Statistica Sinica **20** (2010), 1203-1226

BAYESIAN GENERALIZED PRODUCT PARTITION MODEL

Ju-Hyun Park and David B. Dunson

National Cancer Institute and Duke University

Abstract: Starting with a carefully formulated Dirichlet process (DP) mixture model, we derive a generalized product partition model (GPPM) in which the partition process is predictor-dependent. The GPPM generalizes DP clustering to relax the exchangeability assumption through the incorporation of predictors, resulting in a generalized Pólya urn scheme. In addition, the GPPM can be used for formulating flexible semiparametric Bayes models for conditional distribution estimation, bypassing the need for expensive computation of large numbers of unknowns characterizing priors for dependent collections of random probability measures. A variety of special cases are considered, and an efficient Gibbs sampling algorithm is developed for posterior computation. The methods are illustrated using simulation examples and an epidemiologic application.

Key words and phrases: Clustering, conditional distribution estimation, Dirichlet process, generalized Pólya urn, latent class model, mixture of experts, nonparametric Bayes, product partition.

1. Introduction

In recent years, there has been an increasing need for flexible models for predictor-dependent clustering and conditional distribution estimation. For example, in epidemiologic studies of continuous health outcomes, the primary focus is often in assessing the effect of exposures on the risk of adverse health responses. Adverse responses typically correspond to values in the tails, so it becomes important to allow the response density to change flexibly in location, shape, and variance with predictors. In addition, in interpreting the results, it is useful to cluster individuals based on their health response, with the allocation to clusters depending on exposures and covariates. A similar focus arises in many applications beyond epidemiology.

Predictor-dependent mixture models provide a natural model-based approach for addressing these interests, with

$$f(y \,|\, \mathbf{x}) = \sum_{h=1}^{k} \pi_h(\mathbf{x})\, f_h(y \,|\, \mathbf{x}, \theta_h), \tag{1.1}$$

where y is the response, $\mathbf{x}$ is a predictor, k is the number of mixture components, $\pi_h(\cdot)$ is the probability weight assigned to component h, and $f_h(\cdot \mid \cdot, \theta_h)$ is a distribution in a parametric family characterized by the finite-dimensional θ_h, for $h = 1, \ldots, k$. Hierarchical mixtures-of-experts models (Jordan and Jacobs (1994)) characterize $\pi_h(\mathbf{x})$ using a probabilistic decision tree, while letting $f_h(y \mid \mathbf{x}, \theta_h) = N(y; \mathbf{x}'\beta_h, \tau_h^{-1})$. A number of authors have considered alternative choices of regression models for the weights and experts (e.g., Jiang and Tanner (1999)). For recent articles, refer to Carvalho and Tanner (2005) and Ge and Jiang (2006).

To bypass issues involved in choosing k, we follow a semiparametric Bayes approach and let $k = \infty$. Without the predictor-dependence, this would be straightforward by letting $y_i \sim f(\phi_i)$, with $\phi_i \sim G$ and G assigned a Dirichlet process (DP) prior (Ferguson (1973, 1974)). When a DP prior is used for the mixture distribution, G, one obtains a DP mixture (DPM) model (Lo (1984)), Escobar and West (1995)). In marginalizing out G, one induces a prior on the partition of subjects $\{1, \ldots, n\}$ into clusters, with the cluster-specific parameters consisting of independent draws from G_0, the base distribution in the DP.

As noted by Quintana and Iglesias (2003), this induced prior is a type of product partition model (PPM) (Hartigan (1990), Barry and Hartigan (1992)). When the focus is on clustering or generating a flexible partition model for prediction, as in Holmes et al. (2005), it is appealing to marginalize out G in order to increase efficiency in computation and to simplify interpretation. The DP induces a particular prior on the partition and one can develop alternative classes of PPMs by replacing the DP prior on G with an alternative choice. Quintana (2006) applied this strategy for species sampling models (SSMs) (Pitman (1996), Ishwaran and James (2003)), which are a very broad class of nonparametric priors that include the DP as a special case.

Our interest is in further generalizing PPMs to include predictor-dependence by starting with (1.1) in the $k = \infty$ case, and attempting to obtain a prior that results in a PPM upon marginalization. There has been considerable recent interest in the nonparametric Bayesian literature on developing priors for predictor-dependent collections of random probability measures. Starting with a stick-breaking representation of the DP (Sethuraman (1994)) MacEachern (1999, 2001) proposed a class of dependent DP (DDP) priors. With the probability weights $\boldsymbol{\pi}$ being fixed across predictor $\mathbf{x}$, DDP priors have been successfully implemented in ANOVA modeling (De Iorio et al. (2004)), spatial data analysis (Gelfand, Kottas and MacEachern (2005)), time series (Caron et al. (2006)) and stochastic ordering (Dunson and Peddada (2008)) applications. Unfortunately, the fixed $\boldsymbol{\pi}$ case does not allow predictor-dependent clustering, motivating articles on order-based DDPs (Griffin and Steel (2006)), weighted mixtures of DPs (Dunson, Pillai and Park (2007)) and kernel stick-breaking processes (Dunson and Park (2008)).

To improve computational efficiency, we focus on obtaining a generalized product partition model (GPPM), that would allow large numbers of parameters to be marginalized out before posterior computation. Section 2 reviews the PPM and its relationship with the DP. Section 3 induces predictor-dependence in the PPM through a carefully-specified joint DPM model related to Müller, Erkanli and West (1996). Section 4 describes a simple and efficient Gibbs sampler for posterior computation. Section 5 illustrates the methods through simulation studies, and Section 6 contains an application to an epidemiologic data example. The results are discussed in Section 7.

2. Product Partition Models and Dirichlet Process Mixtures

Let $\mathcal{S}^* = (\mathbf{S}_1^*, \ldots, \mathbf{S}_k^*)$ denote a partition of identification numbers (IDs) $\mathcal{I}_n = \{1, \ldots, n\}$ for n subjects, with the elements of $\mathbf{S}_h^*$ corresponding to the IDs of those subjects in cluster h and the number k of clusters ranging from one to n. PPMs are defined by first expressing the prior probability for $\mathcal{S}^*$ as the product of nonnegative cohesions $c(\mathbf{S}_h^*)$ for $\mathbf{S}_h^* \in \mathcal{S}^*$, $h = 1, \ldots, k$:

$$\pi(\mathcal{S}^*) = c_0 \prod_{h=1}^{k} c(\mathbf{S}_h^*), \tag{2.1}$$

where c_0 is a normalizing constant that sums to one over all possible partitions. Given $\mathcal{S}^*$, let $\mathbf{y}_h = \{y_i : i \in \mathbf{S}_h^*\}$ denote the data for subjects in cluster h, for $h = 1, \ldots, k$. Then the specification of the PPM is completed by specifying the conditional likelihood function for $\mathbf{y} = (y_1, \ldots, y_n)'$,

$$f(\mathbf{y}|\mathcal{S}^*) = \prod_{h=1}^{k} f_h(\mathbf{y}_h), \tag{2.2}$$

where $f_h(\mathbf{y}_h) = \int \prod_{i \in \mathbf{S}_h^*} f(y_i \,|\, \theta_h) dG_0(\theta_h)$, $f(\cdot \,|\, \theta)$ is a likelihood characterized by θ, and the elements of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ are independently and identically distributed with prior G_0. Note that the partitioning is obtained for subject-specific parameters, with the data within each cluster assumed to be conditionally independent given the cluster-specific parameters, while the data for different clusters are marginally independent.

The PPMs in (2.1) and (2.2) have the appealing properties of consistency and conjugacy. Let S_i indicate the partition set subject i belongs to, and superscript (i) on any matrix or vector indicate that the contribution of subject i has been removed. The prior distribution $\pi(\mathcal{S}^{*(n)})$ of a partition of the first $n-1$ IDs $\mathcal{I}_{n-1}$ can be also obtained by integrating $\pi(\mathcal{S}^*)$ out with respect to S_n (consistency with respect to sample size). In addition, the posterior distribution of $\mathcal{S}^*$ given $\mathbf{y}$ has a PPM form, but with the posterior cohesion $c(\mathbf{S}_h^*) f_h(\mathbf{y}_h)$ (conjugacy).

Alternatively, a PPM can be induced through the hierarchical model

$$y_i \,|\, \boldsymbol{\theta}, \mathbf{S} \overset{ind}{\sim} f(\theta^*_{S_i}),$$
$$S_i \overset{i.i.d.}{\sim} \sum_{l=1}^{k} \pi_l \delta_l, \quad \theta^*_l \overset{i.i.d.}{\sim} G_0, \tag{2.3}$$

where $\mathbf{S} = (S_1, \ldots, S_n)'$ is the cluster membership indicator vector, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ is the probability weighting, and taking the number k of components to infinity induces a nonparametric PPM. Equivalently, one can let $y_i \sim f(\phi_i)$ with $\phi_i \sim G$ and $G = \sum_{l=1}^{k} \pi_h \delta_{\theta^*_l}$, with δ_θ a probability measure concentrated at θ. In these hierarchical models, the partition $\mathcal{S}^*$ is induced by grouping subjects based on $\mathbf{S}$ and $\boldsymbol{\theta}$ corresponds to the unique values of $\{\phi_i\}_{i=1}^n$. Therefore, a prior on the weight $\boldsymbol{\pi}$ induces a particular form for $\pi(\mathcal{S}^*)$, and hence the cohesion $c(\cdot)$.

