The multivariate beta process and an extension of the Polya tree model

By LORENZO TRIPPA

Department of Biostatistics and Computational Biology, Dana–Farber Cancer Institute, Boston, Massachusetts 02115, U.S.A.

ltrippa@jimmy.harvard.edu

PETER MÜLLER

Department of Biostatistics, University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, U.S.A.

pmueller@mdanderson.org

AND WESLEY JOHNSON

Department of Statistics, University of California, Irvine, California 92697, U.S.A.

wjohnson@uci.edu

SUMMARY

We introduce a novel stochastic process that we term the multivariate beta process. The process is defined for modelling-dependent random probabilities and has beta marginal distributions. We use this process to define a probability model for a family of unknown distributions indexed by covariates. The marginal model for each distribution is a Polya tree prior. An important feature of the proposed prior is the easy centring of the nonparametric model around any parametric regression model. We use the model to implement nonparametric inference for survival distributions. The nonparametric model that we introduce can be adopted to extend the support of prior distributions for parametric regression models.

Some key words: Dependent random probability measures; Multivariate beta process; Polya tree distribution.

1. INTRODUCTION

We introduce a stochastic process that we call the multivariate beta process; this is not related to the process introduced in Hjort (1990). The multivariate beta process $\{Y_x\}_{x \in X}$ is defined on a Euclidean space $X \subset \mathbb{R}^k$, for every x in X, the random variable Y_x has a beta distribution and, for every pair of points (x_1, x_2) in X, the random variables Y_{x_1} and Y_{x_2} are positively dependent. Moreover, the degree of dependence becomes stronger as x_1 approches x_2. In the first part of this article, the multivariate beta process is studied and a relation with the Dirichlet process (Ferguson, 1974) is illustrated. In the second part, a Bayesian model for a family $\mathcal{P} = \{\mathcal{P}_x; \ x \in X\}$ of dependent random probability measures $\mathcal{P}_x$ indexed by covariates x is defined by using the multivariate beta process. The model is an extension of the Polya tree model (Lavine, 1992).

Many important applications naturally lead to statistical inference based on dependent random distributions. Typical examples are MacEachern (1999), De Iorio et. al. (2004) and

Dunson & Park (2008). We use the proposed model to implement a nonparametric extension of a parametric proportional hazards model.

The stochastic process $\mathcal{P}$ is indexed by the points of the set $\{X \times \mathcal{B}\}$, where X is the covariate space and $\mathcal{B}$ is the usual Borel σ-field of the real line. It is assumed that $X \subset \mathbb{R}^k$, for a given positive integer k: covariates can be continuous or discrete. We summarize a few salient properties of the model. For any $x \in X$, the marginal process $\{\mathcal{P}_x(B)\}_{B \in \mathcal{B}}$ is a random probability measure with a Polya tree distribution. The process can a priori be centred around a given regression model $\{F_x\}_{x \in X}$. The model $\{F_x\}_{x \in X}$ can include unknown hyperparameters θ. For example, in an application to survival data, we use a Weibull model with unknown regression parameters θ for prior centring. Another relevant property of the model is that for every measurable subset $B \in \mathcal{B}$, point $x \in X$ and sequence $\{x_i \in X\}_{i \geqslant 1}$ such that the Euclidean distances $\{\| x_i - x \|\}_{i \geqslant 1}$ vanish, the random sequence $\{\mathcal{P}_{x_i}(B)\}_{i \geqslant 1}$ converges in probability to $\mathcal{P}_x(B)$. We also illustrate that, for every finite subset of the covariate space $\{x_1, \ldots, x_m\}$, the law of the random probability measures $(\mathcal{P}_{x_1}, \ldots, \mathcal{P}_{x_m})$ has full support with respect to the weak convergence topology. We call the outlined extension of the Polya tree model $\mathcal{P}$ the multivariate Polya tree.

2. The multivariate beta process

2·1. *Definition*

We define a random process $\{Y_x\}_{x \in X}$ with beta marginal distributions, $Y_x \sim \text{Be}(\alpha_0, \alpha_1)$ for every $x \in X$. Constructions of dependent beta random variables have been studied in Olkin & Liu (2003), Nadarajah & Kotz (2005) and Nieto-Barajas & Walker (2007). Olkin & Liu (2003) define a bivariate beta distribution by ratios of gamma random variables. We build on a similar construction; the definition of the process $\{Y_x\}_{x \in X}$ is a natural extension of the following specification of a bivariate random vector (Y_1, Y_2) with beta marginal distributions:

$$(Y_1, Y_2) = \left(\frac{G_1 + G_2}{G_1 + G_2 + G_4 + G_5}, \frac{G_2 + G_3}{G_2 + G_3 + G_5 + G_6} \right), \tag{1}$$

where the equality is in distribution and $G_1, \ldots, G_6$ are independent gamma random variables with a fixed common scale parameter. Consider the kernels centred at x_1 and x_2 in Fig. 1(a). The kernels create six non-overlapping regions $A_1, \ldots, A_6$. Let X denote the covariate space, represented as the horizontal axis in Fig. 1(a). We define the gamma random variables $G_1, \ldots, G_6$ as the random weights assigned to A_1 through A_6 by a gamma process on $X \times \mathbb{R}$. The kernels centred at x_1 and x_2 are used to specify the distribution of the random vector (Y_1, Y_2). We extend the construction to $\{Y_x\}_{x \in X}$ by expanding from kernels centred at x_1 and x_2 to a family of kernels with location parameters in X. The extension exploits the fact that a beta random variable can be represented as the ratio of two gamma distributed random variables and the infinite divisibility property of the gamma distribution.

Let X be endowed with a σ-field $\mathcal{X}$ and a σ-finite measure μ. Lebesgue measure and counting measure are natural candidates for μ in continuous and discrete covariate cases, respectively. Let $Q = \{Q_x\}_{x \in X}$ be a location family of probability measures, absolutely continuous with respect to μ and with derivatives $\{q_x\}_{x \in X}$. Throughout the article such derivatives will be assumed to be unimodal; for example Gaussian, $q_x(\cdot) = N(\cdot; x, \sigma)$. Let $\{G(A)\}_{A \in \{\mathcal{X} \times \mathcal{B}\}}$ be a gamma process indexed by the sets of the product σ-field $\mathcal{X} \times \mathcal{B}$. For every $m \geqslant 1$ and non-overlapping $A_1, \ldots, A_m$, the random variables $G(A_1), \ldots, G(A_m)$ are independently gamma distributed with a fixed-scale parameter and shape parameters $\nu(A_j)$, where ν is the product measure on $X \times \mathbb{R}$ of μ and Lebesgue measure. Note that $\{G(A)\}_{A \in \{\mathcal{X} \times \mathcal{B}\}}$ is a positive random measure
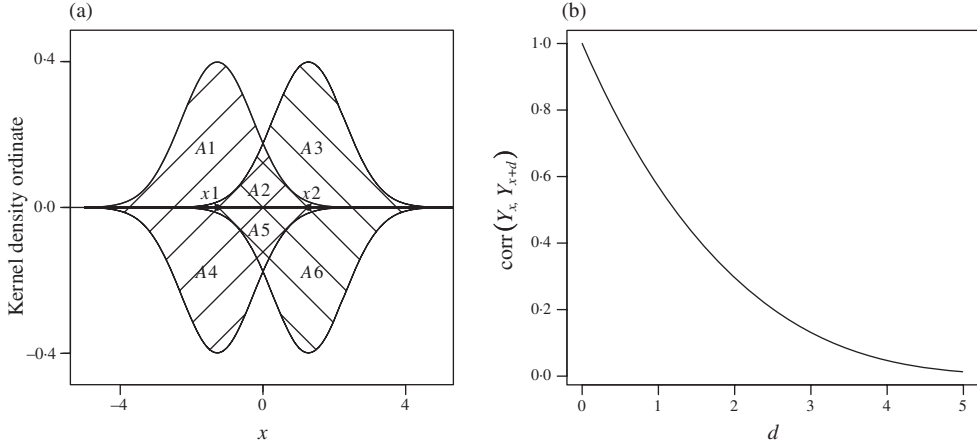
Fig. 1. Construction of a multivariate beta process. (a) Kernels $\alpha_0 q_{x_1}(\cdot)$, $\alpha_0 q_{x_2}(\cdot)$, $-\alpha_1 q_{x_1}(\cdot)$ and $-\alpha_1 q_{x_2}(\cdot)$, centred at two covariate values, x_1 and x_2. The areas indicated by A_1 through A_6 are used to define the dependent random variables Y_1 and Y_2 in (1). (b) Correlation $\mathrm{corr}(Y_x, Y_{x+d})$ as a function of d. Here $X = \mathbb{R}$, $\{Y_x\}_{x \in X} \sim \mathrm{MBP}(\alpha_0, \alpha_1, Q)$, $\alpha_0 = \alpha_1 = 1$, μ is the Lebesgue measure and $\{q_x\}_{x \in \mathbb{R}}$ are Gaussian kernels with mean x and variance equal to 1.

on $X \times \mathbb{R}$. Finally, for every x, let $S_x^0 = \{(z, y) \in X \times \mathbb{R} : 0 < y < \alpha_0 q_x(z)\}$ and symmetrically $S_x^1 = \{(z, y) \in X \times \mathbb{R} : -\alpha_1 q_x(z) < y < 0\}$, where α_0 and α_1 are positive parameters. The sets S_x^0 and S_x^1 are the regions bounded by the kernels $\alpha_0 q_x(\cdot)$ and $-\alpha_1 q_x(\cdot)$. We will use the notation $S_x = S_x^0 \cup S_x^1$. In Fig. 1 we illustrate the case $X = \mathbb{R}$, with $q_x(\cdot) = N(\cdot\,; x, \sigma)$. The regions $S_{x_i}^0$ and $S_{x_i}^1$, for $i \in \{1, 2\}$, are bordered by the two kernels centred at x_i. We can now define, for every $x \in X$, the beta random variable Y_x as a function of the gamma process $\{G(A)\}_{A \in \{X \times \mathcal{B}\}}$:

$$Y_x = \frac{G(S_x^0)}{G(S_x^0) + G(S_x^1)}.$$

We call the constructed process $\{Y_x\}_{x \in X}$ the multivariate beta process with parameters (α_0, α_1, Q) and use the notation $\{Y_x\}_{x \in X} \sim \mathrm{MBP}(\alpha_0, \alpha_1, Q)$.

