Biometrika (2008), **95**, 2, *pp*. 307–323
doi: 10.1093/biomet/asn012
Advance Access publication 30 April 2008

Kernel stick-breaking processes

BY DAVID B. DUNSON

Biostatistics Branch, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, North Carolina 27709, U.S.A.
dunson1@niehs.nih.gov

AND JU-HYUN PARK

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A.
parkj3@niehs.nih.gov

SUMMARY

We propose a class of kernel stick-breaking processes for uncountable collections of dependent random probability measures. The process is constructed by first introducing an infinite sequence of random locations. Independent random probability measures and beta-distributed random weights are assigned to each location. Predictor-dependent random probability measures are then constructed by mixing over the locations, with stick-breaking probabilities expressed as a kernel multiplied by the beta weights. Some theoretical properties of the process are described, including a covariate-dependent prediction rule. A retrospective Markov chain Monte Carlo algorithm is developed for posterior computation, and the methods are illustrated using a simulated example and an epidemiological application.

Some key words: Conditional density estimation; Dependent Dirichlet process; Kernel methods; Nonparametric Bayes; Mixture model; Prediction rule; Random partition.

1. INTRODUCTION

This article focuses on the problem of choosing priors for an uncountable collection of random probability measures, $\mathcal{G}_{\mathcal{X}} = \{G_x : x \in \mathcal{X}\}$, where $\mathcal{X}$ is a Lesbesgue measurable subset of $\Re^p$ and G_x is a probability measure over a measurable Polish space $(\Omega, \mathcal{B})$, with Ω the sample space and $\mathcal{B}$ the corresponding Borel σ-algebra. A motivating application is the problem of estimating the conditional density of a response variable using the mixture specification $f(y \mid x) = \int f(y \mid x, \phi) dG_x(\phi)$, where $f(y \mid x, \phi)$ is a known kernel and G_x an unknown probability measure indexed by the predictor value, $x = (x_1, \ldots, x_p)'$.

The problem of defining priors for dependent random probability measures has received increasing attention in recent years. Most approaches focus on generalizations of the Ferguson (1973, 1974) Dirichlet process prior, with methods varying in how they incorporate dependence. One approach is to include a regression in the base measure (Cifarelli & Regazzini, 1978), which has the disadvantage of capturing dependence only in aspects of the distribution characterized by the base parametric model.

Much of the recent work has instead relied on generalizations of Sethuraman's (1994) stick-breaking representation of the Dirichlet process. If G is assigned a Dirichlet process prior

with precision α and base measure G_0, denoted by $G \sim \text{DP}(\alpha G_0)$, then the stick-breaking representation of G is

$$G = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}, \quad p_h = V_h \prod_{l=1}^{h-1}(1 - V_l), \quad V_h \sim \text{Be}(1, \alpha), \quad \theta_h \sim G_0, \tag{1}$$

where δ_θ is a probability measure concentrated at θ and all V_h's and θ_h's are independent. MacEachern (1999, 2001) proposed the dependent Dirichlet process, which generalizes (1) to allow a collection of unknown distributions indexed by x by allowing the weights $p = (p_h, h = 1, \ldots, \infty)$ and atoms $\theta = (\theta_h, h = 1, \ldots, \infty)$ to vary with x according to a stochastic process. If we assume fixed p, the dependent Dirichlet process has been successfully applied to the analysis of variance (De Iorio et al., 2004), spatial modelling (Gelfand et al., 2005) and time series (Caron et al., 2006).

Noting limited flexibility due to the fixed-p assumption, Griffin & Steel (2006) proposed an order-based dependent Dirichlet process, which incorporates dependence by allowing the ordering of the random variables $\{V_h, h = 1, \ldots, \infty\}$ in the stick-breaking construction to depend on predictors. An alternative is to incorporate dependence through weighted mixtures of independent Dirichlet processes. Müller et al. (2004) used this idea to allow dependence across experiments, while Dunson (2006) and Pennell & Dunson (2006) considered discrete dynamic settings. Dunson et al. (2007) defined a prior for $\mathcal{G}_\mathcal{X}$ through a weighted mixture of independent Dirichlet process random probability measures introduced at the sampled predictor values. This prior is sample-dependent and lacks reasonable marginalization and updating properties.

In developing a prior for $\mathcal{G}_\mathcal{X}$, we would also like to generalize the Dirichlet process prediction rule, commonly referred to as the Blackwell & MacQueen (1973) Pólya urn scheme, to incorporate predictors. Assuming $\phi_i \sim G$, with $G \sim \text{DP}(\alpha G_0)$, one obtains the Dirichlet process prediction rule upon marginalizing over the prior for G:

$$\text{pr}(\phi_1 \in \cdot) = G_0(\cdot), \quad \text{pr}(\phi_i \in \cdot \mid \phi_1, \ldots, \phi_{i-1}) = \left(\frac{\alpha}{\alpha + i - 1}\right) G_0(\cdot) + \sum_{j=1}^{i-1} \left(\frac{1}{\alpha + i - 1}\right) \delta_{\phi_j}(\cdot). \tag{2}$$

This prediction rule forms the basis for commonly used algorithms for efficient posterior computation in Dirichlet process mixture models (MacEachern, 1994).

The Dirichlet process prediction rule induces clustering of the subjects according to a Chinese restaurant process (Aldous, 1985; Pitman, 1996). This clustering behaviour is often exploited as a dimension-reduction device and a tool for exploring latent structure (Dunson et al., 2007; Kim et al., 2006; Medvedovic et al., 2004). The Dirichlet process and related approaches, including product partition models (Barry & Hartigan, 1992; Quintana & Iglesias, 2003) and species sampling models (Pitman, 1996; Ishwaran & James, 2003), assume exchangeability. In many applications, it is appealing to relax the exchangeability assumption to allow predictor-dependent clustering.

Motivated by these issues, this article proposes a class of kernel stick-breaking processes to be used as a sample-free prior for $\mathcal{G}_\mathcal{X}$, which induces a covariate-dependent prediction rule upon marginalization.

2. Predictor-dependent random probability measures

2·1. *Proposed formulation*

Let $\mathcal{G}_{\mathcal{X}} \sim \mathcal{P}$, with $\mathcal{P}$ a probability measure on $(\Psi, \mathcal{C})$, where Ψ is the space of uncountable collections of probability measures on the Polish space $(\Omega, \mathcal{B})$ indexed by $x \in \mathcal{X}$ and $\mathcal{C}$ is a corresponding σ-algebra. Our focus is on choosing $\mathcal{P}$.

We first introduce a countable sequence of mutually independent random components,

$$\{\Gamma_h, V_h, G_h^*, h = 1, \ldots, \infty\},$$

where, for each h, independently, $\Gamma_h \sim H$ is a location, $V_h \sim \text{Be}(a_h, b_h)$ is a probability weight, and $G_h^* \sim \mathcal{Q}$ is a probability measure. Here, H is a probability measure on the Polish space $(\mathcal{L}, \mathcal{A})$, where $\mathcal{A}$ is a Borel σ-algebra of subsets of $\mathcal{L}$, and $\mathcal{L}$ is a Lebesgue-measurable subset of $\Re^p$ that may or may not correspond to $\mathcal{X}$. In addition, $\mathcal{Q}$ is a probability measure on the space of probability measures on $(\Omega, \mathcal{B})$. For example, $\mathcal{Q}$ may correspond to a Dirac measure at a random location, a Dirichlet process or a species sampling model (Pitman, 1996).

The kernel stick-breaking process is defined as follows:

$$\begin{aligned} G_x &= \sum_{h=1}^{\infty} U(x; V_h, \Gamma_h) \prod_{l<h} \{1 - U(x; V_l, \Gamma_l)\} G_h^*, \\ U(x; V_h, \Gamma_h) &= V_h K(x, \Gamma_h), \quad \text{for all} \quad x \in \mathcal{X}, \end{aligned} \tag{3}$$

where $K : \Re^p \times \Re^p \to [0, 1]$ is a bounded kernel function. Note that (3) formulates G_x as a predictor-dependent mixture over an infinite sequence of basis probability measures, with G_h^* located at Γ_h, for $h = 1, \ldots, \infty$. Bases located close to x and having a smaller index, h, will tend to receive higher probability weight.

Starting with the unit probability stick, G_x allocates probability $V_1 K(x, \Gamma_1)$ to the first basis measure, G_1^*. This probability ranges from 0, for x far from Γ_1, to V_1, for x close to Γ_1, depending on the choice of kernel K. A proportion $V_2 K(x, \Gamma_2)$ is then broken off from the remaining stick of length $1 - V_1 K(x, \Gamma_1)$ and allocated to the second basis measure, G_2^*. This proportion ranges from 0 to V_2 depending on the distance from x to Γ_2. This predictor-dependent stick-breaking process continues infinitely many times, using up the unit sticks at all locations $x \in \mathcal{X}$ in the limit. Note that G_x and $G_{x'}$ will allocate similar probabilities to the elements of the basis set $\{G_h^*\}_{h=1}^{\infty}$ if x is close to x'. In this manner, the kernel stick-breaking process accommodates dependence.