As motivated by Quintana and Iglesias (2003), a convenient choice corresponds to the Dirichlet process prior, $G \sim DP(\alpha G_0)$, with α a precision parameter and G_0 a non-atomic base measure. By the Dirichlet process prediction rule (Blackwell and MacQueen (1973)), the conditional prior of ϕ_i given $\boldsymbol{\phi}^{(i)}$, with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)'$, and marginalizing out G, is

$$(\phi_i \,|\, \boldsymbol{\phi}^{(i)}) \sim \left(\frac{\alpha}{\alpha + n - 1}\right) G_0(\phi_i) + \left(\frac{1}{\alpha + n - 1}\right) \sum_{j \neq i} \delta_{\phi_j}(\phi_i), \tag{2.4}$$

which generates new values from G_0 with probability $\alpha/(\alpha+n-1)$ and otherwise sets ϕ_i equal to one of the existing values $\boldsymbol{\phi}^{(i)}$ chosen by sampling from a discrete uniform. Hence, the joint distribution of $\boldsymbol{\phi}$ is

$$\pi(\boldsymbol{\phi}) = \prod_{i=1}^{n} \left\{ \frac{\alpha G_0(\phi_i) + \sum_{j<i} \delta_{\phi_j}(\phi_i)}{\alpha + i - 1} \right\}. \tag{2.5}$$

Let $k = n(\mathcal{S}^*_n)$ denote the number of partition sets, $k^*_h = n(\mathbf{S}^*_h)$ the cardinality of $\mathbf{S}^*_h$, $\boldsymbol{\phi}_h = \{\phi_i : i \in \mathbf{S}^*_h\}$, and $\boldsymbol{\phi}_{h,l}$ the parameter for the lth subject in cluster h, with subjects ordered by IDs. Lo (1984) and Quintana and Iglesias (2003) show that (2.5) is equivalent to

$$\pi(\boldsymbol{\phi}) = \sum_{\mathcal{S}^* \in \mathcal{P}} \frac{1}{\prod_{l=1}^{n}(\alpha + l - 1)} \prod_{h=1}^{k} \alpha(k^*_h - 1)! G_0(\boldsymbol{\phi}_{h,1}) \prod_{j=2}^{k^*_h} \delta_{\boldsymbol{\phi}_{h,1}}(\boldsymbol{\phi}_{h,j})$$
$$= c_0 \sum_{\mathcal{S}^* \in \mathcal{P}} \prod_{h=1}^{k} c(\mathbf{S}^*_h) \pi_h(\boldsymbol{\phi}_h), \tag{2.6}$$

where $\mathcal{P}$ is the set of all partitions of $\{1,\ldots,n\}$, $c_0 = \prod_{l=1}^{n}(\alpha+l-1)^{-1}$, $c(\mathbf{S}_h^*) = \alpha(k_h^*-1)!$, and $\pi_h(\boldsymbol{\phi}_h)$ is the prior on $\boldsymbol{\phi}_h$. The marginal likelihood of $\mathbf{y}$ is then

$$f(\mathbf{y}) = c_0 \sum_{\mathcal{S}^*\in\mathcal{P}} \prod_{h=1}^{k} c(\mathbf{S}_h^*) \int \prod_{i\in\mathbf{S}_h^*} f(y_i|\theta) dG_0(\theta),$$

which is a special case of the form implied by (2.2) corresponding to a PPM with cohesion $c(\mathbf{S}_h^*) = \alpha(k_h^*-1)!$. This implies that simple and efficient Markov Chain Monte Carlo (MCMC) algorithms developed for DPMs can be used for posterior computation in PPMs. However, the class of PPMs induced by the DPM specification above assumes that the subjects are exchangeable, and does not allow for the incorporation of predictors.

3. Predictor Dependent Product Partition Models

3.1 Proposed formulation

Our goal is to incorporate predictor values of $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)'$ into a class of PPMs, so that the prior on the partition $\mathcal{S}^*$ has the form

$$\pi(\mathcal{S}^*|\mathbf{X}) \propto \prod_{h=1}^{k} c(\mathbf{S}_h^*,\, \mathbf{X}_h^*), \tag{3.1}$$

where $\mathbf{X}_h^* = \{\mathbf{x}_i : i \in \mathbf{S}_h^*\}$ for $h = 1,\ldots,k$, and the cohesion $c(\cdot)$ depends on the subjects predictor values. Expression (3.1) has two appealing properties. First, the posterior distribution of the partition $\mathbf{S}_n^*$ updated with the likelihood of response $\mathbf{y} = (y_1,\ldots,y_n)'$ is still in a class of PPMs, but with updated cohesion $c(\mathbf{S}_h^*, \mathbf{X}_h^*) f_h(\mathbf{y}_h)$. Secondly, there is a direct influence of predictor $\mathbf{X}$ on the partition process. Previous incorporation of predictors in PPMs instead relies on replacing $f(y_i \,|\, \theta_h)$ with $f(y_i \,|\, \mathbf{x}_i, \theta_h)$ in (2.2), which allows the predictor effect to vary across clusters, but does not allow the clustering process itself to be predictor dependent.

To specify cohesion $c(\mathbf{S}_h^*, \mathbf{X}_h^*)$, we exploit the connection between PPMs and DPMs. For simplicity of notation, we focus on a univariate response y, though multivariate generalizations are straightforward. Suppose $\mathbf{z}_i = (y_i, \mathbf{x}_i')'$ follows the hierarchical model

$$\begin{aligned} f(\mathbf{z}_i \,|\, \boldsymbol{\phi}_i) = f(y_i, \mathbf{x}_i|\varphi_i, \gamma_i) = f_1(y_i|\mathbf{x}_i, \varphi_i) f_2(\mathbf{x}_i|\gamma_i), \\ \boldsymbol{\phi}_i \sim G, \quad G \sim DP(\alpha G_0), \end{aligned} \tag{3.2}$$

where $G_0 = G_{0\varphi} \bigotimes G_{0\gamma}$ is the product measure of $G_{0\varphi}$ and $G_{0\gamma}$, components inducing a base prior for φ_i and γ_i, respectively. This DPM model will induce

partitioning of the subjects $\{1,\ldots,n\}$ into $k \leq n$ clusters, with $i \in \mathbf{S}_h^*$ denoting that subject i belongs to cluster h, which implies that $\varphi_i = \varphi_h^*$ and $\gamma_i = \gamma_h^*$, where $\boldsymbol{\gamma}^* = (\gamma_1^*,\ldots,\gamma_k^*)'$ and $\boldsymbol{\varphi}^* = (\varphi_1^*,\ldots,\varphi_k^*)'$ denote the unique values of $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_n)'$ and $\boldsymbol{\varphi} = (\varphi_1,\ldots,\varphi_n)'$, respectively.

Under (3.2), we can obtain a joint distribution of $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$ using the same approach used in deriving expression (2.6). If we then multiply by the conditional likelihood $\prod_{i=1}^n f_2(\mathbf{x}_i|\gamma_i)$ and marginalize out $\boldsymbol{\gamma}$, the joint distribution of $\boldsymbol{\varphi}$ and $\mathbf{X}$ is

$$\pi(\boldsymbol{\varphi},\mathbf{X}) = \sum_{\mathcal{S}^*\in\mathcal{P}} \Bigg[c_0 \prod_{h=1}^{k} \alpha(k_h^*-1)! \Bigg\{ \int \prod_{i\in\mathbf{S}_h^*} f_2(\mathbf{x}_i|\gamma_h^*) dG_{0\gamma}(\gamma_h^*) \Bigg\} G_{0\varphi}(\boldsymbol{\varphi}_{h,1}) \times \prod_{j=2}^{k_h^*} \delta_{\boldsymbol{\varphi}_{h,1}}(\boldsymbol{\varphi}_{h,j}) \Bigg], \tag{3.3}$$

where $\boldsymbol{\varphi}_{h,l}$ is the parameter for the response y of the lth subject, ordered by the IDs, in cluster h. Therefore, the conditional distribution of $\boldsymbol{\varphi}$ given $\mathbf{X}$ is

$$\begin{aligned}\pi(\boldsymbol{\varphi}|\mathbf{X}) &= c_0^* \sum_{\mathcal{S}^*\in\mathcal{P}} \Bigg[\prod_{h=1}^{k} \alpha(k_h^*-1)! \Bigg\{ \int \prod_{i\in\mathbf{S}_h^*} f_2(\mathbf{x}_i|\gamma_h^*) dG_{0\gamma}(\gamma_h^*) \Bigg\} G_{0\varphi}(\boldsymbol{\varphi}_{h,1}) \\ &\qquad \times \prod_{j=2}^{k_h^*} \delta_{\boldsymbol{\varphi}_{h,1}}(\boldsymbol{\varphi}_{h,j}) \Bigg] \\ &= c_0^* \sum_{\mathcal{S}^*\in\mathcal{P}} \prod_{h=1}^{k} c(\mathbf{S}_h^*,\mathbf{X}_h^*)\pi_h(\boldsymbol{\varphi}_h), \end{aligned} \tag{3.4}$$

where c_0^* is a normalizing constant so that the sum over $\mathcal{P}$ is unity, $\pi_h(\boldsymbol{\varphi}_h)$ is a prior on partitioned set $\boldsymbol{\varphi}_h$, and $c(\mathbf{S}_h^*,\mathbf{X}_h^*) = \alpha(k_h^*-1)!\int\prod_{i\in\mathbf{S}_h^*} f_2(\mathbf{x}_i|\gamma) dG_{0\gamma}(\gamma)$. Hence, we have induced a GPPM of the form shown in (3.1) starting with a joint DPM model for the response and predictors related to that proposed by Müller, Erkanli and West (1996). A related idea was independently developed by Fernando Quintana and collaborators in recent work (unpublished communication), though our subsequent development differs from theirs.

