Bayesian generalized product partition model

BY JU-HYUN PARK

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

parkj3@niehs.nih.gov

AND DAVID B. DUNSON

Biostatistics Branch, National Institute of Environmental Health Sciences, P.O. Box 12233, RTP, NC 27709, USA

dunson1@niehs.nih.gov

SUMMARY

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. Properties are discussed, 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.

Some key words: Clustering; Conditional distribution estimation; Dirichlet process; Generalized Pólya urn; Latent class model; Mixture of experts; Nonparametric Bayes; Product partition.

1. Introduction

With the increasing need for flexible tools for clustering, density estimation, dimensionality reduction and discovery of latent structure in high dimensional data, mixture models are now used routinely in a wide variety of application areas ranging from genomics to machine learning. Much of this work has focused on finite mixture models of the form:

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

where k is the number of mixture components, π_h is the probability weight assigned to component h, and $f_h(\cdot \,|\, \theta_h)$ is a distribution in a parametric family characterized by the finite-dimensional θ_h, for $h = 1, \ldots, k$. For a review of the use of (1) in clustering and density estimation, refer to Fraley and Raftery (2002).

In order to generalize (1) to incorporate predictors, one can model predictor dependence in $\pi = (\pi_1, \ldots, \pi_k)'$ and/or $f_h(\theta_h)$, $h = 1, \ldots, k$, as follows:

$$f(y \,|\, x) = \sum_{h=1}^{k} \pi_h(x) \, f_h(y \,|\, \theta_h, x). \tag{2}$$

For example, hierarchical mixtures-of-experts models (Jordan and Jacobs, 1994) characterize $\pi_h(x)$ using a probabilistic decision tree, while letting $f_h(y \,|\, \theta_h) = N(y; x'\beta_h, \tau_h^{-1})$ correspond to the conditional density for a normal linear model. The term "expert" corresponds to the choice of $f_h(y \,|\, \theta_h, x)$, as different experts in a field may have different parametric models for the conditional distribution. 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).

In this article, our goal is to develop a flexible semiparametric Bayes framework for predictor-dependent clustering and conditional distribution modeling. Potentially, we could simply rely on (2), as predictor-dependent clustering will naturally arise through the allocation of subjects sampled from (2) to experts. However, a concern is the sensitivity to the

choice of the number of experts, k. A common strategy is to fit mixture models having different numbers of components, with the AIC or BIC used to select the model with the best fit penalized for model complexity. Unfortunately, these criteria are not appropriate for mixture models and other hierarchical models in which the number of parameters is unclear. For this reason, there has been recent interest in defining new model selection criteria that are appropriate for mixture models. Some examples include the DIC (Spiegelhalter et al., 2002) and the MRC (Naik et al., 2007).

Even if an appropriate criteria is defined, it is not clear that a finite mixture model can provide an accurate characterization of the data. For example, suppose that there are k mixture components represented in a current data set having n subjects and one performs model selection based on this data set. Then the assumption is that future subjects will belong to one of these k mixture components. It seems much more realistic to suppose that there are infinitely many components, or latent attributes, in the general population, with finitely many of these components represented in the current data set. Such infinite mixture models would allow a new subject to have a new attribute that is not yet represented, allowing discovery of new components as observations are added.

There is a rich Bayesian literature on infinite mixture models, which let $k \rightarrow \infty$ in expression (1). This is accomplished by letting $y_i \sim f(\phi_i)$, with $\phi_i \sim G$, where $G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}$, with $\pi = \{\pi_h\}_{h=1}^{\infty}$ an infinite sequence of probability weights, δ_θ a probability measure concentrated at θ, and $\theta = \{\theta_h\}_{h=1}^{\infty}$ an infinite sequence of atoms. A wide variety of priors have been proposed for G, with the most common choice being the 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 simplify computation and 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 focus is on further generalizing PPMs to include predictor-dependence by starting with (2) in the $k = \infty$ case, and attempting to obtain a prior which results in a PPM upon marginalization. There has been considerable recent interest in the Bayesian nonparametric literature on developing priors for predictor-dependent collections of random probability measures. Starting with the Sethuraman (1994) stick-breaking representation of the DP, MacEachern (1999; 2001) proposed a class of dependent DP (DDP) priors. In the fixed π case, DDP priors have been successfully implemented in ANOVA modeling (De Iorio et al., 2004), spatial data analysis (Gelfand et al., 2005), time series (Caron et al., 2006) and stochastic ordering (Dunson and Peddada, 2007) applications. Unfortunately, the fixed π 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, 2007).

In order to avoid the need for computation of the very many parameters characterizing these nonparametric priors, we focus instead on obtaining a generalized product partition model (GPPM) through relying on a related specification to Müller et al (1996). 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. Section 4 describes a simple and efficient Gibbs sampler for posterior computation. Section 5 contains an application,

and Section 6 discusses the results.

2. Product Partition Models and Dirichlet Process Mixtures

Let $\mathbf{S}^* = (\mathbf{S}_1^*, \ldots, \mathbf{S}_k^*)$ denote a partition of $\{1, \ldots, n\}$, with the elements of $\mathbf{S}_h^*$ corresponding to the ids of those subjects in cluster h. Letting $\mathbf{y}_h = \{y_i : i \in \mathbf{S}_h^*\}$ denote the data for subjects in cluster h, for $h = 1, \ldots, k$, PPMs are defined as follows:

$$f(\mathbf{y}|\mathbf{S}^*) = \prod_{h=1}^{k} f_h(\mathbf{y}_h), \quad \pi(\mathbf{S}^*) = c_0 \prod_{h=1}^{k} c(\mathbf{S}_h^*), \tag{3}$$

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, the elements of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ are independently and identically distributed with prior G_0, $c(\mathbf{S}_h^*)$ is a nonnegative cohesion and c_0 is a normalizing constant. The posterior distribution of the partition $\mathbf{S}^*$ given $\mathbf{y}$ also has a PPM form, but with the posterior cohesion $c(\mathbf{S}_h^*) f_h(\mathbf{y}_h)$.

