Kernel Stick-Breaking Processes

David B. Dunson[1] **and Ju-Hyun Park**[1,2]

[1] *Biostatistics Branch,*

National Institute of Environmental Health Sciences

U.S. National Institute of Health

P.O. Box 12233, RTP, NC 27709, U.S.A

[2] *Department of Biostatistics*

University of North Carolina at Chapel Hill

dunson1@niehs.nih.gov

Summary. This article proposes a class of kernel stick-breaking processes (KSBP) for uncountable collections of dependent random probability measures. The KSBP 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 KSBP are described, including a covariate-dependent prediction rule. A retrospective MCMC algorithm is developed for posterior computation, and the methods are illustrated using a simulated example and an epidemiologic application.

Keywords: 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, $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, where $\mathcal{X}$ is a Lesbesgue measurable subset of

$\Re^p$ and $G_{\mathbf{x}}$ is a probability measure with respect to $(\Omega, \mathcal{F})$, with Ω the sample space and $\mathcal{F}$ the corresponding Borel σ-algebra. A motivating application is the problem of estimating the conditional density of a response variable using the mixture specification $f(\mathbf{y} \,|\, \mathbf{x}) = \int f(\mathbf{y} \,|\, \mathbf{x}, \phi) dG_{\mathbf{x}}(\phi)$, with $f(\mathbf{y} \,|\, \mathbf{x}, \phi)$ a known kernel and $G_{\mathbf{x}}$ an unknown probability measure indexed by the predictor value, $\mathbf{x} = (x_1, \ldots, x_p)'$.

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 (DP) prior, with methods varying in how they incorporate dependency. One approach is to include a regression in the base measure (Cifarelli and Regazzini, 1978), which has the disadvantage of capturing dependency only in aspects of the distribution characterized by the base parametric model.

Much of the recent work has instead relied on generalizations of Sethuraman (1994)'s stick-breaking representation of the DP. If G is assigned a DP prior with precision α and base measure G_0, denoted $G \sim DP(\alpha G_0)$, then the stick-breaking representation of G is

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

where δ_θ is a probability measure concentrated at θ. MacEachern (1999; 2001) proposed the dependent DP (DDP), which generalizes (1) to allow a collection of unknown distributions indexed by $\mathbf{x}$ by assuming fixed weights $\mathbf{p} = (p_h, h = 1, \ldots, \infty)$ while allowing the atoms $\boldsymbol{\theta} = (\theta_h, h = 1, \ldots, \infty)$ to vary with $\mathbf{x}$ according to a stochastic process. The DDP has been successfully applied to ANOVA (De Iorio et al., 2004), spatial modeling (Gelfand et al., 2005), functional data (Dunson and Herring, 2006), and time series (Caron et al., 2006) applications.

Noting limited flexibility due to the fixed weights assumption, Griffin and Steel (2006) and Duan et al. (2006) have recently developed methods to allow $\mathbf{p}$ to vary with predictors. Griffin and Steel's (2006) approach is based on an innovative order-based DDP, which incorporates dependency by allowing the ordering of the random variables $\{V_h, h = 1, \ldots, \infty\}$ in the stick-breaking construction to depend on predictors. Motivated by spatial applications, Duan et al. (2006) instead propose a multivariate extension of the stick-breaking representation. An alternative to these approaches is to incorporate dependency through weighted mixtures of independent DPs. Müller et al. (2004) used this idea to allow dependency across experiments, while Dunson (2006) and Pennell and Dunson (2006) considered discrete dynamic settings.

A conceptually-related idea was proposed by Dunson, Pillai and Park (2006), who defined a prior for $\mathcal{G}_{\mathcal{X}}$ conditionally on a sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ as follows:

$$\begin{aligned} G_{\mathbf{x}} &= \sum_{j=1}^{n} \Big(\frac{\gamma_j K(\mathbf{x}, \mathbf{x}_j)}{\sum_{l=1}^{n} \gamma_l K(\mathbf{x}, \mathbf{x}_l)} \Big) G_j^*, \\ \gamma_j &\sim \text{gamma}(\kappa, n\kappa), \; G_j^* \sim DP(\alpha G_0), \; j = 1, \ldots, n, \end{aligned} \tag{2}$$

which expresses $G_{\mathbf{x}}$ as a weighted mixture of independent random probability measures introduced at the observed predictor values. Here, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$ is a vector of random weights on the n different bases, located at $\mathbf{x}_j$ for $j = 1, \ldots, n$, $K : \Re^p \times \Re^p \to [0, 1]$ is a bounded kernel function, and G_j^* is a DP random basis measure. Because bases located close to $\mathbf{x}$ are assigned higher weight, expression (2) accommodates spatial dependency.

Although prior (2) has had good performance in several applications, the sample dependence is problematic from a Bayesian perspective and results in some unappealing properties. For example, the specification lacks reasonable marginalization and updating properties. If we define a prior of the form (2) based on a particular sample realization $\mathbf{X}$, then we do not obtain a prior of the same form based on $\mathbf{S} \subset \mathbf{X}$ or $\mathbf{S} = [\mathbf{X}', \mathbf{X}_{new}]'$, with $\mathbf{X}_{new}$ denoting additional samples.

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 and MacQueen (1973) Pólya urn scheme, to incorporate predictors. Assuming $\phi_i \sim G$, with $G \sim DP(\alpha G_0)$, one obtains the

DP prediction rule upon marginalizing over the prior for G:

$$P(\phi_1 \in \cdot) = G_0(\cdot), \quad P(\phi_i \in \cdot \,|\, \phi_1, \ldots, \phi_{i-1}) = \Big(\frac{\alpha}{\alpha + i - 1}\Big) G_0(\cdot) + \sum_{j=1}^{i-1} \Big(\frac{1}{\alpha + i - 1}\Big) \delta_{\phi_j}(\cdot). \quad (3)$$

The DP prediction rule forms the basis for commonly-used algorithms for efficient posterior computation in DP mixture models (MacEachern, 1994).

The DP prediction rule induces clustering of the subjects according to a Chinese Restaurant Process (CRP) (Aldous, 1985; Pitman, 1996). This clustering behavior is often exploited as a dimensionality reduction device and a tool for exploring latent structure (Dunson et al., 2006; Kim et al., 2006; Lau and Green, 2006; Medvedovic et al., 2004). The DP and related approaches, including product partition models (Barry and Hartigan, 1992; Quintana and Iglesias, 2003) and species sampling models (Pitman, 1996; Ishwaran and 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 (KSBP) to be used as a sample-free prior for $\mathcal{G}_\mathcal{X}$, which induces a covariate-dependent prediction rule upon marginalization. Section 2 proposes the formulation and considers basic properties. Section 3 presents the prediction rule. Section 4 develops a retrospective MCMC algorithm (Papaspiliopoulos and Roberts, 2006) for posterior computation. Section 5 applies the approach to simulated examples. Section 6 contains an epidemiologic application, and Section 7 discusses the results. Proofs are included in Appendices.

2. Predictor-Dependent Random Probability Measures

2.1 *Formulation and Special Cases*

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 $(\Omega, \mathcal{F})$ indexed by $\mathbf{x} \in \mathcal{X}$ 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 $\Gamma_h \overset{iid}{\sim} H$ is a location, $V_h \overset{ind}{\sim} \text{beta}(a_h, b_h)$ is a probability weight, and $G_h^* \overset{iid}{\sim} \mathcal{Q}$ is a probability measure. Here, H is a probability measure on $(\mathcal{L}, \mathcal{A})$, where $\mathcal{A}$ is a Borel σ-algebra of subsets of $\mathcal{L}$, and $\mathcal{L}$ is a Lesbesgue 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 $(\Omega, \mathcal{F})$. For example, $\mathcal{Q}$ may correspond to a Dirac measure at a random location, a Dirichlet process, or a species sampling model (Pitman, 1996; Ishwaran and James, 2003).

Using these components, the kernel stick-breaking process (KSBP) is defined as follows:

$$\begin{aligned} G_{\mathbf{x}} &= \sum_{h=1}^{\infty} W(\mathbf{x}; V_h, \Gamma_h) \prod_{l<h} \{1 - W(\mathbf{x}; V_l, \Gamma_l)\} G_h^*, \\ W(\mathbf{x}; V_h, \Gamma_h) &= V_h K(\mathbf{x}, \Gamma_h), \quad \forall \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{4}$$

where $K : \Re^p \times \Re^p \to [0, 1]$ is a bounded kernel function, which is initially assumed to be known. Note that (4) formulates $G_{\mathbf{x}}$ 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 $\mathbf{x}$ and having a smaller index, h, will tend to receive higher probability weight. In this manner, the KSBP accommodates dependency between $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$.

Let $\pi_h(\mathbf{x}; \mathbf{V}_h, \boldsymbol{\Gamma}_h) = W(\mathbf{x}; V_h, \Gamma_h) \prod_{l<h} \{1 - W(\mathbf{x}; V_l, \Gamma_l)\}$, for $h = 1, \ldots, \infty$, with $\mathbf{V}_h = (V_1, \ldots, V_h)'$ and $\boldsymbol{\Gamma}_h = (\Gamma_1, \ldots, \Gamma_h)'$. Then, the following Lemma must hold in order for $G_{\mathbf{x}}$ to be a well defined probability measure for all $\mathbf{x} \in \mathcal{X}$:

Lemma 1. $\sum_{h=1}^{\infty} \pi_h(\mathbf{x}; \mathbf{V}_h, \boldsymbol{\Gamma}_h) = 1$ a.s. for all $\mathbf{x} \in \mathcal{X}$ if

$$\sum_{h=1}^{\infty} \mathrm{E}[\log\{1 - V_h K(\mathbf{x}, \Gamma_h)\}] = -\infty, \quad \text{for all } \mathbf{x} \in \mathcal{X}.$$

For example, when $V_h \overset{iid}{\sim} \text{beta}(a, b)$, $\mathrm{E}[\log\{1 - V_h K(\mathbf{x}, \Gamma_h)\}] = C(\mathbf{x})$, where $C(\mathbf{x}) < 0$ is a constant strictly less than 0, which implies that the conditional of Lemma 1 holds.

To motivate the form chosen in (4), it is useful to first consider the special case in which $K(\mathbf{x},\Gamma) = 1$ for all $(\mathbf{x},\Gamma) \in \mathcal{X} \bigotimes \mathcal{L}$ and $G_h^* \sim DP(\alpha G_0)$. In this case, $G_{\mathbf{x}} = G$, with G assigned a stick-breaking mixture of Dirichlet processes. Specifically, letting $\phi_i \sim G$, ϕ_i is drawn from G_1^* with probability V_1, from G_2^* with probability $V_2(1-V_1)$, and from G_h^* with probability $V_h \prod_{l<h}(1-V_l)$, for $h = 3,\ldots,\infty$. Here, $\prod_{l<h}(1-V_l)$ of the unit probability stick remains to be allocated after assigning probabilities to the first $h-1$ basis locations, and V_h is the proportion of this remaining piece allocated to location Γ_h.