In addition to the appealing properties of (3.1), our specification results in the interesting feature that the GPPM is still consistent with respect to the sample size. Since the prior on the partition $\mathcal{S}^*$ given $\mathbf{X}$ can be expressed as

$$\pi(\mathcal{S}^*|\mathbf{X}) = \pi(S_n,\mathcal{S}^{*(n)}|\mathbf{X}) = \pi(S_n|\mathcal{S}^{*(n)},\mathbf{X})\pi(\mathcal{S}^{*(n)}|\mathbf{X}),$$

by summing over S_n and taking the conditional expectation with respect to $\mathbf{x}_n$ given $\mathbf{X}^{(n)}$, we have

$$\pi(\mathcal{S}^{*(n)}|\mathbf{X}^{(n)}) = \int \sum_{S_n=1}^{k} \pi(S_n|\mathcal{S}^{*(n)}, \mathbf{X})\pi(\mathcal{S}^{*(n)}|\mathbf{X})g(\mathbf{x}_n|\mathbf{X}^{(n)})d\mathbf{x}_n,$$

where

$$g(\mathbf{x}_n|\mathbf{X}^{(n)}) = \frac{\sum_{\mathcal{S}^*\in\mathcal{P}} c_0 \prod_{h=1}^{k} \alpha(k_h^*-1)!\left\{\int \prod_{i\in\mathbf{S}_h^*} f_2(\mathbf{x}_i|\gamma_h^*)dG_{0\gamma}(\gamma_h^*)\right\}}{\sum_{\mathcal{S}^{*(n)}\in\mathcal{P}^{(n)}} c_0^{(n)} \prod_{h=1}^{k^{(n)}} \alpha(k_h^{*(n)}-1)!\left\{\int \prod_{i\in\mathbf{S}_h^{*(n)}} f_2(\mathbf{x}_i|\gamma_h^*)dG_{0\gamma}(\gamma_h^*)\right\}}.$$

3.2. Generalized Pòlya urn scheme

It is not obvious from (3.4) how the predictor and hyperparameter values impact clustering. However, we can show that the proposed GPPM induces a simple predictor-dependent generalization of the Blackwell and MacQueen Pólya urn scheme of the DP in (2.4), which should be useful both in interpretation and posterior computation.

Theorem 1. *The full conditional prior of φ_i given α, $\boldsymbol{\varphi}^{(i)}$, and $\mathbf{X}$, or equivalently given α, $\boldsymbol{\varphi}^{*(i)}$, $\mathbf{S}^{(i)}$, and $\mathbf{X}$, is*

$$\left(\varphi_i \,|\, \alpha, \boldsymbol{\varphi}^{*(i)}, \mathbf{S}^{(i)}, \mathbf{X}\right), \sim w_0(\mathbf{x}_i)G_{0\varphi} + \sum_{h=1}^{k^{(i)}} w_h(\{\mathbf{x}_i, \mathbf{X}_h^{*(i)}\})\delta_{\varphi_h^{*(i)}}, \tag{3.5}$$

with the probability weights

$$w_0(\mathbf{x}_i) = \widetilde{c}\alpha \int f_2(\mathbf{x}_i|\gamma)dG_{0\gamma}(\gamma)$$

$$w_h(\{\mathbf{x}_i, \mathbf{X}_h^{*(i)}\}) = \widetilde{c}k_h^{*(i)} \int f_2(\mathbf{x}_i|\gamma)dG_{0\gamma}^*(\gamma|\mathbf{X}_h^{*(i)}),$$

where $\widetilde{c}$ is a normalizing constant and $G_{0\gamma}^*(\cdot|\mathbf{X}_h^{*(i)})$ is the posterior distribution updated with the likelihood of predictor cluster h excluding the contribution from the ith subject.

The proof is in Appendix A. Theorem 1 implies that subject i is assigned to either a new generated value (creating a new cluster) or one of the existing unique values, with the probability weights being proportional to a product of the DP probability weights and the marginal likelihoods at its predictor value

varying across clusters. Therefore, if the predictor value of subject i is close to values $\mathbf{X}_h^*$ for subjects in cluster h, with the measure of closeness depending on the choice of $f_2(\cdot)$, the contribution of subject i to the marginal likelihood will tend to be highest if that subject is allocated to cluster h.

Conceptually, this idea is related to the Bayesian partition model (BPM) of Holmes et al. (2005) in that subjects close together in the predictor space will tend to have similar response distributions. However, instead of measuring closeness by assuming a particular distance metric, our specification automatically induces a distance metric through a flexible nonparametric model for the joint distribution of the predictors. This allows the measure of closeness to be adaptive depending on location in the predictor space, automatically producing spatially-adaptive bandwidth selection. In the special case of a degenerate distribution for $\mathbf{x}$, $f_2(\mathbf{x}|\gamma) = \delta_\gamma(\mathbf{x})$, (3.5) reduces to the Blackwell and MacQueen Pòlya urn scheme of (2.4).

An apparent disadvantage of our formulation is that, by inducing a prior for the conditional distribution of y_i given $\mathbf{x}_i$ through a prior for the joint distribution of y_i and $\mathbf{x}_i$, we are implicitly assuming that the predictors are random variables. In fact, in many applications one or more of the predictors may be fixed by design, representing spatial location, time of observation, or an experimental condition. The predictor-dependent urn scheme shown in Theorem 1 is still useful and coherent in such cases, as this urn scheme is defined conditionally on the predictor values. This urn scheme clearly results in a coherent joint prior for $\boldsymbol{\varphi}$, conditionally on $\mathbf{X}$, that is invariant to permutations in the ordering of the subjects. It is in general very difficult to define a predictor-dependent urn scheme, that satisfies these conditions.

In order for the weights in (3.5) to be in a closed form, the marginal likelihood $\int f_2(\mathbf{x}_i|\gamma)dG_{0\gamma}(\gamma)$ must be available in closed form. Hence, by using a conjugate base measure $G_{0\gamma}$, computation can be simplified. However, in non-conjugate cases, one can follow the standard strategy of instead using an approximation to the marginal likelihood, such as the Laplace. Among many choices, we focus on two special cases: a normal-Wishart prior and a Poisson-gamma prior. Suppose that a normal-Wishart distribution is assumed for continuous $p \times 1$ predictor $\mathbf{x}$ and parameter $\gamma = (\boldsymbol{\mu}_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x})'$:

$$\begin{aligned} \mathbf{x}|\boldsymbol{\mu}_\mathbf{x}, c_\mathbf{x}, \boldsymbol{\Sigma}_\mathbf{x} &\sim N(\boldsymbol{\mu}_\mathbf{x}, c_\mathbf{x}^{-1}\boldsymbol{\Sigma}_\mathbf{x}), \\ \boldsymbol{\mu}_\mathbf{x}|\boldsymbol{\mu}_{0\mathbf{x}}, c_\mu, \boldsymbol{\Sigma}_{0\mathbf{x}} &\sim N(\boldsymbol{\mu}_{0\mathbf{x}}, c_\mu^{-1}\boldsymbol{\Sigma}_\mathbf{x}), \\ \boldsymbol{\Sigma}_\mathbf{x}^{-1}|\nu_\mathbf{x}, \boldsymbol{\Sigma}_{0\mathbf{x}} &\sim \mathcal{W}(\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1}, \nu_\mathbf{x}), \end{aligned} \tag{3.6}$$

where $c_\mathbf{x}^{-1}$ and c_μ^{-1} are multiplicative constants, and $\mathcal{W}(\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1}, \nu_\mathbf{x})$ is a Wishart with degrees of freedom $\nu_\mathbf{x}$ and expectation $\nu_\mathbf{x}\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1}$. Then the marginal likelihood of $\mathbf{x}_i$,

in probability weight $w_0(\mathbf{x}_i)$ in (3.5), is a non-central multivariate t-distribution with degrees of freedom $\nu = \nu_{\mathbf{x}} - p + 1$, location $\boldsymbol{\mu} = \boldsymbol{\mu}_{0\mathbf{x}}$, and scale $\boldsymbol{\Sigma} = (c_{\mathbf{x}} + c_\mu)/(\nu c_{\mathbf{x}} c_\mu)\boldsymbol{\Sigma}_{0\mathbf{x}}$, denoted by $t_p(\mathbf{x}_i; \nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$f(\mathbf{x}|\boldsymbol{\mu}, \nu, \boldsymbol{\Sigma}) = \frac{\Gamma((\nu+p)/2)}{(\pi\nu)^{p/2}\Gamma(\nu/2)|\boldsymbol{\Sigma}|^{1/2}}\left(1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-(\nu+p)/2},$$

while that in probability weight $w_h(\{\mathbf{x}_i, \mathbf{X}_h^{(i)}\})$, $h = 1, \ldots, k^{(i)}$, is also a noncentral multivariate t-distribution, but with updated hyperparameters:

$$\boldsymbol{\mu}_{0\mathbf{x}}^* = \frac{c_\mu \boldsymbol{\mu}_{0\mathbf{x}} + c_{\mathbf{x}} k_h^{*(i)} \bar{\mathbf{x}}_h^{(i)}}{c_\mu + c_{\mathbf{x}} k_h^{*(i)}}, \quad c_\mu^* = c_\mu + c_{\mathbf{x}} k_h^{*(i)}, \quad \nu_{\mathbf{x}}^* = \nu_{\mathbf{x}} + k_h^{*(i)},$$

$$\boldsymbol{\Sigma}_{0\mathbf{x}}^* = \left\{\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1} + k_h^{*(i)} D_h^{(i)} + \frac{k_h^{*(i)} c_{\mathbf{x}} c_\mu}{c_\mu + c_{\mathbf{x}} k_h^{*(i)}}(\bar{\mathbf{x}}_h^{(i)} - \boldsymbol{\mu}_{0x})(\bar{\mathbf{x}}_h^{(i)} - \boldsymbol{\mu}_{0x})'\right\}^{-1}, \quad (3.7)$$

where $\bar{\mathbf{x}}_h^{(i)} = \sum_{j:S_j^{(i)}=h} \mathbf{x}_j / k_h^{*(i)}$ and $D_h^{(i)} = \sum_{j:S_j^{(i)}=h} (\mathbf{x}_j - \bar{\mathbf{x}}_h^{(i)})(\mathbf{x}_j - \bar{\mathbf{x}}_h^{(i)})' / k_h^{*(i)}$.