For later reference, we note an alternative construction for the finite-dimensional distributions of $\{Y_x\}_{x \in X}$. Consider a set of covariate values $x_1, \dots x_n$. Define $\nu_{x_1, \dots, x_n}$ as the ν measure restricted to

$$\left[(z, y) \in X \times \mathbb{R} : \min_i \{-\alpha_1 q_{x_i}(z)\} < y < \max_i \{\alpha_0 q_{x_i}(z)\} \right].$$

If $D_{x_1, \dots, x_n}$ is a Dirichlet process parameterized by the positive measure $\nu_{x_1, \dots, x_n}$, then

$$\left(Y_{x_1}, \dots, Y_{x_n} \right) = \left\{ \frac{D_{x_1, \dots, x_n}(S_{x_1}^0)}{D_{x_1, \dots, x_n}(S_{x_1}^0 \cup S_{x_1}^1)}, \dots, \frac{D_{x_1, \dots, x_n}(S_{x_n}^0)}{D_{x_1, \dots, x_n}(S_{x_n}^0 \cup S_{x_n}^1)} \right\}, \tag{2}$$

where the equality is in distribution. The identity between the two definitions of the process $\{Y_x\}_{x \in X}$ follows from the fact that a Dirichlet process can be represented as a normalized gamma process.

2·2. *Sampling from a multivariate beta process prior*

We exploit (2) for sampling binary variables when their joint distribution is specified by means of a multivariate beta process. Consider a vector of binary random variables $Z = (Z_1, \ldots, Z_N)$, with covariates x_i $(i = 1, \ldots, N)$, and assume

$$p(Z_1, \ldots, Z_N \mid Y_{x_1}, \ldots, Y_{x_N}) = \prod_{i=1}^{N} Y_{x_i}^{Z_i} (1 - Y_{x_i})^{1-Z_i}, \quad \{Y_x\}_{x \in X} \sim \text{MBP}(\alpha_0, \alpha_1, Q). \quad (3)$$

The urn scheme in Blackwell & MacQueen (1973) can be used to sample Z, marginalizing over $\{Y_x\}_{x \in X}$. Let $\{(l_{ij}, h_{ij}) \in X \times \mathbb{R}; \ i = 1, \ldots, N, \ j \geqslant 1\}$ be an array of exchangeable random vectors with random distribution $D_{x_1, \ldots, x_N}$. Let $(j_1, \ldots, j_N)$ be a vector of integers defined as $j_i = \inf\{j : (l_{ij}, h_{ij}) \in S_{x_i}\}$. The vector $(j_i; \ i = 1, \ldots, N)$ selects N pairs $\{(l_{ij}, h_{ij}); \ (i, j) = (1, j_1), \ldots, (N, j_N)\}$ that belong to the areas contoured by the curves $\alpha_0 q_{x_i}$ and $-\alpha_1 q_{x_i}$. The set of variables $\{(l_{ij}, h_{ij}); \ i = 1, \ldots, N, \ j = 1, \ldots, j_i\}$ can be generated by the Polya urn. The binary random vector $\{I(h_{ij} > 0); \ (i, j) = (1, j_1), \ldots, (N, j_N)\}$ has the same distribution as Z. The equality in distribution follows from the fact that the random probabilities of the binary vector $\{I(h_{ij} > 0); \ (i, j) = (1, j_1), \ldots, (N, j_N)\}$ are exactly the right-hand side of (2) and have the same distribution as the random variables $(Y_{x_i}; \ i = 1, \ldots, N)$ that define the probability distribution of Z.

We will use this construction for the implementation of posterior simulation in §4. The augmented model with the latent variables $\{(l_{ij}, h_{ij}); \ i = 1, \ldots, N+1, \ j \geqslant 1\}$ can be used to sample from the predictive distribution for a future Z_{N+1} with covariate x_{N+1} and, more generally, to implement posterior inference.

2·3. *Assessing the correlation function of a multivariate beta process*

The correlation function of the process $\{Y_x\}_{x \in X} \sim \text{MBP}(\alpha_0, \alpha_1, Q)$ can be computed using a result in Olkin & Liu (2003), who studied the bivariate beta distribution of a random vector (Y_1, Y_2) defined by the equality

$$(Y_1, Y_2) = \left(\frac{G_1}{G_1 + G_3}, \frac{G_2}{G_2 + G_3} \right),$$

where G_1, G_2 and G_3 are independent gamma random variables with shape parameters a, b and c. Olkin & Liu (2003) show that

$$E(Y_1 Y_2) = \frac{ab \, \Gamma(a+b)\Gamma(b+c)\Gamma(a+b+c+1)}{\Gamma(c)\Gamma(a+1)\Gamma(b+1)}$$
$$\times \sum_{j=0}^{\infty} \frac{\Gamma(a+j+1)}{\Gamma(a+b+c+j+1)} \frac{\Gamma(b+j+1)}{\Gamma(a+b+c+j+1)} \frac{1}{j!}. \quad (4)$$

Consider $\{Y_x\}_{x \in X} \sim \text{MBP}(\alpha_0, \alpha_1, Q)$ and two points x_1 and x_2 in X. For $i \in \{1, 2\}$ and $j \in \{0, 1\}$, let $S_i^j = S_{x_i}^j$, $S_i = S_i^0 \cup S_i^1$, $S_{12}^j = S_1^j \cap S_2^j$, $S_{12} = S_{12}^1 \cup S_{12}^0$ and $D = D_{x_1, x_2}$. We note that

$$E(Y_{x_1} Y_{x_2}) = E\left[\left\{ \frac{D(S_{12}^0)}{D(S_1)} + \frac{D(S_1^0 \setminus S_2^0)}{D(S_1)} \right\} \left\{ \frac{D(S_{12}^0)}{D(S_2)} + \frac{D(S_2^0 \setminus S_1^0)}{D(S_2)} \right\} \right]$$

$$= E\left\{\frac{D(S_{12})}{D(S_1)}\frac{D(S_{12})}{D(S_2)}\right\} E\left[\left\{\frac{D(S_{12}^0)}{D(S_{12})}\right\}^2\right] + E\left\{\frac{D(S_{12})}{D(S_1)}\frac{D(S_2\backslash S_1)}{D(S_2)}\right\}$$

$$\times \left(\frac{\alpha_0}{\alpha_0+\alpha_1}\right)^2 + E\left\{\frac{D(S_1\backslash S_2)}{D(S_1)}\frac{D(S_{12})}{D(S_2)}\right\}\left(\frac{\alpha_0}{\alpha_0+\alpha_1}\right)^2$$

$$+ E\left\{\frac{D(S_1\backslash S_2)}{D(S_1)}\frac{D(S_2\backslash S_1)}{D(S_2)}\right\}\left(\frac{\alpha_0}{\alpha_0+\alpha_1}\right)^2. \tag{5}$$

The equalities follow from representation (2) and from the tail-free property of the Dirichlet process (Doksum, 1974). The right-hand side of the equation allows us to compute $E(Y_{x_1}Y_{x_2})$ using the identity (4). This is possible since the law of the random vector

$$\left\{\frac{D(S_1\backslash S_2)}{D(S_1)}, \frac{D(S_2\backslash S_1)}{D(S_2)}\right\}$$

belongs to the family studied in Olkin & Liu (2003). Figure 1(b) illustrates the correlation function of a specific multivariate beta process.

The parameters have a clear interpretation: α_0 and α_1 characterize the univariate marginal distributions while the correlation function can be flexibly chosen through a suitable specification of Q. A simple example helps to clarify the relationship between Q and the correlation function. Let Q be the location family of normal distributions on the real line with common variance σ^2. The correlation function of the process $\{Y_x\}_{x\in\mathbb{R}}$ depends on the choice of σ^2; the larger the variance the larger the correlations of the bivariate marginal distributions. The triple $\alpha_0, \alpha_1, \nu(S_{x_1}\cap S_{x_2})$ parameterizes the joint distribution of (Y_{x_1}, Y_{x_2}). For fixed values of α_0 and α_1, the correlation between Y_{x_1} and Y_{x_2} is an increasing function of $\nu(S_{x_1}\cap S_{x_2})$. See Proposition A1 in the Appendix. It follows that, in the example, the map $(\sigma, x_1, x_2) \mapsto \text{corr}_{\sigma^2}(Y_{x_1}, Y_{x_2})$ is decreasing with respect to $|x_1 - x_2|/\sigma$. The elicitation of the multivariate beta process includes the choice of the parameter Q; this family of distributions defines the positive definite function

$$(x_1, x_2) \longmapsto \nu\left[(z, y) \in X \times \mathbb{R} : 0 < y < \min\{q_{x_1}(z), q_{x_2}(z)\}\right]. \tag{6}$$

The right-hand of (6) is identical to the correlations $\text{corr}\{G(S_{x_1}^0), G(S_{x_2}^0)\}$ and $\text{corr}\{G(S_{x_1}^1), G(S_{x_2}^1)\}$. The mapping in (6) provides an easily interpretable representation of the dependence between the random variables Y_{x_1} and Y_{x_2}. Moreover, for fixed values of α_1 and α_2, expression (5) allows us to evaluate the correlation function $(x_1, x_2) \mapsto \text{corr}(Y_{x_1}, Y_{x_2})$ corresponding to any specific choice of the map (6). Positive definite functions as in (6) arise from integrating suitably chosen indicator functions with respect to infinitely divisible processes and have been studied in the literature (Mittal, 1976; Berman, 1978; Gneiting, 1999).