By choosing a kernel, such as $K(\mathbf{x},\Gamma) = \exp(-\psi||\mathbf{x}-\Gamma||)$ for $\psi > 0$, we allow these stick-breaking probabilities to depend on predictors. In particular, ϕ_i is drawn from G_1^* with probability $V_1 K(\mathbf{x},\Gamma_1)$, which decreases monotonically from V_1 to 0 as the distance between $\mathbf{x}$ and Γ_1 increases. Hence, if $\mathbf{x}$ is far from Γ_1, the location of the first basis, then more of the stick will remain to be allocated to other basis locations. The resulting kernel stick breaking process represents a fundamentally different approach than the order-based approach of Griffin and Steel (2006) or the multivariate idea of Duan et al. (2006).

Although the kernel stick-breaking formulation is new, a number of previously proposed formulations arise as special cases when $K(\mathbf{x},\Gamma) = 1$ for all $(\mathbf{x},\Gamma) \in \mathcal{X} \bigotimes \mathcal{L}$. When $G_h^* = \delta_{\theta_h}, \theta_h \overset{iid}{\sim} G_0$, for $h = 1,\ldots,\infty$, we obtain $G_{\mathbf{x}} \equiv G$, with G having a stick-breaking prior in the class considered by Ishwaran and James (2001). In the further special case in which $a_h = 1-a$ and $b_h = b + h\,a$, we obtain a Pitman-Yor (1997) process for G, with $G \sim DP(\lambda G_0)$ when $a = 0$ and $b = \lambda$. If we instead let $G_h^* \sim DP(\alpha G_0)$, $a_h = 1, b_h = \lambda$, G is assigned a DP mixture of DPs, which is a two parameter generalization of the Dirichlet process. The DP mixture of DPs reduces to a DP in the limiting case as either $\lambda \rightarrow 0$ or $\alpha \rightarrow 0$.

2.2 *Conditional Properties*

Returning to the general case, we first derive moments of $G_{\mathbf{x}}$ conditionally on the random weights $\mathbf{V}$ and random locations $\mathbf{\Gamma}$, but marginalizing out the random basis measures,

$\{G_h^*, h = 1, \dots, \infty\}$. Letting $G_0(\mathcal{B}) = \mathrm{E}_{\mathcal{Q}}\{G_h^*(\mathcal{B})\}$, for all $\mathcal{B} \in \mathcal{F}$, we obtain

$$\mathrm{E}\{G_{\mathbf{x}}(\mathcal{B}) \,|\, \mathbf{V}, \mathbf{\Gamma}\} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}; \mathbf{V}_h, \mathbf{\Gamma}_h)\mathrm{E}_{\mathcal{Q}}\{G_h^*(\mathcal{B})\} = G_0(\mathcal{B}), \quad \forall \mathcal{B} \in \mathcal{F}. \tag{5}$$

Due to the lack of dependency on $\mathbf{V}$ and $\mathbf{\Gamma}$, we also have $\mathrm{E}_{\mathcal{P}}\{G_{\mathbf{x}}(\mathcal{B})\} = G_0(\mathcal{B})$, so that the prior is centered on the base measure G_0. In addition,

$$\begin{aligned}
\mathrm{E}\{G_{\mathbf{x}}(\mathcal{B})^2 \,|\, \mathbf{V}, \mathbf{\Gamma}\} &= \Big[\sum_{h=1}^{\infty} \pi_h(\mathbf{x}; \mathbf{V}_h, \mathbf{\Gamma}_h)^2 \mathrm{E}_{\mathcal{Q}}\{G_h^*(\mathcal{B})^2\}\Big] \\
&\quad + \Big[\sum_{h=1}^{\infty}\sum_{l \neq h} \pi_h(\mathbf{x}; \mathbf{V}_h, \mathbf{\Gamma}_h)\pi_l(\mathbf{x}; \mathbf{V}_l, \mathbf{\Gamma}_l)\mathrm{E}_{\mathcal{Q}}\{G_h^*(\mathcal{B})\}\mathrm{E}_{\mathcal{Q}}\{G_l^*(\mathcal{B})\}\Big] \\
&= \Big(\sum_{h=1}^{\infty} \pi_h(\mathbf{x}; \mathbf{V}_h, \mathbf{\Gamma}_h)^2 \Big[\mathrm{E}_{\mathcal{Q}}\{G_h^*(\mathcal{B})^2\} - G_0(\mathcal{B})^2\Big]\Big) + G_0(\mathcal{B})^2 \\
&= ||\pi(\mathbf{x}; \mathbf{V}, \mathbf{\Gamma})||^2 \mathrm{V}_{\mathcal{Q}(\mathcal{B})} + G_0(\mathcal{B})^2,
\end{aligned} \tag{6}$$

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

$$\begin{aligned}
\mathrm{corr}\{G_{\mathbf{x}}(\mathcal{B}), G_{\mathbf{x}'}(\mathcal{B}) \,|\, \mathbf{V}, \mathbf{\Gamma}\} &= \frac{\sum_{h=1}^{\infty} \pi_h(\mathbf{x}; \mathbf{V}_h, \mathbf{\Gamma}_h)\pi_h(\mathbf{x}'; \mathbf{V}_h, \mathbf{\Gamma}_h)}{\Big\{\sum_{h=1}^{\infty} \pi_h(\mathbf{x}; \mathbf{V}_h, \mathbf{\Gamma}_h)^2\Big\}^{1/2}\Big\{\sum_{h=1}^{\infty} \pi_h(\mathbf{x}'; \mathbf{V}_h, \mathbf{\Gamma}_h)^2\Big\}^{1/2}} \\
&= \frac{< \pi(\mathbf{x}; \mathbf{V}, \mathbf{\Gamma}), \pi(\mathbf{x}'; \mathbf{V}, \mathbf{\Gamma}) >}{||\pi(\mathbf{x}; \mathbf{V}, \mathbf{\Gamma})|| \cdot ||\pi(\mathbf{x}'; \mathbf{V}, \mathbf{\Gamma})||} = \rho(\mathbf{x}, \mathbf{x}'; \mathbf{V}, \mathbf{\Gamma}).
\end{aligned} \tag{7}$$

From the Cauchy-Schwarz inequality, $\rho(\mathbf{x}, \mathbf{x}'; \mathbf{V}, \mathbf{\Gamma}) \leq 1$, with the value → 1 in the limit as $\mathbf{x} \to \mathbf{x}'$, assuming $\lim_{\mathbf{x} \to \mathbf{x}'} K(\mathbf{x}, \Gamma) = K(\mathbf{x}', \Gamma)$, 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 $\mathcal{B}$, so that a single quantity can be reported for each $\mathbf{x}, \mathbf{x}'$ 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 KSBP, it is interesting to marginalize out the random weights, $\mathbf{V}$, and random locations, $\mathbf{\Gamma}$. Let $K_h(\mathbf{x}) \sim F_{\mathbf{x}}$ denote the random variable obtained in the transformation from $\Gamma_h \sim H$ to $K(\mathbf{x}, \Gamma_h)$. Because the random locations are iid, we have $K_h(\mathbf{x}) \overset{iid}{\sim} F_{\mathbf{x}}$, for $h = 1, \ldots, \infty$. In addition, the random variables $K_h(\mathbf{x})$ and $K_h(\mathbf{x}')$ are dependent, while $K_h(\mathbf{x})$ and $K_l(\mathbf{x}')$ are independent, for $h \neq l$.

Letting $U_h(\mathbf{x}) = V_h K_h(\mathbf{x})$, for $h = 1, \ldots, \infty$, and $P_h(\mathbf{x}) = U_h(\mathbf{x}) \prod_{l<h}\{1 - U_l(\mathbf{x})\}$, we obtain the following alternative representation of the KSBP:

$$G_{\mathbf{x}} \;=\; \sum_{h=1}^{\infty} P_h(\mathbf{x}) G_h^*, \quad G_h^* \sim \mathcal{Q}, \tag{8}$$

with dependence in the random weights, $\mathbf{P}(\mathbf{x}) = \{P_h(\mathbf{x}), h = 1, \ldots, \infty\}$ and $\mathbf{P}(\mathbf{x}') = \{P_h(\mathbf{x}'), h = 1, \ldots, \infty\}$, arising through dependence between the components $U_h(\mathbf{x})$ and $U_h(\mathbf{x}')$, for $h = 1, \ldots, \infty$. In the sequel, we focus on the case in which $V_h \overset{iid}{\sim} \text{beta}(a, b)$, for $h = 1, \ldots, \infty$.

Theorem 1. Let $\mu(\mathbf{x}) = \mathrm{E}\{U_h(\mathbf{x})\}$ and $\mu(\mathbf{x}, \mathbf{x}') = \mathrm{E}\{U_h(\mathbf{x})U_h(\mathbf{x}')\}$. Then, for any Borel set $\mathcal{B}$, we have

$$\mathrm{E}\{G_{\mathbf{x}}(\mathcal{B})G_{\mathbf{x}'}(\mathcal{B})\} = \frac{\mu(\mathbf{x}, \mathbf{x}')\mathrm{V}_{\mathcal{Q}(\mathcal{B})}}{\mu(\mathbf{x}) + \mu(\mathbf{x}') - \mu(\mathbf{x}, \mathbf{x}')} + G_0(\mathcal{B})^2$$

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

$$\mathrm{V}\{G_{\mathbf{x}}(\mathcal{B})\} = \frac{\mu^{(2)}(\mathbf{x})\mathrm{V}_{\mathcal{Q}(\mathcal{B})}}{2\mu(\mathbf{x}) - \mu^{(2)}(\mathbf{x})}, \tag{9}$$

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

$$\mathrm{corr}\{G_{\mathbf{x}}(\mathcal{B}), G_{\mathbf{x}'}(\mathcal{B})\} = \Big[\frac{\mu(\mathbf{x}, \mathbf{x}')}{\mu(\mathbf{x}) + \mu(\mathbf{x}') - \mu(\mathbf{x}, \mathbf{x}')}\Big]\Big[\frac{\{2\mu(\mathbf{x}) - \mu^{(2)}(\mathbf{x})\}\{2\mu(\mathbf{x}') - \mu^{(2)}(\mathbf{x}')\}}{\mu^{(2)}(\mathbf{x})\mu^{(2)}(\mathbf{x}')}\Big]^{1/2}. \tag{10}$$