Note that the structure in (3.6) is slightly different from the commonly used normal-Wishart priors in that a constant is multiplied not only to the variance of the expectation of $\mathbf{x}$ but also to the variance of $\mathbf{x}$. We find that the additional flexibility provided by the additional multiplier is useful in avoiding over-clustering problems that sometimes arise using the standard formulation. In the typical normal-Wishart prior, when the prior is updated with the data likelihood, the posterior variance of the expected value $\boldsymbol{\mu}_{\mathbf{x}}$ is smaller than that of $\mathbf{x}$, because the updated multiplicative constant c_μ^* is always greater than 1. This can lead to clustering of subjects with dissimilar predictors in some cases.

In the case of discrete predictors, we can also obtain a closed form marginal likelihood of $\mathbf{x}$. In order to simplify calculations, we assume a priori independence for the different predictors, while dependence among continuous predictors is allowed through $\boldsymbol{\Sigma}_{\mathbf{x}}$ in (3.6). Suppose that x_j, $j = 1, \ldots, p$, follows a Poisson distribution with mean Γ_j, which is assigned a Gamma prior with mean a_j/b_j, $\mathcal{G}(a_j, b_j)$, as the base measure $G_{0\gamma}$. The marginal distribution of $\mathbf{x}$ in w_0 is a product of negative binomials with the number of successes $r_j = a_j$ and success probability $p_j = b_j/(1 + b_j)$:

$$\Pr(X_j = k) = \frac{\Gamma(r_j + k)}{k!\Gamma(r_j)} p_j^{r_j}(1 - p_j)^k, \qquad j = 1, \ldots, p.$$

The marginal distribution in w_h, $h = 1, \ldots, k^{(i)}$, is also a product of negative binomials, but with hyperparameters $a_j^* = a_j + \sum_{j:S_j^{(i)}=h} x_j$ and $b_j^* = b_j + k_h^{*(i)}$. For bounded discrete predictors, we can instead use a multinomial likelihood

with a Dirichlet prior for the category probabilities. The case of mixed discrete and continuous predictors can also be easily dealt with.

4. Posterior Computation

One of the appealing features of our predictor-dependent urn scheme is that we can rely on efficient Pólya urn Gibbs sampling algorithms developed for computation in marginalized DPMs, with minimal modifications. In addition, although we focus here on posterior computation through MCMC, our predictor-dependent urn scheme could similarly be used to develop sequential importance sampling (SIS) algorithms (MacEachern, Clyde and Liu (1999), Quintana and Newton (2000)). modified weighted Chinese restaurant (WCR) sampling algorithms (Ishwaran and James (2003), as well as fast variational Bayes approximations (Kurihara, Welling and Vlassis (2006)).

For DPMs the algorithm of Bush and MacEachern (1996) is one of the most widely-used approaches due to the combination of simplicity and computational efficiency. Their approach first updates the configuration of subjects to clusters based on the Pólya urn scheme in (2.4), and then separately updates cluster specific parameters given the cluster configuration. This separate updating process makes their algorithm distinguishable from that in Escobar (1994) and Escobar and West (1995), and helps to improve rates of mixing and convergence. Our proposed approach relies on a direct generalization of Bush and MacEachern (1996). Although their algorithm and our generalization require the use of conjugate priors, extension to non-conjugate priors can proceed as proposed by MacEachern and Müller (1998) and Neal (2000) for the DP case relying on Metropolis-Hastings (Hastings (1970)).

From Theorem 1, the full conditional posterior distribution of φ_i can be derived as

$$\big(\varphi_i \,|\, \alpha, \boldsymbol{\varphi}^{*(i)}, \mathbf{S}^{(i)}, \mathbf{X}, \mathbf{y}\big), \sim q_{i,0} G_{0\varphi,i} + \sum_{h=1}^{k^{(i)}} q_{i,h} \delta_{\varphi_h^{*(i)}}, \tag{4.1}$$

where the posterior obtained by updating the prior $G_{0\varphi}$ with the likelihood of y_i is

$$G_{0\varphi,i}(\varphi_i) = \frac{G_{0\varphi}(\varphi_i) f_1(y_i|\mathbf{x}_i, \varphi_i)}{\int f_1(y_i|\mathbf{x}_i, \varphi_i) dG_{0\varphi}(\varphi_i)} = \frac{G_{0\varphi}(\varphi_i) f_1(y_i|\mathbf{x}_i, \varphi_i)}{h_i(y_i|\mathbf{x}_i)},$$

$q_{i,0} = \widetilde{c} w_0(\mathbf{x}_i) h_i(y_i|\mathbf{x}_i)$, $q_{i,h} = \widetilde{c} w_h(\{\mathbf{x}_i, \mathbf{X}^{(i)}\}) f_1(y_i|\mathbf{x}_i, \varphi_h^{*(i)})$, and $\widetilde{c}$ is a normalizing constant. Instead of sampling directly from (4.1) in implementing the Gibbs sampling, we first sample S_i, $i = 1, \ldots, n$, from its multinomial conditional posterior distribution with

$$\Pr(S_i = h|\boldsymbol{\varphi}^{*(i)}, \mathbf{S}^{(i)}, \mathbf{X}, \mathbf{y}) = q_{i,h}, \qquad h = 0, 1, \ldots, k^{(i)}. \tag{4.2}$$

When $S_i = 0$, φ_i is set to a new value generated from $G_{0\varphi,i}$. As a result of updating $\mathbf{S}$, the number k of clusters is automatically updated. As a next step, we update $\boldsymbol{\varphi}^*$ conditional on $\mathbf{S}$ and k from

$$\big(\varphi_h^*|\boldsymbol{\varphi}^{*(h)}, \mathbf{S}, k, \mathbf{X}, \mathbf{y}\big) \propto \Big\{ \prod_{i:S_i=h} f_1(y_i|\mathbf{x}_i, \varphi_h) \Big\} G_{0\varphi}(\varphi_h^*). \tag{4.3}$$

When there are some unknown parameters ψ characterizing the base measure $G_{0\varphi}$, we include an additional step for updating ψ based on the full conditional posterior distribution

$$\big(\psi|\boldsymbol{\varphi}, \mathbf{y}\big) \propto \pi(\psi) \Big\{ \prod_{h=1}^{k} G_{0\varphi}(\varphi_h^*|\psi) \Big\}. \tag{4.4}$$

We have found this algorithm to be simple to implement and efficient in those cases we have considered. The full conditional posterior distributions for the model considered in Sections 5 and 6 are in Appendix B.

5. Simulation Examples

5.1. Model specification

In this section, we illustrate the proposed method with simulations focusing on conditional density regression. Predictor-dependent partitioning will be evaluated with a data example in Section 6. We consider the following infinite mixture model:

$$f(y_i|\widetilde{\mathbf{x}}_i) = \sum_{h=1}^{\infty} \pi_h(\widetilde{\mathbf{x}}_i)\, f_1(y_i \,|\, \widetilde{\mathbf{x}}_i, \varphi_h^*),$$

where $\widetilde{\mathbf{x}}_i = (1, \mathbf{x}_i')' = (1, x_{i1}, \ldots, x_{ip})'$ and $f_1(y_i|\,\widetilde{\mathbf{x}}_i, \varphi_h^*) = N(y_i; \mu_h, \sigma_{y,h}^2)$ with $\varphi_h^* = (\mu_h, \sigma_{y,h}^2)'$ for the first simulation, and $f_1(y_i|\,\widetilde{\mathbf{x}}_i, \varphi_h^*) = N(y_i; \widetilde{\mathbf{x}}_i'\boldsymbol{\beta}_h, \sigma_{y,h}^2)$ with $\varphi_h^* = (\boldsymbol{\beta}_h, \sigma_{y,h}^2)'$ for the second simulation. The GPPM proposed in Section 3 is used to place a prior on the partition $\mathcal{S}^*$ and atoms $\boldsymbol{\varphi}^*$. Although there are $k \leq n$ mixture components represented in the sample of n subjects under the GPPM, there are conceptually infinitely many components, since the number of components increases stochastically as subjects are added.

In the absence of prior knowledge about the scale, it is recommended that continuous predictors be standardized to simplify prior elicitation. We require G_0 to correspond to a proper distribution, since marginal likelihoods will be used in calculating conditional posterior probabilities for partitioning. To simplify updating of the scale parameter, $c_{\mathbf{x}}$, we assume a discrete uniform prior on $(0, 1]$. For discrete predictors, we fix $a_j = b_j = 1$, $j = 1, \ldots, p-1$. In addition, let

$\sigma_{y,h}^{-2} \sim \mathcal{G}(a_y, b_y)$, $\mu_h \sim N(\mu, \kappa^{-1}\sigma_{y,h}^2)$, $\boldsymbol{\beta}_h \sim N(\boldsymbol{\beta}, \sigma_{y,h}^2 \mathbf{V})$ with $\mathbf{V} = \kappa^{-1} n(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}$ and $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_n)'$, $\mu \sim N(\mu_0, \kappa^{-1}\sigma_\mu^2)$, $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \kappa^{-1}\mathbf{V}_0)$, and $\kappa \sim \mathcal{G}(a_\kappa, b_\kappa)$. The last three prior distributions on $\psi = (\mu, \kappa)'$ or $\psi = (\boldsymbol{\beta}, \kappa)'$ are for additional flexibility. In the implementation, we let $\alpha = 1$, $\boldsymbol{\mu}_{0\mathbf{x}} = \mathbf{0}$, $\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1} = 4I_{p\times p}$, $\nu_{\mathbf{x}} = p$, $\mu_0 = 0$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{V}_0 = n(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}$, and $a_y = b_y = a_\kappa = b_\kappa = 1$. Other choices of these parameters were also considered to check sensitivity of models to our primary choice.