We conclude this section with a proposition on the distribution of (Y_{x_1}, Y_{x_2}) when the distance between x_1 and x_2 becomes infinitesimal.

PROPOSITION 1. *For every $x \in X$, if $\{x_i \in X\}_{i\geqslant 1}$ is a sequence such that $\sup_{B\in\mathcal{B}} |Q_{x_i}(B) - Q_x(B)| \to 0$, then the random variables $\{Y_{x_i}\}_{i\geqslant 1}$ converge in probability to Y_x.*

The proofs of this and subsequent propositions are given in the Appendix.

3. Multivariate Polya trees

3·1. *The Polya tree model*

For later reference, we recall the definition of the Polya tree prior (Lavine, 1992, 1994). Let $\Pi = (B_\oslash = \Omega; B_0, B_1; B_{00}, B_{01}, \ldots)$ be a binary tree of nested partitions of a separable measurable space Ω such that $(B_\oslash; B_0, B_1; B_{00}, \ldots)$ generate the measurable sets. Let $\mathcal{A} = (\alpha_0, \alpha_1, \alpha_{00}, \ldots)$ be a sequence of nonnegative numbers. Finally, let $\mathcal{E} = \bigcup_{m=1}^\infty \{0, 1\}^m$ denote the index set of $\mathcal{A}$.

DEFINITION 1. *A random probability measure $\mathcal{P}$ on Ω is said to have a Polya tree distribution, with parameter $(\Pi, \mathcal{A})$, written $\mathcal{P} \sim PT(\Pi, \mathcal{A})$, if there exist independent random variables $\mathcal{Y} = (Y_\varepsilon; \ \varepsilon \in \mathcal{E} \cup \oslash)$ such that, for every $\varepsilon \in \mathcal{E} \cup \oslash$, $Y_\varepsilon \sim \mathrm{Be}(\alpha_{\varepsilon 1}, \alpha_{\varepsilon 0})$ and $\mathcal{P}(B_{\varepsilon 1})/\mathcal{P}(B_\varepsilon) = Y_\varepsilon$.*

For extensive discussion of the properties of the Polya tree model, we refer to Mauldin et al. (1992).

3·2. *The multivariate Polya tree*

We introduce the multivariate Polya tree distribution. The idea formalized in the following definition is to replace the beta random variables that characterize the Polya tree model with random processes indexed by the points of a covariate space. We will use multivariate beta processes $\{Y_{\varepsilon,x}\}_{x \in X}$, independent across ε, to generate beta-distributed random variables that define Polya tree random measures $\mathcal{P}_x$ for each x.

In the sequel $(B_0, B_1; B_{00}, \ldots) = \{(0, 1/2], (1/2, 1]; (0, 1/4], \ldots\}$ is fixed as the standard dyadic tree of partitions of the unit interval. The proposed model is parameterized by $(\mathcal{A}, Q, F)$ defined as follows. Let $\mathcal{A} = (\alpha_\varepsilon; \ \varepsilon \in \mathcal{E} \cup \oslash)$ denote a sequence of positive numbers. For a σ-finite measure μ on $(X, \mathcal{X})$, let $Q = \{Q_x\}_{x \in X}$ denote a location family of probability measures absolutely continuous with respect to μ. Finally, $F = \{F_x\}_{x \in X}$ is a class of continuous distribution functions.

DEFINITION 2. *A class of random probability measures on the real line $\{\mathcal{P}_x\}_{x \in X}$ has a multivariate Polya tree distribution, with parameters $(\mathcal{A}, Q, F)$, if there exist random processes $\mathcal{Y} = \{\{Y_{\varepsilon,x}\}_{x \in X}; \ \varepsilon \in \mathcal{E} \cup \oslash\}$ such that the following hold: the random processes in $\mathcal{Y}$ are independent across ε; for every $\varepsilon \in \mathcal{E} \cup \oslash$, $\{Y_{\varepsilon,x}\}_{x \in X} \sim \mathrm{MBP}(\alpha_\varepsilon, \alpha_\varepsilon, Q)$; for every $m = 1, 2, \ldots$ and every $\varepsilon \in \{0, 1\}^m$*

$$\mathcal{P}_x \left\{ F_x^{-1}(B_{\varepsilon_1, \ldots \varepsilon_m}) \right\} = \prod_{j=1; \ \varepsilon_j=1}^m Y_{\varepsilon_1, \ldots, \varepsilon_{j-1}, x} \prod_{j=1; \ \varepsilon_j=0}^m (1 - Y_{\varepsilon_1, \ldots, \varepsilon_{j-1}, x}) \tag{7}$$

where the first factor in the products is interpreted as $Y_{\oslash,x}$ or as $1 - Y_{\oslash,x}$ and, for every $a \in (0, 1]$, $F^{-1}(a) = \inf\{b \in \mathbb{R} : F(b) > a\}$.

We write $\{\mathcal{P}_x\} \sim \mathrm{MPT}(\mathcal{A}, Q, F)$. The parameter F defines the prior mean, $\mathcal{A}$ specifies the variability, and Q defines the strength of dependence.

The following proposition asserts that the defined class of random probability measures $\{\mathcal{P}_x\}_{x \in X}$ can be centred on an arbitrary model $\{F_x\}_{x \in X}$.

PROPOSITION 2. *For every $x \in X$ and every Borel subset B, if $\{\mathcal{P}_x\}_{x \in X} \sim \mathrm{MPT}(\mathcal{A}, Q, F)$, then the expected value of the random variable $\mathcal{P}_x(B)$ is equal to $F_x(B)$.*

For data analysis, it is important that the prior can be centred on a prior guess and that the degree of variability of the random probability measure can be controlled to reflect the strength of the available prior information. The Polya tree prior is a flexible probability model that can achieve both of these objectives. See for example Lavine (1992).

In the multivariate Polya tree model, a third aspect enters the prior elicitation. The investigator can specify the degree of dependence between the jointly modelled random probability measures $\{\mathcal{P}_x\}_{x\in X}$. For example, consider a covariate space $X = \{x_1, x_2\}$ consisting of two points. The higher the degree of dependence between the two random probability measures, the more borrowing of strength will occur between the two groups defined by the two covariates. In the extreme cases, if the random measures are independent, the observations with covariate x_2 are not used to estimate the unknown distribution $\mathcal{P}_{x_1}$, while if the random measures are almost surely identical, then all data are pooled to carry out inference about the one common random probability measure.

The degree of dependence between jointly modelled random distributions is most easily characterized by the correlations of the random probabilities. For an exhaustive motivation of this approach, we refer to Walker & Muliere (2003). They study the joint law of two Dirichlet processes, D_1 and D_2, such that the quantity $\mathrm{corr}\{D_1(B), D_2(B)\}$ is a fixed positive constant independent of the specific Borel subset B. In our case, the correlations of interest can easily be evaluated for a rich class of Borel subsets. Given the covariate points x_1 and x_2, for every ordered pair of integers (i_1, i_2), consider

$$\mathrm{corr}\left\{ \mathcal{P}_{x_1}\left(F_{x_1}^{-1}\left(\frac{i_1}{2^j}\right), F_{x_1}^{-1}\left(\frac{i_2}{2^j}\right) \right], \mathcal{P}_{x_2}\left(F_{x_2}^{-1}\left(\frac{i_1}{2^j}\right), F_{x_2}^{-1}\left(\frac{i_2}{2^j}\right) \right] \right\}.$$

Using (7), the computation of the correlation coefficients can be reduced to the previously discussed problem of evaluating second-order moments of bivariate marginal distributions of independent multivariate beta processes. Figure 2 illustrates, for a specific multivariate Polya tree parametrization, some of the correlations that can be computed following the outlined procedure. As desired, the closer the two covariates, the stronger the dependence between the corresponding random measures. The parameters $(\mathcal{A}, F)$ characterize the marginal distribution of a single random measure $\mathcal{P}_x$. If $(\mathcal{A}, F)$ is fixed, then the choice of Q allows flexible modelling of the strength of dependence between the random measures. Consider, for example, the parametrization adopted in Fig. 2. Following the same arguments used earlier to discuss the relationship between Q and the correlation function of $\{Y_x\}_{x\in X} \sim \mathrm{MBP}(\alpha_0, \alpha_1, Q)$, it can be shown that the ratio $|x_1 - x_2|/\sigma$ determines the degree of dependence between $\mathcal{P}_{x_1}$ and $\mathcal{P}_{x_2}$.

The next proposition asserts that, for a sequence of covariate points $\{x_i\}_{i\geqslant 1}$ converging to x, the differences between the associated random distributions, $\mathcal{P}_{x_i}$ and $\mathcal{P}_x$, become negligible.