Note that a PPM can be induced through the hierarchical specification:

$$\begin{aligned} y_i \,|\, \boldsymbol{\theta}, \mathbf{S} &\overset{ind}{\sim} f(\theta_{S_i}), \\ S_i &\overset{iid}{\sim} \sum_{h=1}^{k} \pi_h \delta_h, \quad \theta_h \overset{iid}{\sim} G_0, \end{aligned} \tag{4}$$

where $S_i = h$ if $i \in \mathbf{S}_h^*$ indexes membership of subject i in cluster h, with $\mathbf{S} = (S_1, \ldots, S_n)'$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ are probability weights, and taking $k \to \infty$ induces a nonparametric PPM. Equivalently, one can let $y_i \sim f(\phi_i)$ with $\phi_i \sim G$ and $G = \sum_{h=1}^{k} \pi_h \delta_{\theta_h}$. A prior on the weights $\boldsymbol{\pi}$ induces a particular form for $\pi(\mathbf{S}^*)$, 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 $\boldsymbol{\phi}^{(i)} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)'$ and marginalizing out G is

$$(\phi_i \,|\, \boldsymbol{\phi}^{(i)}) \sim \left(\frac{\alpha}{\alpha + n - 1}\right) G_0(\phi_i) + \left(\frac{1}{\alpha + n - 1}\right) \sum_{j \neq i} \delta_{\phi_j}(\phi_i), \tag{5}$$

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 $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)'$ is obtained as

$$\pi(\boldsymbol{\phi}) = \prod_{i=1}^{n} \Big\{ \frac{\alpha G_0(\phi_i) + \sum_{j<i} \delta_{\phi_j}(\phi_i)}{\alpha + i - 1} \Big\}. \tag{6}$$

Let $k = n(\mathbf{S}^*)$ denote the number of partition sets, with $k_h = n(\mathbf{S}^*_h)$ the cardinality of $\mathbf{S}^*_h$. Letting $\boldsymbol{\phi}_h = \{\phi_i : i \in \mathbf{S}^*_h\}$, with $\boldsymbol{\phi}_{h,l}$ being the parameter for the lth subject, ordered by the ids, in cluster h, Quintana and Iglesias (2003) show that (6) is equivalent to

$$\begin{aligned} \pi(\boldsymbol{\phi}) &= \sum_{\mathbf{S}^* \in \mathcal{P}} \frac{1}{\prod_{l=1}^{n}(\alpha + l - 1)} \prod_{h=1}^{k} \alpha(k_h - 1)! G_0(\boldsymbol{\phi}_{h,1}) \prod_{j=2}^{k_h} \delta_{\boldsymbol{\phi}_{h,1}}(\boldsymbol{\phi}_{h,j}) \\ &= c_0 \sum_{\mathbf{S}^* \in \mathcal{P}} \prod_{h=1}^{k} c(\mathbf{S}^*_h) \pi_h(\boldsymbol{\phi}_h), \end{aligned} \tag{7}$$

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 obtained as

$$f(\mathbf{y}) = c_0 \sum_{\mathbf{S}^* \in \mathcal{P}} \prod_{h=1}^{k} c(\mathbf{S}^*_h) \int \prod_{i \in \mathbf{S}^*_h} f(y_i|\theta) dG_0(\theta), \tag{8}$$

which is a special case of the form implied by (3) corresponding to a PPM with cohesion $c(\mathbf{S}^*_h) = \alpha(n(\mathbf{S}^*_h) - 1)!$. This implies that simple and efficient 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 $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ into a class of PPMs, so that the prior on the partition $\mathbf{S}^*$ has the form

$$\pi(\mathbf{S}^*|\mathbf{X}) \propto \prod_{h=1}^{k} c(\mathbf{S}^*_h, \mathbf{X}_h), \tag{9}$$

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 (9) has two appealing properties. First, the posterior distribution of the partition $\mathbf{S}^*$ 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 an direct influence of predictors $\mathbf{X}$ 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 expression (3), 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 PPM and DPMs. For simplicity of notation, we focus on univariate response y, though multivariate generalizations are straightforward. Suppose $\mathbf{z}_i = (y_i, \mathbf{x}_i')'$ follows the hierarchical model:

$$\begin{aligned} f(\mathbf{z}_i \,|\, \boldsymbol{\phi}_i) &= f(y_i, \mathbf{x}_i | \varphi_i, \gamma_i) = f_1(y_i | \mathbf{x}_i, \varphi_i) f_2(\mathbf{x}_i | \gamma_i), \\ \boldsymbol{\phi}_i &\sim G, \quad G \sim DP(\alpha G_0), \end{aligned} \tag{10}$$

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 (10), we can obtain a joint distribution of $\boldsymbol{\phi} = (\boldsymbol{\varphi}, \boldsymbol{\gamma})$ using the same approach used in deriving expression (7). 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 given by

$$\pi(\boldsymbol{\varphi}, \mathbf{X}) = \sum_{\mathbf{S}^* \in \mathcal{P}} c_0 \prod_{h=1}^{k} \alpha (k_h - 1)! \Big\{ \int \prod_{i \in \mathbf{S}_h^*} f_2(\mathbf{x}_i | \gamma_h^*) dG_{0\gamma}(\gamma_h^*) \Big\} G_{0\varphi}(\boldsymbol{\varphi}_{h,1}) \prod_{j=2}^{k_h} \delta_{\boldsymbol{\varphi}_{h,1}}(\boldsymbol{\varphi}_{h,j}), \tag{11}$$

where $\boldsymbol{\varphi}_{h,l}$ is the parameter for the response y of the lth subject, ordered by the ids, in

cluster h, and therefore the conditional distribution of $\boldsymbol{\varphi}$ given $\mathbf{X}$ is

$$\begin{aligned}\pi(\boldsymbol{\varphi}|\mathbf{X}) &= c_0^* \sum_{\mathbf{S}^* \in \mathcal{P}} \prod_{h=1}^{k} \alpha(k_h - 1)! \Big\{ \int \prod_{i \in \mathbf{S}_h^*} f_2(x_i|\gamma_h^*) dG_{0\gamma}(\gamma_h^*) \Big\} G_{0\varphi}(\boldsymbol{\varphi}_{h,1}) \prod_{j=2}^{k_h} \delta_{\boldsymbol{\varphi}_{h,1}}(\boldsymbol{\varphi}_{h,j}) \\ &= c_0^* \sum_{\mathbf{S}^* \in \mathcal{P}} \prod_{h=1}^{k} c(\mathbf{S}_h^*, \mathbf{X}_h) \pi_h(\boldsymbol{\varphi}_h), \end{aligned} \tag{12}$$