5.2. Implementation and results

We consider two cases in which $n = 500$, $p = 1$, and x_{i1} is generated from a uniform distribution over $(0, 1)$. We first simulated data from a normal distribution with mean x_{i1}^2 and variance 0.04, $N(y_i; x_{i1}^2, 0.04)$. The data were analyzed using a mixture of normals with the prior specification of Section 5.1, and with the MCMC algorithm of Section 4 implemented for 10,000 iterations, discarding the initial 1,000 iterations as a burn-in. After the burn-in period, it took the algorithm 0.51 second per iteration on Matlab, performed on Windows XP® with Intel$^{\text{TM}}$ Core®2 Duo E7300 2.66GHz/1066MHz/3MB L2. Figure 1 shows selected results. The algorithm converged rapidly and mixing was good based on trace plots of μ, the number of clusters, and $f(y = 1.5|\widetilde{\mathbf{x}} = (1,\ 0.25)')$, where the data point for y was randomly selected among possible values (the left panel of Figure 1). As shown in the right panel of Figure 1, the predictive densities and mean function of y (solid lines) well approximated the true values (dotted lines), that are completely embedded within pointwise 99% credible intervals (dashed lines). The posterior mean of the number of clusters was 2.4 with a 95% credible interval of $[2, 4]$ and the estimated normal means were almost equally spaced over $(0, 1)$.

As a more challenging second simulation case, we simulated data to approximately mimic the data in the reproductive epidemiology study considered in Section 6. In particular, we generated data from the mixture of two linear models

$$f(y_i|\mathbf{x}_i) = (1 - x_{i1}^4)N(y_i; 1, 0.04) + x_{i1}^4 N(y_i; , 1 - x_{i1}^2, 0.01),$$

where a secondary peak appears in the left tail of the response distribution, moving closer to zero as x_{i1} increases. This behavior, in which the tail of the distribution, corresponding to those subjects with the most extreme response, is particularly sensitive to changes in an exposure variable, is common in toxicology and epidemiology studies. We analyzed the data using a mixture of regression models with the GPPM approach specified in Section 5.1, and also using the DPM-based PPM described in Section 2. These two approaches result in mixtures

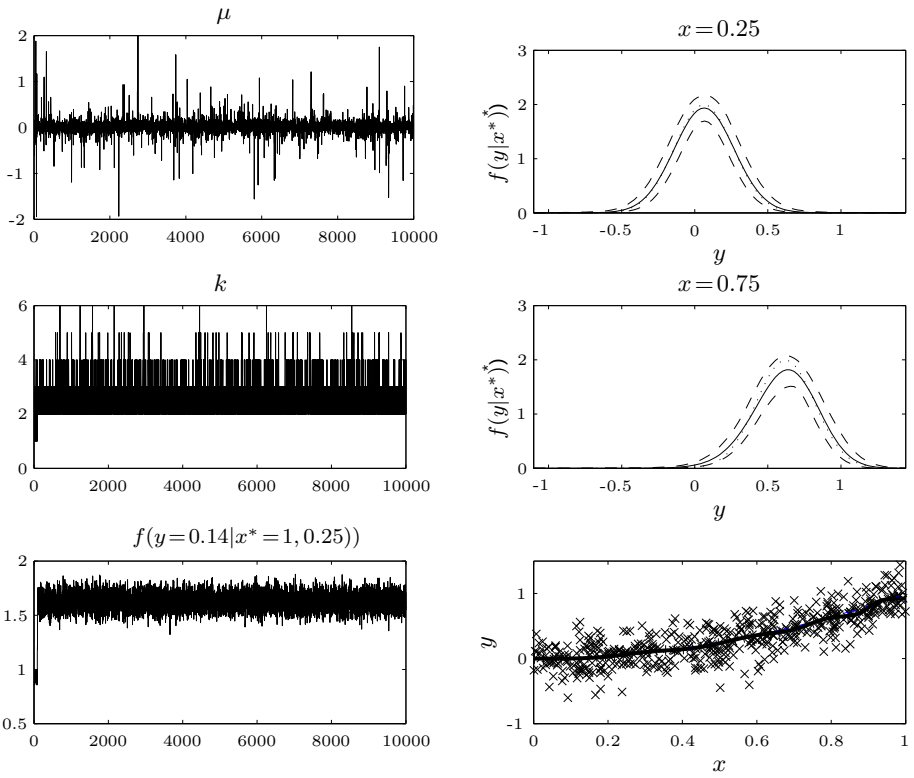

Figure 1. Results for the first simulation example. The left column provides trace plots for representative quantities, while the right panel shows the conditional distributions for two different values of x, as well as the mean function estimation along with the raw data. Posterior means are solid lines, pointwise 99% credible intervals are dashed lines, and true values are dotted lines.

of normal linear regressions, but the first approach allows the mixture weights to be predictor-dependent, while the second does not. The precision parameter α and the base measure G_0 for the DPM-based PPM were set to be the same as those used in the GPPM approach. Both analyses were run for 30,000 iterations with a 10,000 iteration burn-in; there were good mixing and convergence rates in both cases, based on examination of trace plots and diagnostics.

From Figure 2, it is clear that the GPPM provided a more flexible model capturing the rapid changes in the distribution across local regions of the predictor space, even for the somewhat small sample size of $n = 500$. In our experience, GPPMs based on mixing linear regressions with variances varying across components tend to do a very good job in sparsely characterizing complex changes in

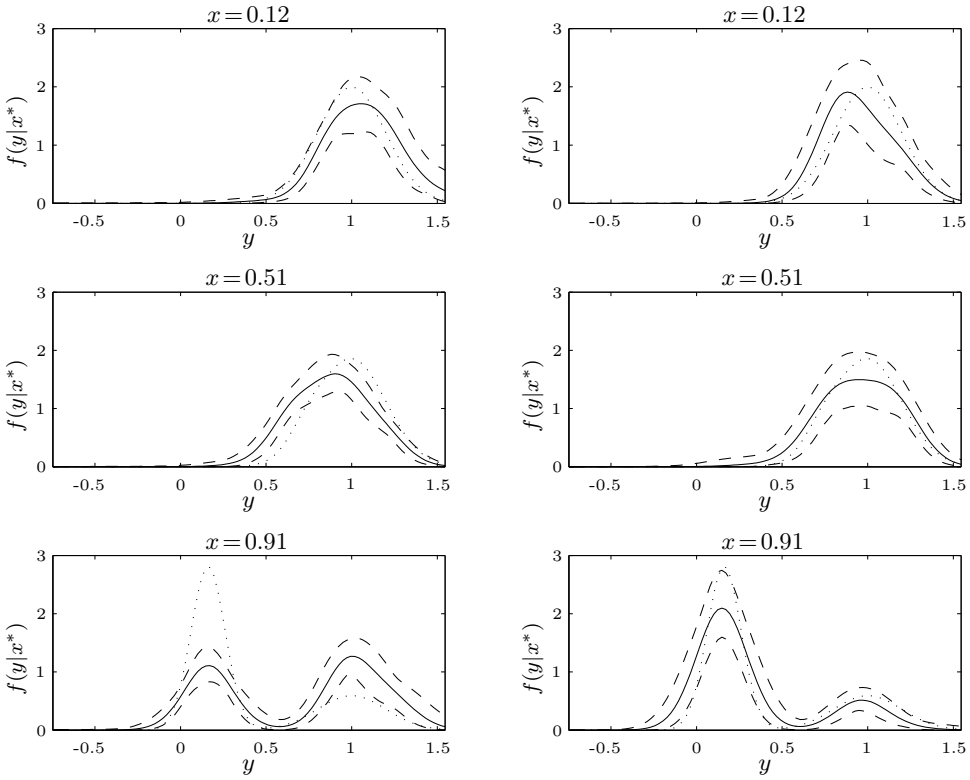

Figure 2. Estimated predictive densities from the PPM (left panel) and the GPPM (right panel) at the 10th, 50th and 90th percentiles of the empirical distribution of x: posterior means (solid lines), pointwise 99% credible intervals (dashed lines), and true values (dotted lines).

conditional densities with predictors, with sparsity corresponding to allocation of subjects to a small number of clusters. In simulation case 2, the posterior mean number of components used was 3.2 (95% credible interval=$[3, 6]$). We repeated the analysis of the second simulation including a discrete predictor, obtained by truncating the continuous predictor into l groups. It was observed that the proposed method worked well for a variety choices of l (results are not shown).

6. Epidemiologic Application

We applied the proposed method to the data used in Longnecker et al. (2001) and Dunson and Park (2008). A synthetic pesticide DDT (Dichloro-diphenyl-trichloroethane) has been widely used and shown to be effective against malaria-transmitting mosquitoes, but several health-threatening effects of DDT have been

reported. Longnecker et al. (2001) used the data from the US Collaborative Perinatal Project to investigate the association between DDT and preterm birth, defined as delivery before 37 weeks of complete gestation. The authors showed that adjusted for other covariates, increasing concentrations of maternal serum DDE, a persistent metabolite of DDT, led to high rate of preterm birth by fitting a logistic regression model with categorized DDE levels. Dunson and Park (2008) applied a kernel stick-breaking process mixture of linear regression models to the same data with a focus on the predictive density of gestational age at delivery (GAD), concluding strong evidence of a steadily increasing left tail with DDE dose. For more information on the study design and data structure, refer to Longnecker et al. (2001).