PROPOSITION 3. *For every $x \in X$, if $\{x_i \in X\}_{i\geqslant 1}$ is a sequence such that $\sup_{B\in\mathcal{B}}|Q_{x_i}(B) - Q_x(B)| \to 0$ and $\sup_{B\in\mathcal{B}}|F_{x_i}(B) - F_x(B)| \to 0$, then, for every Borel subset B of the real line, the random variables $\{\mathcal{P}_{x_i}(B)\}_{i\geqslant 1}$ converge in probability to $\mathcal{P}_x(B)$.*

Proposition 4 describes the support of the prior $\{\mathcal{P}_x\}_{x\in X}$. For every $x \in X$, the support of the random probability measure $\mathcal{P}_x$ includes all probability distributions that are absolutely continuous with respect to F_x. If the support of F_x is the real line, then the law has full weak support, which remains the same even under restrictions on neighbouring random probability measures. Specifically, let $x_1, \ldots, x_m$ denote distinct covariate points. The random probability measure $\mathcal{P}_x$ still has full support, even conditional on $\mathcal{P}_{x_1}, \ldots, \mathcal{P}_{x_m}$ belonging to m, arbitrarily chosen, open sets $(\Delta_1, \ldots, \Delta_m)$ with respect to the weak convergence topology.
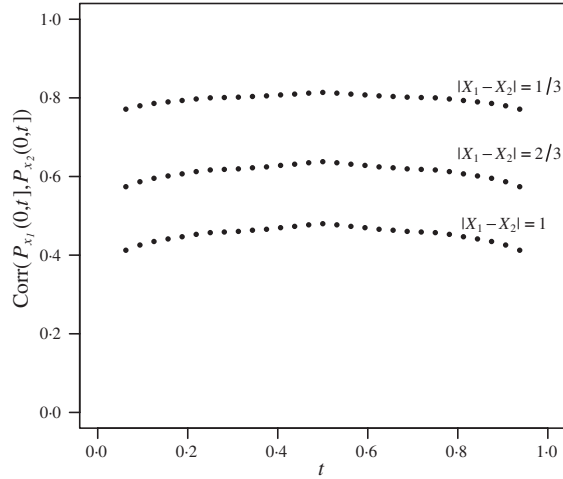
Fig. 2. Correlation between $\mathcal{P}_{x_1}(0, t]$ and $\mathcal{P}_{x_2}(0, t]$. Here
$\{\mathcal{P}_x\}_{x \in X} \sim \text{MPT}(\mathcal{A}, Q, F)$, $X = \mathbb{R}$, $\alpha_\varepsilon = 2$ for every ε, F_x
is the uniform distribution function on $[0, 1]$ for every x and
$(q_x; \, x \in X)$ is the family of Gaussian kernels with variance
$$\sigma^2 = 2.$$

This property distinguishes the multivariate Polya tree from many other Bayesian models for heterogeneous populations. Consider, for example, a Bayesian proportional hazards model F. If the covariate space X is a subset of the real line, then, for any monotone bounded continuous function f, the map $x \mapsto \int f \, dF_x$ is monotone. In contrast, assume that a multivariate Polya tree $\mathcal{P}$ is centred on a proportional hazards model; if $x_1 < x_2 < x_3$ belong to the covariate space, then the probability of the event $\int f \, d\mathcal{P}_{x_2} > \max(\int f \, d\mathcal{P}_{x_1}, \int f \, d\mathcal{P}_{x_3})$ is strictly positive.

PROPOSITION 4. *For every strictly positive δ, for every $m = 1, 2, \ldots$, distinct $x_1, \ldots, x_m \in X$, and for any family of probability measures $\mathcal{S}_1, \ldots, \mathcal{S}_m$ such that the absolute-continuity relations $\mathcal{S}_1 \ll F_{x_1}, \ldots, \mathcal{S}_m \ll F_{x_m}$ are satisfied, the following holds. If $\{V_1, \ldots, V_k\}$ is a partition of the real line into intervals V_j, then the event*

$$\left\{ |\mathcal{P}_{x_i}(V_j) - \mathcal{S}_i(V_j)| < \delta; \;\; i = 1, \ldots, m, \;\; j = 1, \ldots, k \right\}$$

has strictly positive probability.

The proposition guarantees that for any finite sequence of continuous and bounded real functions $\{f_1, \ldots, f_l\}$, and for every positive ξ, the event

$$\left\{ \bigcap_{i=1}^{m} \bigcap_{j=1}^{l} \left| \int f_j \, d\mathcal{P}_{x_i} - \int f_j \, d\mathcal{S}_i \right| < \xi \right\}$$

has a priori strictly positive probability.

3·3. *Mixtures of multivariate Polya trees*

We introduce to hierarchical extension to the multivariate Polya tree model. The extension is similar to mixtures of Polya tree models. Mixtures of Polya trees have been studied by authors

including Hanson (2006), Hanson & Johnson (2002) and Berger & Guglielmi (2001). An important advantage of the mixture of Polya tree models compared with the Polya tree prior is discussed in Lavine (1994). The awkward sensitivity of the inference with respect to the choice of the partitions is mitigated, and predictive distributions are smoother than under a Polya tree prior; see Hanson (2006) and Hanson & Johnson (2002).

A mixture of multivariate Polya trees is a natural extension to the mixture of Polya tree models. We consider a class of random probability measures $\{\mathcal{P}_x\}_{x \in X}$ indexed by covariates x such that conditionally on a random parameter θ the process $\{\mathcal{P}_x\}_{x \in X}$ is a multivariate Polya tree centred on F_θ,

$$\{\mathcal{P}_x\}_{x \in X} \mid \theta \sim \text{MPT}(\mathcal{A}, Q, F_\theta), \quad \theta \sim \lambda. \tag{8}$$

The parameter θ indexes a regression model F_θ. For example, in the following discussion, $\{\mathcal{P}_x\}_{x \in X}$ will be centred on a parametric survival regression model. Simple modifications allow one to adapt the framework to other regression problems.

In §5 we will use the following nonparametric Bayesian model for event time data. Let $\{W_i\}_{i=1}^n$ denote survival times, possibly censored from the right, of a sample of subjects with covariates $\{x_i\}_{i=1}^n$. We assume that the censoring times are independently distributed with respect to the survival times. Let $H_{\theta_1}(\cdot)$ be a continuous cumulative hazard function with parameter $\theta_1 \in \Theta_1$. Let $L_{\theta_2}(\cdot)$ denote a link function with parameter $\theta_2 \in \Theta_2$. Define the cumulative distribution functions $F_{\theta_1, \theta_2, x}(\cdot) = 1 - \exp\{-H_{\theta_1}(\cdot)L_{\theta_2}(x)\}$ and $F_\theta = \{F_{\theta_1, \theta_2, x}(\cdot)\}_{x \in X}$. We use the prior (8) with $\theta = (\theta_1, \theta_2)$ and conditionally independent survival times, $W_i \mid \{\mathcal{P}_x\}_{x \in X} \sim \mathcal{P}_{x_i}$ for $i = 1, \ldots, n$. We will specify F_θ in the hierarchical prior as a Weibull proportional hazards model: $\theta_1 = (\theta_{11}, \theta_{12})$ and $L_{\theta_2}(x)H_{\theta_1}(t) = \exp(\theta_2 x)\theta_{11}\theta_{12}t^{\theta_{12}}$. A vague prior is adopted for the regression parameter θ_2, which is normally distributed with zero mean and large variance, while θ_{11} and θ_{12} are exponentially distributed.

4. POSTERIOR INFERENCE

4·1. *Posterior inference with multivariate Polya trees*

Posterior inference in the proposed model can be implemented by Markov chain Monte Carlo simulation.

Consider posterior inference conditional on data $W = \{W_i, x_i\}_{i=1}^n$, under the sampling model $W_i \sim \mathcal{P}_{x_i}$ and $\{\mathcal{P}_x\}_{x \in X} \sim \text{MPT}(\mathcal{A}, Q, F)$. We explain two key concepts that allow the construction of posterior Markov chain Monte Carlo simulation, the independence of the processes $\{Y_{\varepsilon,x}\}_{x \in X}$ across ε, both a priori and a posteriori, and the Polya urn scheme with latent variables $\{(l_{ij}, h_{ij}); i = 1, \ldots, N, j \geqslant 1\}$ introduced in §2·2. Throughout the sequel, we use finite Polya trees as approximations. Let m be a positive integer, $M = 2^m$ and let $\varepsilon_1, \ldots, \varepsilon_M$ denote the indices of the partitioning subsets B_ε at level m of the multivariate Polya tree construction. Posterior inference on $\mathcal{P}_x\{F_x^{-1}(B_{\varepsilon_j})\}$, $j = 1, \ldots, M$, provides an approximation of the conditional law of the random probability measure $\mathcal{P}_x$. Let $\mathcal{Y}_m = \{Y_{\varepsilon,x}; \varepsilon \in \bigcup_{\ell=0}^{m-1}\{0, 1\}^\ell, x \in X\}$ denote all random branching probabilities up to level $(m - 1)$. Given the data W, a sufficient statistic for updating $\mathcal{Y}_m$ consists of the array of indicators

$$I_m = \left[I \left\{ F_{x_i}^{-1}\left(\frac{j}{M}\right) < W_i \leqslant F_{x_i}^{-1}\left(\frac{j+1}{M}\right) \right\}; \ i = 1, \ldots, n, \ j = 1, \ldots, M-1 \right].$$

The sampling model for I_m, conditionally on $\mathcal{Y}_m$, can be conveniently factored using (7). We rearrange the likelihood such that all factors related to a single multivariate beta process $\{Y_{\varepsilon,x}\}_{x \in X}$ are grouped together and observe that the multivariate beta processes that

define the multivariate Polya tree are both a priori and a posteriori independent. It thus suffices to discuss posterior inference of $\{Y_{\varepsilon,x}\}_{x\in X}$ for a fixed ε. Let $N = \sum_{i=1}^{n} I\{F_{x_i}(W_i) \in B_\varepsilon\}$ and, without loss of generality, assume that $F_{x_i}(W_i) \in B_\varepsilon$ for $i = 1, \ldots, N$. The next paragraph describes a Gibbs sampler algorithm that allows posterior inference for $\{Y_{\varepsilon,x}\}_{x\in X}$ given a sample $Z = (Z_1, \ldots, Z_N)$ of binary variables with covariates $x_1, \ldots, x_N$, characterized as in (3). The variables Z_i, for $i = 1, \ldots, N$, coincide with the indicator functions $I\{F_{x_i}(W_i) \in B_{\varepsilon 1}\}$.