where c_0^* is a normalizing constant, so that the sum over $\mathcal{P}$ is unity, $c(\mathbf{S}_h^*, \mathbf{x}_h) = \alpha(k_h - 1)! \int \prod_{i \in \mathbf{S}_h^*} f_2(\mathbf{x}_i|\gamma) dG_{0\gamma}(\gamma)$, and $\pi_h(\boldsymbol{\varphi}_h)$ is a prior on partitioned set $\boldsymbol{\varphi}_h$. Hence, we have induced a generalized PPM (GPPM) of the form shown in (9) starting with a joint DPM model for the response and predictors related to that proposed by Müller et al. (1996). A related idea was independently developed by Fernando Quintana and collaborators in recent work (unpublished communication), though our subsequent development differs from theirs.

3·2. *Generalized Pòlya urn scheme*

It is not obvious from expression (12) how the predictor and hyperparameter values impact clustering. However, as shown in Theorem 1, we can show that the proposed GPPM induces a simple predictor-dependent generalization of the Blackwell and MacQueen (1973) Pólya urn scheme, which should be useful both in interpretation and posterior computation.

Theorem 1. Let superscript (i) on any matrix or vector indicate that the contribution of subject i has been removed. 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 $\mathbf{X}$, has the form

$$(\varphi_i \,|\, \alpha, \boldsymbol{\varphi}^{*(i)}, \mathbf{S}^{(i)}, \mathbf{X}), \sim w_0(\mathbf{x}_i) G_{0\varphi} + \sum_{h=1}^{k^{(i)}} w_h(\{\mathbf{x}_i, \mathbf{X}_h^{(i)}\}) \delta_{\varphi_h^{*(i)}}, \tag{13}$$

with the probability weights

$$w_0(\mathbf{x}_i) = c\alpha \int f_2(\mathbf{x}_i|\gamma) dG_{0\gamma}(\gamma), \quad w_h(\{\mathbf{x}_i, \mathbf{X}_h^{(i)}\}) = c k_h^{(i)} \int f_2(\mathbf{x}_i|\gamma) dG_{0\gamma}^*(\gamma|\mathbf{X}_h^{(i)}),$$

where 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, subject i is more likely to be grouped into cluster h if the predictor value of subject i, $\mathbf{x}_i$, is close to those of other subjects in the hth cluster, $\mathbf{X}_h$, with the measure of closeness depending on the scale of the data through the choice of $f_2(\cdot)$.

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 through 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 $\mathbf{x}$, $f_2(\mathbf{x}|\gamma) = \delta_\gamma(\mathbf{x})$, formulation (13) reduces to the Blackwell and MacQueen Pòlya urn scheme of expression (5).

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 $\mathbf{X}$, which is invariant to permutations in the ordering of the subjects. It is in general very difficult to define a predictor-dependent urn scheme, which satisfies these conditions.

The use of the conjugacy simplifies the weights in (13), resulting in a closed and simple form for computation. 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 $p \times 1$ predictors $\mathbf{x}$ and parameter $\gamma = (\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}})'$:

$$\begin{aligned} \mathbf{x}|\boldsymbol{\mu}_{\mathbf{x}}, c_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}} &\sim N(\boldsymbol{\mu}_{\mathbf{x}}, c_{\mathbf{x}}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}}), \\ \boldsymbol{\mu}_{\mathbf{x}}|\mu_{\mathbf{x}}, c_{\mu}, \boldsymbol{\Sigma}_{0\mathbf{x}} &\sim N(\boldsymbol{\mu}_{0\mathbf{x}}, c_{\mu}^{-1}\boldsymbol{\Sigma}_{\mathbf{x}}) \\ \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}|\nu_{\mathbf{x}}, \boldsymbol{\Sigma}_{0\mathbf{x}} &\sim \mathcal{W}(\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1}, \nu_{\mathbf{x}}), \end{aligned} \tag{14}$$

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 (13) is a noncentral multivariate t-distribution with degrees of freedom $\nu = \nu_{\mathbf{x}} - p + 1$, mean $\boldsymbol{\mu} = \boldsymbol{\mu}_{0\mathbf{x}}$, and scale $\boldsymbol{\Sigma} = (c_{\mathbf{x}} + c_{\mu})/(\nu c_{\mathbf{x}} c_{\mu})\boldsymbol{\Sigma}_{0\mathbf{x}}$:

$$f(\mathbf{x}|\boldsymbol{\mu}, \nu, \boldsymbol{\Sigma}) = \frac{\Gamma((\nu+p)/2)}{(\pi\nu)^{p/2}\Gamma(\nu/2)|\boldsymbol{\Sigma}|^{1/2}}\left(1 + \frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-(\nu+p)/2}, \tag{15}$$

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

$$\begin{aligned} \boldsymbol{\mu}_{0\mathbf{x}}^{*} &= \frac{c_{\mu}\boldsymbol{\mu}_{0\mathbf{x}} + c_{\mathbf{x}}k_h^{(i)}\bar{\mathbf{x}}_h^{(i)}}{c_{\mu} + c_{\mathbf{x}}k_h^{(i)}}, \quad c_{\mu}^{*} = c_{\mu} + c_{\mathbf{x}}k_h^{(i)}, \quad \nu_{\mathbf{x}}^{*} = \nu_{\mathbf{x}} + k_h^{(i)} \\ \boldsymbol{\Sigma}_{0\mathbf{x}}^{*} &= \left\{\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1} + k_h^{(i)}\sum_{j:S_j^{(i)}=h}(\mathbf{x}_j - \bar{\mathbf{x}}_h^{(i)})(\mathbf{x}_j - \bar{\mathbf{x}}_h^{(i)})' + \frac{k_h^{(i)}c_{\mathbf{x}}c_{\mu}}{c_{\mu} + c_{\mathbf{x}}k_h^{(i)}}(\bar{\mathbf{x}}_h^{(i)} - \boldsymbol{\mu}_{0x})(\bar{\mathbf{x}}_h^{(i)} - \boldsymbol{\mu}_{0x})'\right\}^{-1}, \end{aligned}$$