Note that this expression is free of $\mathcal{B}$ and only depends on the expectation of $U_h(\mathbf{x})$ and $U_h(\mathbf{x})U_h(\mathbf{x}')$. Recalling that $U_h(\mathbf{x}) = V_h K_h(\mathbf{x})$ and focusing on $V_h \sim \text{beta}(1, \lambda)$, we obtain the modified expression:

$$\mathrm{corr}\{G_{\mathbf{x}}(\mathcal{B}), G_{\mathbf{x}'}(\mathcal{B})\} = \frac{\kappa(\mathbf{x}, \mathbf{x}')\Big\{(2+\lambda)\frac{\kappa(\mathbf{x})}{\kappa_2(\mathbf{x})} - 1\Big\}^{1/2}\Big\{(2+\lambda)\frac{\kappa(\mathbf{x}')}{\kappa_2(\mathbf{x}')} - 1\Big\}^{1/2}}{(1 + \lambda/2)\{\kappa(\mathbf{x}) + \kappa(\mathbf{x}')\} - \kappa(\mathbf{x}, \mathbf{x}')}, \tag{11}$$

where $\kappa(\mathbf{x}) = \mathrm{E}\{K_h(\mathbf{x})\}$, $\kappa_2(\mathbf{x}) = \mathrm{E}\{K_h(\mathbf{x})^2\}$ and $\kappa(\mathbf{x}, \mathbf{x}') = \mathrm{E}\{K_h(\mathbf{x})K_h(\mathbf{x}')\}$. 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 $U_h(\mathbf{x})$ are free from $\mathbf{x}$, expression (11) reduces to

$$\operatorname{corr}\{G_{\mathbf{x}}(\mathcal{B}), G_{\mathbf{x}'}(\mathcal{B})\} = \left[\frac{(2+\lambda)\kappa}{\kappa_2} - 1\right] \Big/ \left[\frac{(2+\lambda)\kappa}{\kappa(\mathbf{x},\mathbf{x}')} - 1\right], \tag{12}$$

dropping the dependency on $\mathbf{x}$ in κ and κ_2.

For some special cases of H and K, the moments of $U_h(\mathbf{x})$ and $U_h(\mathbf{x})U_h(\mathbf{x}')$ can be calculated in closed form, so that the above expressions are also available in closed form. Theorem 2 focuses on rectangular kernels, with $V_h \sim \text{beta}(1,\lambda)$ and the predictor space $\mathcal{X}$ assumed to be bounded, focusing on the unit hypercube without loss of generality. Similar closed form results can be obtained for Gaussian kernels (details available from the authors upon request).

Theorem 2. Suppose $\mathcal{X} = [0,1]^p$ and $\mathcal{L} = \bigotimes_{j=1}^{p}[-\psi_j, 1+\psi_j]$, with $\psi_j > 0$ for $j = 1,\ldots,p$ so that $\mathcal{X} \subset \mathcal{L}$. Assume H is a uniform probability measure and let $K(\mathbf{x},\Gamma_h) = 1\big(|x_j - \Gamma_{hj}| < \psi_j, j = 1,\ldots,p\big)$. Then, for any $\mathbf{x} \in \mathcal{X}$

$$\begin{aligned} \mathrm{E}\{U_h(\mathbf{x})^m\} &= \Big(\prod_{l=1}^{m} \frac{l}{\lambda + l}\Big) \prod_{j=1}^{p} \Big(\frac{2\psi_j}{1+2\psi_j}\Big) \\ \mathrm{E}\{U_h(\mathbf{x})U_h(\mathbf{x}')\} &= \Big(\frac{1}{1+\lambda}\Big)\Big(\frac{2}{2+\lambda}\Big) \prod_{j=1}^{p} \Big(\frac{\Delta_j(x_j,x_j')}{1+2\psi_j}\Big), \end{aligned}$$

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 Theorem 2, it is apparent that the moments of $U_h(\mathbf{x})$ are free from $\mathbf{x}$, while the expectation of $U_h(\mathbf{x})U_h(\mathbf{x}')$ depends only on the distance between $\mathbf{x}$ and $\mathbf{x}'$. Calculating the variance, we obtain the simple expression

$$\mathrm{V}\{G_{\mathbf{x}}(\mathcal{B})\} = \frac{\mathrm{V}_{\mathcal{Q}(\mathcal{B})}}{1+\lambda}. \tag{13}$$

In addition, the correlation coefficient takes the form:

$$\rho(\mathbf{x}-\mathbf{x}'; \lambda, \boldsymbol{\psi}) = \operatorname{corr}\{G_{\mathbf{x}}(\mathcal{B}), G_{\mathbf{x}'}(\mathcal{B})\} = \frac{1+\lambda}{(2+\lambda)\prod_{j=1}^{p} \frac{2\psi_j}{\Delta_j(x_j,x_j')} - 1}, \tag{14}$$

which is a function of the distance between $\mathbf{x}$ and $\mathbf{x}'$. When $(\mathbf{x}-\mathbf{x}') \notin C_{\boldsymbol{\psi}} = \otimes_{j=1}^{p}[-2\psi_j, 2\psi_j]$, $\rho(\mathbf{x}-\mathbf{x}'; \lambda, \boldsymbol{\psi}) = 0$. In addition, in the limit as $\mathbf{x} \to \mathbf{x}'$, $\Delta_j(x_j, x_j') \to 2\psi_j$ and $\rho(\mathbf{x}-\mathbf{x}'; \lambda, \boldsymbol{\psi}) \to 1$. Hence, the correlation coefficient is bounded between 0 and 1, depending on the distance between the predictor values.

2.4 *Truncations*

It is often useful to consider finite approximations to infinite stick-breaking processes. For example, truncation approximations form the basis for commonly-used computational algorithms for DP mixture models. For previous work on truncations of stick-breaking random measures, refer to Muliere and Tardella (1998) and Ishwaran and James (2001) among others. Our development follows that of Ishwaran and James (2001).

We focus on the following truncation approximation to (8):

$$G_{\mathbf{x}} = \sum_{h=1}^{N} P_h(\mathbf{x}) G_h^* + \Big(1 - \sum_{h=1}^{N} P_h(\mathbf{x})\Big) G_0^*, \quad G_h^* \sim \mathcal{Q}, \quad h = 0, \ldots, N, \tag{15}$$

resulting in $\mathcal{G}_{\mathcal{X}} \sim \mathcal{P}^N$, where $\lim_{N\to\infty} \mathcal{P}^N \to \mathcal{P}$, with $\mathcal{P}$ a KSBP. Letting $P_0(\mathbf{x})$ denote the probability mass on G_0^*, we have $\sum_{h=0}^{N} P_h(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{X}$.

Theorem 3. Let $P_h(\mathbf{x}), h = 1, \ldots, \infty$, denote the infinite sequence of random weights in expression (8). For each positive integer $N \geq 1$ and positive integer $m \geq 1$, let

$$T_N(m, \mathbf{x}) = \left(\sum_{h=N}^{\infty} P_h(\mathbf{x})\right)^m, \quad W_N(m, \mathbf{x}) = \sum_{h=N}^{\infty} P_h(\mathbf{x})^m.$$

It follows that

$$\mathrm{E}\{T_N(m, \mathbf{x})\} = \mathrm{E}[\{1-U_1(\mathbf{x})\}^m]^{N-1}, \quad \mathrm{E}\{W_N(m, \mathbf{x})\} = \frac{\mathrm{E}\{T_N(m, \mathbf{x})\}\mathrm{E}\{U_1(\mathbf{x})^m\}}{1 - \mathrm{E}[\{1 - U_1(\mathbf{x})\}^m]}.$$

Refer to the appendix for a proof. Both expectations decrease exponentially fast in N suggesting that an accurate approximation may be obtained for moderate N.

To compare this rate to that corresponding to the N-truncation of a Dirichlet process prior, we recall that $U_h(\mathbf{x}) = V_h K_h(\mathbf{x})$, with $V_h \sim \text{beta}(1, \lambda)$, the stick-breaking random weights from a $DP(\lambda H)$ random measure. It is straightforward to show that

$$\mathrm{E}[\{1 - U_1(\mathbf{x})\}^m] > \mathrm{E}\{(1 - V_1)^m\}, \quad \text{for any } \mathbf{x} \in \mathcal{X}.$$

It follows that $\mathrm{E}\{T_N(m, \mathbf{x})\}$ for a $KSBP(\alpha, G_0, \lambda, H)$ measure is bounded below by the respective value for a $DP(\lambda H)$ measure. We use the $KSPB(\alpha, G_0, \lambda, H)$ notation to refer to the KSBP with $V_h \sim \text{beta}(1, \lambda)$ and $G_h^* \sim DP(\alpha G_0)$. The tightness of the bound depends on the measure H and the kernel $K(\cdot)$. In the case in which $K(\mathbf{x}, \mathbf{x}') = \exp(-\psi||\mathbf{x}-\mathbf{x}'||^2)$ it is straightforward to show that $\mathrm{E}\{T_N(m, \mathbf{x})\}$ increases monotonically with increasing precision ψ. This is intuitive, as high values of ψ imply little borrowing of information across $\mathcal{X}$, necessitating a moderate to large number of random basis measures.

3. Clustering and Prediction Rules

As mentioned in Section 1, one of the most appealing and widely utilized properties of the Dirichlet process is the simple prediction rule shown in expression (3). In this section, we obtain a predictor-dependent prediction rule derived by marginalizing over the KSBP prior for $\mathcal{G}_{\mathcal{X}}$ shown in expression (4). For tractability, we focus on the special case in which $G_h^* = \delta_{\Theta_h}$, with $\Theta_h \sim G_0$, for $h = 1, \ldots, \infty$. In this case, 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:

$$\begin{aligned} (\phi_i \,|\, \mathbf{x}_i) &\overset{ind}{\sim} G_{\mathbf{x}_i}, \quad i = 1, \ldots, n \\ \mathcal{G}_{\mathcal{X}} &\sim \mathcal{P}, \end{aligned} \tag{16}$$

where $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, $\mathcal{P}$ is a KSBP characterized in terms of a precision parameter, λ, a kernel, K, and a base measure, G_0, focusing on the case in which $\Gamma_h \overset{iid}{\sim} \text{beta}(1, \lambda)$, for

$h = 1, \ldots, \infty$. Note that (16) can be equivalently expressed as:

$$
\begin{aligned}
(\phi_i \,|\, Z_i, \mathbf{x}_i, \mathbf{\Theta}) &\overset{ind}{\sim} \delta_{\Theta_{Z_i}}, \quad i = 1, \ldots, n \\
(Z_i \,|\, \mathbf{x}_i, \mathbf{V}, \mathbf{\Gamma}) &\overset{ind}{\sim} \sum_{h=1}^{\infty} \pi_h(\mathbf{x}_i; \mathbf{V}_h, \mathbf{\Gamma}_h)\delta_h \\
V_h &\overset{iid}{\sim} \text{beta}(1, \lambda) \\
\Gamma_h &\overset{iid}{\sim} H \\
\Theta_h &\overset{iid}{\sim} G_0,
\end{aligned}
\tag{17}
$$

where Z_i indexes the (unobserved) location for subject i. It follows that $\Pr(\phi_i \in \cdot \,|\, \mathbf{x}_i) = G_0(\cdot)$. As a notation to aid in describing marginal properties, we let

$$
\mu_{\mathcal{I}} = \mathrm{E}\Big\{ \prod_{i \in \mathcal{I}} U_h(\mathbf{x}_i) \Big\}, \tag{18}
$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is an $n \times p$ matrix and $\mathcal{I} \subset \{1, \ldots, n\}$ 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 using a straightforward generalization of Theorem 2.

