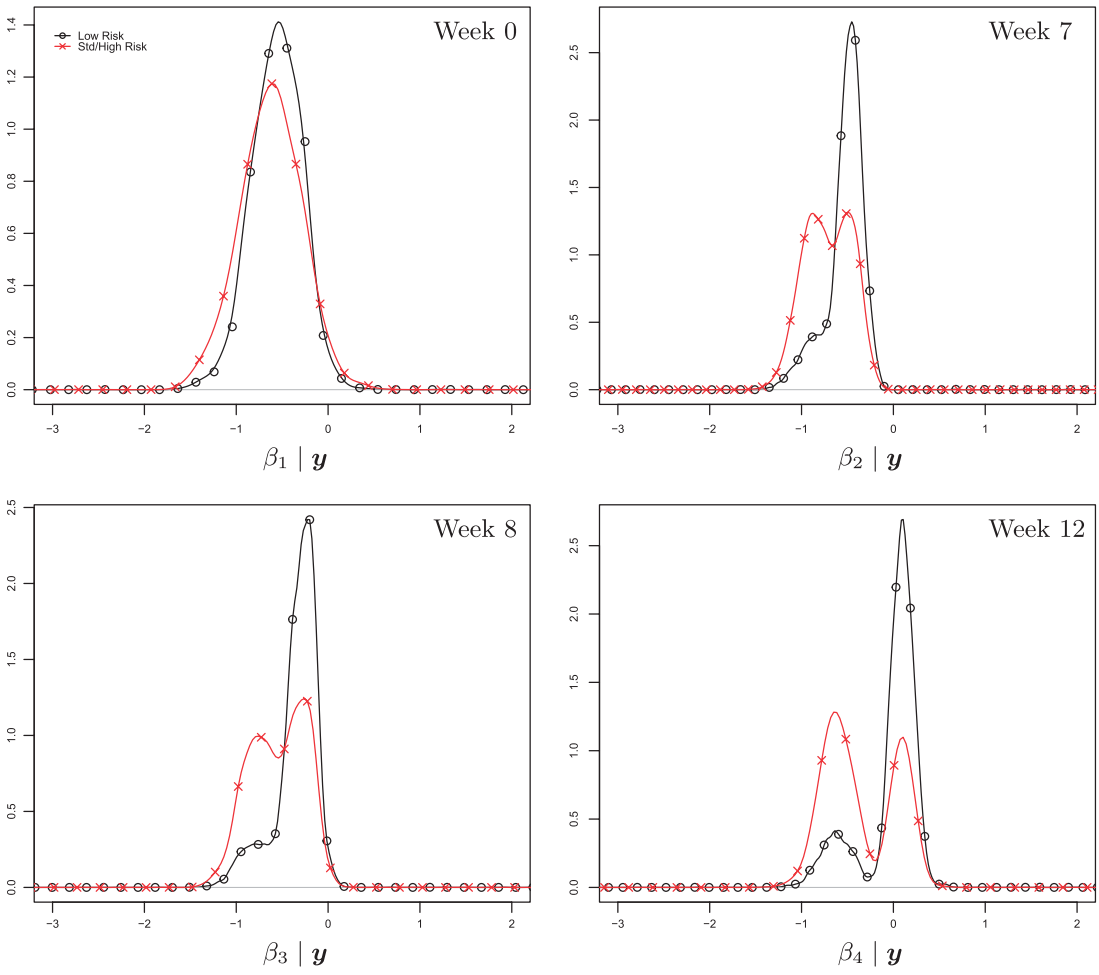

Fig. 5. Posterior densities of the regression coefficients related to albumin at times $t = 0, 7, 8$, and 12, for SHR patients and LR patients.

densities, often called conditional predictive ordinates, have been approximated using MCMC samples. PSBF is preferred to the common Bayes factor or posterior Bayes factor (Aitkin, 1991) because it is less sensitive to prior choices and simpler from a computational point of view.

The results of the comparison provide evidence in favor of proposed DGDP mixture model against the two competitor models. In particular, log(PSBF) of the DGDP mixture model relative to the mixed-effect model is equal to 39.16. The same quantity calculated using the DGDP versus the DDP mixture model is 10.85. Thus, there is strong evidence that the DGDP provides a superior fit to these data, when compared to a mixed-effects model and to a DDP mixture model.

6. Discussion

In this article, we introduce the DGDP, a stochastic process over discrete random probability measures. The DGDP has GDP-distributed marginals. This process directly generalizes the well-known DDP, which

instead has DP-distributed marginals. The generalization allows more flexibility at the marginal level, as well as better interpretability of the parameters. The DGDP can be constructed using sequences of correlated stick-breaking weights indexed by covariate levels. Random functions of the covariate levels can be included in the means of these Beta random variables. When probit or logit regression models are employed, the DGDP can be seen as a stochastic version of the probit stick-breaking or logistic stick-breaking priors, respectively.