where $\bar{\mathbf{x}}_h^{(i)} = \sum_{j:S_j^{(i)}=h}\mathbf{x}_j/k_h^{(i)}$. Note that the structure in expression (14) is slightly different from a commonly used normal-Wishart prior in that a multiplicative constant is multiplied not only to the variance of the expectation of $\mathbf{x}$ but also to the variance of $\mathbf{x}$. The reasoning for this is to induce local clustering by making the distribution of $\mathbf{x}$ denser around its expected value, while the expected value can be drawn over the range of $\mathbf{x}$, with $c_{\mathbf{x}}^{-1}$ restricted to be in $(0, 1]$ and $c_{\mu}^{-1} = 1$. Allowing $c_{\mathbf{x}}$ to vary across clusters gives us additional flexibility.

In the case of discrete predictors, we can also obtain a closed form marginal likelihood of $\mathbf{x}$. In order to simplify calculations in the discrete case, we assume a priori independence for the different predictors. Suppose that x_j for $j = 1, \ldots, p$ follow 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 $p_j = b_j/(1 + b_j)$:

$$Pr(X_j = k) = \frac{\Gamma(r_j + k)}{k!\Gamma(r_j)} p_j^{r_j}(1 - p_j)^k \qquad j = 1, \ldots, p. \tag{16}$$

The marginal distribution in w_h, for $h = 1, \ldots, k^{(i)}$, 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 dealt with easily.

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 (Bush and MacEachern, 1996) 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 et al., 1999; Quintana and Newton, 2000), modified weighted Chinese restaurant (WCR) sampling algorithms (Ishwaran and James, 2003), as well as fast variational Bayes approximations (Kurihara et al., 2006).

Following Bush and MacEachern (1996), our algorithm updates the cluster specific parameters $\boldsymbol{\varphi}^*$ separately from the cluster membership indicators $\mathbf{S}$. From Theorem 1, the full

conditional posterior distribution of φ_i can be derived as follows:

$$(\varphi_i \,|\, \alpha, \boldsymbol{\varphi}^{*(i)}, \mathbf{S}^{(i)}, \mathbf{X}, \mathbf{y}), \sim q_{i,0} G_{0\varphi,i} + \sum_{h=1}^{k^{(i)}} q_{i,h} \delta_{\varphi_h^{*(i)}}, \tag{17}$$

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

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

$q_{i,0} = c w_0(\mathbf{x}_i) h_i(y_i|\mathbf{x}_i)$, $q_{i,h} = c w_h(\{\mathbf{x}_i, \mathbf{X}^{(i)}\}) f_1(y_i|\mathbf{x}_i, \varphi_h^{*(i)})$, and c is a normalizing constant. Instead of sampling directly from expression (17) in implementing the Gibbs sampling, we first sample S_i, for $i = 1, \ldots, n$, from its multinomial conditional posterior distribution with:

$$\Pr(S_i = h|\boldsymbol{\varphi}^{*(i)}, \mathbf{S}^{(i)}, \mathbf{X}, \mathbf{y}) = q_{i,h}, \qquad h = 1, \ldots, k^{(i)}, \tag{18}$$

and 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 of clusters, k is automatically updated. As a next step, we update $\boldsymbol{\varphi}^*$ conditional on $\mathbf{S}$ and k from

$$(\varphi_h|\boldsymbol{\varphi}^{*(h)}, \mathbf{S}, k, \mathbf{y}, \mathbf{X}) \propto \Big\{ \prod_{i:S_i=h} f_1(y_i|\mathbf{x}_i, \varphi_h) \Big\} G_{0\varphi}(\varphi_h). \tag{19}$$

In a case that 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

$$(\psi|\boldsymbol{\varphi}, \mathbf{y}, \mathbf{x}) \propto \pi(\psi) \Big\{ \prod_{h=1}^{k} G_{0\varphi}(\varphi_h^*|\psi) \Big\}. \tag{20}$$

We have found this algorithm to be both simple to implement and efficient in cases we have considered, as will be described in the subsequent sections.

5. Simulation Examples

5·1. *Model specification*

In this section, we illustrate the proposed method with simulations focusing on conditional density regression. We consider the following infinite mixtures-of-experts model:

$$f(y_i|\mathbf{x}_i) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}_i)\, f_1(y_i \,|\, \mathbf{x}_i, \varphi_h^*), \tag{21}$$

where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip-1})'$, $f_1(y_i|\mathbf{x}_i, \varphi_h^*) = N(y_i; \mathbf{x}_i'\boldsymbol{\beta}_h, \sigma_{y,h}^2)$, and $\varphi_h^* = (\boldsymbol{\beta}_h', \sigma_{y,h}^2)'$. The GPPM proposed in Section 3 is used to place a prior on the partition $\mathbf{S}^*$ 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 $a_j = b_j = 1$, for $j = 1, \ldots, p-1$. In addition, let $\sigma_{y,i}^{-2}|a_y, b_y \sim \mathcal{G}(a_y, b_y)$, $\boldsymbol{\beta}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta)$, $\boldsymbol{\beta}|\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0})$, and $\boldsymbol{\Sigma}_\beta^{-1}|\nu_0, \boldsymbol{\Sigma}_0 \sim \mathcal{W}(\boldsymbol{\Sigma}_0^{-1}, \nu_0)$. The last two prior distributions on $\psi = (\boldsymbol{\beta}, \boldsymbol{\Sigma}_\beta)'$ are for additional flexibility. In the implementation, we let $\alpha = 1$, $\boldsymbol{\mu}_{0\mathbf{x}} = \mathbf{0}$, $\boldsymbol{\Sigma}_{0\mathbf{x}}^{-1} = 4I_{(p-1)\times(p-1)}$, $\nu_{\mathbf{x}} = p - 1$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $\boldsymbol{\Sigma}_{\beta_0} = n(\mathbf{X}'\mathbf{X})^{-1}$, $\nu_0 = p$, $\boldsymbol{\Sigma}_0^{-1} = I_{p\times p}$, and $a_y = b_y = 0.1$. Other choices of these parameters are also considered to check sensitivity of models to our primary choice.