> **Lemma 2**. The probability that subjects i and j belong to the same cluster conditionally on the subjects predictor values, but marginalizing out $\mathcal{P}$, is
>
> $$\Pr(\phi_i = \phi_j \,|\, \mathbf{x}_i, \mathbf{x}_j) = \frac{\mu_{ij}}{\mu_i + \mu_j - \mu_{ij}}, \quad \forall i, j \in \{1, \ldots, n\},$$
>
> with μ_i, μ_j, μ_{ij} defined in (18).

Under the conditions of Theorem 2, the expression in Lemma 2 takes the form:

$$
\Pr(\phi_i = \phi_j \,|\, \mathbf{x}_i, \mathbf{x}_j) = \frac{\prod_{j=1}^{p} \Delta_j(x_j, x_j')}{(2 + \lambda) \prod_{j=1}^{p} (2\psi_j) - \prod_{j=1}^{p} \Delta_j(x_j, x_j')}, \tag{19}
$$

which reduces to 0 if $\mathbf{x}_i - \mathbf{x}_j \notin C_{\boldsymbol{\psi}} = \bigotimes_{j=1}^{p} [-2\psi_j, 2\psi_j]$, as $\mathbf{x}_i$ and $\mathbf{x}_j$ are not in the same *neighborhood* in that case. In addition, as $\mathbf{x}_i \to \mathbf{x}_j$, $\Pr(\phi_i = \phi_j \,|\, \mathbf{x}_i, \mathbf{x}_j) \to 1/(1 + \lambda)$, which corresponds to the clustering probability for the DP prediction rule when $\phi_i \sim DP(\lambda G_0)$.

Theorem 4. 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, the following prediction rule is obtained on marginalizing out $\mathcal{P}$:

$$P(\phi_i \in \cdot \,|\, \phi_1,\ldots,\phi_{i-1},\mathbf{x}_1,\ldots,\mathbf{x}_{i-1}) =$$
$$\Big(1-\sum_{r=2}^{i}(-1)^r\sum_{\mathcal{I}\in\mathcal{N}_i^{(r,i)}}\omega_{\mathcal{I}}\Big)G_0(\cdot)+\sum_{j=1}^{i-1}\Big(\sum_{r=2}^{i}(-1)^r\sum_{\mathcal{I}\in\mathcal{N}_{i,j}^{(r,i)}}\frac{\omega_{\mathcal{I}}}{r-1}\Big)\delta_{\phi_j}(\cdot).$$

Note that the resulting law of $(\phi_1,\ldots,\phi_n)$ is not dependent on the ordering of the subjects, and one can obtain equivalent expressions to that shown in Theorem 4 for any ordering. For example, this allows one to obtain full conditional prior distributions for ϕ_i given $\boldsymbol{\phi}^{(i)} = \{\phi_1,\ldots,\phi_{i-1},\phi_{i+1},\ldots,\phi_n\}$ and $\mathbf{X}$. Updating these conditional priors with the data, the collapsed Gibbs sampler of MacEachern (1994) can be applied directly for posterior computation.

Under the conditions of Theorem 2, we obtain the following simple expression for $\mu_{\mathcal{I}}$:

$$\begin{aligned}\mu_{\mathcal{I}} &= \mathrm{E}\Big\{\prod_{i\in\mathcal{I}}U_h(\mathbf{x}_i)\Big\} = \mathrm{E}(V_h^{\#\mathcal{I}})\int\prod_{i\in\mathcal{I}}\prod_{j=1}^{p}1(|x_{ij}-\Gamma_{hj}|<\psi_j)dH(\Gamma_h)\\ &= \Big\{\prod_{l=1}^{\#\mathcal{I}}\frac{l}{l+\lambda}\Big\}\Big\{\prod_{j=1}^{p}\frac{\Delta_j(\mathbf{X}_{\mathcal{I}})}{1+2\psi_j}\Big\}, \qquad (20)\end{aligned}$$

where $\Delta_j(\mathbf{X}_{\mathcal{I}}) = \max[0, 2\psi_j + \min_{i\in\mathcal{I}}\{x_{ij}\} - \max_{i\in\mathcal{I}}\{x_{ij}\}]$. From this result, one can show that the prediction rule from Theorem 4 reduces to the DP prediction rule in the special case in which $\mathbf{x}_i = \mathbf{x}$, for $i = 1,\ldots,i$.

4. Posterior Computation

4.1 *Background*

For DP mixture (DPM) models, there are two main strategies that have been used in developing algorithms for posterior computation: (1) the marginal approach; and (2) the conditional approach. Letting $\phi_i \sim G$, with $G \sim DP(\alpha G_0)$, the marginal approach avoids computation for the infinite-dimensional G by relying on the Pólya urn scheme, which is obtained marginalizing over the DP prior. The most widely used marginal algorithm is the generalized Pólya urn Gibbs sampler of MacEachern (1994) and West, Müller and Escobar (1994). Ishwaran and James (2001) extend this approach to a general class of stick-breaking measures, while Ishwaran and James (2003) consider species sampling priors.

The conditional approach avoids marginalizing over the prior, resulting in greater flexibility in computation and inferences (Ishwaran and Zarepour, 2000). Conditional algorithms typically rely on a truncation approximation to the stick-breaking representation in (1). In particular, to avoid the impossibility of conducting posterior computation for the infinitely-many parameters in (1), one approximates (1) by letting $V_N = 1$ and discarding the $N+1, \ldots, \infty$ terms. Refer to Ishwaran and James (2001) for a formal justification. Although the approximation can be shown to be highly accurate for DPM models for N sufficiently large, one must be conservative in choosing N, which may lead to unnecessary computation. In addition, by using a finite approximation, one is essentially fitting a finite mixture model. To avoid these problems, Papaspiliopoulos and Roberts (2006) recently proposed a retrospective MCMC algorithm, which avoids truncation.

In this section, we propose a conditional approach to posterior computation for KSBP models, relying on a combined MCMC algorithm that utilizes retrospective sampling and generalized Pólya urn sampling steps.

4.2 *MCMC Algorithm*

We focus on the general case in which $V_h \sim \text{beta}(a_h, b_h)$, K and H have arbitrary forms, and $\mathcal{Q}$ corresponds to a species sampling model. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ denote the $k \leq n$ unique values of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)'$, let $\mathcal{S}_i = h$ if $\phi_i = \theta_h$ denote that subject i is allocated to the hth unique value, with $\mathbf{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_n)'$, and let $\mathcal{C}_h = j$ denote that θ_h is an atom from G_j^*, with $\mathbf{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_k)'$. Let $\boldsymbol{\phi}^{(i)}, \boldsymbol{\theta}^{(i)}$, $\mathbf{S}^{(i)}$, $\mathbf{C}^{(i)}$, and $\mathbf{Z}^{(i)}$ correspond to the vectors $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, $\mathbf{S}$, $\mathbf{C}$, and $\mathbf{Z}$ 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: *occupied* locations, $\mathcal{I}_{oc} = \{j \in \mathcal{I} : n_j > 0\}$, and *vacant* locations, $\mathcal{I}_{vc} = \{j \in \mathcal{I} : n_j = 0\}$, so that $\mathcal{C}_h \in \mathcal{I}_{oc}$, for $h = 1, \ldots, k$.

Letting $\mathcal{N}_h = \{i : Z_i = h, i = 1, 2, \ldots, \infty\}$ denote 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 and James (2003) that

$$P(\phi_i \in \cdot \,|\, Z_i = h, \mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \boldsymbol{\theta}^{(i)}, \mathbf{X}) = l_{ih0} G_0(\cdot) + \sum_{j \in \mathcal{N}_h^{(i)}} l_{ihj} \delta_{\phi_j}(\cdot), \tag{21}$$

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 DP 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 (21) by marginalizing out Z_i, noting $\Pr(Z_i = h \,|\, \mathbf{x}_i, \mathbf{V}, \boldsymbol{\Gamma}) = \pi_h(\mathbf{x}_i; \mathbf{V}_h, \boldsymbol{\Gamma}_h) = \pi_{ih}$ for $h = 1, \ldots, \infty$, and grouping together the subjects with the same unique value:

$$P(\phi_i \in \cdot \,|\, \mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \boldsymbol{\theta}^{(i)}, \mathbf{X}) = w_{i0} G_0(\cdot) + \sum_{j=1}^{k^{(i)}} w_{ij} \delta_{\theta_j^{(i)}}(\cdot) + w_{i,k^{(i)}+1} G_0(\cdot), \tag{22}$$

with $k^{(i)}$ the length of $\boldsymbol{\theta}^{(i)}$ and the weights defined as follows:

$$w_{i0} = \sum_{h \in \mathcal{I}_{oc}^{(i)}} \pi_{ih}\, l_{ih0}, \;\; w_{ij} = \pi_{i,\mathcal{C}_j^{(i)}} \sum_{g : \mathcal{S}_g^{(i)} = j} l_{i\mathcal{C}_j^{(i)} g}, \;\; j = 1, \ldots, k^{(i)}, \;\; w_{i,k^{(i)}+1} = \sum_{h \in \mathcal{I}_{vc}^{(i)}} \pi_{ih}\, l_{ih0}. \tag{23}$$

Assuming the likelihood contribution for subject i is $f(y_i \,|\, \mathbf{x}_i, \phi_i)$, expression (23) can be updated to obtain a conditional posterior distribution for ϕ_i. From this posterior, we obtain

$$P(\mathcal{S}_i = j \,|\, \mathbf{y}, \mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \boldsymbol{\theta}^{(i)}, \mathbf{X}) = q_{ij}, \tag{24}$$

where $q_{ij} = c_i\, w_{ij}\, f_0(y_i \,|\, \mathbf{x}_i)$ for $j = 0, k^{(i)}+1$, and $q_{ij} = c_i\, w_{ij}\, f(y_i \,|\, \mathbf{x}_i, \theta_j^{(i)})$ for $j = 1, \ldots, k^{(i)}$, with $f_0(y_i \,|\, \mathbf{x}_i) = \int f(y_i \,|\, \mathbf{x}_i, \phi)\, dG_0(\phi)$ and c_i a normalizing constant. We update $\mathcal{S}_i$ by sampling based on (24). 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. Because there are infinitely many possibilities for this new location, we use a retrospective sampling approach, which follows along similar lines to Papaspiliopoulos and Roberts (2006).