Let $\pi_h(x; V_h, \Gamma_h) = U(x; V_h, \Gamma_h) \prod_{l<h} \{1 - U(x; V_l, \Gamma_l)\}$, for $h = 1, \ldots, \infty$, with $V_h = (V_1, \ldots, V_h)'$ and $\Gamma_h = (\Gamma_1, \ldots, \Gamma_h)'$. By replicating the arguments in Lemma 1 in Ishwaran & James (2001), one can show that G_x is well defined, with $\sum_{h=1}^{\infty} \pi_h(x; V_h, \Gamma_h) = 1$ almost surely for all $x \in \mathcal{X}$.

In order for the kernel stick-breaking process to be useful in applications with limited data in any particular small local region of $\mathcal{X}$, it is important to favour sparseness and borrowing of information. Sparseness can be represented in the process by choosing (i) a sequence $\{a_h, b_h\}_{h=1}^{\infty}$ that favours values of $\{V_h\}_{h=1}^{\infty}$ close to one, (ii) a kernel K that decreases to 0 slowly with increasing separation between predictors, and (iii) a random measure $\mathcal{Q}$ that tends to assign high mass to few atoms. Property (i) leads to dominant basis locations that are assigned high probability by all G_x in a local region, property (ii) allows borrowing of information broadly across $\mathcal{X}$, and property (iii) ensures that few parameters are needed to characterize each basis measure.

A number of interesting special cases arise when $K(x, \Gamma) = 1$ for all $(x, \Gamma) \in \mathcal{X} \otimes \mathcal{L}$, so that $G_x \equiv G = \sum_{h=1}^{\infty} V_h \prod_{l<h} (1 - V_l) G_h^*$. Then if $G_h^* \sim \text{DP}(\alpha G_0)$, independently for each h, we obtain a stick-breaking mixture of Dirichlet processes as a prior for G. Under the additional

conditions $a_h = 1$ and $b_h = \lambda$, the prior for G is a Dirichlet process mixture of Dirichlet processes, which is a two-parameter extension of the Dirichlet process. This mixture reduces to a Dirichlet process in the limiting case as either λ or $\alpha \to 0$. Finally, when $G_h^* = \delta_{\theta_h}$, $\theta_h \sim G_0$, independently for $h = 1, \ldots, \infty$, without any constraints on $\{a_h, b_h\}_{h=1}^{\infty}$, G is assigned a stick-breaking prior in the class considered by Ishwaran & James (2001).

2·2. *Conditional properties*

Returning to the general case, we first derive moments of G_x conditionally on the random weights V and random locations Γ, but marginalizing out the random basis measures, $\{G_h^*, h = 1, \ldots, \infty\}$. Letting $G_0(B) = E_{\mathcal{Q}}\{G_h^*(B)\}$, for all $B \in \mathcal{B}$, we obtain

$$E\{G_x(B) \mid V, \Gamma\} = \sum_{h=1}^{\infty} \pi_h(x; V_h, \Gamma_h) E_{\mathcal{Q}}\{G_h^*(B)\} = G_0(B), \quad \text{for all } B \in \mathcal{B}. \tag{4}$$

As a result of the lack of dependence on V and Γ, we also have $E_{\mathcal{P}}\{G_x(B)\} = G_0(B)$, so that the prior is centred on the base measure G_0. In addition,

$$\begin{aligned} E\{G_x(B)^2 \mid V, \Gamma\} &= \sum_{h=1}^{\infty} \pi_h(x; V_h, \Gamma_h)^2 E_{\mathcal{Q}}\{G_h^*(B)^2\} \\ &\quad + \sum_{h=1}^{\infty} \sum_{l \neq h} \pi_h(x; V_h, \Gamma_h)\pi_l(x; V_l, \Gamma_l) E_{\mathcal{Q}}\{G_h^*(B)\} E_{\mathcal{Q}}\{G_l^*(B)\} \\ &= \left(\sum_{h=1}^{\infty} \pi_h(x; V_h, \Gamma_h)^2 [E_{\mathcal{Q}}\{G_h^*(B)^2\} - G_0(B)^2] \right) + G_0(B)^2 \\ &= ||\pi(x; V, \Gamma)||^2 \mathrm{V}_{\mathcal{Q}(B)} + G_0(B)^2, \end{aligned} \tag{5}$$

where $\mathrm{V}_{\mathcal{Q}(B)} = \mathrm{V}_{\mathcal{Q}}\{G_h^*(B)\}$ and $\mathrm{var}\{G_x(B) \mid V, \Gamma\} = ||\pi(x; V, \Gamma)||^2 \mathrm{V}_{\mathcal{Q}(B)}$. By a similar route, the correlation coefficient is

$$\begin{aligned} \mathrm{corr}\{G_x(B), G_{x'}(B) \mid V, \Gamma\} &= \frac{\sum_{h=1}^{\infty} \pi_h(x; V_h, \Gamma_h)\pi_h(x'; V_h, \Gamma_h)}{\{\sum_{h=1}^{\infty} \pi_h(x; V_h, \Gamma_h)^2\}^{1/2} \{\sum_{h=1}^{\infty} \pi_h(x'; V_h, \Gamma_h)^2\}^{1/2}} \\ &= \frac{\langle \pi(x; V, \Gamma), \pi(x'; V, \Gamma) \rangle}{||\pi(x; V, \Gamma)||\, ||\pi(x'; V, \Gamma)||} = \rho(x, x'; V, \Gamma), \end{aligned} \tag{6}$$

where $\langle \cdot, \cdot \rangle$ is the inner product. The correlation coefficient $\rho(x, x'; V, \Gamma) \leqslant 1$, with the limiting value of 1 as $x \to x'$, if we assume that $\lim_{x \to x'} K(x, \Gamma) = K(x', \Gamma)$, for all $\Gamma \in \mathcal{L}$. This expression is quite intuitive, being a simple normed cross-product of the weight functions. An appealing property is that the correlation coefficient is free from the set B, so that a single quantity can be reported for each (x, x') pair. Interestingly, the correlation coefficient does not depend on the choice of $\mathcal{Q}$, the probability measure generating the bases at each of the locations.

2·3. *Marginal properties*

To obtain additional insight into the properties of the kernel stick-breaking process, it is interesting to marginalize out the random weights, V, and random locations, Γ. Let $K_h(x) \sim F_x$ denote the random variable obtained in the transformation from $\Gamma_h \sim H$ to $K(x, \Gamma_h)$. Since the random locations are independent and identically distributed, we have $K_h(x) \sim F_x$, independently for $h = 1, \ldots, \infty$. In addition, the random variables $K_h(x)$ and $K_h(x')$ are dependent, while $K_h(x)$ and $K_l(x')$ are independent, for $h \neq l$.

Let $U_h(x) = V_h K_h(x)$ and $\pi_h(x) = U_h(x)\prod_{l<h}\{1 - U_l(x)\}$, for $h = 1, \ldots, \infty$. Dependence in the random weights, $\pi(x) = \{\pi_h(x), h = 1, \ldots, \infty\}$ and $\pi(x') = \{\pi_h(x'), h = 1, \ldots, \infty\}$, arises through dependence between the components $U_h(x)$ and $U_h(x')$, for $h = 1, \ldots, \infty$.

THEOREM 1. *Let $\mu(x) = E\{U_h(x)\}$ and $\mu(x, x') = E\{U_h(x)U_h(x')\}$. Then, if we assume that $V_h \sim \mathrm{Be}(a, b)$, independently for each h, for any $B \in \mathcal{B}$, we have*

$$E\{G_x(B)G_{x'}(B)\} = \frac{\mu(x, x')V_{\mathcal{Q}(B)}}{\mu(x) + \mu(x') - \mu(x, x')} + G_0(B)^2.$$

The derivation is in the Appendix. From this expression, it is straightforward to show that

$$\mathrm{V}\{G_x(B)\} = \frac{\mu^{(2)}(x)\mathrm{V}_{\mathcal{Q}(B)}}{2\mu(x) - \mu^{(2)}(x)}, \tag{7}$$

where $\mu^{(2)}(x) = \mu(x, x)$. In addition, the correlation coefficient has the simple form

$$\mathrm{corr}\{G_x(B), G_{x'}(B)\} = \frac{\mu(x, x')}{\mu(x) + \mu(x') - \mu(x, x')} \times \left[\frac{\{2\mu(x) - \mu^{(2)}(x)\}\{2\mu(x') - \mu^{(2)}(x')\}}{\mu^{(2)}(x)\mu^{(2)}(x')}\right]^{1/2}. \tag{8}$$

Note that this expression is free of B and only depends on the expectations of $U_h(x)$ and $U_h(x)U_h(x')$.

If $V_h \sim \mathrm{Be}(1, \lambda)$, independently for each h, we obtain the modified expression

$$\mathrm{corr}\{G_x(B), G_{x'}(B)\} = \frac{\kappa(x, x')\{(2 + \lambda)\frac{\kappa(x)}{\kappa_2(x)} - 1\}^{1/2}\{(2 + \lambda)\frac{\kappa(x')}{\kappa_2(x')} - 1\}^{1/2}}{(1 + \lambda/2)\{\kappa(x) + \kappa(x')\} - \kappa(x, x')}, \tag{9}$$

where $\kappa(x) = E\{K_h(x)\}$, $\kappa_2(x) = E\{K_h(x)^2\}$ and $\kappa(x, x') = E\{K_h(x)K_h(x')\}$. This expression is useful in considering the correlation structure induced for different choices of H and K, as well as the impact of the hyperparameter λ. For example, note that, when the first two moments of $U_h(x)$ are free from x, expression (9) reduces to

$$\mathrm{corr}\{G_x(B), G_{x'}(B)\} = \left\{\frac{(2 + \lambda)\kappa}{\kappa_2} - 1\right\} \Big/ \left\{\frac{(2 + \lambda)\kappa}{\kappa(x, x')} - 1\right\}, \tag{10}$$

with the dependence on x in κ and κ_2 dropped. For some special cases of H and K, the moments of $U_h(x)$ and $U_h(x)U_h(x')$ can be calculated in closed form, so that the above expressions are also available in closed form.