5·2. *Implementation and Results*

We consider two cases in which $n = 500$, $p = 2$, 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 the model with $f_1(y_i|\mathbf{x}_i, \varphi_h^*) = N(y_i; \mu_h, \sigma_{y,h}^2)$ and $\mu_h \equiv \boldsymbol{\beta}_h$ and prior specification of Section 5.1, with the MCMC algorithm of Section 4 implemented for 10,000 iterations, discarding the initial 1,000 iterations as a burn-in. Figure 1 shows selected results. The algorithm converged rapidly and mixing was good based on trace plots of $E(\mu_i)$, σ_y^{-2}, and the number of clusters (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 approximate the true values (dotted lines), which are completely embedded

within pointwise 99% credible intervals (dashed lines). The posterior mean of the number of clusters was 8.06 with a 95% credible interval of $[5, 12]$ 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 following 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 the GPPM approach specified in Section 5.1, and also using the DPM-based PPM described in Section 2. This second approach results in a mixture of normal linear regressions in which the weights are not predictor-dependent. Both analyses were run for 30,000 iterations with a 10,000 iteration burn-in, with good mixing and convergence rates in both cases based on examination of trace plots and diagnostics.

From Figure 2, it is clear that the proposed approach provides a more flexible model capturing a rapid changes in the distribution across local regions of the predictor space even for the somewhat small sample size of $n = 500$. We also repeated the analysis of the second simulation including a discrete predictor, which was obtained by truncating the continuous predictor into k groups. It was observed that the proposed method worked well for a variety choices of k (results are not shown).

6. Epidemiologic Application

We apply the proposed method to the data used in Longnecker et al. (2001) and Dunson and Park (2007). DDT has been widely used and shown to be effective against malaria-

transmitting mosquitoes, but several health-threatening effects of DDT have been also 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 (2007) 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, which are suspected as unrealistic values in reproductive epidemiology. By running the algorithm of the GPPM approach applied to the second simulation example for 30,000 iterations with a 10,000 iteration burn-in, we obtained the estimated predictive densities of GAD at selected percentiles $(10, 30, 60, 90)$ of the empirical distribution of DDE (Figure 3). The results also show that the left tail of the distribution increases for high DDE dose with the credible intervals wider at high DDE values due to the relatively few observations in this region. It is observed in Figure 4 that the conditional predictive mean of GAD had a slightly decreasing nonlinear trend over DDE level.

Although the results of the analysis were similar to Dunson and Park (2007), the proposed computational algorithm was considerably less complex and simpler to implement. Dunson and Park (2007) relied on a retrospective MCMC algorithm (Papaspiliopoulos and Roberts, 2007), which involved updating of random basis locations, stick-breaking weights, atoms and kernel parameters. In contrast, by using the GPPM proposed in the present paper, we bypass the need to perform computation for the very 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 Dunson and Park (2007), but the computational time was substantially reduced, as fewer computations were needed at each step of the MCMC algorithm.

7. Discussion

There has been increasing interest in the use of partitioning to generate flexible classes of models and to identify interesting clusters of observations for further exploration. 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 and useful of our results is the proposed class of predictor-dependent urn schemes, which generalize the Blackwell and MacQueen (1973) 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, the proposed class of predictor-dependent urn schemes are nonetheless useful and are defined conditionally on the predictor values. In this sense, the use of a joint distribution on the predictors in inducing the urn scheme can be viewed simply as a tool for proving that a coherent joint prior exists in cases in which the predictors

are not random.

ACKNOWLEDGMENT

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

Appendix A: Proof of theorem 1

The Pólya urn scheme in expression (5) can be reexpressed with a vector of unique values $\boldsymbol{\theta}^{(i)}$ and configuration $\mathbf{S}^{(i)}$:

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

Then, using expression (7), the joint distribution of $\boldsymbol{\phi}$ is

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

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 expression (11), we can obtain the joint distribution of $\boldsymbol{\varphi}$ and $\mathbf{X}$:

$$\begin{aligned}
&\pi(\boldsymbol{\varphi}, \mathbf{X}) \\
&= \alpha c_0 G_{0\varphi}(\varphi_i)\int f_2(\mathbf{x}_i|\gamma)dG_{0\gamma}(\gamma)
\end{aligned}$$

$$\times\Big\{\prod_{m=1}^{k^{(i)}} c(\mathbf{S}_m^{*(i)})\Big[\int\prod_{i\in\mathbf{S}_m^{*(i)}} f_2(\mathbf{x}_i|\gamma)dG_{0\gamma}(\gamma)\Big]G_{0\varphi}(\boldsymbol{\varphi}_{m,1})\prod_{j=2}^{k_m^{(i)}}\delta_{\boldsymbol{\varphi}_{m,1}^{(i)}}(\boldsymbol{\varphi}_{m,j}^{(i)})\Big\}$$

$$+c_0\sum_{h=1}^{k^{(i)}} k_h^{(i)}\delta_{\varphi_h^{*(i)}}(\varphi_i)$$

$$\times\Big\{\prod_{m=1}^{k^{(i)}} c(\mathbf{S}_m^{*(i)})\Big[\int f_2(\mathbf{x}_i|\gamma)^{1(m=h)}\prod_{l\in\mathbf{S}_m^{*(i)}} f_2(\mathbf{x}_l|\gamma)dG_{0\gamma}(\gamma)\Big]G_{0\varphi}(\boldsymbol{\varphi}_{m,1})\prod_{j=2}^{k_m^{(i)}}\delta_{\boldsymbol{\varphi}_{m,1}^{(i)}}(\boldsymbol{\varphi}_{m,j}^{(i)})\Big\}.$$

By Bayes rule the square bracket in the second term of the last equation can be reexpressed as follows:

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

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

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

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

REFERENCES

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

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

BUSH, C. A. & MACEACHERN, S. N. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika* **83**, 175-85.

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.

CARVALHO, A. X. & TANNER, M. A. (2005). Modeling nonlinear time series with local mixtures of generalized linear models. *Can. J. Stat.* **33**, 97-113.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. & MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *J. Am. Statist. Assoc.* **99**, 205-15.