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

$$(\theta_h \,|\, \mathbf{y}, \mathbf{S}, \mathbf{C}, k, \mathbf{X}) \;\propto\; \{\prod_{i:S_i = h} f(y_i | \mathbf{x}_i, \theta_h)\} G_0(\theta_h). \tag{25}$$

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 $A_{ih} \overset{ind}{\sim} Bernoulli(V_h)$ and $B_{ih} \overset{ind}{\sim} Bernoulli(K(\mathbf{x}_i, \Gamma_h))$, with $Z_i = \mathcal{C}_{\mathcal{S}_i} = \min\{h : A_{ih} = B_{ih} = 1\}$. Then, alternate between (i) sampling (A_{ih}, B_{ih}) from their conditional distribution given Z_i; (ii) updating V_h by sampling from the conditional posterior distribution:

$$\mathrm{beta}\Big(a_h + \sum_{i:Z_i \geq h} A_{ih}, b_h + \sum_{i:Z_i \geq h} (1 - A_{ih})\Big).$$

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

5. Simulation Examples

In this section, we use a KSBP mixture of normal linear regression models for conditional density estimation. In particular, let $f(y_i \,|\, \mathbf{x}_i, \phi_i) = (2\pi\tau^{-1})^{-1/2} \exp\{-\tau/2(y_i - \mathbf{x}_i'\boldsymbol{\beta}_i)\}$, with $\phi_i = \boldsymbol{\beta}_i \sim G_{\mathbf{x}_i}$ and $\mathcal{G}_{\mathcal{X}} \sim \mathcal{P}$, with $\mathcal{P}$ a KSBP 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 $\boldsymbol{\beta}$ and variance $\mathbf{\Sigma}_\beta$. In addition, we let $K(\mathbf{x}, \mathbf{x}') = \exp(-\psi ||\mathbf{x} - \mathbf{x}'||^2)$ and choose priors: $\pi(\tau) = \mathrm{gamma}(\tau; a_\tau, b_\tau)$, $\pi(\boldsymbol{\beta}) = \mathrm{N}(\boldsymbol{\beta}; \boldsymbol{\beta}_0, \mathbf{V}_{\beta_0})$, $\pi(\mathbf{\Sigma}_\beta^{-1}) = \mathcal{W}(\mathbf{\Sigma}_\beta^{-1}; (\nu_0 \mathbf{\Sigma}_0)^{-1}, \nu_0)$, the Wishart density with degrees of freedom ν_0 and $E(\mathbf{\Sigma}_{\beta}^{-1}) = \mathbf{\Sigma}_0^{-1}$, and $\pi(\psi) = \text{log-N}(\psi; \mu_\psi, \sigma_\psi^2)$.

Following Dunson et al., (2006), we simulate data for $n = 500$ subjects from a mixture of two normal linear regression models as follows:

$$f(y_i|\mathbf{x}_i) = e^{-2x_i} N(y_i; x_i, 0.01) + (1 - e^{-2x_i}) N(y_i; x_i^4, 0.04),$$

where $\mathbf{x}_i = (1, x_i)'$, with $p = 2$ and $x_i \sim$ unifom$(0, 1)$. We let $\lambda = 1$ and $\alpha = 1$ to favor few occupied basis locations and few clusters per location, $\mu_\psi = 2.5$, $\sigma_\psi^2 = 0.5$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{V}_{\beta_0} = (\mathbf{X}'\mathbf{X})^{-1}/n$, $\nu_0 = p$, $\Sigma_0^{-1} = I_{p\times p}$, and $a_\tau = b_\tau = 0.1$, and choose every point from 0 to 1 with increment of 0.02 as $\mathbf{\Gamma}^*$, with $T = 51$ and probability weight $a_l = 1/T$. The MCMC algorithm described in Section 4 was run for 30,000 iterations, with a burn-in of 8,000 iterations discarded. Based on examination of trace plots, convergence and mixing were good. Figure 1 plots the true density (dotted line) and estimated predictive density (solid line), along with pointwise 99% credible intervals (dashed lines). An x-y plot of the data along with the estimated and true mean curve is also provided. Even though the sample size was only 500 the estimates are good.

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. DDT is effective against malaria-transmitting mosquitoes, 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 Figure 2 for the 2313 children in the study, excluding the children having GAD$>$ 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 cutoff, so that deliveries occurring prior to 37 weeks of completed gestation were classified as preterm. Categorizing DDE into 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 < 37 week interval have greater risk of mortality and morbidity. Hence, from a public health and clinical 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:

$$\begin{aligned} f(y_i \,|\, \mathbf{x}_i) &= \int N(y_i; \mathbf{x}_i' \boldsymbol{\beta}_i, \tau^{-1}) dG_{\mathbf{x}_i}(\boldsymbol{\beta}_i) \\ \mathcal{G}_{\mathcal{X}} &\sim \mathcal{P}, \end{aligned} \tag{26}$$

where $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \Re^p\}$, y_i is the normalized gestational age at delivery, $\mathbf{x}_i = (1, dde_i, age_i)'$, dde_i is the normalized DDE dose for child i, age_i is the normalized age of the mother, and $\mathcal{P}$ is a KSBP, with a Gaussian kernel and $\mathcal{Q}$ corresponding to a $DP(\alpha G_0)$. Prior specification and other details are as described in Section 5 for the simulation examples.

As in the simulation examples, convergence was rapid and mixing was good based on examination of trace plots. Figure 3 shows the trace plots for mean parameters in G_0, the number of occupied locations, the total number of clusters, and the smoothing parameter ψ. 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 4 shows the estimated conditional densities of gestational age at delivery for a range of DDE values. Based on these plots, there is some suggestion of an increasing left tail with dose, representing increasing risk of premature delivery at higher exposure values. At very high exposures, data are sparse and the credible intervals are much wider. To more directly assess impact of DDE on the left tail, Figure 5 shows dose response curves for the probability $Y < T$ for different choices of cutoff T. For early preterm birth at < 33 weeks, the dose response curve is flat except at high doses where the credible interval is wide. As the cutoff increases, the dose response becomes more significant. Hence, the Longnecker et al. (2001) result can be attributed to an increasing risk of late preterm births with dose of DDE.

Although model (26) is extremely flexible, a potential concern is that the prior structure may lead to lack of fit. To assess model adequacy, we implemented a recently developed pivotal statistic-based approach (Johnson, 2006), which showed excellent fit.

7. Discussion

The article proposed 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 modeling and spatial data analysis.

The KSBP should provide a useful alternative to recently developed generalized stick-breaking processes, which allow predictors and spatial dependence (Griffin and Steel, 2006; Duan et al., 2005). An advantage of the KSBP 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. We also obtained a predictor-dependent urn scheme, which generalizes the Pólya urn scheme (Blackwell and MacQueen, 1973). In future work, it will be interesting to use this urn scheme for computation and clustering without need to explicitly consider the random weights and locations in the stick-breaking representation.

Acknowledgments

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

Proof of Lemma 1.

Following a related approach to Ishwaran and James (2001), we first note that

$$1-\sum_{h=1}^{N-1}\pi_h(\mathbf{x};\mathbf{V}_h,\boldsymbol{\Gamma}_h)=\{1-V_1K(\mathbf{x},\Gamma_1)\}\cdots\{1-V_{N-1}K(\mathbf{x},\Gamma_{N-1})\}.$$

Then, taking logs on both sides and $N\rightarrow\infty$,

$$\sum_{h=1}^{\infty}\pi_h(\mathbf{x};\mathbf{V}_h,\boldsymbol{\Gamma}_h)=1 \quad \text{a.s} \quad \text{iff} \quad \sum_{h=1}^{\infty}\log\{1-V_hK(\mathbf{x},\Gamma_h)\}=-\infty \quad \text{a.s.}$$

The summation on the right is over independent random variables and by the Kolmogorov three series theorem equals $-\infty$ a.s. iff $\sum_{h=1}^{\infty}\mathrm{E}[\log\{1-V_hK(\mathbf{x},\Gamma_h)\}]=-\infty$.

Proof of Theorem 1.

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

$$\begin{aligned}
&\mathrm{E}\{G_{\mathbf{x}}(\mathcal{B})G_{\mathbf{x}'}(\mathcal{B})\}\\
&\quad=\mathrm{E}\Bigg(\Big[\sum_{h=1}^{\infty}U_h(\mathbf{x})\Big\{\prod_{l=1}^{h-1}(1-U_l(\mathbf{x}))\Big\}Q_h\Big]\Big[\sum_{h=1}^{\infty}U_h(\mathbf{x}')\Big\{\prod_{l=1}^{h-1}(1-U_l(\mathbf{x}'))\Big\}Q_h\Big]\Bigg)\\
&\quad=\mathrm{E}\Bigg(\sum_{h=1}^{\infty}U_h(\mathbf{x})U_h(\mathbf{x}')\prod_{l=1}^{h-1}\{1-U_h(\mathbf{x})\}\{1-U_h(\mathbf{x}')\}Q_h^2\Bigg)\\
&\qquad+\mathrm{E}\Bigg(\sum_{h=1}^{\infty}\sum_{l=1}^{h-1}U_h(\mathbf{x})\Big[\prod_{r=1}^{l-1}\{1-U_r(\mathbf{x})\}\{1-U_r(\mathbf{x}')\}\Big]\Big[U_l(\mathbf{x}')-U_l(\mathbf{x})U_l(\mathbf{x}')\Big]\\
&\qquad\quad\times\Big[\prod_{s=l+1}^{h-1}\{1-U_s(\mathbf{x})\}\Big]Q_hQ_l\Bigg)+\mathrm{E}\Bigg(\sum_{h=1}^{\infty}\sum_{l=h+1}^{\infty}U_l(\mathbf{x}')\Big[\prod_{r=1}^{h-1}\{1-U_r(\mathbf{x})\}\{1-U_r(\mathbf{x}')\}\Big]\\
&\qquad\quad\times\Big[U_h(\mathbf{x})-U_h(\mathbf{x})U_h(\mathbf{x}')\Big]\Big[\prod_{s=h+1}^{l-1}\{1-U_s(\mathbf{x}')\}\Big]Q_hQ_l\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{h=1}^{\infty} \mu(\mathbf{x},\mathbf{x}')\{1-\mu(\mathbf{x})-\mu(\mathbf{x}')+\mu(\mathbf{x},\mathbf{x}')\}^{h-1}\mathrm{E}(Q_h^2) \\
&\quad +\mu(\mathbf{x})\{\mu(\mathbf{x}')-\mu(\mathbf{x},\mathbf{x}')\}\sum_{l=1}^{\infty}\sum_{h=l+1}^{\infty}\{1-\mu(\mathbf{x})-\mu(\mathbf{x}')+\mu(\mathbf{x},\mathbf{x}')\}^{l-1}\{1-\mu(\mathbf{x})\}^{h-l-1}Q_0^2 \\
&\quad +\mu(\mathbf{x}')\{\mu(\mathbf{x})-\mu(\mathbf{x},\mathbf{x}')\}\sum_{h=1}^{\infty}\sum_{l=h+1}^{\infty}\{1-\mu(\mathbf{x})-\mu(\mathbf{x}')+\mu(\mathbf{x},\mathbf{x}')\}^{h-1}\{1-\mu(\mathbf{x}')\}^{l-h-1}Q_0^2 \\
&= \frac{\mu(\mathbf{x},\mathbf{x}')\mathrm{V}_{\mathcal{Q}(\mathcal{B})}}{\mu(\mathbf{x})+\mu(\mathbf{x}')-\mu(\mathbf{x},\mathbf{x}')}+Q_0^2,
\end{aligned}
$$