2·4. *Example* 1

Suppose that $V_h \sim \mathrm{Be}(1, \lambda)$, independently for all h, $K(x, \Gamma_h) = 1(|x_j - \Gamma_{hj}| < \psi_j, j = 1, \ldots, p)$ is a rectangular kernel, with $\psi_j > 0$ for $j = 1, \ldots, p$, and H corresponds to a uniform probability measure on $\mathcal{L}$. Focusing on the unit hypercube, $\mathcal{X} = [0, 1]^p$, and letting $\mathcal{L} = \bigotimes_{j=1}^{p}[-\psi_j, 1 + \psi_j]$, for any $x \in \mathcal{X}$, we have

$$E\{U_h(x)^m\} = \left(\prod_{l=1}^{m}\frac{l}{\lambda + l}\right)\prod_{j=1}^{p}\left(\frac{2\psi_j}{1 + 2\psi_j}\right),$$

$$E\{U_h(x)U_h(x')\} = \left(\frac{1}{1 + \lambda}\right)\left(\frac{2}{2 + \lambda}\right)\prod_{j=1}^{p}\left(\frac{\Delta_j(x_j, x'_j)}{1 + 2\psi_j}\right), \tag{11}$$

where $\Delta_j(x_j, x'_j) = \max\{0, \min(x_j + \psi_j, x'_j + \psi_j) - \max(x_j - \psi_j, x'_j - \psi_j)\}$.

From expression (11), it is apparent that the moments of $U_h(x)$ are free from x, while the expectation of $U_h(x)U_h(x')$ depends only on the distance between x and x'. Calculating the variance, we obtain the simple expression

$$\text{var}\{G_x(B)\} = \frac{V_{\mathcal{Q}(B)}}{1+\lambda}. \tag{12}$$

In addition, the correlation coefficient takes the form

$$\rho(x-x';\lambda,\psi) = \text{corr}\{G_x(B), G_{x'}(B)\} = \frac{1+\lambda}{(2+\lambda)\prod_{j=1}^{p}\{2\psi_j/\Delta_j(x_j,x'_j)\}-1}, \tag{13}$$

which is a function of the distance between x and x'. When $(x-x') \notin C_\psi = \bigotimes_{j=1}^{p}[-2\psi_j, 2\psi_j]$, $\rho(x-x';\lambda,\psi) = 0$. In addition, in the limit as $x \to x'$, $\Delta_j(x_j,x'_j) \to 2\psi_j$ and $\rho(x-x';\lambda,\psi) \to 1$. Hence, the correlation coefficient is bounded between 0 and 1, depending on the distance between the predictor values. The results for Example 1 are easily generalizable to arbitrary bounded predictor spaces and to Gaussian kernels.

3. CLUSTERING AND PREDICTION RULES

As mentioned in § 1, one of the most appealing and widely used properties of the Dirichlet process is the simple prediction rule shown in expression (2). In this section, we obtain a predictor-dependent prediction rule derived by marginalizing over the kernel stick-breaking process prior for $\mathcal{G}_\mathcal{X}$ shown in expression (3) with three additional assumptions: (i) $G_h^* = \delta_{\Theta_h}$, with $\Theta_h \sim G_0$, independently for $h = 1,\ldots,\infty$; (ii) G_0 is nonatomic; and (iii) $V_h \sim \text{Be}(1,\lambda)$, independently for $h = 1,\ldots,\infty$. Assumption (i) implies that there is a single atom Θ_h, located at Γ_h, so that all subjects allocated to a given location will belong to the same cluster.

Consider the following hierarchical model:

$$\begin{aligned}(\phi_i \mid x_i) &\sim G_{x_i}, \quad \text{independently for } i = 1,\ldots,n,\\ \mathcal{G}_\mathcal{X} &\sim \mathcal{P},\end{aligned} \tag{14}$$

where $\mathcal{G}_\mathcal{X} = \{G_x : x \in \mathcal{X}\}$, $\mathcal{P}$ is a kernel stick-breaking process characterized in terms of a precision parameter, λ, a kernel, K, and a base measure, G_0. Note that (15) can be expressed equivalently as

$$\begin{aligned}(\phi_i \mid Z_i, x_i, \Theta) &\sim \delta_{\Theta_{Z_i}},\\ (Z_i \mid x_i, V, \Gamma) &\sim \sum_{h=1}^{\infty}\pi_h(x_i; V_h, \Gamma_h)\delta_h, \quad \text{independently for } i = 1,\ldots,n,\\ V_h \sim \text{Be}(1,\lambda), \quad \Gamma_h &\sim H, \quad \Theta_h \sim G_0, \quad \text{independently for each } h,\end{aligned} \tag{15}$$

where Z_i indexes the unobserved location for subject i. It follows that $\text{pr}(\phi_i \in \cdot \mid x_i) = G_0(\cdot)$. As a notation to aid the description of marginal properties, we let

$$\mu_\mathcal{I} = E\left\{\prod_{i\in\mathcal{I}} U_h(x_i)\right\}, \tag{16}$$

where $\mathcal{I} \subset \{1,\ldots,n\}$ is a subset of the integers between 1 and n. In some important special cases, including rectangular and Gaussian kernels, these moments can be calculated in closed form, if we use a straightforward generalization of results shown for Example 1.

LEMMA 1. *Under the prior structure* (15)*, the probability that subjects i and j belong to the same cluster conditionally on the subjects' predictor values, but marginalizing out* $\mathcal{P}$*, is*

$$\text{pr}(\phi_i = \phi_j \mid x_i, x_j) = \frac{\mu_{ij}}{\mu_i + \mu_j - \mu_{ij}}, \quad \textit{for all } i, j \in \{1, \ldots, n\},$$

with μ_i, μ_j, μ_{ij} *defined in* (16)*.*

Under the conditions of Example 1, the expression in Lemma 1 takes the form

$$\text{pr}(\phi_i = \phi_j \mid x_i, x_j) = \frac{\prod_{j=1}^{p} \Delta_j(x_j, x_j')}{(2+\lambda)\prod_{j=1}^{p}(2\psi_j) - \prod_{j=1}^{p} \Delta_j(x_j, x_j')}, \tag{17}$$

which reduces to 0 if $x_i - x_j \notin C_\psi = \bigotimes_{j=1}^{p}[-2\psi_j, 2\psi_j]$, as x_i and x_j are not in the same neighbourhood in that case. In addition, as $x_i \to x_j$, $\text{pr}(\phi_i = \phi_j \mid x_i, x_j) \to 1/(1+\lambda)$, which corresponds to the clustering probability for the Dirichlet process prediction rule when $G_{x_i} = G \sim \text{DP}(\lambda G_0)$.

THEOREM 2. *Let* $\mathcal{N}_i^{(r,s)}$ *denote the set of possible r-dimensional subsets of* $\{1, \ldots, s\}$ *that include i, let* $\mathcal{N}_{i,j}^{(r,s)}$ *denote the set of possible r-dimensional subsets of* $\{1, \ldots, s\}$ *including i and j, and let*

$$\omega_{\mathcal{I}} = \frac{\mu_{\mathcal{I}}}{\sum_{t=1}^{\#\mathcal{I}}(-1)^{t-1}\sum_{\mathcal{J}\in\mathcal{I}_t}\mu_{\mathcal{J}}},$$

where $\#\mathcal{I}$ *is the cardinality of set* $\mathcal{I}$ *and* $\mathcal{I}_t$ *is the set of length-t subsets of* $\mathcal{I}$*. Then, under expression* (15)*, the following prediction rule is obtained on marginalizing out* $\mathcal{P}$*:*

$$\begin{aligned}&\text{pr}(\phi_i \in \cdot \mid \phi_1, \ldots, \phi_{i-1}, x_1, \ldots, x_{i-1})\\ &\quad= \left\{1 - \sum_{r=2}^{i}(-1)^r \sum_{\mathcal{I}\in\mathcal{N}_i^{(r,i)}} \omega_{\mathcal{I}}\right\} G_0(\cdot) + \sum_{j=1}^{i-1}\left\{\sum_{r=2}^{i}(-1)^r \sum_{\mathcal{I}\in\mathcal{N}_{i,j}^{(r,i)}} \frac{\omega_{\mathcal{I}}}{r-1}\right\}\delta_{\phi_j}(\cdot).\end{aligned}$$

Under the conditions of Example 1, we obtain the following simple expression for $\mu_{\mathcal{I}}$:

$$\begin{aligned}\mu_{\mathcal{I}} &= E\left\{\prod_{i\in\mathcal{I}} U_h(x_i)\right\} = E\left(V_h^{\#\mathcal{I}}\right)\int \prod_{i\in\mathcal{I}}\prod_{j=1}^{p} 1(|x_{ij} - \Gamma_{hj}| < \psi_j) dH(\Gamma_h)\\ &= \left\{\prod_{l=1}^{\#\mathcal{I}} \frac{l}{l+\lambda}\right\}\left\{\prod_{j=1}^{p}\frac{\Delta_j(x_{\mathcal{I}})}{1+2\psi_j}\right\},\end{aligned} \tag{18}$$

where $x = (x_1, \ldots, x_n)'$ is an $n \times p$ matrix and $\Delta_j(x_{\mathcal{I}}) = \max[0, 2\psi_j + \min_{i\in\mathcal{I}}\{x_{ij}\} - \max_{i\in\mathcal{I}}\{x_{ij}\}]$. From this result, one can show that the prediction rule from Theorem 2 reduces to the Dirichlet process prediction rule in the special case in which $x_j = x$, for $j = 1, \ldots, i$.