We let x_{i1} and x_{i2} be the DDE dose for child i and the mother's age after normalization, respectively. There were $2,313$ children left in the study after removing children with GAD > 45 weeks, taken as unrealistic values in reproductive epidemiology. By running the algorithm of the GPPM approach applied to the first simulation example for 30,000 iterations with a 10,000 iteration burn-in, we obtained the estimated predictive densities of GAD at selected percentiles $(10, 30, 70, 90)$ of the empirical distribution of DDE (Figure 3), with the maternal age being fixed at its mean. The shape and location of the estimated densities did not change much at different values of the maternal age. The results also showed the left tail of the distribution increasing for high DDE dose, with the credible intervals wider at high DDE values due to relatively few observations in this region. Observe from Figure 4 that the conditional predictive mean of GAD had a slightly decreasing nonlinear trend over DDE level, while the maternal age was fixed at its mean.

In using the GPPM for conditional density estimation and quantile regression estimation, the predictor-dependent partitioning is used as a tool for flexible modeling of the conditional response distribution given the predictors through Theorem 1. However, in some cases, there may be interest in using the methodology for identifying clusters of subjects. Because the meaning of the clusters varies across the MCMC iterations, known as the label switching problem, there have been some contributions on post-processing approaches for clustering (Celeux, Hurn and Robert (2000), Stephens (2000), Dahl (2006), Lau and Green (2007)). We followed the Lau and Green (2007) approach to estimate an optimal partition.

Figure 5 contains a symmetric heatmap presenting the pairwise marginal probabilities of being grouped with another subject in the given data. There were 13 clusters as a result of the obtained optimal partition, and some summary statistics within these clusters are arranged in Table 1. All preterm births except one were grouped into four clusters. Most of the preterm births were assigned to

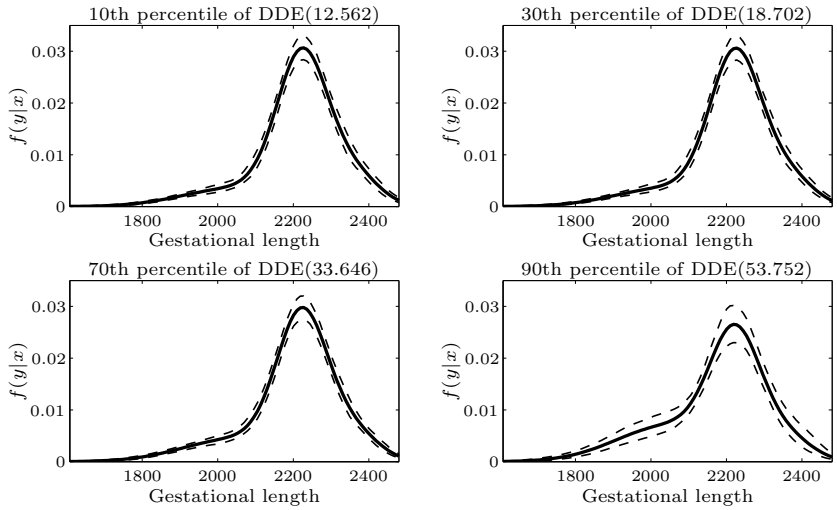

Figure 3. Estimated predictive densities (solid lines) for gestational age at delivery at preselected values of DDE with 99% pointwise credible intervals (dashed lines).

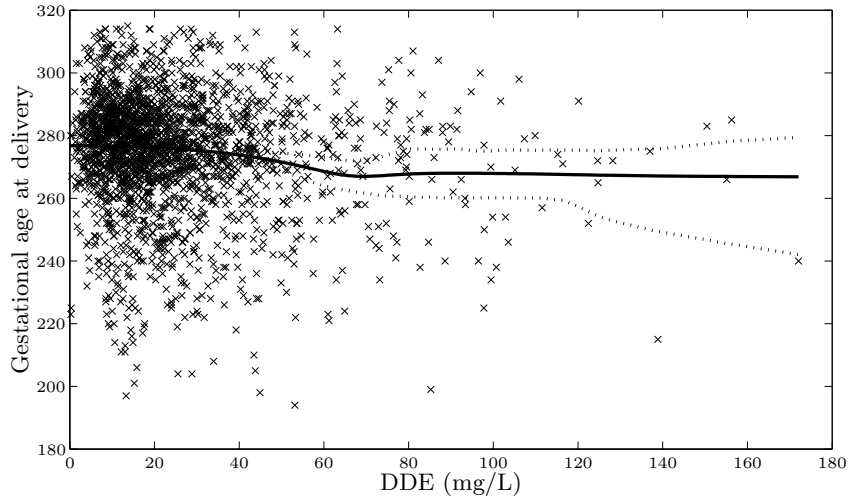

Figure 4. The conditional predictive mean of gestational age at delivery (solid line) with 99% pointwise credible intervals (dotted lines).

Cluster 6, the mean DDE level of which was about the 80th percentile of observed DDE values. Preterm births in Cluster 2 were characterized by both high DDE dose and old maternal age, while those in Clusters 11 and 13 had extreme DDE levels beyond the 98th and 99th percentiles, respectively. It is observed that most

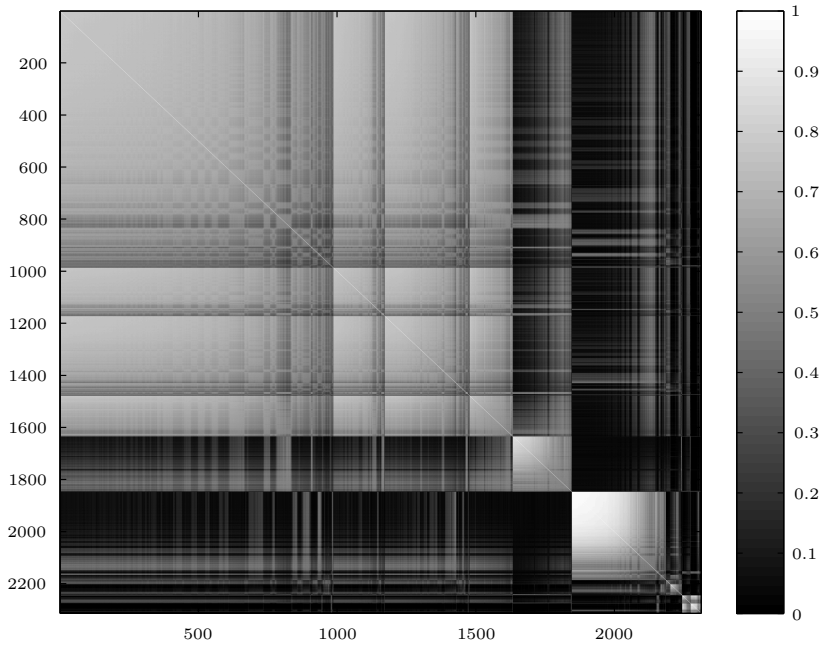

Figure 5. Pairwise marginal probabilities of being grouped with another subject in the CPP data.

of normal births in these four clusters had GAD values close to 37 weeks. Hence, the clustering result also strongly supports the contention that preterm births were more likely to be observed with high DDE dose. Note that the order of clusters is arbitrary and that some of clusters have similar mean values of GAD, but they are separately grouped due to different predictor values.

Although the results of the analysis for conditional density estimation are similar to Dunson and Park (2008), our proposed computational algorithm was considerably simpler to implement. The kernel stick-breaking process (KSBP) proposed by Dunson and Park (2008) relied on a retrospective MCMC algorithm (Papaspiliopoulos and Roberts (2008)), that involved updating of random basis locations, stick-breaking weights, atoms, and kernel parameters. In contrast, by using the present GPPM, we bypass the need to perform computation for the many unknowns characterizing the collection of predictor-dependent mixture distributions. Instead, through marginalization, relying on the simple predictor-dependent urn scheme shown in Theorem 1, we obtain a simple and efficient Gibbs sampling algorithm. We found the mixing and convergence rates to be similar to those for the MCMC algorithm of the KSBP, but the computational

Table 1. Table 1. Summary statistics by clusters.

Cluster	n^2		GAD[1] mean (SD[3])	DDE mean (SD[3])	AGE mean (SD[3])
1	985	(1)	39.7 (1.21)	26.8 (14.49)	24.0 (5.72)
2	185	(0)	40.2 (1.04)	26.6 (14.05)	23.5 (5.57)
3	306	(0)	39.2 (1.10)	28.1 (13.78)	25.7 (6.19)
4	156	(0)	41.1 (1.21)	25.5 (13.80)	24.1 (4.85)
5	212	(0)	43.3 (0.78)	26.8 (14.89)	22.9 (5.37)
6	339	(309)	34.8 (1.94)	32.4 (16.55)	22.3 (5.09)
7	16	(0)	40.1 (0.91)	30.8 (18.13)	42.3 (1.53)
8	38	(30)	35.3 (2.05)	33.2 (14.93)	38.8 (2.69)
9	4	(0)	43.4 (0.88)	39.7 (14.39)	40.8 (0.50)
10	31	(0)	40.7 (1.53)	93.1 (9.62)	23.7 (5.20)
11	33	(19)	36.2 (2.38)	101.4 (13.95)	24.2 (6.65)
12	6	(0)	39.4 (1.03)	148.0 (13.88)	23.5 (4.51)
13	2	(2)	32.5 (2.53)	161.5 (23.48)	25.0 (1.41)

[1]in weeks, [2]preterm births in parenthesis, [3]SD=standard deviation

time was substantially reduced as fewer computations were needed at each step of the MCMC algorithm.