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

Proof of Theorem 2.

Under the assumption that $V_h \sim \text{beta}(1,\lambda)$, we have

$$
\begin{aligned}
\mathrm{E}\{U_h(\mathbf{x})^m\} &= \mathrm{E}(V_h^m)\mathrm{E}\{K_h(\mathbf{x})^m\} = \Big(\prod_{l=1}^{m}\frac{l}{\lambda+l}\Big)\int_{\mathcal{L}}\prod_{j=1}^{p}1(|x_j-\Gamma_{hj}|<\psi_j)dH(\Gamma_h) \\
&= \Big(\prod_{l=1}^{m}\frac{l}{\lambda+l}\Big)\prod_{j=1}^{p}\int_{-\psi_j}^{1+\psi_j}1(|x_j-\Gamma_{hj}|<\psi_j)\,\frac{1}{1+2\psi_j}\,d\Gamma_{hj} \\
&= \Big(\prod_{l=1}^{m}\frac{l}{\lambda+l}\Big)\prod_{j=1}^{p}\Big(\frac{2\psi_j}{1+2\psi_j}\Big),
\end{aligned}
$$

as required. The expression for $\mathrm{E}\{U_h(\mathbf{x})U_h(\mathbf{x}')\}$ follows trivially using the same approach.

Proof of Theorem 3.

The random measure $G_{\mathbf{x}}$ can be expressed as

$$
\begin{aligned}
G_{\mathbf{x}}(\cdot) &= U_1(\mathbf{x})G_1^*(\cdot)+\{1-U_1(\mathbf{x})\}[\widetilde{U}_1(\mathbf{x})\widetilde{G}_1(\cdot)+\{1-\widetilde{U}_1(\mathbf{x})\}\widetilde{U}_2(\mathbf{x})\widetilde{G}_2(\cdot)+\dots], \\
&\overset{\mathcal{D}}{=} U_1(\mathbf{x})G_1^*(\cdot)+\{1-U_1(\mathbf{x})\}\widetilde{G}_{\mathbf{x}}(\cdot),
\end{aligned}
$$

where $\widetilde{G}_{\mathbf{x}}$ is a KSBP measure, with $U_1(\mathbf{x})$, G_1^* and $\widetilde{G}_{\mathbf{x}}$ mutually independent. Using the same trick, it can be shown that $W_1(m,\mathbf{x}) \overset{\mathcal{D}}{=} U_1(\mathbf{x})^m + \{1-U_1(\mathbf{x})\}^m W_1(m,\mathbf{x})$, with $U_1(\mathbf{x})$ and $W_1(m,\mathbf{x})$ mutually independent. Taking expectations with $\mathbf{x}$ fixed,

$$
\mathrm{E}\{W_1(m,\mathbf{x})\} = \frac{\mathrm{E}\{U_1(\mathbf{x})^m\}}{1-\mathrm{E}[\{1-U_1(\mathbf{x})\}^m]}.
$$

For $N \geq 2$, we can obtain the expression

$$W_N(m, \mathbf{x}) \overset{\mathcal{D}}{=} \Big[\prod_{h=1}^{N-1} \{1 - U_h(\mathbf{x})\}^m\Big] W_1(m, \mathbf{x}),$$

with $U_1(\mathbf{x}), \ldots, U_{N-1}(\mathbf{x}), W_1(m, \mathbf{x})$ mutually independent, and the term in $[\cdot]$ equal to $T_N(m, \mathbf{x})$. Thus, taking expectations with $\mathbf{x}$ fixed

$$\mathrm{E}\{W_N(m, \mathbf{x})\} = \mathrm{E}[\{1 - U_1(\mathbf{x})\}^m]^{N-1}\mathrm{E}\{W_1(m, \mathbf{x})\} = \frac{\mathrm{E}\{T_N(m, \mathbf{x})\}\mathrm{E}\{U_1(\mathbf{x})^m\}}{1 - \mathrm{E}[\{1 - U_1(\mathbf{x})\}^m]}.$$

Proof of Lemma 2.

Under formulation (17), we have

$$\begin{aligned}
\Pr(\phi_i = \phi_j \,|\, \mathbf{x}_i, \mathbf{x}_j) &= \int \Pr(Z_i = Z_j \,|\, \mathbf{x}_i, \mathbf{x}_j, \mathbf{V}, \mathbf{\Gamma}) d\pi(\mathbf{V})\, d\pi(\mathbf{\Gamma}) \\
&= \mathrm{E}\bigg(\sum_{h=1}^{\infty}\Big[U_h(\mathbf{x}_i)U_h(\mathbf{x}_j)\prod_{l<h}\Big\{1 - U_l(\mathbf{x}_i)\Big\}\Big\{1 - U_l(\mathbf{x}_j)\Big\}\Big]\bigg) \\
&= \mu_{ij}\sum_{h=1}^{\infty}\prod_{l<h}\Big(1 - \mu_i - \mu_j + \mu_{ij}\Big) \\
&= \mu_{ij}\sum_{h=0}^{\infty}\Big(1 - \mu_i - \mu_j + \mu_{ij}\Big)^h \\
&= \frac{\mu_{ij}}{\mu_i + \mu_j - \mu_{ij}}.
\end{aligned}$$

Proof of Theorem 4.

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

$$\begin{aligned}
\mathrm{E}\Big(\sum_{h=1}^{\infty}\prod_{j\in\mathcal{I}} P(Z_j = h)\Big) &= \sum_{h=1}^{\infty}\mathrm{E}\Big(\prod_{j\in\mathcal{I}} U_h(\mathbf{x}_j)\prod_{l=1}^{h-1}\{1 - U_l(\mathbf{x}_j)\}\Big) \\
&= \sum_{h=1}^{\infty}\mathrm{E}\Big(\prod_{j\in\mathcal{I}} U_h(\mathbf{x}_j)\Big)\prod_{l=1}^{h-1}\mathrm{E}\Big(\prod_{j\in\mathcal{I}}\{1 - U_l(\mathbf{x}_j)\}\Big) \\
&= \mu_{\mathcal{I}}\sum_{h=1}^{\infty}\mathrm{E}\Big(\prod_{j\in\mathcal{I}}\{1 - U_l(\mathbf{x}_j)\}\Big)^{h-1} \\
&= \frac{\mu_{\mathcal{I}}}{1 - \sum_{t=0}^{\#\mathcal{I}}(-1)^t\sum_{\mathcal{J}\in\mathcal{I}_t}\mu_{\mathcal{J}}} = \frac{\mu_{\mathcal{I}}}{\sum_{t=1}^{\#\mathcal{I}}(-1)^{t-1}\sum_{\mathcal{J}\in\mathcal{I}_t}\mu_{\mathcal{J}}} = \omega_{\mathcal{I}},
\end{aligned}$$

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 $\overline{\mathcal{K}} = \{1,\ldots,i\}\setminus\mathcal{K}$. Then, letting $Z_i = Z_j$ for all $j \in \mathcal{K}\setminus\{i\}$ 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:

$$\begin{aligned}
&\mathrm{E}\bigg(\sum_{h=1}^{\infty}\prod_{k\in\mathcal{K}}P(Z_k=h)\prod_{k\in\overline{\mathcal{K}}}\{1-P(Z_k=h)\}\bigg)\\
&= \mathrm{E}\bigg(\sum_{h=1}^{\infty}\prod_{k\in\mathcal{K}}P(Z_k=h)\bigg\{\sum_{s=0}^{\#\overline{\mathcal{K}}}(-1)^s\sum_{\mathcal{L}\in\overline{\mathcal{K}}_s}\prod_{l\in\mathcal{L}}P(Z_l=h)\bigg\}\bigg)\\
&= \sum_{s=0}^{\#\overline{\mathcal{K}}}(-1)^s\sum_{\mathcal{L}\in\overline{\mathcal{K}}_s}\mathrm{E}\bigg(\sum_{h=1}^{\infty}\prod_{k\in\mathcal{L}\bigcup\mathcal{K}}P(Z_k=h)\bigg) = \sum_{s=0}^{\#\overline{\mathcal{K}}}(-1)^s\sum_{\mathcal{L}\in\overline{\mathcal{K}}_s}\omega_{\mathcal{L}\bigcup\mathcal{K}},
\end{aligned}$$

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 4 through normalization.

References

Aldous, D. J. (1985) Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XII* , Ed. P.L. Hennequin, Springer Lecture Notes Math. **1117**.

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. *International Conference on Information Fusion*, Florence, Italia, July 10-13.

Cifarelli, D. M. & Regazzini, E. (1978) Nonparametric statistical problems under partial exchangeability: The use of associative means. *Annali del'Instituto di Matematica Finianziara dell'Universitá di Torino, Serie 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.

Duan, J., Guindani M. & Gelfand A. E. (2005) Generalized spatial dirichlet process models. *ISDS Discussion Paper* **23**, Duke University, Durham, NC, USA.

Dunson, D. B. (2006) Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, **7**, 551-568.