4. POSTERIOR COMPUTATION

4·1. *Background and overview*

For Dirichlet process mixture models, two main strategies have been used in developing algorithms for posterior computation, namely the marginal approach and the conditional approach. If $\phi_i \sim G$, with $G \sim \text{DP}(\alpha G_0)$, the marginal approach avoids computation for the infinite-dimensional G by relying on the Pólya urn scheme, which is obtained if we marginalize over the

Dirichlet process prior. The most widely used marginal algorithm is the generalized Pólya urn Gibbs sampler of MacEachern (1994) and West et al. (1994). Ishwaran & James (2001) extend this approach to a general class of stick-breaking measures.

The conditional approach avoids marginalizing over the prior, resulting in greater flexibility in computation and inference (Ishwaran & Zarepour, 2000). To avoid the need for a truncation approximation, Papaspiliopoulos & Roberts (2008) recently proposed a retrospective Markov chain Monte Carlo algorithm. We propose a conditional approach to posterior computation for kernel stick-breaking process models, relying on a combined Markov chain Monte Carlo algorithm that uses retrospective sampling and generalized Pólya urn sampling steps.

We focus on the kernel stick-breaking process model of expression (3), with $V_h \sim \text{Be}(a_h, b_h)$, independently for each h, K and H having arbitrary forms, and $\mathcal{Q}$ corresponding to a stick-breaking prior. Although the algorithm can in principle deal with any model in this class, some models are more tractable than others. In particular, as for stick-breaking priors and Dirichlet process mixture models, simplifications result when the base G_0 is conjugate to the likelihood.

4·2. *Details*

Let $\theta = (\theta_1, \ldots, \theta_k)'$ denote the $k \leqslant n$ unique values of $\phi = (\phi_1, \ldots, \phi_n)'$, let $\mathcal{S}_i = h$ if $\phi_i = \theta_h$ denotes that subject i is allocated to the hth unique value, with $S = (\mathcal{S}_1, \ldots, \mathcal{S}_n)'$, and let $\mathcal{C}_h = j$ denote that θ_h is an atom from G_j^*, with $C = (\mathcal{C}_1, \ldots, \mathcal{C}_k)'$. Let $\phi^{(i)}, \theta^{(i)}$, $S^{(i)}$, $C^{(i)}$ and $Z^{(i)}$ correspond to the vectors ϕ, θ, S, C and Z that would have been obtained without subject i's contribution. The number of subjects allocated to the jth location is $n_j = \sum_{i=1}^{n} 1(Z_i = j)$, with $\sum_{j=1}^{\infty} n_j = n$. The index set for locations, $\mathcal{I} = \{1, 2, \ldots, \infty\}$, consists of two mutually exclusive subsets, namely occupied locations, $\mathcal{I}_{\text{oc}} = \{j \in \mathcal{I} : n_j > 0\}$, and vacant locations, $\mathcal{I}_{\text{vc}} = \{j \in \mathcal{I} : n_j = 0\}$, so that $\mathcal{C}_h \in \mathcal{I}_{\text{oc}}$, for $h = 1, \ldots, k$.

If $\mathcal{N}_h = \{i : Z_i = h, i = 1, 2, \ldots, \infty\}$ denotes the subset of the positive integers indexing subjects allocated to location h, $\{\phi_j, j \in \mathcal{N}_h\}$ is a species sampling sequence (Pitman, 1996). Hence, it follows from Pitman (1996) and Ishwaran & James (2003) that

$$\text{pr}(\phi_i \in \cdot \mid Z_i = h, S^{(i)}, C^{(i)}, \theta^{(i)}, x) = l_{ih0} G_0(\cdot) + \sum_{j \in \mathcal{N}_h^{(i)}} l_{ihj} \delta_{\phi_j}(\cdot), \tag{19}$$

where $\mathcal{N}_h^{(i)} = \mathcal{N}_h \bigcap \{1, \ldots, n\} \backslash \{i\}$ and $\{l_{ihj}\}$ are the probability weights implied by the species sampling prediction rule. For example, in the Dirichlet process special case, we have $l_{ih0} = \alpha/(\alpha + \#\mathcal{N}_h^{(i)})$ and $l_{ihj} = 1/(\alpha + \#\mathcal{N}_h^{(i)})$. We obtain the following from (19) by marginalizing out Z_i, noting that $\text{pr}(Z_i = h \mid x_i, V, \Gamma) = \pi_h(x_i; V_h, \Gamma_h) = \pi_{ih}$ for $h = 1, \ldots, \infty$, and grouping together the subjects with the same unique value:

$$\text{pr}(\phi_i \in \cdot \mid S^{(i)}, C^{(i)}, \theta^{(i)}, x) = w_{i0} G_0(\cdot) + \sum_{j=1}^{k^{(i)}} w_{ij} \delta_{\theta_j^{(i)}}(\cdot) + w_{i,k^{(i)}+1} G_0(\cdot), \tag{20}$$

with $k^{(i)}$ the length of $\theta^{(i)}$ and the weights defined by

$$w_{i0} = \sum_{h \in \mathcal{I}_{oc}^{(i)}} \pi_{ih} \, l_{ih0}, \; w_{ij} = \pi_{i,\mathcal{C}_j^{(i)}} \sum_{g : \mathcal{S}_g^{(i)} = j} l_{i\mathcal{C}_j^{(i)} g}, \; j = 1, \ldots, k^{(i)}, \; w_{i,k^{(i)}+1} = \sum_{h \in \mathcal{I}_{vc}^{(i)}} \pi_{ih} \, l_{ih0}. \tag{21}$$

If the likelihood contribution for subject i is $f(y_i \mid x_i, \phi_i)$, expression (21) can be updated to obtain a conditional posterior distribution for ϕ_i. From this posterior, we obtain

$$\text{pr}(\mathcal{S}_i = j \mid y, S^{(i)}, C^{(i)}, \theta^{(i)}, x) = q_{ij}, \tag{22}$$

where $q_{ij} = c_i\, w_{ij}\, f_0(y_i \mid x_i)$, for $j = 0, k^{(i)} + 1$, and $q_{ij} = c_i\, w_{ij}\, f(y_i \mid x_i, \theta_j^{(i)})$, for $j = 1, \ldots, k^{(i)}$, with $f_0(y_i \mid x_i) = \int f(y_i \mid x_i, \phi)\, dG_0(\phi)$ and c_i a normalizing constant. We update $\mathcal{S}_i$ by sampling based on (22). Sampling $\mathcal{S}_i = 0$ corresponds to assigning subject i to a new atom at an occupied location, with $\mathcal{C}_{\mathcal{S}_i} \sim \sum_{h \in \mathcal{I}_{oc}^{(i)}} \pi_{ih}^* \delta_h$, where $\pi_{ih}^* = \pi_{ih} / \sum_{l \in \mathcal{I}_{oc}^{(i)}} \pi_{il}$. When $\mathcal{S}_i = k^{(i)} + 1$, subject i is assigned to an atom at a new location. Since there are infinitely many possibilities for this new location, we use a retrospective sampling approach, which follows along similar lines to Papaspiliopoulos & Roberts (2008).

After updating S and C, we update θ_h, for $h = 1, \ldots, k$, from

$$(\theta_h \mid y, S, C, k, x) \propto \left\{ \prod_{i:S_i = h} f(y_i | x_i, \theta_h) \right\} G_0(\theta_h). \tag{23}$$

Let $M^{(t)}$ correspond to the maximum element of $\mathcal{I}_{oc}$ across the first t iterations of the sampler. To update V_h, for $h = 1, \ldots, M^{(t)}$, we use a data augmentation approach. Let $A_{ih} \sim \text{Ber}(V_h)$ and $B_{ih} \sim \text{Ber}\{K(x_i, \Gamma_h)\}$, independently for each h, with $Z_i = \mathcal{C}_{\mathcal{S}_i} = \min\{h : A_{ih} = B_{ih} = 1\}$. Then, alternate between sampling (A_{ih}, B_{ih}) from their conditional distribution given Z_i and updating V_h by sampling from the conditional posterior distribution

$$\text{Be}\left(a_h + \sum_{i:Z_i \geqslant h} A_{ih}, b_h + \sum_{i:Z_i \geqslant h} (1 - A_{ih}) \right).$$

Updating of Γ_h, for $h = 1, \ldots, M^{(t)}$, can proceed by a Metropolis–Hastings step or a Gibbs step if $H(\cdot) = \sum_{l=1}^{T} a_l \delta_{\Gamma_l^*}(\cdot)$, with $\Gamma^* = (\Gamma_1^*, \ldots, \Gamma_T^*)'$ a grid of potential locations.