For predictive purposes, the KSBP may be more efficient in introducing only those clusters that are needed to flexibly characterize changes with predictors in the response distribution. However, in utilizing information in the predictor distribution, the GPPM may be particularly useful in semi-supervised learning settings when there are missing predictors, and when interest focuses on inverse regression problems. Also, in many clustering applications, one would prefer to have subjects with very different predictor values, but the same response, allocated to different clusters.

7. Discussion

Model-based clustering and mixture modeling have become routine tools in a wide variety of application areas, including machine learning and biomedical research. With a few notable exceptions, the literature on these topics has focused on finite mixture models that do not allow the weights on the mixture components, and hence the probability of allocation to clusters, to depend on predictors. This leads to some notable problems. First, in clustering there are often predictors available that inform the clustering process. For example, in clustering individuals based on their drinking behavior, information can be obtained on predictors of alcohol dependence and behavioral factors (Caetano and Cunradi (2002)). By including this information, one can obtain more interpretable clusters, while also obtaining the insight into the role of predictors, that is often of

primary interest. In the alcohol dependence application, information on predictors of allocation to a heavy drinking cluster can be used in targeting interventions. Although finite latent class mixture models can be used in such settings, infinite mixture models, such as the model underlying our proposed GPPM, are more realistic in allowing clusters to be slowly introduced without bound as the sample size increases.

In addition, partitioning is used to generate flexible classes of models. Much of the recent literature has relied on Dirichlet process-based clustering, an approach closely related to product partition models (PPMs). Our contribution is to develop a simple modification to PPMs to allow predictor dependent clustering, while bypassing the need for consideration of complex nonparametric Bayes methods for collections of predictor-dependent random probability measures. The resulting class of generalized PPMs (GPPMs) should be widely useful as a tool for generating new classes of models and for efficient computation in existing models, such as hierarchical mixtures-of-experts models.

Perhaps the most interesting result is the proposed class of predictor-dependent urn schemes, that generalize the Blackwell and MacQueen Pólya urn scheme in a natural manner to include weights that depend on the distances between subjects predictor values. The distance metric is induced through a flexible nonparametric joint model for the predictors. Although this approach may be viewed as unnatural when the predictors are not random variables but fixed by design, (3.6) can be viewed as an auxiliary model that is only defined to induce a coherent probability model on the conditional distribution of the response y given the fixed $\mathbf{x}$, and therefore the proposed class of predictor-dependent urn schemes is still valid.

Acknowledgement

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. The authors thank Shyamal Peddada, Liaming Wang, and Yeonseung Chung for their helpful comments on a draft of this manuscript. The authors also greatly appreciate critical comments from an associated editor and the referees, that led to a number of improvements in the exposition.

Appendix A: Proof of Theorem 1

The Pólya urn scheme in (2.4) can be reexpressed with a vector of unique values $\boldsymbol{\theta}^{(i)}$ and configuration $\mathbf{S}^{(i)}$ as

$$\left(\phi_i \,|\, \boldsymbol{\phi}^{(i)}\right) \sim \left(\frac{\alpha}{\alpha+n-1}\right) G_0(\phi_i) + \left(\frac{1}{\alpha+n-1}\right) \sum_{h=1}^{k^{(i)}} k_h^{*(i)} \delta_{\theta_h^{(i)}}(\phi_i).$$

Then, using (2.6), the joint distribution of $\boldsymbol{\phi}$ is

$$\begin{aligned}
\pi(\boldsymbol{\phi}) &= \pi(\phi_i|\boldsymbol{\phi}^{(i)})\pi(\boldsymbol{\phi}^{(i)}) \\
&= \left\{\left(\frac{\alpha}{\alpha+n-1}\right)G_0(\phi_i) + \left(\frac{1}{\alpha+n-1}\right)\sum_{h=1}^{k^{(i)}} k_h^{*(i)}\delta_{\theta_h^{(i)}}(\phi_i)\right\} \\
&\quad \times\left\{\sum_{\mathcal{S}^{*(i)}} \frac{1}{\prod_{l=1}^{n-1}(\alpha+l-1)}\prod_{m=1}^{k^{(i)}}\alpha(k_m^{(i)}-1)!G_0(\boldsymbol{\phi}_{m,1}^{(i)})\prod_{j=2}^{k_m^{(i)}}\delta_{\boldsymbol{\phi}_{m,1}^{(i)}}(\boldsymbol{\phi}_{m,j}^{(i)})\right\}, \\
&= \alpha c_0 G_0(\phi_i)\left\{\sum_{\mathcal{S}^{*(i)}\in\mathcal{P}^{(i)}}\prod_{m=1}^{k^{(i)}} c(\mathbf{S}_m^{*(i)})G_0(\boldsymbol{\phi}_{m,1}^{(i)})\prod_{j=2}^{k_m^{(i)}}\delta_{\boldsymbol{\phi}_{m,1}^{(i)}}(\boldsymbol{\phi}_{m,j}^{(i)})\right\} \\
&\quad + c_0\sum_{h=1}^{k^{(i)}} k_h^{*(i)}\left\{\sum_{\mathcal{S}^{*(i)}}\prod_{m=1}^{k^{(i)}} c(\mathbf{S}_m^{*(i)})G_0(\boldsymbol{\phi}_{m,1}^{(i)})\{\delta_{\boldsymbol{\phi}_{m,1}^{(i)}}(\phi_i)\}^{1(m=h)}\prod_{j=2}^{k_m^{(i)}}\delta_{\boldsymbol{\phi}_{m,1}^{(i)}}(\boldsymbol{\phi}_{m,j}^{(i)})\right\},
\end{aligned}$$

where $c_0 = \prod_{i=1}^{n}(\alpha+l-1)^{-1}$, $c(\mathbf{S}_h^{*(i)}) = \alpha(k_h^{*(i)}-1)!$, and $1(\cdot)$ is an indicator function. By setting $\phi = (\gamma,\varphi)'$ and doing the same thing to obtain (3.3), we can obtain the joint distribution of $\boldsymbol{\varphi}$ and $\mathbf{X}$ as

$$\begin{aligned}
\pi(\boldsymbol{\varphi},\mathbf{X}) &= \alpha c_0 G_{0\varphi}(\varphi_i)\int f_2(\mathbf{x}_i|\gamma)dG_{0\gamma}(\gamma)\sum_{\mathcal{S}^{*(i)}}\Bigg[\prod_{m=1}^{k^{(i)}} c(\mathbf{S}_m^{*(i)}) \\
&\quad \times\left\{\int\prod_{i\in\mathbf{S}_m^{*(i)}} f_2(\mathbf{x}_i|\gamma)dG_{0\gamma}(\gamma)\right\}G_{0\varphi}(\boldsymbol{\varphi}_{m,1})\prod_{j=2}^{k_m^{(i)}}\delta_{\boldsymbol{\varphi}_{m,1}^{(i)}}(\boldsymbol{\varphi}_{m,j}^{(i)})\Bigg] \\
&\quad + c_0\sum_{h=1}^{k^{(i)}} k_h^{*(i)}\delta_{\varphi_h^{*(i)}}(\varphi_i)\sum_{\mathcal{S}^{*(i)}}\Bigg[\prod_{m=1}^{k^{(i)}} c(\mathbf{S}_m^{*(i)}) \\
&\quad \times\left\{\int f_2(\mathbf{x}_i|\gamma)^{1(m=h)}\prod_{l\in\mathbf{S}_m^{*(i)}} f_2(\mathbf{x}_l|\gamma)dG_{0\gamma}(\gamma)\right\}G_{0\varphi}(\boldsymbol{\varphi}_{m,1})\prod_{j=2}^{k_m^{(i)}}\delta_{\boldsymbol{\varphi}_{m,1}^{(i)}}(\boldsymbol{\varphi}_{m,j}^{(i)})\Bigg].
\end{aligned}$$

By Bayes rule the curly bracket in the second term of the last equation can be reexpressed as

$$\begin{aligned}
&\int f_2(\mathbf{x}_i|\gamma)^{1(m=h)}\prod_{l\in\mathbf{S}_m^{*(i)}} f_2(\mathbf{x}_l|\gamma)dG_{0\gamma}(\gamma) \\
&= \int\prod_{l\in\mathbf{S}_m^{*(i)}} f_2(\mathbf{x}_l|\gamma)dG_{0\gamma}(\gamma)\int f_2(\mathbf{x}_i|\gamma)^{1(m=h)}\frac{\prod_{l\in\mathbf{S}_m^{*(i)}} f_2(\mathbf{x}_l|\gamma)}{\int\prod_{l\in\mathbf{S}_m^{*(i)}} f_2(\mathbf{x}_l|\gamma)dG_{0\gamma}(\gamma)}dG_{0\gamma}(\gamma)
\end{aligned}$$

$$= \int \prod_{l \in \mathbf{S}_m^{*(i)}} f_2(\mathbf{x}_l|\gamma) dG_{0\gamma}(\gamma) \int f_2(\mathbf{x}_i|\gamma)^{1(m=h)} dG^*_{0\gamma}(\gamma|\mathbf{X}_m^{(i)}),$$

where $\mathbf{X}_m^{(i)} = \{\mathbf{x}_i | i \in \mathbf{S}_m^{*(i)}\}$, and $G^*_{0\gamma}(\gamma|\mathbf{X}_m^{(i)})$ is the posterior distribution of γ updated with the likelihood of $\mathbf{X}_m^{(i)}$. Therefore, the joint distribution of $\boldsymbol{\varphi}$ and $\mathbf{X}$ is simplified as

$$\begin{aligned}
\pi(\boldsymbol{\varphi}, \mathbf{X}) = \Big\{ & \alpha \int f_2(\mathbf{x}_i|\gamma) dG_{0\gamma}(\gamma) G_{0\varphi}(\varphi_i) \\
& + \sum_{h=1}^{k^{(i)}} k_h^{*(i)} \int f_2(\mathbf{x}_i|\gamma) dG^*_{0\gamma}(\gamma|\mathbf{X}_m^{(i)}) \delta_{\varphi_h^{*(i)}}(\varphi_i) \Big\} \\
& \times c_0 \sum_{\mathcal{S}^{*(i)}} \prod_{m=1}^{k^{(i)}} c(\mathbf{S}_m^{*(i)}) \Big\{ \int \prod_{i \in \mathbf{S}_m^{*(i)}} f_2(\mathbf{x}_i|\gamma) dG_{0\gamma}(\gamma) \Big\} \\
& \times G_{0\varphi}(\boldsymbol{\varphi}_{m,1}) \prod_{j=2}^{k_m^{(i)}} \delta_{\boldsymbol{\varphi}_{m,1}^{(i)}}(\boldsymbol{\varphi}_{m,j}^{(i)}).
\end{aligned}$$

Marginalizing the above equation over φ_i and dividing it by $\pi(\boldsymbol{\varphi}^{(i)}, \mathbf{X})$ completes the proof.