Dunson, D. B. & Herring, A. H. (2006) Semiparametric Bayesian latent trajectory models. *ISDS Discussion Paper* **16**, Duke University, Durham, NC, USA.

Dunson, D. B., Herring, A. H. & Engel, S. M. (2006) Bayesian selection and clustering of polymorphisms in functionally-related genes. *J. Am. Statist. Assoc.*, under revision.

Dunson, D. B., Pillai, N. & Park, J-H. (2006) Bayesian density regression. *J. R. Statist. Soc.* B, in press.

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 modeling 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.

Johnson V. E. (2006) Bayesian model assessment using pivotal quantities. UT MD Anderson Cancer Center Department of Biostatistics and Applied Mathematics Working Paper Series, **25**.

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

Lau, J. W. & Green, P. J. (2006) Bayesian model based clustering procedures. submitted.

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 Nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, 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, Creta: ISBA, pp551-60.

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

Muliere, P. & Tardella, L. (1998) Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Can. J. Statist.*, **26**, 283-97.

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 & Robers (2006) Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Technical Report, Lancaster University, Lancaster LA1 4YF, UK.

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

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. IMS Lecture Notes-Monograph series, **30**.

Pitman, J. & Yor, M. (1997) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855-900.

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) Hierarchial 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. London: John Wiley.

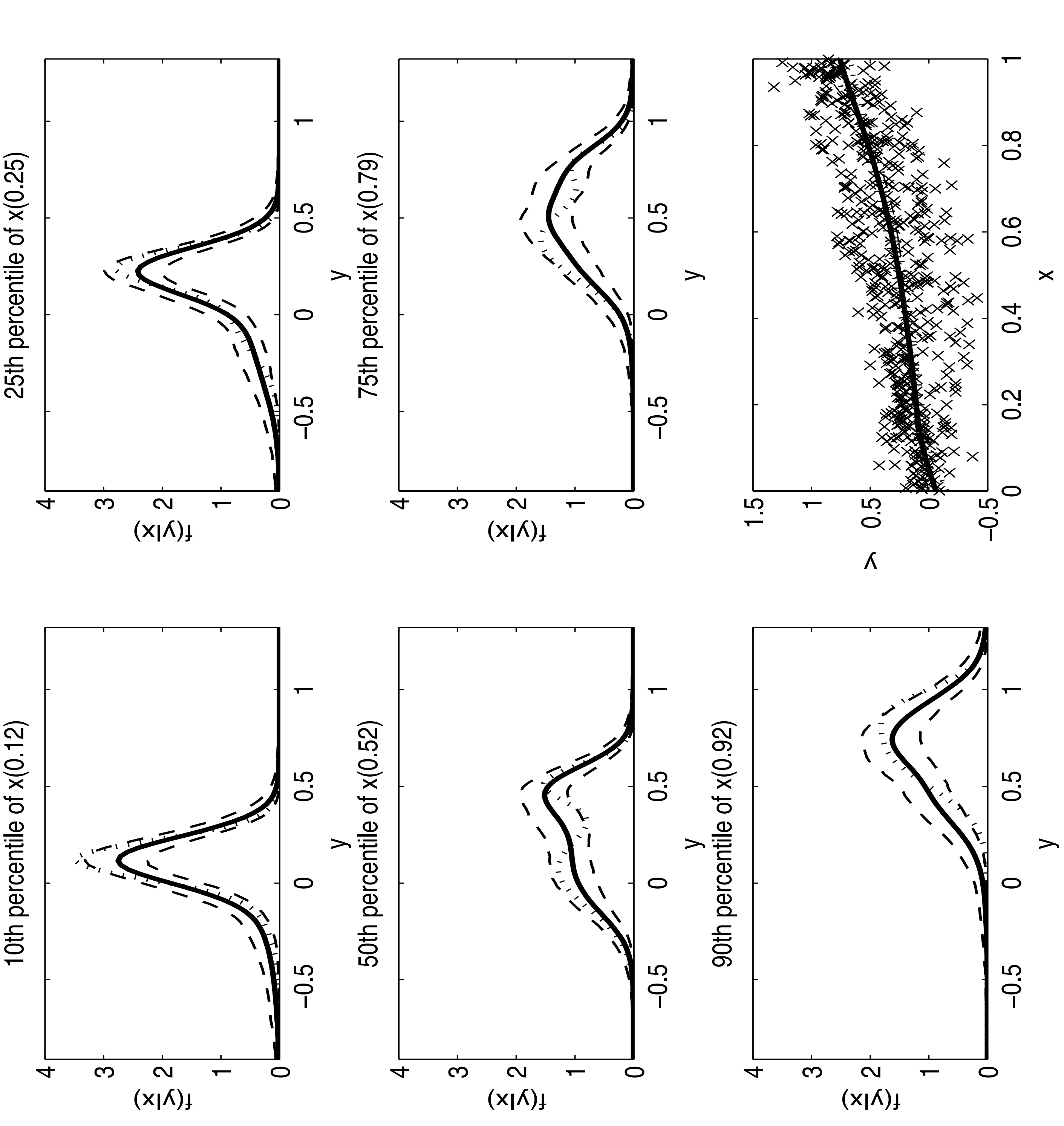

Figure 1: Results for the simulation example.

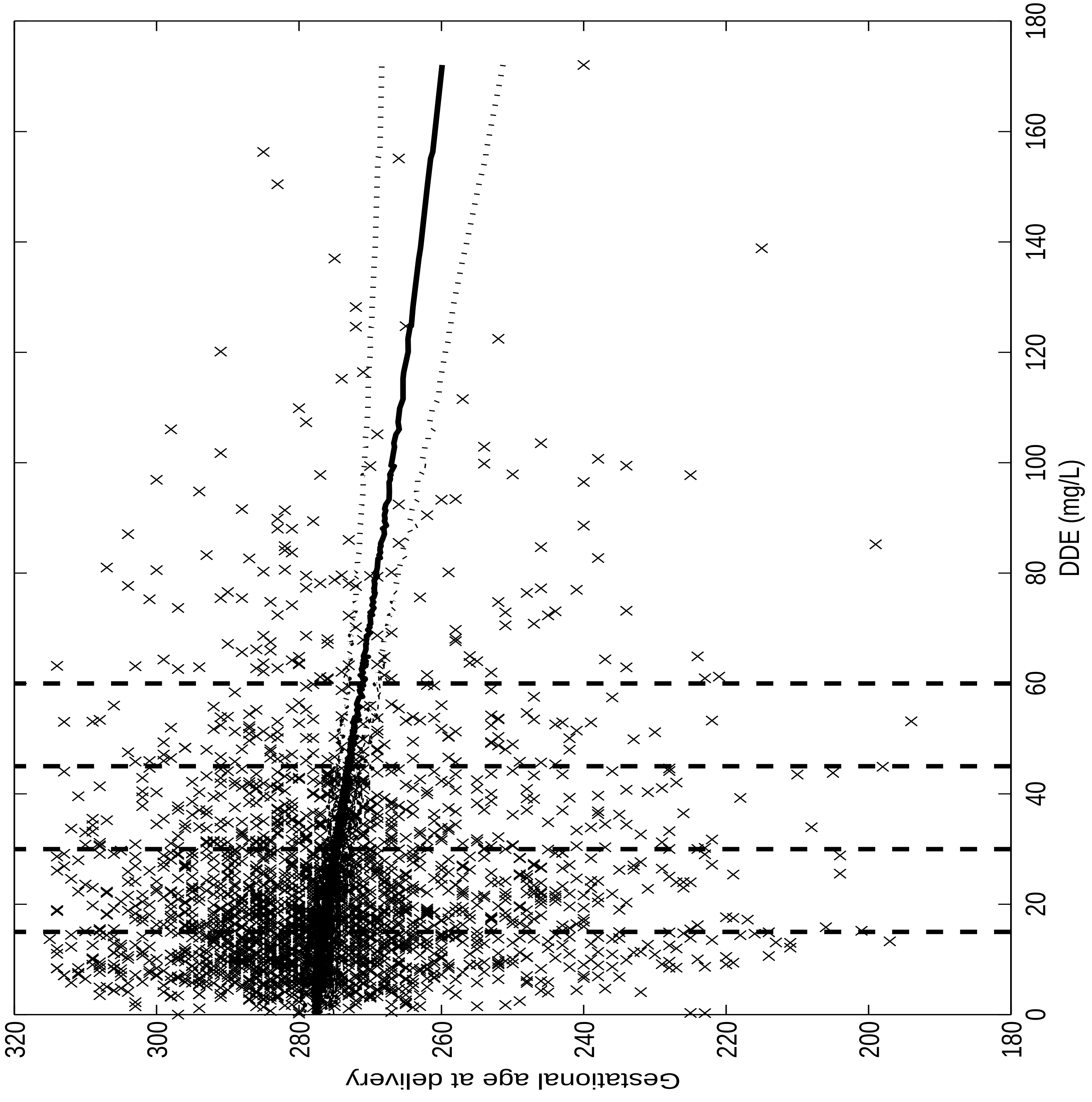

Figure 2: DDE vs gestational age at delivery in days for 2313 women in the Longnecker et al. (2001) study. The solid line is the conditional predictive mean, while the dotted lines are 99% pointwise credible intervals. Vertical dashed lines are DDE quintiles.