5. Simulation example

In this section, we illustrate the proposed method in a mixture of normal linear regression models for conditional density estimation, assessing sensitivity to the kernel and hyperparameters in the kernel stick-breaking process. Let $f(y_i \mid x_i, \phi_i) = (2\pi\tau^{-1})^{-1/2} \exp\{-\tau(y_i - x_i'\beta_i)^2/2\}$, with $\phi_i = \beta_i \sim G_{x_i}$ and $\mathcal{G}_{\mathcal{X}} \sim \mathcal{P}$, in which $\mathcal{P}$ is a kernel stick-breaking process chosen so that $a_h = 1$, $b_h = \lambda$, $\mathcal{Q}$ is a DP(αG_0) random measure, and G_0 follows a Gaussian law with mean β and variance matrix Σ_β. In addition, we choose priors $\pi(\tau) = \text{Ga}(\tau; a_\tau, b_\tau)$, $\pi(\beta) = N(\beta; \beta_0, V_{\beta_0})$, and $\pi(\Sigma_\beta^{-1}) = \mathcal{W}\{\Sigma_\beta^{-1}; (\nu_0 \Sigma_0)^{-1}, \nu_0\}$, the Wishart density with ν_0 degrees of freedom and $E(\Sigma_\beta^{-1}) = \Sigma_0^{-1}$. We let $\beta_0 = 0$, $V_{\beta_0} = (x'x)^{-1}/n$, $\nu_0 = p$, $\Sigma_0^{-1} = I_{p \times p}$, and $a_\tau = b_\tau = 0{\cdot}1$, and choose every point from 0 to 1 with increment of $0{\cdot}02$ as Γ^*, with $T = 51$ and probability weight $a_l = 1/T$.

Following Dunson et al. (2007), we simulate data for $n = 500$ subjects from a mixture of two normal linear regression models as follows:

$$f(y_i | x_i) = e^{-2x_i} N(y_i; x_i, 0{\cdot}01) + (1 - e^{-2x_i}) N(y_i; x_i^4, 0{\cdot}04),$$

where $x_i = (1, x_i)'$, with $p = 2$ and $x_i \sim \text{Un}(0, 1)$. This case was chosen as a challenging example as the shape of the conditional density changes rapidly, with limited sample size in any particular local region. The performance is at least as good in other examples we have considered, and is excellent when the base parametric linear regression model provides an adequate approximation.

In a reference analysis, we chose $\alpha = 1$ and $\lambda = 1$ to favour few occupied basis locations and few clusters per location. In addition, we chose a Gaussian kernel $K(x, x') = \exp(-\psi ||x - x'||^2)$, letting the kernel precision ψ be unknown through use of a lognormal prior on ψ with $\mu_\psi = 2{\cdot}5$ and $\sigma_\psi^2 = 0{\cdot}5$. As the choice of ψ has a strong impact on borrowing of information across the

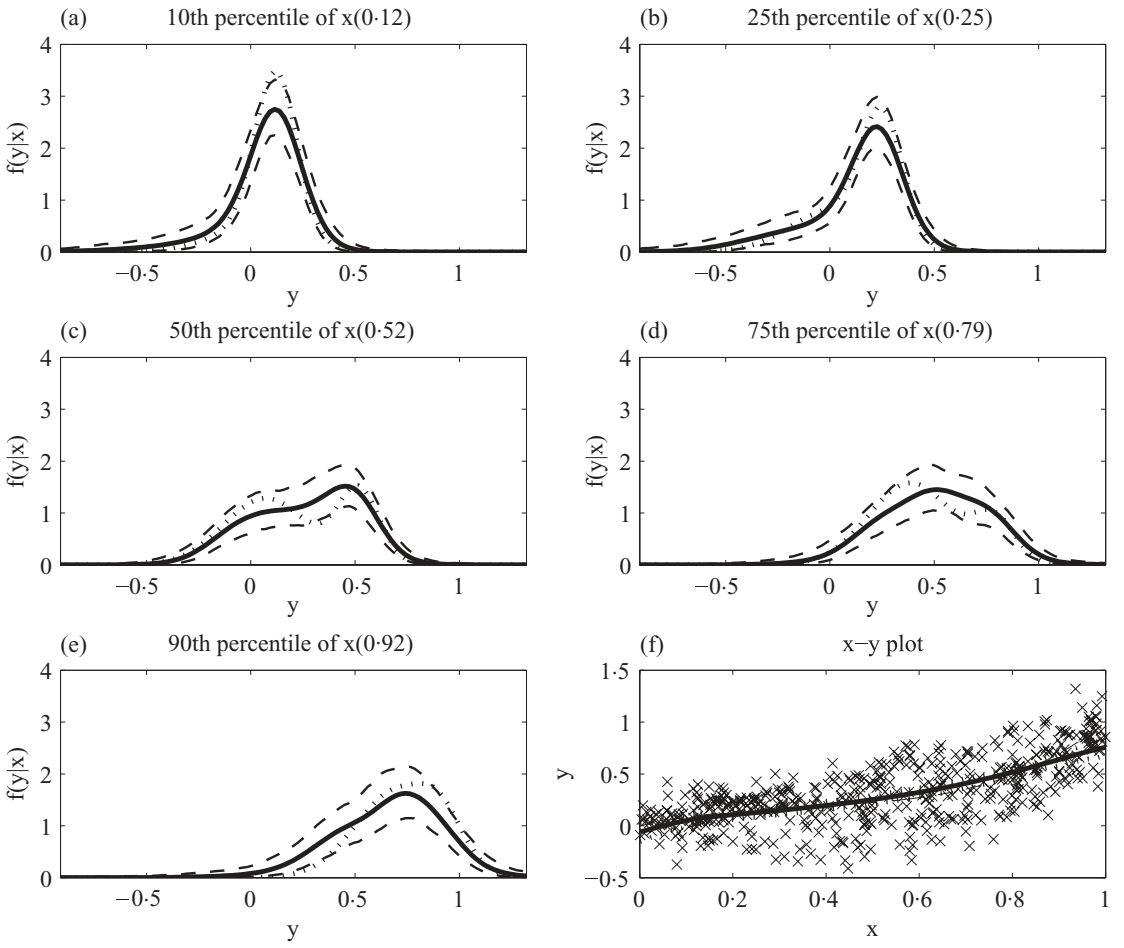

Fig. 1. Results for the kernel stick-breaking reference analysis in the simulation example. Estimated conditional response densities are shown for different percentiles of the predictor, including (a) 10th, (b) 25th, (c) 50th, (d) 75th, (e) 90th. The raw data and mean regression estimator are shown in (f). The solid lines are the posterior means, the dashed lines are pointwise 99% credible intervals, and the dotted lines are the true values.

predictor space, we expect the results to be sensitive to ψ and recommend choosing a hyperprior to allow the data to inform about the choice. In simulated and real data examples, we find the data are highly informative about ψ.

Results were obtained by running the Markov chain Monte Carlo algorithm described in § 4 for 30 000 iterations, with a burn-in of 8000 iterations discarded. Based on examination of trace plots, apparent convergence occurred quickly and there was efficient mixing. Figure 1 plots the true density, dotted line, and estimated predictive density, solid line, along with pointwise 99% credible intervals, dashed lines, at five selected percentiles (10, 25, 50, 75, 90) of the sampled x_i. The true density is contained in the pointwise 99% credible intervals. A plot of the data along with the estimated and true mean curve is also provided, showing they are indistinguishable.

To assess sensitivity to the hyperparameter specification, we repeated the analysis with ψ fixed at 0·2, 1 or 5, $K(x, x') = 1(||x - x'|| < \psi)$ or $K(x, x') = \exp(-\psi||x - x'||)$, $\alpha = 10$ and $\lambda = 10$. In each of these cases, the other hyperparameters were chosen as in the reference analysis. For small values of ψ we expect that local features of the conditional densities will be poorly estimated, while for large α or λ we expect the conditional densities to be strongly shrunk toward

Table 1. *Kullback–Leibler divergence between the true and estimated densities*

	Percentile of empirical distribution of x				
	10th	25th	50th	75th	90th
KSBP					
Reference setting	1·653	1·105	1·105	1·051	3·131
$\alpha = 10, \lambda = 1$	6·268	1·519	4·706	2·795	9·514
$\alpha = 1, \lambda = 10$	6·085	1·515	4·442	2·638	9·096
$\alpha = 10, \lambda = 10$	6·693	1·694	4·718	2·820	9·883
Exponential kernel	2·107	1·525	1·244	1·559	3·528
Rectangular kernel	2·133	0·954	1·187	1·819	4·082
$\psi = 0{\cdot}2$	4·828	1·915	4·040	2·003	7·403
$\psi = 1$	2·892	1·930	1·709	1·555	4·787
$\psi = 5$	2·358	0·630	1·544	1·192	2·791
WMDP	1·489	1·625	0·686	2·610	0·782

KSBP, kernel stick-breaking process; WMDP, Dunson, Pillai & Park (2007) method.

the baseline normal linear regression model. Table 1 shows the Kullback–Leibler divergence between the true and estimated conditional densities in each case, also reporting results for the method proposed by Dunson et al. (2007) using their recommended hyperparameter values.