Recall the definitions of $D_{x_1 \ldots x_N}$, $\{(l_{ij}, h_{ij}); \ i = 1, \ldots N, \ j \geqslant 1\}$ and $\{j_1, \ldots, j_N\}$ in the sampling algorithm described in §2·2. The posterior distribution of the truncated sequences $\{(l_{ij}, h_{ij}); \ i = 1, \ldots, N, \ j = 1, \ldots, j_i\}$, conditionally on the binary random variables $Z = \{I(h_{1j_1} > 0), \ldots, I(h_{Nj_N} > 0)\}$, with regression function $\{Y_x\}_{x\in X} \sim \text{MBP}(\alpha_0, \alpha_1, Q)$, can be approximated by iteratively sampling from the full conditionals

$$\{l_{uj}, h_{uj}\}_{j=1}^{j_u} \mid \{(l_{ij}, h_{ij}); \ i = 1, \ldots, u-1, u+1, \ldots, N, \ j = 1, \ldots, j_i\}, Z. \qquad (9)$$

Without loss of generality we assume that $u = N$. A minor modification to the algorithm in §2·2 can be used for sampling from (9). We sample $\{(l_{Nj}, h_{Nj}); \ j = 1, \ldots, j_N\}$ from $D_{x_1 \ldots x_N}$ conditional on $\{(l_{ij}, h_{ij}); i = 1, \ldots, N-1, \ j = 1, \ldots, j_i\}$ using the updated Polya urn. So far we are following the sampling scheme illustrated in §2·2. An additional accept–reject step is introduced. If $h_{j_N} > 0$ and $Z_N = 1$ or, symmetrically, $h_{i_N} < 0$ and $Z_N = 0$, then the realization is saved. Otherwise, sampling is repeated until the condition is satisfied. Using the latent random variables $\{(l_{ij}, h_{ij}); \ i = 1, \ldots, N, \ j = 1, \ldots, j_i\}$, we can generate a predictive sample for Z_{N+1}. It suffices to exploit the conjugacy property of the Dirichlet process in (2) with respect to the random variables $\{(l_{ij}, h_{ij}); \ i = 1, \ldots, N, \ j = 1, \ldots, j_i\}$ and, if x_{N+1} does not belong to $\{x_1, \ldots, x_N\}$, the tail free property of the Dirichlet process.

4·2. *Posterior inference with mixtures of multivariate Polya trees*

In this subsection we consider mixtures of truncated multivariate Polya trees: $Y_{\varepsilon_1, \ldots, \varepsilon_j, x} = \frac{1}{2}$ for $j \geqslant m$. We write $\{\mathcal{P}_x\}_{x\in X} \sim \text{MPT}_m(\mathcal{A}, Q, F)$ to denote a truncated multivariate Polya tree, truncated at level m. Consider the model

$$\theta \sim \lambda, \qquad \{\mathcal{P}_x\}_{x\in X}|\theta \sim \text{MPT}_m(\mathcal{A}, Q, F_\theta), \qquad W_i|\{\mathcal{P}_x\}_{x\in X} \sim \mathcal{P}_{x_i} \qquad (i = 1, \ldots, n).$$

Below we give Markov chain transition probabilities that can be used to define Markov chain Monte Carlo posterior simulation for the model.

Consider again the array $\{(l_{ij}, h_{ij}); i = 1, \ldots, n, \ j \geqslant 1\}$ defined in §2, $(j_1, \ldots, j_n)$, $D_{x_1, \ldots, x_n}$ and the binary vector $(Z_1, \ldots, Z_n)$ with $Z_i = I(h_{ij_i} > 0)$. Let k be an arbitrarily chosen positive integer, say $k = 30$. We define two more latent quantities, k_i and H_i. Let $k_i = \max(k, j_i)$. We will later condition on $\{(l_{ij}, h_{ij}); \ j = 1, \ldots, k_i\}$. Also, we define the unsorted set $H_i = \{(l_{ij}, h_{ij}), \ j = 1, \ldots, k_i\}$. It will be important that H_i is an unsorted set rather than a sequence. We repeat this construction for each of the $\{Y_{\varepsilon,x}\}_{x\in X}$ in the construction of $\{\mathcal{P}_x\}_{x\in X}$; we define arrays $\{(l_{ij\varepsilon}, h_{ij\varepsilon}); i = 1, \ldots, n, \ j \geqslant 1\}$, vectors $(j_{1\varepsilon}, \ldots, j_{n\varepsilon})$, binary variables $(Z_{1\varepsilon}, \ldots, Z_{n\varepsilon})$ and sets $(H_{1\varepsilon}, \ldots, H_{n\varepsilon})$. The use of the random sets $H_{i\varepsilon}$ is motivated by the following fact: if the parameter θ and the arrays $\{(l_{ij\varepsilon}, h_{ij\varepsilon}); i = 1, \ldots, n, \ j \geqslant 1\}$ are iteratively sampled from the full conditional distributions, then we only obtain a random sequence of θ values that preserve

the classes B_ε to which $\{F_{\theta,x_1}(W_1), \ldots, F_{\theta,x_n}(W_n)\}$ belong. That is, the random sets $H_{i\varepsilon}$ are introduced in order to define an irreducible Markov chain. The Markov chain Monte Carlo algorithm exploits the identities

$$\mathrm{pr}\{F_{\theta,x_i}(W_i) \in B_{\varepsilon_1,\ldots,\varepsilon_m}\} = \mathrm{pr}(Z_{i\varepsilon_1\cdots\varepsilon_j} = \varepsilon_{j+1}; \ j = 0, \ldots, m-1),$$

$$F_{\theta,x_i}(W_i) \mid F_{\theta,x_i}(W_i) \in B_{\varepsilon_1,\ldots,\varepsilon_m} \sim \mathrm{Un}(B_{\varepsilon_1,\ldots,\varepsilon_m}).$$

We use three transition probabilities and adopt the notation $(x \mid y, z)$ to indicate that the random quantity x is updated conditional on currently imputed values for y and z. In the following description $L_{i\varepsilon}$ is the ith row of $L_\varepsilon = \{(l_{ij\varepsilon}, h_{ij\varepsilon}); \ i = 1, \ldots, n, \ j = 1, \ldots, j_{i\varepsilon}\}$ and H denotes the random sets $H_{i\varepsilon}$.

The first transition is $(L_{i\varepsilon} \mid \theta, W, H)$. This step is carried out separately for all i and all ε, exploiting independence across i and ε, conditional on θ, W and H. Conditioning on (θ, W) could determine $Z_{i\varepsilon}$ and the indicator $I(h_{j_{i\varepsilon}} \geqslant 0)$. A simple example clarifies that sampling from the conditional distribution reduces to a combinatorial exercise. Consider $A_1, \ldots, A_6$ and D_{x_1,x_2} as in Fig. 1(a). Let $k = 3$, $\varepsilon = \oslash$ and $i = 1$. Assume $H_{1\oslash}$ is constituted by three points, respectively, in A_1, A_5 and A_6. If $F_{\theta,x_1}(W_1) > 0.5$ and $Z_{1\oslash} = 1$, then, simply counting the permutations of the elements in $H_{1\oslash}$ that satisfy $h_{j_1} > 1$ we find $\mathrm{pr}(L_{1\oslash} \in A_6 \times A_1 \mid \theta, W, H) = 1/3$. By similar arguments, we generate $L_{i\varepsilon}$.

The second transition is $(H \mid L_\varepsilon, \theta, W)$. This step is carried out separately for each ε, exploiting conditional independence. It is implemented by sampling, from the updated Polya urn, the random variables $\{(l_{ij\varepsilon}, h_{ij\varepsilon}); \ i = 1, \ldots, n, \ j = j_{i\varepsilon} + 1, \ldots, k_{i\varepsilon}\}$ conditional on L_ε.

Finally, the third transition is $(\theta \mid H, \theta, W)$. We include θ also in the conditioning subset since the proposed Metropolis–Hastings transition probability also depends on the currently imputed value of θ. Propose a transition $\theta \to \theta'$ and accept or reject the proposed value on the basis of the standard Metropolis–Hastings rule. The acceptance probability requires the evaluation of

$$p(\theta \mid W, H) \propto p(\theta, W \mid H) \propto p(\theta) \prod_{i=1}^{n} p\{F_{\theta,x_i}(W_i) \mid H\} \prod_{i=1}^{n} F'_{\theta,x_i}(W_i), \tag{10}$$

where $p\{F_{\theta,x_i}(W_i) \mid H\}$ denotes the conditional density function of $F_{\theta,x_i}(W_i)$ and F'_{θ,x_i} is the derivative of F_{θ,x_i}. The right-hand side of (10) is evaluated using a simple combinatorial argument, which is best explained with an example. For simplicity, assume $m = 1$ and focus on $i = 1$. Consider $A_1, \ldots, A_6$ and D_{x_1,x_2} as in Fig. 1(a), $k = 3$ and $H_{1\oslash}$ constituted by three points, respectively, in A_1, A_2 and A_5. Then

$$p\{F_{\theta,x_i}(W_i) \mid H\} = \begin{cases} 2/3, & 0 < F_{\theta,x_1}(W_1) < 0.5, \\ 4/3, & 0.5 \leqslant F_{\theta,x_1}(W_1) < 1. \end{cases}$$

Only two out of the six possible permutations of A_1, A_2, A_5 imply $(l_{j_1}, h_{j_1}) \in A_5$ and $Z_1 = 0$.

Censored data are handled by imputing the missing event times and sampling them conditionally on θ and H; a similar approach is discussed in Walker & Mallick (1999).
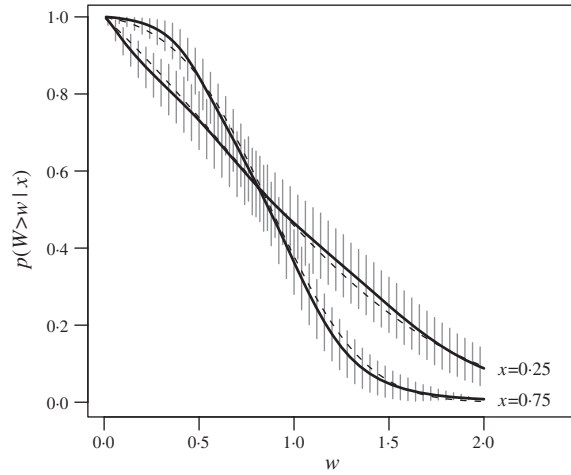
Fig. 3. Simulation example. The dashed lines represent the
true survival functions, $\mathrm{pr}(W_i > w_i \mid x_i)$, for $x_i = 0\cdot25$ and
$x_i = 0\cdot75$. The solid lines illustrate the posterior estimates
of the two survival functions. The shaded bands show 90%
credible intervals.