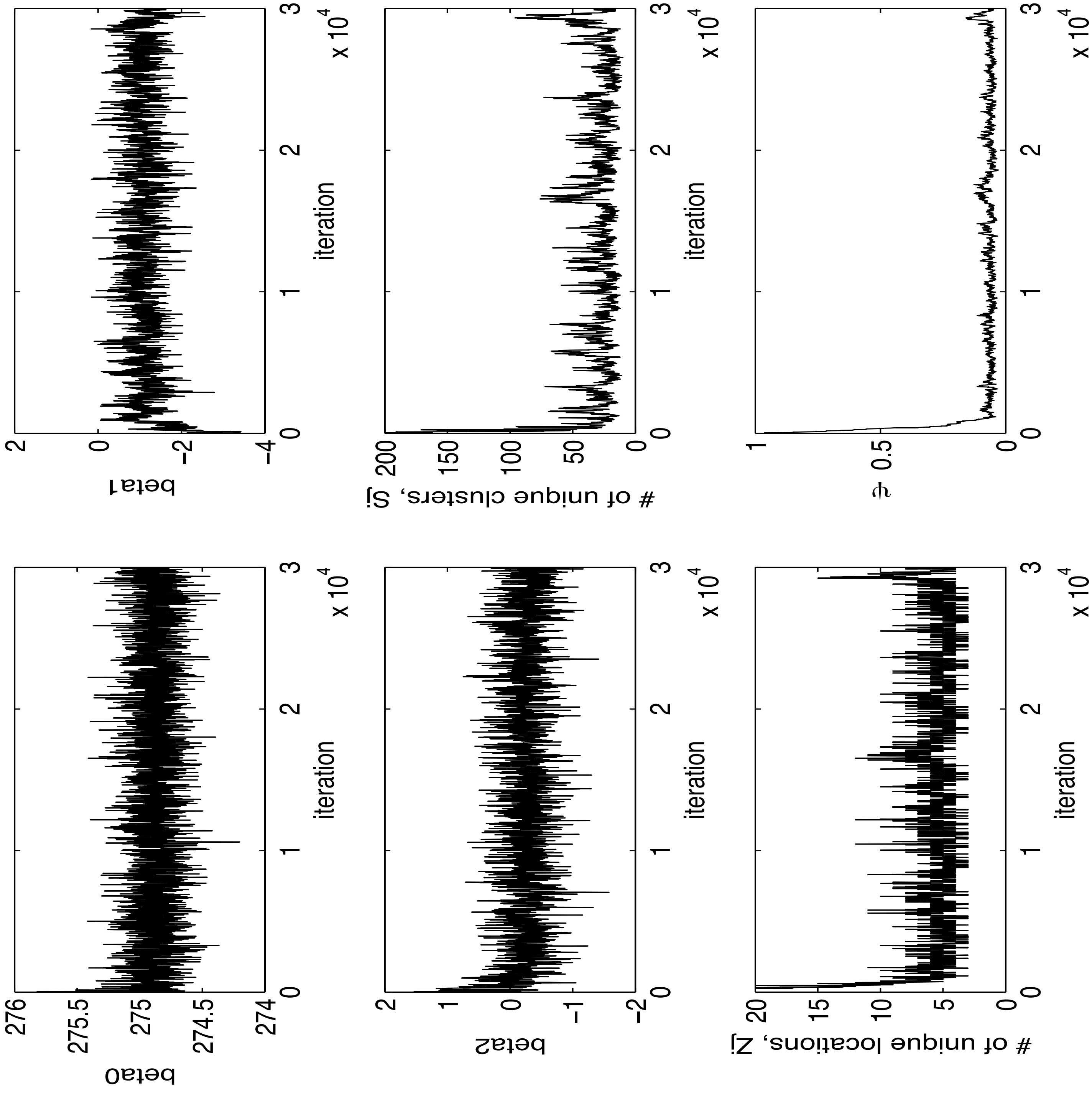

Figure 3: Trace plots for mean parameters in G_0, the number of unique values over all loations, the number of occupied locations, and the kernel precision parameter.

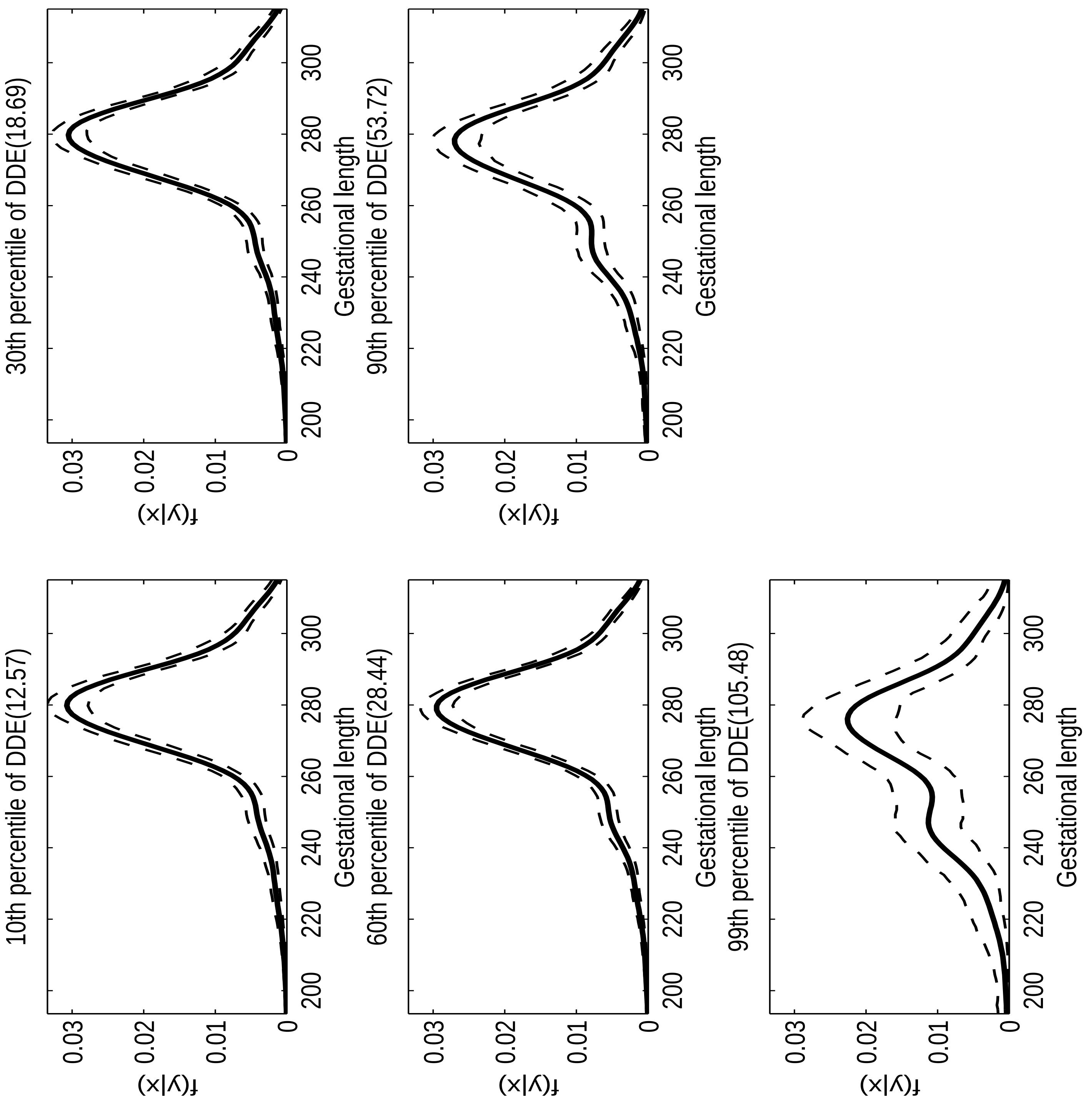

Figure 4: Estimated gestational age at delivery (in days) densities and pointwise 99% credible intervals conditional on DDE.

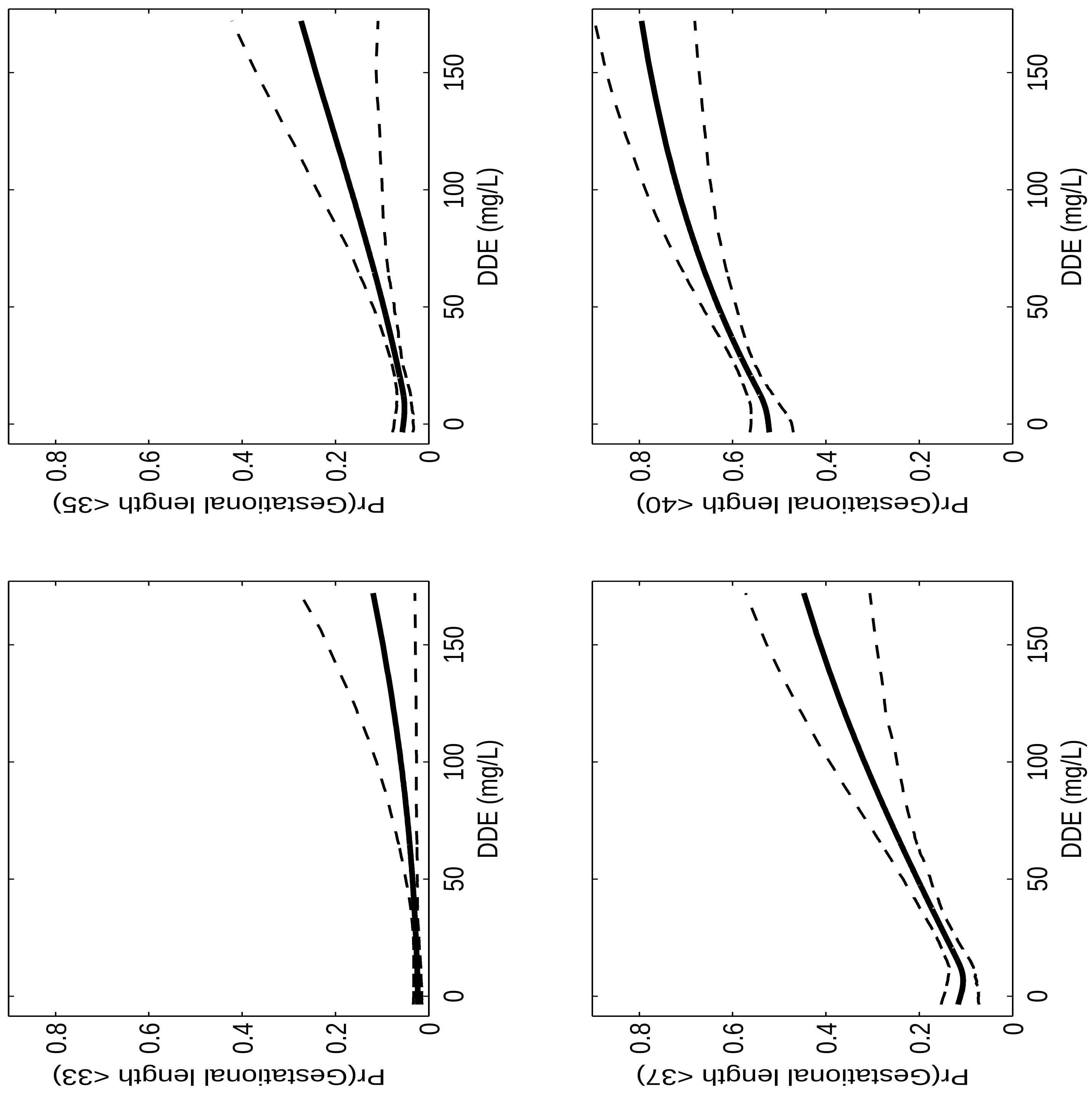

Figure 5: Estimated probability that gestational age at delivery is less than T (33,35,37,40) weeks versus DDE dose. Solid lines are posterior means, while dashed lines are pointwise 99% credible intervals.