We find that the results are robust to the choice of kernel as long as a hyperprior is chosen for the kernel precision, though the Gaussian kernel gave the best performance. In addition, as expected the Kullback–Leibler divergence increased for small values of ψ and for large values of α or λ. In these cases, there was a tendency to underestimate local peaks and oversmooth the conditional densities. The estimates for the kernel stick-breaking process and Dunson et al. (2007) approaches were similar, although the former has clear conceptual advantages over the latter because of the coherent updating property and lack of sample dependence. In addition, by not including basis distributions at every data-point, computational speed is increased considerably. We repeated the simulation example for $n = 1000$ for the reference analysis and Dunson et al. (2007) approach. The average Kullback–Leibler divergence was reduced by 50% in this case.

6. Epidemiology application

6·1. *Background and motivation*

In epidemiology studies, a common focus is on assessing changes in a response distribution with a continuous exposure, adjusting for covariates. For example, Longnecker et al. (2001) studied the relationship between the DDT metabolite DDE and preterm delivery. The substance DDT is effective against malaria-transmitting mosquitoes, and so is widely used in malaria-endemic areas in spite of growing evidence of health risks. The Longnecker et al. (2001) study measured DDE in mother's serum during the third trimester of pregnancy, while also recording the gestational age at delivery, GAD, and demographic factors, such as age. Data on DDE and GAD are shown in Fig. 2 for the 2313 children in the study, excluding the children for whom GAD exceeded 45 weeks, unrealistically high values attributable to measurement error.

Following standard practice in reproductive epidemiology, Longnecker et al. (2001) dichotomized GAD using a 37-week cut-off, so that deliveries occurring prior to 37 weeks of completed gestation were classified as preterm. Categorizing DDE using quintiles based on the empirical distribution, they fitted a logistic regression model, reporting evidence of a highly significant dose–response trend. Premature deliveries occurring earlier in the period before 37 weeks have greater risk of mortality and morbidity. Hence, from a public health and clinical

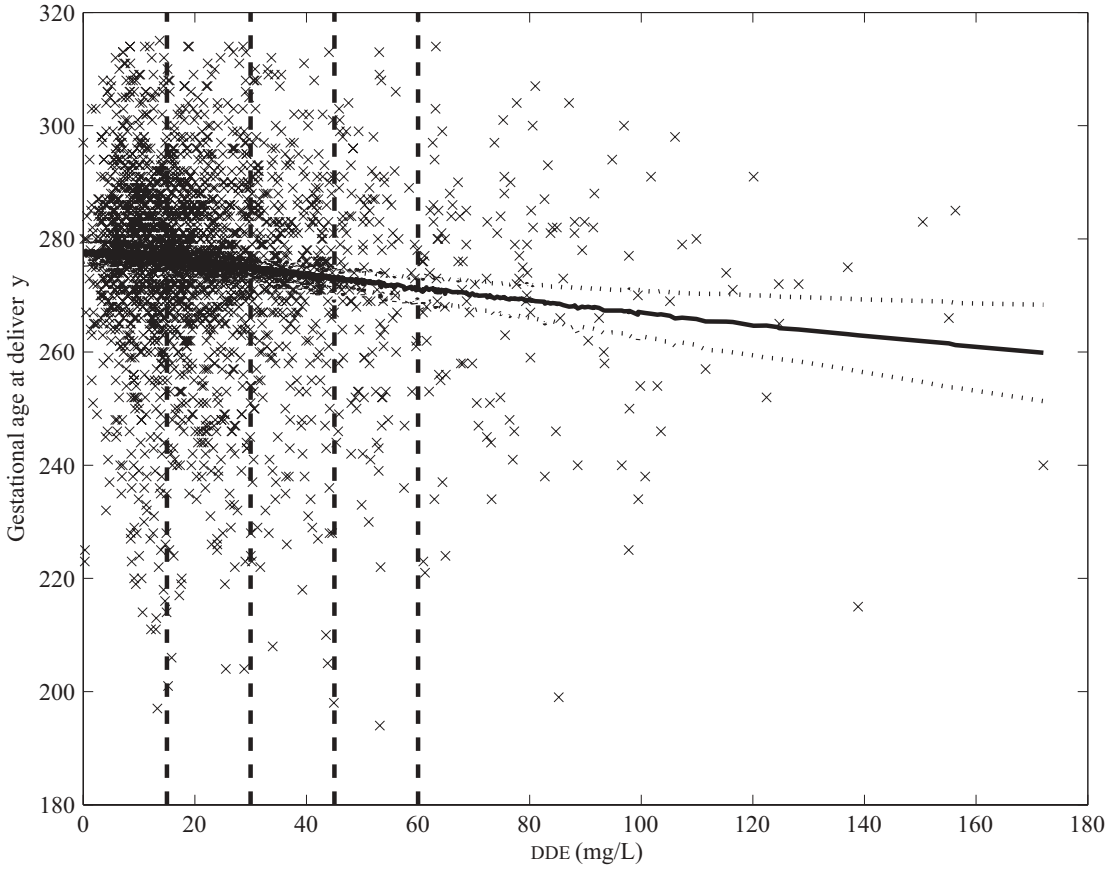

Fig. 2. DDE vs gestational age at delivery in days for 2313 women in the Longnecker et al. (2001) study. The solid line is the conditional predictive mean, while the dotted lines are 99% pointwise credible intervals. Vertical dashed lines are DDE quintiles.

perspective, it is of interest to assess how the entire left tail of the GAD distribution changes with DDE dose, with effects earlier in gestation more important.

6·2. *Analysis and results*

We analyzed the Longnecker et al. data using the following semiparametric Bayes model:

$$f(y_i \mid x_i) = \int N(y_i; x_i'\beta_i, \tau^{-1}) dG_{x_i}(\beta_i) \\ \mathcal{G}_{\mathcal{X}} \sim \mathcal{P}, \tag{24}$$

where $\mathcal{G}_{\mathcal{X}} = \{G_x : x \in \Re^p\}$, y_i is the normalized gestational age at delivery, $x_i = (1, \text{DDE}_i, \text{age}_i)'$, DDE_i is the normalized DDE dose for woman i, age_i is her normalized age, and $\mathcal{P}$ is a kernel stick-breaking process, with a Gaussian kernel and $\mathcal{Q}$ corresponding to a DP(αG_0). Prior specification and other details are as described in § 5 for the reference analysis.

Convergence was rapid and mixing was good based on examination of trace plots, not shown. Even though the sample size was 2313, the posterior mean number of occupied locations was only 5·4, while the posterior mean number of clusters was 28·1.

Figure 3 shows the estimated conditional densities of gestational age at delivery for a range of DDE values. There is some suggestion of an increasing left tail with dose, representing increasing