5. Examples

5·1. *Simulation example*

The first example illustrates the flexibility of the multivariate Polya tree model. We consider
a simulated dataset generated from the following model:

$$X = [0, 1], \qquad p(W_i \mid x_i) \propto \exp\left\{ -\frac{(W_i - \mu_{x_i})^2}{2\sigma_{x_i}^2} \right\} I(W_i \geqslant 0),$$

$$\mu_{x_i} = 0\cdot5x_i + 0\cdot5, \qquad \sigma_{x_i} = -1\cdot08x_i + 1\cdot2.$$

The density $p(W_i \in dw \mid x_i)$ is a truncated normal density. The covariates x_i are uniformly sam-
pled from the unit interval. The simulation model includes censoring times $\{C_i\}_{i \geqslant 1}$ that are inde-
pendent of $\{W_i\}_{i \geqslant 1}$. The censoring times are exponentially distributed with mean equal to $5\cdot7$;
the resulting censoring probability $\mathrm{pr}(C_i < W_i)$ is approximately equal to $0\cdot15$. We generated a
sample of $n = 300$ observations and estimated the underlying distributions. We use a mixture of
multivariate Polya trees truncated at $m = 5$. The hierarchical prior (8) is centred on a Weibull
model. The dashed lines in Fig. 3 show the simulation truth for two covariate values $x_1 = 0\cdot25$
and $x_2 = 0\cdot75$. The distribution functions associated with x_1 and x_2 cross. The sampling model
violates the proportional hazards assumption. The simulation results show how the dependent
branching probabilities allow posterior inference to adapt to the data when the assumptions of
the centring model are violated. Figure 3 shows the corresponding posterior estimates and point-
wise credibility intervals. Posterior inference appears to achieve a balance between borrowing
strength across covariate levels versus reporting meaningful covariate effects. These two goals are
in conflict with each other. Excessive prior dependence diminishes the flexibility of the posterior
inference to adapt to the model structure. On the other hand, in the absence of prior dependence
we would not borrow any strength across x.

The estimates are based on 30 000 iterations of the Markov chain Monte Carlo algorithm
described in §4·2 after a burn in of 10 000 iterations. Convergence of the Markov chain Monte

Table 1. *Monte Carlo approximations of mean absolute errors for point estimates of* $\mathrm{pr}(W_{n+1} > w \mid x_{n+1})$ *for some values of* x_{n+1} *and* w: *a comparison between mixtures of multivariate Polya trees and dependent mixtures of log-normal densities specified with kernel stick breaking processes* (*Dunson,* 2008). *Covariate values are* 0·25, 0·50 *and* 0·75. *For each value of* x_{n+1} *we consider* w *equal to the 25th, 50th and 75th percentile of the corresponding truncated normal distribution.*

		Mean absolute errors		
	x_{n+1}	$w = $ 25th percentile	$w = $ 50th percentile	$w = $ 75th percentile
MMPT	0·25	0·044	0·049	0·041
	0·50	0·042	0·046	0·040
	0·75	0·047	0·049	0·042
KSBP	0·25	0·042	0·050	0·045
	0·50	0·043	0·051	0·042
	0·75	0·045	0·047	0·046

MMPT, mixtures of multivariate Polya trees; KSBP, kernel stick breaking process.

Carlo simulations has been validated by simulating six independent Markov chains starting from distinct values of the hyperparameter θ. We also evaluated the convergence diagnostic proposed by Raftery & Lewis (1992). This method requires a single Markov chain realization, which is used to verify the accuracy of the computed predictive probabilities of the events $\{W_{n+1} > w\}$ for a grid of values of w. We found no evidence of lack of convergence. The computational time required for computing the estimates in Fig. 3 with a laptop computer was less than 9 minutes.

We compare the proposed mixture of multivariate Polya trees with a recent proposal for modelling dependent random distributions (Dunson & Park, 2008). This approach extends the Dirichlet mixture model by introducing a kernel stick breaking process which defines dependent discrete random distributions. The random distributions are then used for specifying dependent mixtures indexed by covariates. In our comparison we used mixtures of lognormal densities. This modelling approach extends the density regression method introduced in Dunson et al. (2007). We add a simple imputation step for censored data to the algorithm discussed in Dunson & Park (2008). We carried out 100 simulations of the inferential procedure described in the previous paragraphs. In Table 1 we summarize the simulation results by reporting mean absolute errors of the posterior estimates of $\mathrm{pr}(W_{n+1} > w \mid x_{n+1})$ for several values of w and x_{n+1}. The results show the overall comparable errors under the two models.

5·2. *A lung cancer trial*

We apply the proposed mixture of multivariate Polya tree models to the analysis of survival data from a clinical trial performed by the Lung Cancer Study Group (Lad et. al., 1980). The objective of the trial was to determine the potential benefit of adjuvant chemotherapy for patients with incompletely resected non-small-cell lung cancer. Patients were randomly assigned to either radiotherapy, or radiotherapy plus chemotherapy. The trial demonstrated the benefit of adjuvant chemotherapy in addition to the radiotherapy. The survival data are published in Piantadosi (1997); 164 patients were enrolled in the trial and 28 were alive at the end of the follow-up period.

The most relevant prognostic factors are cancer histology and performance status at enrollment. Histology is coded as $X_1 = 0$ for squamous and $X_1 = 1$ for non squamous. Performance
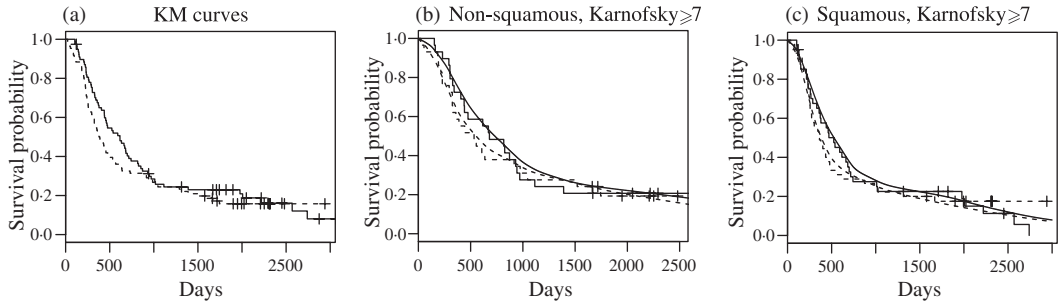
Fig. 4. Lung cancer trial. (a) Kaplan–Meier estimates of survival functions. Solid lines refer to treatment and dashed lines refer to the control group. (b) and (c) The model-based estimates of survival curves arranged by cancer histology. For comparison, the corresponding Kaplan–Meier estimates are also included.

status is coded as $X_2 = 0$ for Karnofsky indicator $\geqslant 7$ and $X_2 = 1$ otherwise; we use the dichotomized Karnofsky measurements provided in Piantadosi (1997). The third covariate is the treatment assignment; we use $X_0 = 1$ for treatment and $X_0 = 0$ for control.

Observed survival times in the treatment and control arms are shown in Fig. 4(a). The Kaplan–Meier survival functions for treatment and control cross at around $w = 1000$ days. For a while after the crossing, the differences are negligible. At the end of the third year, the Kaplan–Meier estimate of the survival function for the treatment group is above the corresponding curve for controls.

The prior is parameterized by $\alpha_\varepsilon = 4$ for every ε. The kernels q_x in this case are multivariate normal densities with diagonal covariance matrix and the three standard deviations equal to 3.

As this example demonstrates the proposed multivariate Polya tree can easily be implemented for a moderate number of covariates. In fact, the computational effort of posterior inference scales only with the sample size, not with the number of covariates. Here the estimates of the survival curves have been obtained on the basis of 25 000 iterations of the proposed algorithm, after a burn in of 7000 iterations. The only concern limiting the number of covariates is the ability to define meaningful kernels to induce the desired correlation. For practical implementations, we see no problem with using the model with up to ten or so covariates.

Figure 4 summarizes some aspects of posterior inference under the proposed model. The model-based estimates identify crossing survival functions. The corresponding Kaplan–Meier curves are shown for comparison. Despite prior centring around the proportional hazards structure, posterior inference correctly identifies the intersecting curves. In contrast to the proportional hazards model, the proposed MPT model allows posterior inference to violate the structure of the prior centring family if the data so indicate.

6. Discussion

The construction of probability models for dependent random measures has been studied by several authors in recent years. Among these, MacEachern (1999) and De Iorio et. al. (2004) explore the idea of defining dependent random probability measures by means of an extension of the Dirichlet process mixture model while Dunson & Park (2008) discuss an extension of the stick-breaking representation.

The two main guiding principles in the proposed construction of dependent random probability measures are the need to specify prior distributions with a clear interpretation and the aim of constructing a probability model with large support.

One of the main strengths of the proposed approach is the use of marginal Polya tree priors, which allows one to restrict the model to continuous random distributions. Posterior inference under the finite multivariate Polya tree model allows easy interpretation as essentially a random histogram method. Some of the remaining limitations are computational challenges of higher dimensional extensions, requiring the manipulation of higher dimensional partitioning subsets, and the sensitivity of posterior inference on partitioning boundaries. The latter is greatly mitigated by the hierarchical extension of the multivariate Polya tree model to a mixture of multivariate Polya trees with a hyperprior for the centring model.

The multivariate beta process has interesting applications beyond the multivariate Polya tree model. The multivariate beta process provides a flexible construction to define dependent random probabilities indexed by covariates; the proposed approach can be generalized for defining dependent random vectors on a finite-dimensional simplex or dependent discrete random distributions. For example, it could be used to define priors for finite mixtures by formalizing a regression of mixture weights on covariates.