Appendix B: Gibbs Sampling Steps for Mixture of Normals

Here we provide the full conditional posterior distributions used in implementing the Gibbs sampler for the predictor-dependent mixture of normals model used in Sections 5.2 and 6. The Gibbs sampler alternates between the following sampling steps.

1. Sample S_i, $i = 1, \ldots, n$, from a multinomial distribution with weights

$$q_{i,h} = \begin{cases} \widetilde{c} t_p(\mathbf{x}_i; m_{\mathbf{x},0}, E_{\mathbf{x},0}, V_{\mathbf{x},0}) t_1(y_i; m_{y,0}, E_{y,0}, V_{y,0}) & \text{for} \quad h = 0 \\ \widetilde{c} t_p(\mathbf{x}_i; m_{\mathbf{x},h}, E_{\mathbf{x},h}, V_{\mathbf{x},h}) N(y_i; \mu_h, \sigma^2_{y,h}) & \text{for} \quad h = 1, \ldots, k^{(i)}, \end{cases}$$

where $m_{\mathbf{x},0} = \nu_{\mathbf{x}} - p + 1$, $E_{\mathbf{x},0} = \boldsymbol{\mu}_{0\mathbf{x}}$, $V_{\mathbf{x},0} = (c_{\mathbf{x}} + c_\mu)/(\nu c_{\mathbf{x}} c_\mu)\boldsymbol{\Sigma}_{0\mathbf{x}}$, $m_{y,0} = 2a_y$, $E_{y,0} = \mu$, $V_{y,0} = b_y(\kappa + 1)/(a_y \kappa)$, $m_{\mathbf{x},h} = \nu^*_{\mathbf{x}} - p + 1$, $E_{\mathbf{x},h} = \boldsymbol{\mu}^*_{0\mathbf{x}}$, $V_{\mathbf{x},h} = (c_{\mathbf{x}} + c^*_\mu)/\{(\nu^*_{\mathbf{x}} - p + 1) c_{\mathbf{x}} c^*_\mu\} \boldsymbol{\Sigma}^*_{0\mathbf{x}}$, with $\nu^*_{\mathbf{x}}$, $\boldsymbol{\mu}^*_{0\mathbf{x}}$, c^*_μ, and $\boldsymbol{\Sigma}^*_{0\mathbf{x}}$ being defined in (3.7), and $\widetilde{c}$ a normalizing constant. On the completion of sampling $\mathbf{S}$, $\{\varphi_i\}_{i:S_i=0}$ with $\varphi_i = (\mu_i, \sigma^2_{y,i})$ are assigned to an *i.i.d.* sample from

$$G_{0\varphi,i}(\mu_i, \sigma^{-2}_{y,i}) = N\Big(\mu_i; \frac{(y_i + \kappa\mu_0)}{\kappa + 1}, \frac{\sigma^2_{y,h}}{\kappa + 1}\Big) \times \mathcal{G}\Big(\sigma^{-2}_{y,i}; a_y + \frac{1}{2}, b_y + \frac{\kappa(y_i - \mu_0)^2}{2(\kappa + 1)}\Big).$$

2. Sample $\sigma_{y,h}^{-2}$ and μ_h given $\sigma_{y,h}^2$, respectively, from

$$\sigma_{y,h}^{-2} \sim \mathcal{G}\left(a_y + \frac{k_h^*}{2},\, b_y + \frac{1}{2}\sum_{i:S_i=h}(y_i - \bar{y}_h)^2 + \frac{k_h^*\kappa}{2(k_h^*+\kappa)}(\bar{y}_h - \mu)^2\right),$$

$$\mu_h \sim N\left(\frac{k_h^*\bar{y}_h + \kappa\mu}{k_h^*+\kappa},\, \frac{\sigma_{y,h}^2}{k_h^*+\kappa}\right),$$

where $\bar{y}_h = \sum_{i:S_i=h} y_i/k_h^*$.

3. Sample $\psi = (\mu, \kappa)$ from

$$\mu \sim N\left(\left\{\sigma_\mu^{-2} + \kappa\sum_{h=1}^k \sigma_{y,h}^{-2}\right\}^{-1}\left\{\sigma_\mu^{-2}\mu_0 + \kappa\sum_{h=1}^k \sigma_{y,h}^{-2}\mu_h\right\},\, \sigma_\mu^{-2} + \kappa\sum_{h=1}^k \sigma_{y,h}^{-2}\right),$$

$$\kappa \sim \mathcal{G}\left(a_\kappa + \frac{k}{2},\, b_\kappa + \frac{1}{2}\sum_{h=1}^k \sigma_{y,h}^{-2}(\mu_h - \mu)^2\right).$$

Note that all the distributions above are conditional on the rest of parameters. The conditional part is omitted due to limited space. A Matlab program is available upon request by sending an email to parkj3@mail.nih.gov.

References

Barry, D. and Hartigan, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20**, 260-279.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353-355.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika* **83**, 275-285.

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2006). Bayesian inference for dynamic models with Dirichlet process mixtures. *International Conference on Information Fusion*, Florence, Italia.

Caetano, R. and Cunradi, C. (2002). Alcohol dependence: a public health perspective. *Addiction* **97**, 633-645.

Carvalho, A. X. and Tanner, M. A. (2005). Modeling nonlinear time series with local mixtures of generalized linear models. *Canad. J. Statist.* **33**, 97-113.

Celeux, G., Hurn, M. and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95**, 957-979.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics* (Kim-Anh Do, Peter Müller, Marina Vannucci Eds.), Cambridge University Press.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99**, 205-215.

Dunson, D. B., Pillai, N. and Park, J.-H. (2007). Bayesian density regression. *J. Roy. Statist. Soc. Ser. B* **69**, 163-183.

Dunson, D. B. and Peddada, S. D. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrika* **95**, 859-874.

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

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89**, 268-277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577-588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615-629.

Ge, Y. and Jiang, W. (2006). On consistency of Bayesian inference with mixtures of logistic regression. *Neural Comp.* **18**, 224-243.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100**, 1021-1035.

Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 179-194.

Hartigan, J. A. (1990). Partition models. *Commun. Statist.* **A 19**, 2745-2756.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.

Holmes, C. C., Denison, D. G. T., Ray, S. and Mallick, B. K. (2005). Bayesian Prediction via Partitioning. *J. Comput. Graph. Statist.* **14**, 811-830.

Ishwaran, H. and James, L. F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica* **13**, 1211-1235.

Jiang, W. X. and Tanner, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood. *Ann. Statist.* **27**, 987-1011.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comp.* **6**, 181-214.

Kurihara, K., Welling, M. and Vlassis, N. (2006). Accelerated variational Dirichlet mixture models. *Advances in Neural Information Processing Systems* **19**.

Lau, J. W. and Green, P. J. (2007). Bayesian model-based clustering procedures. *J. Comput. Graph. Statist.* **16**, 526-558.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351-357.

Longnecker, M. P., Klebanoff, M. A., Zhou, H. B. and 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**, 110-114.

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

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), pp551-560 Creta: ISBA.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models, *J. Comput. Graph. Statist.* **7**, 223-238.

MacEachern, S. N., Clyde, M. and Liu, J. (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *Canad. J. Statist.* **27**, 251-267.

Müller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67-79.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9**, 249-265.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169-186.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory* (Edited by T. S. Ferguson, L. S. Shapley and J. B. MacQueen). IMS Lecture Notes-Monograph series **30**.

Quintana, F. A. (2006). A predictive view of Bayesian clustering. *J. Statist. Plann. Inference* **136**, 2407-2429.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *J. Roy. Statist. Soc. Ser. B*, **65**, 557-574.

Quintana, F. A. and Newton, M. A. (2000). Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences. *J. Comput. Graph. Statist.* 9, 711-737.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4**, 639-650.

Stephens, M. (2000). Dealing with label switching in mixture models. *J. Roy. Statist. Soc. Ser. B* **62**, 795-809.

Biostatistics Branch, National Cancer Institute, EPS 8049, 6120 Executive Blvd, Rockville, MD 20852, U.S.A.

E-mail: parkj3@mail.nih.gov

Department of Statistical Science, Duke University, 218 Old Chemistry Building, Durham, NC 27708-0251, U.S.A.

E-mail: dunson@stat.duke.edu

(Received July 2008; accepted March 2009)