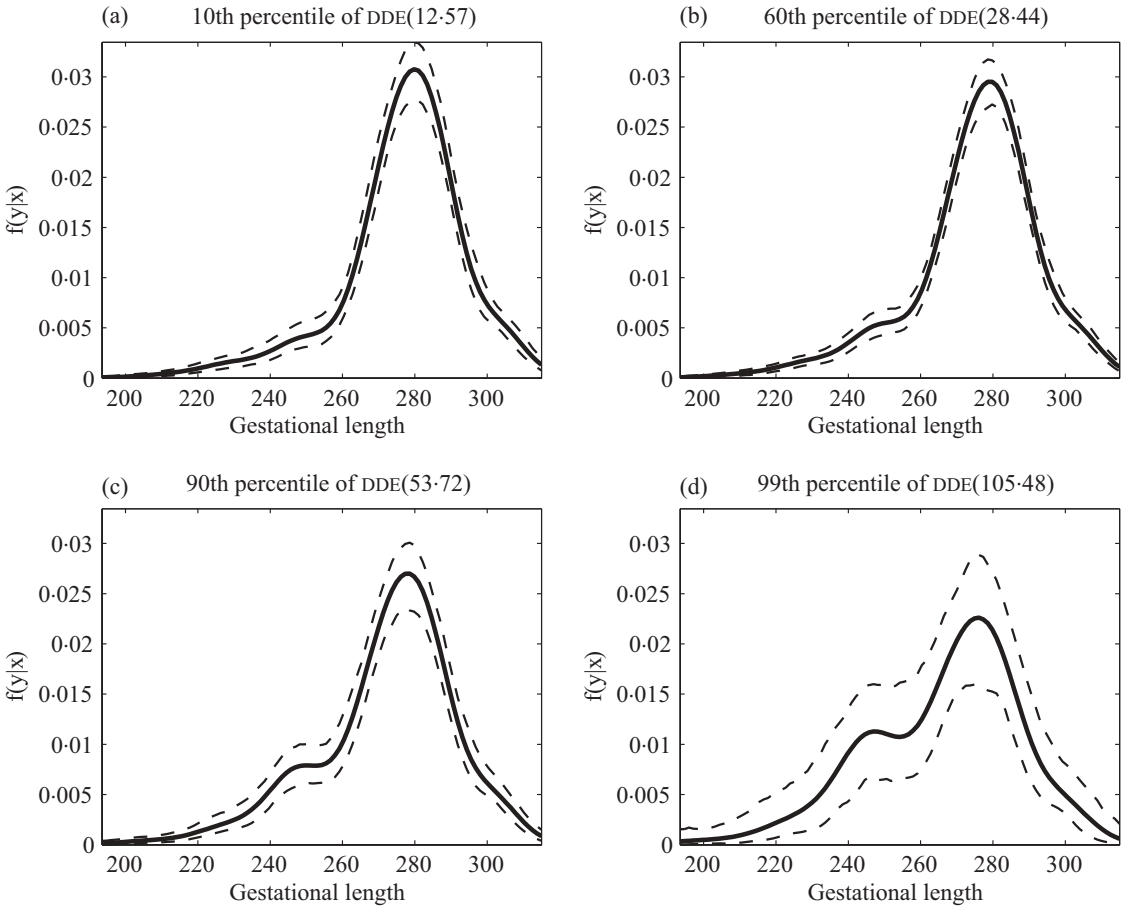

Fig. 3. Estimated densities of gestational age at delivery (in days) conditionally on DDE, $f(y|x)$, for the kernel stick-breaking reference analysis. Estimates correspond to different percentiles of the predictor distribution, including (a) 10th, (b) 60th, (c) 90th and (d) 99th. Solid lines represent posterior means, and dashed lines represent 99% credible intervals.

risk of premature delivery at higher exposure values. At very high exposures, data are sparse and the credible intervals are much wider. To assess more directly the impact of DDE on the left tail, Fig. 4 shows dose–response curves for $\text{pr}(Y < T)$ for different choices of cut-off T. For early preterm birth before 33 weeks, the dose–response curve is flat except at high doses where the credible interval is wide. As the cut-off increases, the dose–response becomes more significant. Hence, the increased risk of preterm birth with increasing DDE dose reported by Longnecker et al. (2001) can be attributed to more deliveries late in the interval before 37 weeks.

7. Discussion

The article proposes a class of kernel stick-breaking processes, which should be widely useful in settings in which there is uncertainty in an uncountable collection of probability measures. We have focused on a density regression application in which one is interested in studying how a response density changes with predictors. However, there are many other applications that can be considered, including predictor-dependent clustering, dynamic modelling and spatial data analysis.

The kernel stick-breaking process should provide a useful alternative to dependent Dirichlet process methods, such as the order-based dependent Dirichlet process (Griffin & Steel, 2006). An

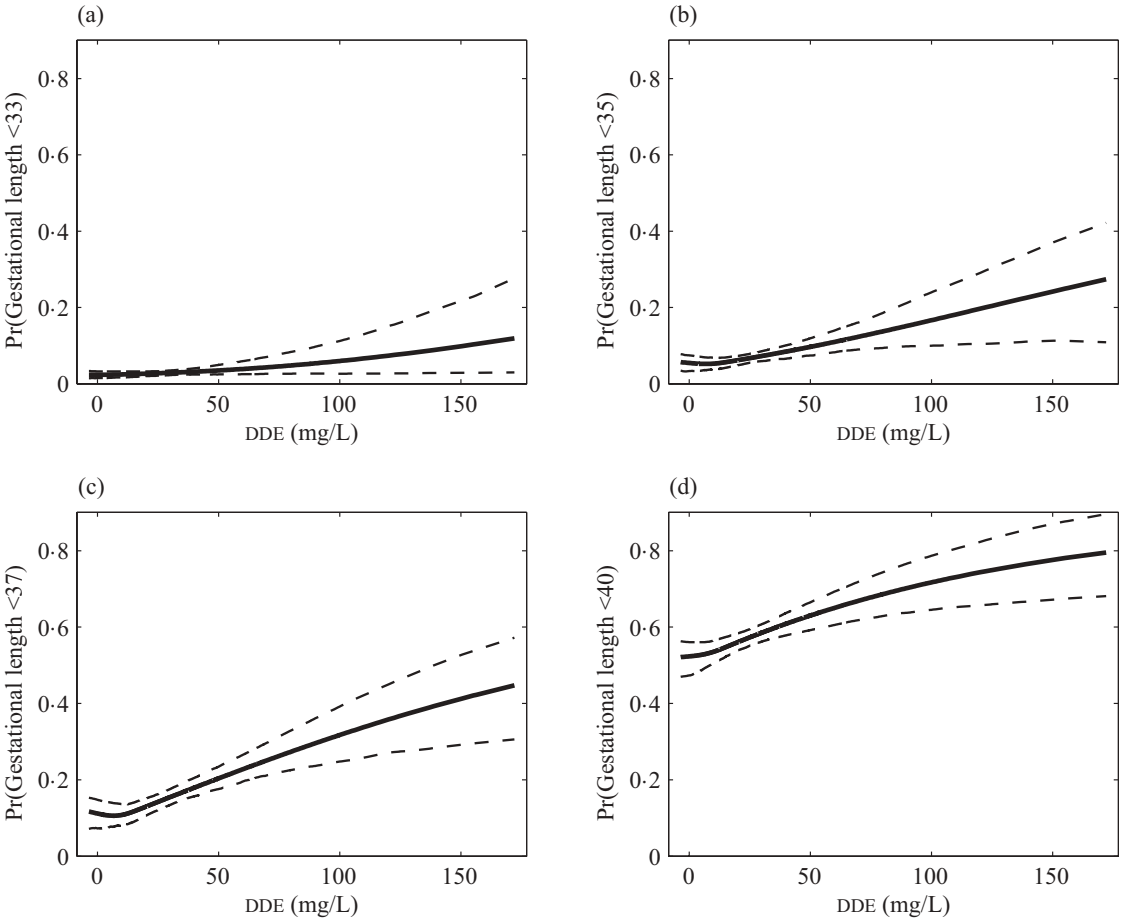

Fig. 4. Estimated probability that gestational age at delivery is less than T weeks versus DDE dose, for (a) $T = 33$, (b) $T = 35$, (c) $T = 37$, (d) $T = 40$. Solid lines are posterior means and dashed lines are pointwise 99% credible intervals.

advantage of the kernel stick-breaking process formulation is that many of the tools developed for exchangeable stick-breaking processes, such as the Dirichlet process, can be applied with minimal modification. This has allowed us to obtain some insight into theoretical properties and to develop computational algorithms, which are straightforward to implement in cases in which stick-breaking basis measures are used having base distributions conjugate to the likelihood. In future work, it will be interesting to consider algorithms for efficient posterior computation in nonconjugate models and in cases with many predictors.

We also obtained a predictor-dependent urn scheme, which generalizes the Pólya urn scheme (Blackwell & MacQueen, 1973). It will be interesting to apply this urn scheme for computation and clustering without the need to consider explicitly the random weights and locations in the stick-breaking representation.

ACKNOWLEDGEMENT

This research was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences of the U.S. National Institutes of Health. The authors would like to thank Yeonseung Chung, Rongheng Lin, Shyamal Peddada and an anonymous referee for helpful comments.

APPENDIX

Proofs

Proof of Theorem 1. As shorthand notation, we let $Q_h = G_h^*(B)$ and $Q_0 = E\{G_h^*(B)\}$. Then we have

$$
\begin{aligned}
E\{G_x(B)G_{x'}(B)\} &= E\left\{\left(\sum_{h=1}^{\infty} U_h(x)\left[\prod_{l=1}^{h-1}\{1-U_l(x)\}\right]Q_h\right)\times\left(\sum_{h=1}^{\infty} U_h(x')\left[\prod_{l=1}^{h-1}\{1-U_l(x')\}\right]Q_h\right)\right\}\\
&= E\left(\sum_{h=1}^{\infty} U_h(x)U_h(x')\prod_{l=1}^{h-1}\{1-U_h(x)\}\{1-U_h(x')\}Q_h^2\right)\\
&\quad + E\left(\sum_{h=1}^{\infty}\sum_{l=1}^{h-1} U_h(x)\left[\prod_{r=1}^{l-1}\{1-U_r(x)\}\{1-U_r(x')\}\right]\{U_l(x')-U_l(x)U_l(x')\}\right.\\
&\qquad \left.\times\left[\prod_{s=l+1}^{h-1}\{1-U_s(x)\}\right]Q_hQ_l\right)\\
&\quad + E\left(\sum_{h=1}^{\infty}\sum_{l=h+1}^{\infty} U_l(x')\left[\prod_{r=1}^{h-1}\{1-U_r(x)\}\{1-U_r(x')\}\right]\right.\\
&\qquad \left.\times\{U_h(x)-U_h(x)U_h(x')\}\left[\prod_{s=h+1}^{l-1}\{1-U_s(x')\}\right]Q_hQ_l\right)\\
&= \sum_{h=1}^{\infty}\mu(x,x')\{1-\mu(x)-\mu(x')+\mu(x,x')\}^{h-1}E\left(Q_h^2\right)\\
&\quad + \mu(x)\{\mu(x')-\mu(x,x')\}\sum_{l=1}^{\infty}\sum_{h=l+1}^{\infty}\{1-\mu(x)-\mu(x')+\mu(x,x')\}^{l-1}\\
&\quad \times\{1-\mu(x)\}^{h-l-1}Q_0^2+\mu(x')\{\mu(x)-\mu(x,x')\}\\
&\quad \times\sum_{h=1}^{\infty}\sum_{l=h+1}^{\infty}\{1-\mu(x)-\mu(x')+\mu(x,x')\}^{h-1}\{1-\mu(x')\}^{l-h-1}Q_0^2\\
&= \frac{\mu(x,x')\mathrm{V}_{\mathcal{Q}(B)}}{\mu(x)+\mu(x')-\mu(x,x')}+Q_0^2,
\end{aligned}
$$

with linearity of expectation and reordering justified as the series is absolutely convergent. □