A natural continuation to the above line of research is the study of the asymptotic properties of Bayesian inferential procedures based on the proposed class of dependent random probability measures. We hope that the strict similarities of the multivariate Polya tree model with the tail free prior distributions could prove helpful in future research for exploring its asymptotic behaviour.

APPENDIX

PROPOSITION A1. *Let α_0 and α_1 be two strictly positive real values. If*

$$r_2^0 = r_1^0 = \alpha_0(1-a), \ \ r_c^0 = \alpha_0 a, \ \ r_2^1 = r_1^1 = \alpha_1(1-a), \ \ r_c^1 = \alpha_1 a,$$

where $a \in (0, 1)$, and $(G_1^0, G_2^0, G_1^1, G_2^1, G_c^0, G_c^1)$ are independent gamma random variables with same scale parameter and shape parameters $(r_1^0, r_2^0, r_1^1, r_2^1, r_c^0, r_c^1)$, then the function

$$\rho(a) = \mathrm{corr}\left(\frac{G_1^0 + G_c^0}{G_1^0 + G_c^0 + G_1^1 + G_c^1}, \frac{G_2^0 + G_c^0}{G_2^0 + G_c^0 + G_2^1 + G_c^1}\right)$$

is monotone increasing.

Proof. The vector

$$\left(\frac{G_1^0}{G_1^0 + G_2^0 + G_1^1 + G_2^1 + G_c^0 + G_c^1}, \dots, \frac{G_c^0}{G_1^0 + G_2^0 + G_1^1 + G_2^1 + G_c^0 + G_c^1}\right)$$

is Dirichlet distributed with parameter $(r_1^0, r_2^0, r_1^1, r_2^1, r_c^0, r_c^1)$. Consider a Blackwell–McQueen random sequence $\{l_k\}_{k \geqslant 1}$ from an urn with initial weights $(r_1^0, r_2^0, r_1^1, r_2^1, r_c^0, r_c^1)$ and colours $(C_1^0, C_2^0, C_1^1, C_2^1, C_c^0, C_c^1)$. Let S be the first ball from the Blackwell–McQeen urn whose colour belongs to the set $(C_1^0, C_1^1, C_c^0, C_c^1)$ and let V be the first ball, subsequent to S, whose colour belongs to

the set $(C_2^0, C_2^1, C_c^0, C_c^1)$. Then

$$E\left(\frac{G_1^0 + G_c^0}{G_1^0 + G_c^0 + G_1^1 + G_c^1} \frac{G_2^0 + G_c^0}{G_2^0 + G_c^0 + G_2^1 + G_c^1}\right)$$

$$= \text{pr}\{S \in (C_1^0 \cup C_c^0), V \in (C_2^0 \cup C_c^0)\}$$

$$= \sum_{m=1}^{\infty} \text{pr}\left\{\bigcap_{k=1}^{m-1} l_k \in (C_2^0 \cup C_2^1), l_m \in (C_1^0 \cup C_c^0)\right\}$$

$$\times \text{pr}\left\{V \in (C_2^0 \cup C_c^0) \mid \bigcap_{k=1}^{m-1} l_k \in (C_2^0 \cup C_2^1), l_m \in (C_1^0 \cup C_c^0)\right\}. \tag{A1}$$

For each component of the sum, the first product term equals

$$p'(a, m) = \frac{\alpha_0}{(2-a)(\alpha_0 + \alpha_1) + m - 1} \prod_{k=1}^{m-1} \frac{(1-a)(\alpha_0 + \alpha_1) + k - 1}{(2-a)(\alpha_0 + \alpha_1) + k - 1}$$

and the second equals

$$p''(a, m) = \left(\frac{\alpha_0}{\alpha_0 + \alpha_1}\right)\left\{1 - a\frac{a(\alpha_0 + \alpha_1) + 1}{(\alpha_0 + \alpha_1) + m}\right\} + \left\{a\frac{a\alpha_0 + 1}{(\alpha_0 + \alpha_1) + m}\right\}.$$

The following inequalities can be verified with simple algebra. If a, a^* and a^{**} belong to the interval $(0, 1)$ and $a^{**} > a^*$, then $\sum_{m \geqslant j} p'(a^{**}, m) \leqslant \sum_{m \geqslant j} p'(a^*, m)$ for every j and, for every m, $p''(a^{**}, m) > p''(a^*, m)$ and $p''(a, m) > p''(a, m+1)$. These inequalities and the equality $\sum_{m \geqslant 1} p'(a^{**}, m) = \sum_{m \geqslant 1} p'(a^*, m)$ imply that

$$\sum_{m=1}^{\infty} p'(a^{**}, m) p''(a^{**}, m) > \sum_{m=1}^{\infty} p'(a^*, m) p''(a^*, m).$$

The last inequality completes the proof, indeed the two terms are equivalent to the expected value in (A1) under different parameterizations of $G_1^0, \ldots, G_c^1$. □

Proof of Proposition 1.   For every i, let

$$G_{x,x_i}^0 = G[(z, y) \in X \times \mathbb{R} : 0 < y < \alpha_0 \min\{q_{x_i}(z), q_x(z)\}],$$

$$G_{x \setminus x_i}^0 = G\{(z, y) \in X \times \mathbb{R} : 0 < y < \alpha_0 q_x(z), \ y \geqslant \alpha_0 q_{x_i}(z)\},$$

$$G_{x_i \setminus x}^0 = G\{(z, y) \in X \times \mathbb{R} : 0 < y < \alpha_0 q_{x_i}(z), \ y \geqslant \alpha_0 q_x(z)\}.$$

We observe that $G(S_x^0) = G_{x,x_i}^0 + G_{x \setminus x_i}^0$, $G(S_{x_i}^0) = G_{x,x_i}^0 + G_{x_i \setminus x}^0$ and

$$E[\{G(S_x^0) - G(S_{x_i}^0)\}^2] = E\{(G_{x \setminus x_i}^0 - G_{x_i \setminus x}^0)^2\} = 2\text{var}(G_{x_i \setminus x}^0).$$

The hypotheses imply that $\nu\{(z, y) \in X \times \mathbb{R} : 0 < y < q_{x_i}(z), \ q_x(z) > y\} \to 0$, and it follows that $\text{var}(G_{x_i \setminus x}^0) \to 0$. We can similarly define G_{x,x_i}^1, $G_{x \setminus x_i}^1$ and $G_{x_i \setminus x}^1$, and verify that $E[\{G(S_x^1) - G(S_{x_i}^1)\}^2] \to 0$. Finally, the convergences in mean square of $\{G(S_{x_i}^0)\}_{i \geqslant 1}$ and $\{G(S_{x_i}^1)\}_{i \geqslant 1}$ imply that, for every $\varepsilon > 0$, $\text{pr}(|Y_{x_i} - Y_x| > \varepsilon) \to 0$. □

Proof of Proposition 2.   For every $\varepsilon \in \cup_{i=1}^{\infty}\{0, 1\}^i$, $E[\mathcal{P}_x\{F_x^{-1}(B_\varepsilon)\}] = \lambda(B_\varepsilon)$, where λ denotes the Lebesgue measure. The same equality holds for the elements of the algebra of finite unions of B_ε sets. The class $(B \subseteq (0, 1] : \lambda(B) = E[\mathcal{P}_x\{F_x^{-1}(B)\}])$ is monotone. The monotone class theorem implies that for every Borel subset B of $(0, 1]$, $E[\mathcal{P}_x\{F_x^{-1}(B)\}] = \lambda(B)$. Finally, if a distribution p satisfies the equality $p\{F_x^{-1}(B)\} = \lambda(B)$ for every Borel subset B of $(0, 1]$ then $p = F_x$. □

Proof of Proposition 3. Proposition 1 implies that, for every $\varepsilon \in \cup_{i=1}^{\infty}\{0,1\}^i$, the sequence $\mathcal{P}_{x_i}\{F_{x_i}^{-1}(B_\varepsilon)\}$ $(i = 1, 2 \ldots)$ converges in probability to $\mathcal{P}_x\{F_x^{-1}(B_\varepsilon)\}$. Similarly, if a measurable set B belongs to the algebra generated by $\{B_\varepsilon; \varepsilon \in \cup_{i=1}^{\infty}\{0,1\}^i\}$, then the random variables $\mathcal{P}_{x_i}\{F_{x_i}^{-1}(B)\}$ converge in probability to $\mathcal{P}_x\{F_x^{-1}(B)\}$. Let $B \subset [0,1]$ be a measurable set. Let $B_1, B_2, \ldots$ be an increasing sequence such that $\cup_{j \geqslant 1} B_j = B$ and assume that, for every $j \geqslant 1$, the random variables $\mathcal{P}_{x_i}\{F_{x_i}^{-1}(B_j)\}$ converge in probability to $\mathcal{P}_x\{F_x^{-1}(B_j)\}$. We observe that, for every pair (i, j), $\mathcal{P}_{x_i}\{F_{x_i}^{-1}(B)\} - \mathcal{P}_{x_i}\{F_{x_i}^{-1}(B_j)\} \geqslant 0$ almost surely. We also observe that $E[\mathcal{P}_{x_i}\{F_{x_i}^{-1}(B)\} - \mathcal{P}_{x_i}\{F_{x_i}^{-1}(B_j)\}] = \lambda(B \setminus B_j)$, where λ denotes the Lebesgue measure. It follows, using the Markov inequality, that, for every $\delta > 0$ and $\xi > 0$, there exists m such that

$$\text{pr}\left[|\mathcal{P}_x\{F_x^{-1}(B_m)\} - \mathcal{P}_x\{F_x^{-1}(B)\}| > \frac{\xi}{3}\right] < \frac{\delta}{3},$$

$$\text{pr}\left[|\mathcal{P}_{x_i}\{F_{x_i}^{-1}(B_m)\} - \mathcal{P}_{x_i}\{F_{x_i}^{-1}(B)\}| > \frac{\xi}{3}\right] < \frac{\delta}{3}, \qquad i \geqslant 1.$$

Moreover, there exists $k \geqslant 1$ such that, for every $j > k$,

$$\text{pr}\left[|\mathcal{P}_{x_j}\{F_{x_j}^{-1}(B_m)\} - \mathcal{P}_x\{F_x^{-1}(B_m)\}| > \frac{\xi}{3}\right] < \frac{\delta}{3}.$$

For every $j > k$, $\text{pr}[|\mathcal{P}_{x_j}\{F_{x_j}^{-1}(B)\} - \mathcal{P}_x\{F_x^{-1}(B)\}| > \xi] < \delta$. Similarly, if the sequence $B_1, B_2, \ldots$ decreases to B and, for every $j \geqslant 1$, the random variables $\mathcal{P}_{x_i}\{F_{x_i}^{-1}(B_j)\}$ converge in probability to $\mathcal{P}_x\{F_x^{-1}(B_j)\}$, then the sequence $\mathcal{P}_{x_i}\{F_{x_i}^{-1}(B)\}$ converges in probability to $\mathcal{P}_x\{F_x^{-1}(B)\}$. It follows, using the monotone class theorem, that for every Borel subset $B \subset [0,1]$ the sequence $\mathcal{P}_{x_i}\{F_{x_i}^{-1}(B)\}$ converges in probability to $\mathcal{P}_x\{F_x^{-1}(B)\}$.