DUNSON, D. B. & PARK, J-H. (2007). Kernel stick-breaking processes. *Biometrika*, in press.

DUNSON, D. B., PILLAI, N. & PARK, J-H. (2007). Bayesian density regression. *J. R. Statist. Soc.* **B 69**, 163-83.

DUNSON, D. B. & PEDDADA, S. D. (2007). Bayesian nonparametric inference on stochastic ordering. *Biometrika*, revision submitted.

ESCOBAR, M. D. & WEST, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *J. Am. Statist. Assoc.* **90**, 577-88.

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.

FRALEY, C. & RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Am. Statist. Assoc.* **97**, 611-31.

GE, Y. & JIANG, W. (2006) On consistency of Bayesian inference with mixtures of logistic regression. *Neural Computation* **18**, 224-43.

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.

HARTIGAN, J. A. (1990). Partition models. *Commun. Statist.* **A 19**, 2745-56.

HOLMES, C. C., DENISON, D. G. T., RAY, S., & MALLICK, B. K. (2005). Bayesian Prediction via Partitioning. *J. Comp. Graph. Statist.* **14**, 811-30.

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

JIANG, W. X. & 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. & JACOBS, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* **6**, 181-214.

KURIHARA, K., WELLING, M. & VLASSIS, N. (2006). Accelerated Variational Dirichlet Mixture Models, *Advances in Neural Information Processing Systems* **19**.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351-7.

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.* **A 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.

MACEACHERN, S. N., CLYDE, M. & LIU, J. (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *Can. J. Statist.* **27**, 251-67.

MÜLLER, P., ERKANLI, A. & WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67-79.

NAIK, P. A., SHI, P. & TSAI, C. L. (2007). Extending the Akaike information criterion to mixture regression models. *J. Am. Statist. Assoc.* **102**, 244-54.

PAPASPILIOPOULOS, O. & ROBERTS, G.O. (2007). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, under revision.

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**.

QUINTANA, F. A (2006). A predictive view of Bayesian clustering. *J. Statist. Planning and Inference*, **136**, 2407-29.

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

QUINTANA, F. A. & NEWTON, M. A. (2000). Computational aspects of nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences. *J. Comp. Graph. Statist.* 9, 711-37.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. R. & VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* **B 64**, 585-616.

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

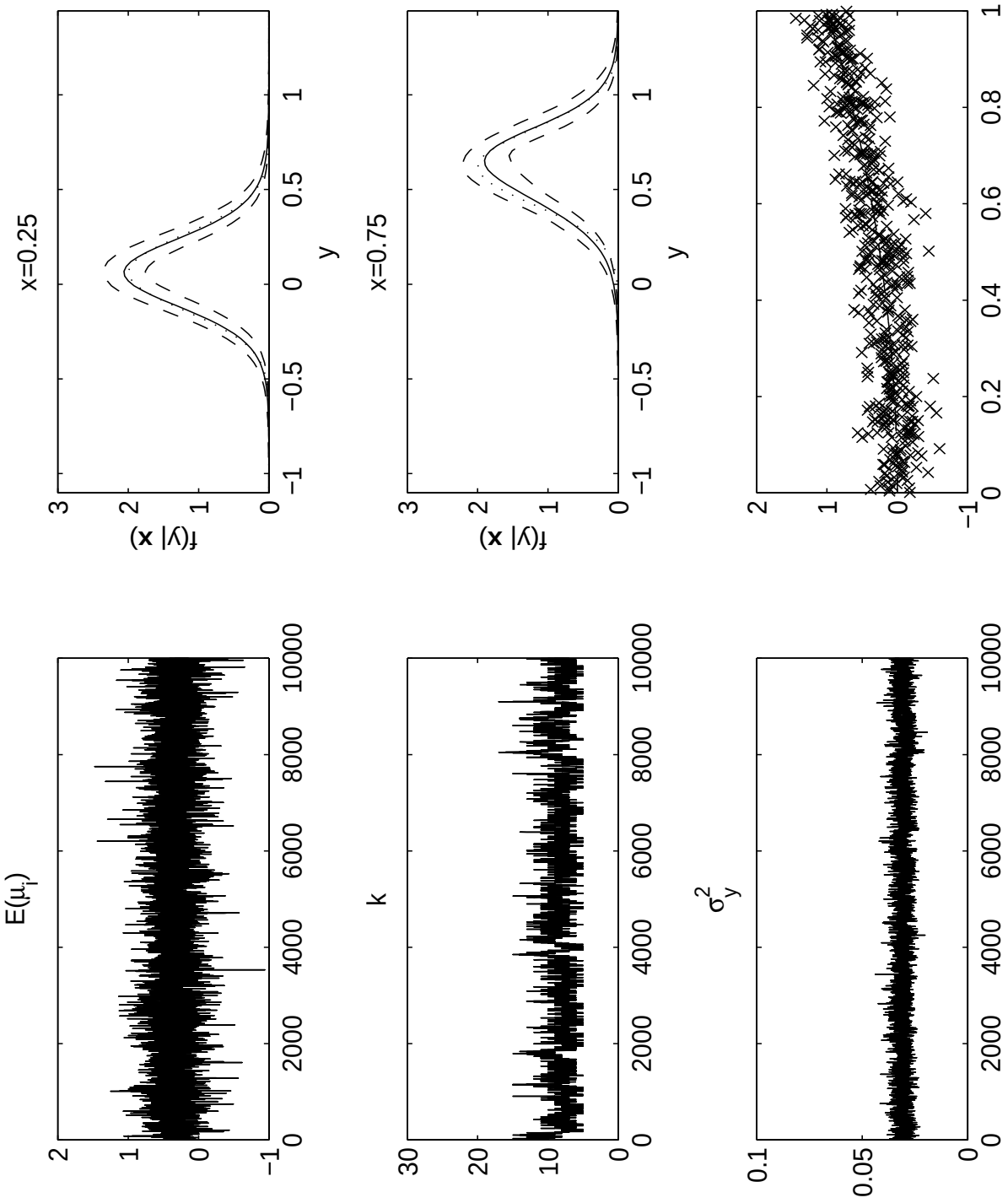

Fig. 1. Results for the first simulation example. The left column provides trace plots for representative parameters, while the right panel shows the conditional distributions for two different values of x, as well as the mean function estimation along with the raw data. Posterior means are solid lines,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