The first part of this article described the main properties of the GDP and introduced new distributional properties of samples generated by realizations from a DGDP, along with results about random truncation of the process. In the second part, we defined the DGDP and employed different criteria for assessing the strength of dependence between DGDP marginals that are indexed by different levels in the covariate space. We discussed different MCMC algorithms for posterior inference with DGDP mixture models and gave details for two of them (contained in Section S.4 of supplementary material available at *Biostatistics* online). The last part of this article illustrated an application of the DGDP for modeling longitudinal data to assess the effect of asparaginase activity on triglyceride levels when treating ALL patients with this drug. Inference was based on albumin levels, which served as surrogates for asparaginase activity.

The method proposed in this article is tailored to handle longitudinal data, i.e., a series of observations for each subject. Multiple series of observations for each subject can also be modeled by extending the covariate space to which the DGDP indexes its realizations. In particular, one can include a vector that indicates for each observation the corresponding series of which it is a part. This would allow one to flexibly account for interactions among the different series while also borrowing strength across groups of patients.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors acknowledge Dr. Mary Relling and St. Jude Children's Research Hospital for providing with the data employed for the analysis in Section 5. *Conflict of Interest*: None declared.

FUNDING

S.F. is supported by the European Research Council (ERC) through StG N-BNP 306406; and G.L.R. is partially supported by U.S. National Institutes of Healths (NIH) through GM092666 and P30CA006973.

REFERENCES

AITKIN, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **53**, 111–142.

BLACKWELL, D. AND MACQUEEN, J. B. (1973). Ferguson distributions via pólya urn schemes. *The Annals of Statistics* **1**, 353–355.

CONNOR, R. J AND MOSIMANN, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association* **64**, 194–206.

CRUZ-MARCELO, A., ROSNER, G. L., MÜLLER, P. AND STEWART, C. F. (2013). Effect on prediction when modeling covariates in Bayesian nonparametric models. *Journal of Statistical Theory and Practice* **7**, 204–218.

DE IORIO, M., JOHNSON, W. O., MÜLLER, P. AND ROSNER, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics* **65**, 762–771.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

Dunson, D. B. and Park, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95**, 307–323.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

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

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153–160.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **56**, 501–514.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

Griffin, J. E. and Leisen, F. (2017). Compound random measures and their use in Bayesian nonparametrics. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **79**, 525–545.

Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association* **101**, 179–194.

Hatjispyros, S. J., Nicoleris, T. and Walker, S. G. (2016). Dependent random density functions with common atoms and pairwise dependence. *Computational Statistics & Data Analysis* **101**(C), 236–249.

Hjort, N. L. (2000). Bayesian analysis for a generalised Dirichlet process prior. *Technical Report*. Matematisk Institutt, Universitetet i Oslo.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models. *Biometrika* **87**, 371–390.

Karabatsos, G., Walker, S. G. (2012). Adaptive-modal Bayesian nonparametric regression. *Electronic Journal of Statistics* **6**, 2038–2068.

Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *Journal of Applied Mathematics and Mechanics* **17**, 373–401.

Kawedia, J. D., Kaste, S. C., Pei, D., Panetta, J. C., Cai, X., Cheng, C., Neale, G., Howard, S. C., Evans, W. E., Pui, C.-H. *and others*. (2011). Pharmacokinetic, pharmacodynamic, and pharmacogenetic determinants of osteonecrosis in children with acute lymphoblastic leukemia. *Blood* **117**, 2340–2347.

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

Lijoi, A., Nipoti, B., Prünster, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20**, 1260–1291.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics* **12**, 351–357.

MacEachern, S. N. (1999). Dependent nonparametric processes. In: *ASA Proceedings of the Section on Bayesian Statistical Science*. Alexandria, VA: American Statistical Association. pp. 50–55.

MacEachern, S. N. (2000). Dependent Dirichlet processes. *Technical Report*. Department of Statistics, The Ohio State University.

Muliere, P. and Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics* **26**, 283–297.

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.

PAPAGEORGIOU, G., RICHARDSON, S. AND BEST, N. (2015). Bayesian non-parametric models for spatially indexed data of mixed type. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **77**, 973–999.

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

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

QUINTANA, F. A., JOHNSON, W. O., WAETJEN, E. AND GOLD, E. (2016). Bayesian nonparametric longitudinal data analysis. *Journal of the American Statistical Association* **111**, 1168–1181.

REN, L., DU, L., CARIN, L. AND DUNSON, D. (2011). Logistic stick-breaking process. *The Journal of Machine Learning Research* **12**, 203–239.

RODRIGUEZ, A. AND DUNSON, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6**, 145–177.

RODRIGUEZ, A. AND DUNSON, D. B. (2014). Functional clustering in nested designs: modeling variability in reproductive epidemiology studies. *The Annals of Applied Statistics* **8**, 1416–1442.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

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

[*Received November 19, 2016; revised July 29, 2017; accepted for publication August 2, 2017*]