Assume that the pair (a, b) satisfies $c = F_x(a)$. The cumulative distribution functions F_{x_i} and F_x are continuous. Therefore, the variables $|\mathcal{P}_{x_i}\{(-\infty, F_{x_i}^{-1}(c)]\} - \mathcal{P}_x\{(-\infty, a]\}|$ and $|\mathcal{P}_{x_i}\{F_{x_i}^{-1}(0, c]\} - \mathcal{P}_x\{F_x^{-1}(0, c]\}|$ are almost surely identical. The random sequence $|\mathcal{P}_{x_i}\{F_{x_i}^{-1}(0, c]\} - \mathcal{P}_x\{F_x^{-1}(0, c]\}|$ converges in probability to zero. Note that $E[|\mathcal{P}_{x_i}\{(-\infty, F_{x_i}^{-1}(c)]\} - \mathcal{P}_{x_i}\{(-\infty, a]\}|] = |c - F_{x_i}(a)|$ and $|c - F_{x_i}(a)| \to 0$. It follows that the random sequences $|\mathcal{P}_{x_i}\{(-\infty, F_{x_i}^{-1}(c)]\} - \mathcal{P}_{x_i}\{(-\infty, a]\}|$ and $|\mathcal{P}_{x_i}\{(-\infty, a]\} - \mathcal{P}_x\{(-\infty, a]\}|$ converge in probability to zero. Finally, through monotone class arguments and exploiting the convergence in total variation of $\{F_{x_i}\}_{i \geqslant 1}$ to F_x, we verify that, for every Borel subset B, the sequence $\mathcal{P}_{x_i}(B)$ converges in probability to $\mathcal{P}_x(B)$. □

Proof of Proposition 4. Assume that $\{Y_x\}_{x \in X} \sim \text{MBP}(\alpha_0, \alpha_1, Q)$. We prove that for every $\xi > 0$, for every $m = 1, 2, \ldots$, for every distinct $x_1, \ldots, x_m \in X$ and $a_1, \ldots, a_m \in [0,1]$ the event $\{|Y_{x_i} - a_i| < \xi; i = 1, \ldots m\}$ has strictly positive probability.

There exist strictly positive numbers (g_1^0, g_1^1) such that $|\{g_1^1/(g_1^0 + g_1^1)\} - a_1| < \xi/2$. Given $G(S_{x_1}^0) = g_1^0$ and $G(S_{x_1}^1) = g_1^1$, there exist strictly positive numbers (g_2^0, g_2^1) such that, if $G(S_{x_2}^0 \setminus S_{x_1}^0) = g_2^0$ and $G(S_{x_2}^1 \setminus S_{x_1}^1) = g_2^1$, then $|[G(S_{x_2}^1)/\{G(S_{x_2}^0) + G(S_{x_2}^1)\}] - a_2| < \xi/2$. Similarly, for any fixed value of $G(S_{x_1}^1 \cup S_{x_2}^1 \cup \ldots \cup S_{x_i}^1)$ and $G(S_{x_1}^0 \cup S_{x_2}^0 \cup \ldots \cup S_{x_i}^0)$, there exist positive numbers (g_{i+1}^0, g_{i+1}^1) such that, if $G\{S_{x_{i+1}}^0 \setminus (S_{x_1}^0 \cup \ldots \cup S_{x_i}^0)\} = g_{i+1}^0$ and $G\{S_{x_{i+1}}^1 \setminus (S_{x_1}^1 \cup \ldots \cup S_{x_i}^1)\} = g_{i+1}^1$, then $|[G(S_{x_{i+1}}^1)/\{G(S_{x_{i+1}}^0) + G(S_{x_{i+1}}^1)\}] - a_{i+1}| < \xi/2$. The quantities $\nu(S_{x_1}^0), \nu(S_{x_1}^1), \nu(S_{x_2}^0 \setminus S_{x_1}^0), \ldots, \nu\{S_{x_m}^1 \setminus (S_{x_1}^1 \cup \ldots \cup S_{x_{m-1}}^1)\}$ are strictly positive, indeed $\{q_x\}_{x \in X}$ is a location family of unimodal densities. It follows that the density function of $[G(S_{x_1}^0), G(S_{x_1}^1), G(S_{x_2}^0 \setminus S_{x_1}^0), \ldots, G\{S_{x_m}^1 \setminus (S_{x_1}^1 \cup \ldots \cup S_{x_{m-1}}^1)\}]$ in an adequate neighbourhood of $(g_1^0, g_1^1, \ldots, g_m^1)$ is strictly positive. This fact proves our assertion.

For every positive ξ and every integer j the event

$$\left\{\left|\mathcal{S}_i\left(F_{x_i}^{-1}\left(\frac{l}{2^j}, \frac{l+1}{2^j}\right]\right) - \mathcal{P}_{x_i}\left(F_{x_i}^{-1}\left(\frac{l}{2^j}, \frac{l+1}{2^j}\right]\right)\right| < \xi; \; l \leqslant (2^j - 1), \; i = 1, \ldots, m\right\}$$

has strictly positive probability. Finally, for any positive δ and partition $(V_1, \ldots, V_k)$ constituted of intervals, the absolute continuity hypothesis $S_i \ll F_{x_i} \ll \lambda$, where λ denotes the Lebesgue measure, implies that, if j and ξ are suitably chosen, then this event becomes a subset of $\{|\mathcal{S}_i(V_l) - \mathcal{P}_{x_i}(V_l)| < \delta; \ l = 1, \ldots, k, \ i = 1, \ldots, m\}$. $\qquad\square$

References

Berger, J. O. & Guglielmi, A. (2001). Bayesian testing of a parametric model versus nonparametric alternatives. *J. Am. Statist. Assoc.* **96**, 174–84

Berman, M. (1978). A class of isotropic distributions in R^n and their characteristic functions. *Pac. J. Math.* **78**, 1–9.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Polya urn schemes. *Ann. Statist.* **1**, 353–5.

De Iorio, M., Muller, P., Rosner, G. L. & MacEachern, S. (2004). An ANOVA model for dependent random measures. *J. Am. Statist. Assoc.* **99**, 205–15.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Prob.* **2**, 183–201.

Dunson, D. B. & Park, B. K. (2008). Kernel stick-breaking processes. *Biometrika* **95**, 307–23.

Dunson, D. B., Pillai, N. & Park, B. K. (2007). Bayesian density regression. *J. R. Statist. Soc.* B **69**, 163–83.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **4**, 615–29.

Gneiting, T. (1999). Radial positive definite functions generated by Euclid's hat. *J. Mult. Anal.* **69**, 88–19.

Hanson, T. (2006). Inference for mixtures of finite Polya tree models. *J. Am. Statist. Assoc.* **101**, 1548–65.

Hanson, T. & Johnson, W. (2002). Modeling regression error with a mixture of Polya trees. *J. Am. Statist. Assoc.* **97**, 1020–33.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Ann. Statist.* **18**, 1259–94.

Lad, T., Rubinstein, L. & Sadeghi, A. (1988). The benefit of adjuvant treatment for resected locally advanced non-small-cell lung cancer. *J. Clin. Oncol.* **6**, 9–17.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **20**, 1222–35.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **22**, 1161–76.

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA Proc. Sec. Bayesian Statist. Sci.*, pp. 50–5. Alexandria, VA: American Statistical Association.

Mauldin, R. D., Sudderth, W. D. & Williams, S. C. (1992). Polya trees and random distributions. *Ann. Statist.* **20**, 1203–21.

Mittal, Y. (1976). A class of isotropic covariance functions. *Pac. J. Math* **64**, 517–38.

Nadarajah, S. & Kotz, S. (2005) Some bivariate beta distributions. *Statistics* **39**, 457–66.

Nieto-Barajas, L. E. & Walker, S. G. (2007). Gibbs and autoregressive processes. *Statist. Prob. Lett.* **77**, 1479–85.

Olkin, I. & Liu, R. (2003). A bivariate beta distribution *Statist. Prob. Lett.* **62**, 407–12.

Piantadosi, S. (1997). *Clinical Trials: A Methodologic Perspective*, New York: Wiley.

Raftery, A. E. & Lewis, S. M. (1992). How many iterations in the Gibbs sampler? In *Bayesian Statist. 4*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 763–73. Oxford: Oxford University Press.

Walker, S. & Mallick, B. K. (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics* **55**, 477–83.

Walker, S. & Muliere, P. (2003). A bivariate Dirichlet process. *Statist. Prob. Lett.* **64**, 1–7.