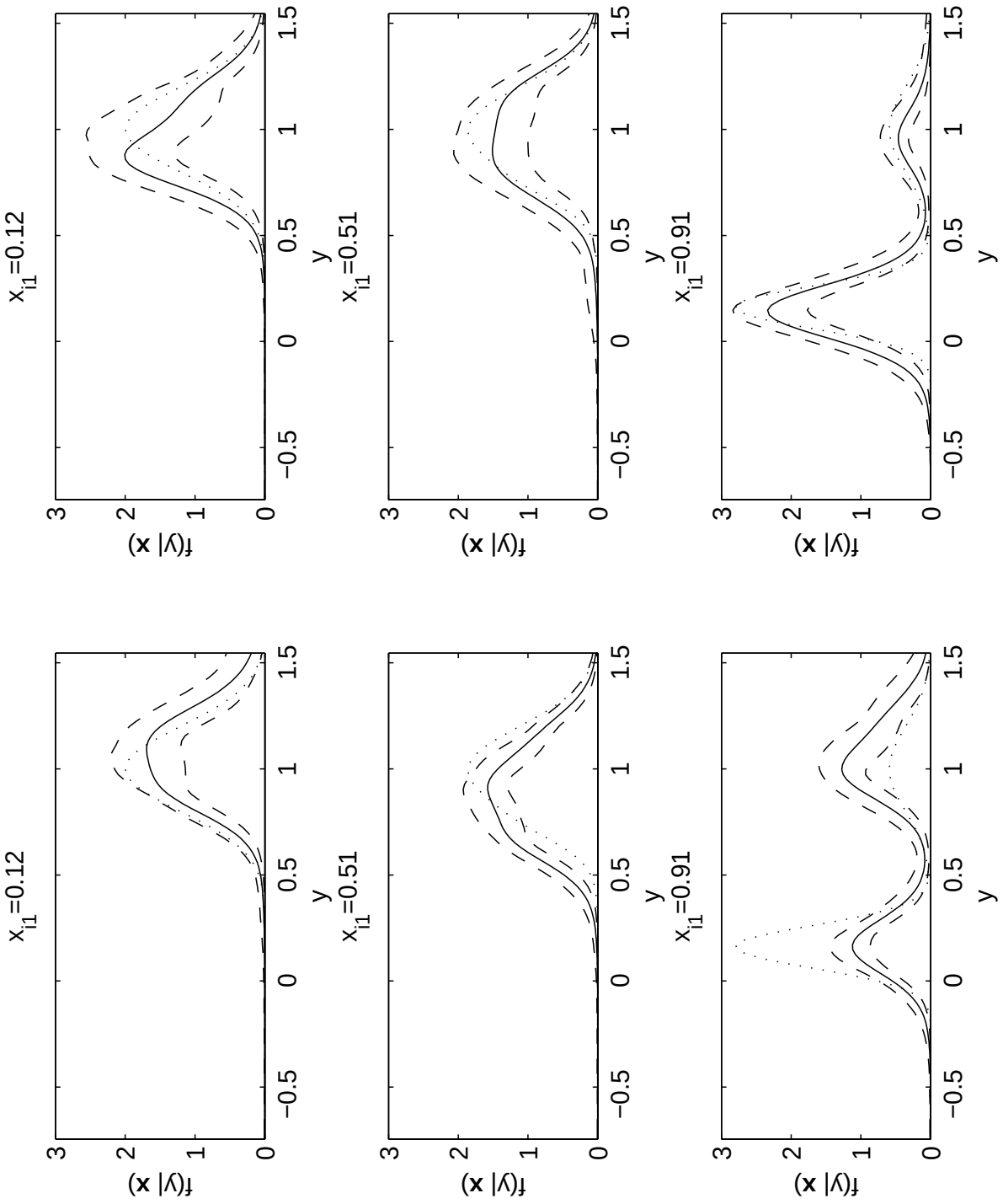

Fig. 2. Estimated predictive densities from the PPM (left panel) and the GPPM (right panel) at the 10th, 50th and 90th percentiles of the empirical distribution of x_{i1}: posterior means (solid lines), pointwise 99% credible intervals (dashed lines), and true values (dotted lines).