Proof of Lemma 1. Under formulation (15), we have

$$
\begin{aligned}
\mathrm{pr}(\phi_i=\phi_j\mid x_i,x_j) &= \int \mathrm{pr}(Z_i=Z_j\mid x_i,x_j,V,\Gamma)d\pi(V)\,d\pi(\Gamma)\\
&= E\left(\sum_{h=1}^{\infty}\left[U_h(x_i)U_h(x_j)\prod_{l<h}\{1-U_l(x_i)\}\{1-U_l(x_j)\}\right]\right)\\
&= \mu_{ij}\sum_{h=1}^{\infty}\prod_{l<h}(1-\mu_i-\mu_j+\mu_{ij})
\end{aligned}
$$

$$= \mu_{ij} \sum_{h=0}^{\infty} (1 - \mu_i - \mu_j + \mu_{ij})^h$$
$$= \frac{\mu_{ij}}{\mu_i + \mu_j - \mu_{ij}}.$$

□

Proof of Theorem 2. Letting $\mathcal{I}$ denote an arbitrary subset of $\{1, \ldots, i\}$ that includes $i \in \mathcal{I}$, we have

$$E\left\{\sum_{h=1}^{\infty} \prod_{j \in \mathcal{I}} \mathrm{pr}(Z_j = h)\right\} = \sum_{h=1}^{\infty} E\left[\prod_{j \in \mathcal{I}} U_h(x_j) \prod_{l=1}^{h-1} \{1 - U_l(x_j)\}\right]$$
$$= \sum_{h=1}^{\infty} E\left\{\prod_{j \in \mathcal{I}} U_h(x_j)\right\} \prod_{l=1}^{h-1} E\left[\prod_{j \in \mathcal{I}} \{1 - U_l(x_j)\}\right]$$
$$= \mu_{\mathcal{I}} \sum_{h=1}^{\infty} E\left[\prod_{j \in \mathcal{I}} \{1 - U_l(x_j)\}\right]^{h-1}$$
$$= \frac{\mu_{\mathcal{I}}}{1 - \sum_{t=0}^{\#\mathcal{I}} (-1)^t \sum_{\mathcal{J} \in \mathcal{I}_t} \mu_{\mathcal{J}}} = \frac{\mu_{\mathcal{I}}}{\sum_{t=1}^{\#\mathcal{I}} (-1)^{t-1} \sum_{\mathcal{J} \in \mathcal{I}_t} \mu_{\mathcal{J}}} = \omega_{\mathcal{I}},$$

where $\mathcal{I}_t$ denotes the set of all possible subsets of $\mathcal{I}$ of length t.

Let $\mathcal{K}$ denote a arbitrary subset of $\{1, \ldots, i\}$ that includes i and j, and let $\overline{\mathcal{K}} = \{1, \ldots, i\} \backslash \mathcal{K}$. Then, if we let $Z_i = Z_j$ for all $j \in \mathcal{K} \backslash \{i\}$ and $Z_i \neq Z_j$ for all $j \in \overline{\mathcal{K}}$, the probability of observing $\mathcal{K}$ and $\overline{\mathcal{K}}$ in a sample from the prior is

$$E\left[\sum_{h=1}^{\infty} \prod_{k \in \mathcal{K}} \mathrm{pr}(Z_k = h) \prod_{k \in \overline{\mathcal{K}}} \{1 - \mathrm{pr}(Z_k = h)\}\right]$$
$$= E\left[\sum_{h=1}^{\infty} \prod_{k \in \mathcal{K}} \mathrm{pr}(Z_k = h) \left\{\sum_{s=0}^{\#\overline{\mathcal{K}}} (-1)^s \sum_{\mathcal{L} \in \overline{\mathcal{K}}_s} \prod_{l \in \mathcal{L}} \mathrm{pr}(Z_l = h)\right\}\right]$$
$$= \sum_{s=0}^{\#\overline{\mathcal{K}}} (-1)^s \sum_{\mathcal{L} \in \overline{\mathcal{K}}_s} E\left\{\sum_{h=1}^{\infty} \prod_{k \in \mathcal{L} \bigcup \mathcal{K}} \mathrm{pr}(Z_k = h)\right\} = \sum_{s=0}^{\#\overline{\mathcal{K}}} (-1)^s \sum_{\mathcal{L} \in \overline{\mathcal{K}}_s} \omega_{\mathcal{L} \bigcup \mathcal{K}},$$

where $\overline{\mathcal{K}}_s$ is the set of subsets of $\overline{\mathcal{K}}$ of length s. The probability of $Z_i = Z_j$ is then

$$\sum_{t=2}^{i} \sum_{\mathcal{K} \in \mathcal{N}_{i,j}^{(t,i)}} \sum_{s=0}^{\#\overline{\mathcal{K}}} (-1)^s \sum_{\mathcal{L} \in \overline{\mathcal{K}}_s} \omega_{\mathcal{L} \bigcup \mathcal{K}} = \sum_{r=2}^{i} (-1)^r \sum_{\mathcal{I} \in \mathcal{N}_{i,j}^{(r,i)}} \omega_{\mathcal{I}}.$$

Here, $r - 1$ indexes the cardinality of the set $\{j : \phi_i = \phi_j\}$, and we obtain the expression in Theorem 2 through normalization. □

REFERENCES

ALDOUS, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XII*, Lecture Notes in Mathematics **1117**, Ed. P. L. Hennequin, pp. 1–198. Berlin: Springer.

BLACKWELL, D. & MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353–5.

BARRY, D. & HARTIGAN, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20**, 260–79.

Caron, F., Davy, M., Doucet, A., Duflos, E. & Vanheeghe, P. (2006). Bayesian inference for dynamic models with Dirichlet process mixtures. In *International Conference on Information Fusion*, pp. 1–8. Florence, Italy: INRIA–CCSd–CNRS.

Cifarelli, D. M. & Regazinni, E. (1978). Nonparametric statistical problems under partial exchangeability: the use of associative means. *Ann. Inst. Mat. Finian. Univ. Torino*, II **12**, 1–36.

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

Dunson, D. B. (2006). Bayesian dynamic modelling of latent trait distributions. *Biostatistics* **7**, 551–68.

Dunson, D. B., Herring, A. H. & Engel, S. M. (2007). Bayesian selection and clustering of polymorphisms in functionally-related genes. *J. Am. Statist. Assoc.*, forthcoming.

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

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–30.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615–29.

Gelfand, A. E., Kottas, A. & Maceachern, S. N. (2005). Bayesian nonparametric spatial modelling with Dirichlet process mixing. *J. Am. Statist. Assoc.* **100**, 1021–35.

Griffin, J. E. & Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *J. Am. Statist. Assoc.* **101**, 179–94.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Assoc.* **96**, 161–73.

Ishwaran, H. & James, L. F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica* **13**, 1211–35.

Ishwaran, H. & Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87**, 371–90.

Kim, S., Tadesse, M. G. & Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93**, 877–93.

Longnecker, M. P., Klebanoff, M. A., Zhou, H. B. & Brock, J. W. (2001). Association between maternal serum concentration of the ddt metabolite dde and preterm and small-for-gestational-age babies at birth. *Lancet* **358**, 110–4.

Maceachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Commun. Statist.* B **23**, 727–41.

Maceachern, S. N. (1999). Dependent onparametric processes. In *Proc. Bayesian Statist. Sci. Sect.*, pp. 50–5. Alexandria, VA: American Statistical Association.

Maceachern, S. N. (2001). Decision theoretic aspects of dependent nonparametric processes. In *Bayesian Methods With Applications to Science, Policy, and Official Statistics*, Ed. E. George, pp. 551–60. Crete: International Society for Bayesian Analysis.

Medvedovic, M., Yeung, K. Y. & Bumgarner, R. E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics* **20**, 1222–32.

Müller, P., Quintana, F. & Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *J. R. Statist. Soc.* B **66**, 735–49.

Papaspiliopoulos, O. & Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169–86.

Pennell, M. L. & Dunson, D. B. (2006). Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics* **62**, 1044–52.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory*, Ed. T. S. Ferguson, L. S. Shapley and J. B. MacQueen. pp. 245–67. IMS Lecture Notes–Monograph Series, **30**. Hayward, CA: Inst. Math. Statist.

Quintana, F. A. & Iglesias, P. L. (2003). Bayesian clustering and product partition models. *J. R. Statist. Soc.* B **65**, 557–74.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica*, **4**, 639–50.

West, M., Müller, P. & Escobar, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *Aspects of Uncertainty: A Tribute to D.V. Lindley*, Ed. A. F. M. Smith and P. R. Freeman, pp. 363–86. New York: John Wiley.

[*Received November* 2006. *Revised August* 2007]