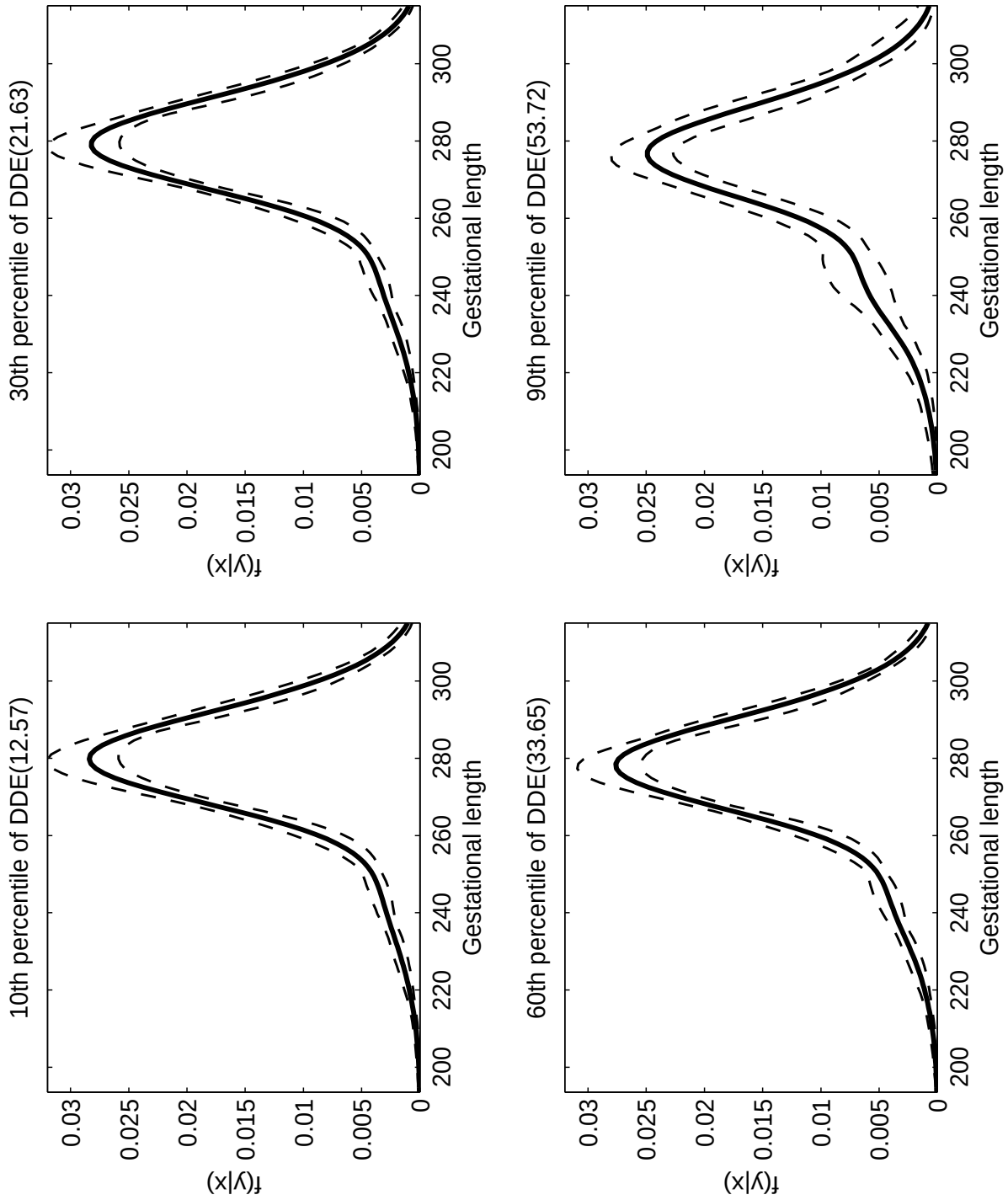

Fig. 3. Estimated predictive densities (solid lines) for gestational age at delivery at preselected values of DDE with 99% pointwise credible intervals (dashed lines).

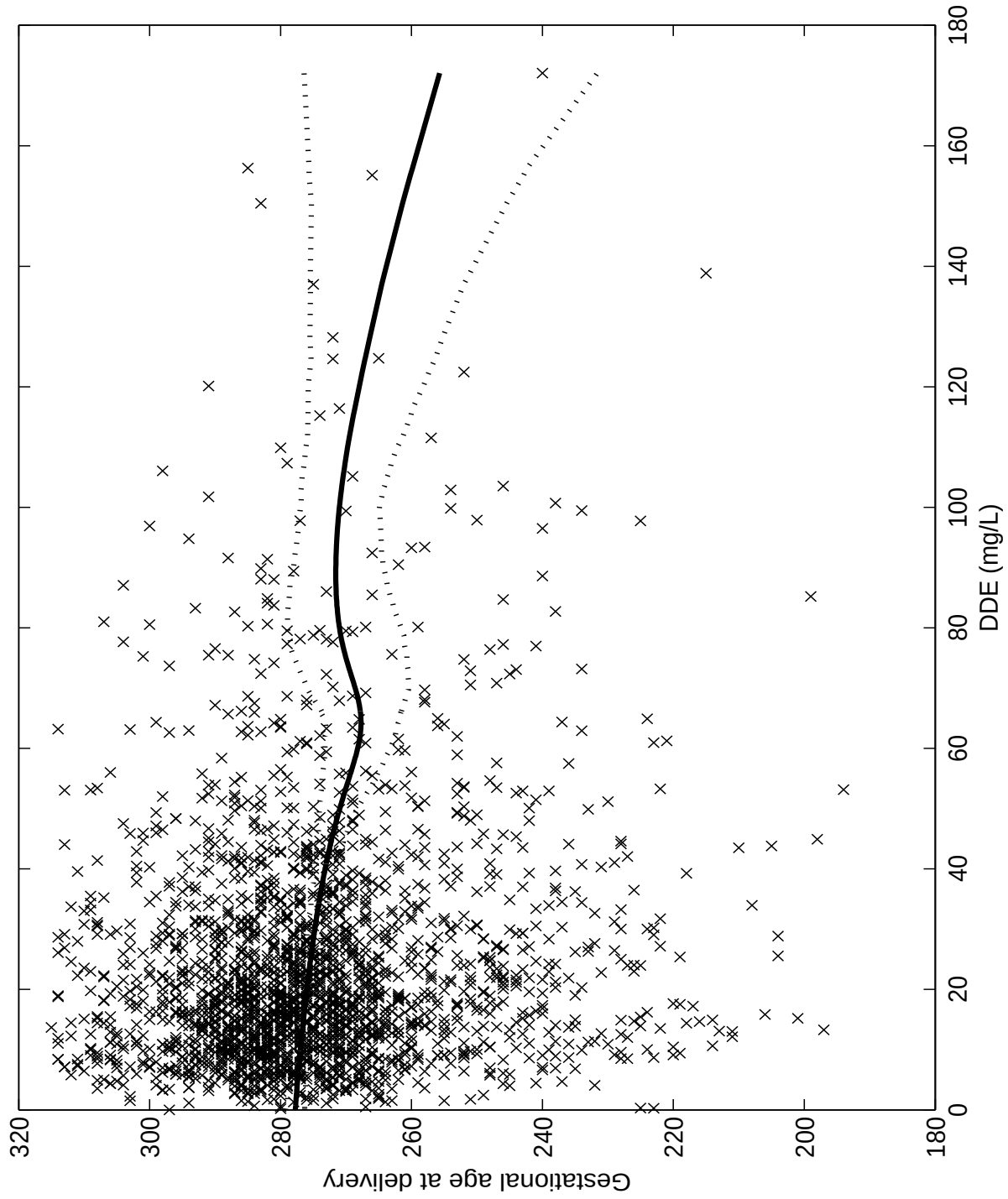

Fig. 4. The conditional predictive mean of gestational age at delivery (solid line) with 99% pointwise credible intervals (dotted lines).