

# Single Factor Transformation Priors for Density Regression

Suprateek Kundu<sup>1</sup> and David B. Dunson<sup>2</sup>

**Abstract:** Although discrete mixture modeling has formed the backbone of the literature on Bayesian density estimation incorporating covariates, the use of discrete mixtures leads to some well known disadvantages. Avoiding discrete mixtures, we propose a flexible class of priors based on random nonlinear functions of a uniform latent variable with an additive residual. These priors are related to Gaussian process latent variable models proposed in the machine learning literature. For density regression, we model the response and predictor means as distinct nonlinear functions of the same latent variable, thus inducing dependence through a single factor. The induced prior is shown to have desirable properties including large support and posterior consistency. We demonstrate advantages over Dirichlet process mixture models in a variety of simulations, and apply the approach to an epidemiology application.

*Keywords:* Nonparametric Bayes; Kernel estimation; Density regression; Gaussian process; Latent variable model; Dirichlet process; Posterior consistency; Latent factor regression.

---

<sup>1</sup>Suprateek Kundu is a doctoral candidate in the Dept. of Biostatistics, UNC Chapel Hill, Chapel Hill, NC 27599, USA (skundu@email.unc.edu).

<sup>2</sup>David B. Dunson is professor in Dept. Statistical Science, Duke University, Durham, NC 27708, USA (dunson@stat.duke.edu).

## 1. INTRODUCTION

Nonparametric kernel mixture models are increasingly popular in density estimation, density regression and high dimensional data modeling. Kernel mixture models for density estimation have the form:

$$F(y; G) = \int \mathcal{K}(y; \theta) G(d\theta), \quad (1)$$

where  $G(\cdot)$  is a mixing distribution and  $\mathcal{K}(\cdot)$  is a probability kernel. The majority of the nonparametric Bayesian development in this area relies on Dirichlet process (DP) priors (Ferguson, 1973; 1974) for  $G$ . These models have been generalized to density regression by defining dependence on the covariates  $x$  in various ways. Müller, Elkanli and West (1996) used a DP mixture of multivariate normals to jointly model the density of the response and predictors to induce a prior on  $f(y|x)$ . In order to let the parameters of the DP vary over the predictor space  $\mathcal{X}$ , MacEachern (1999) defined dependent Dirichlet processes (DDP) by assigning stochastic processes on the components in Sethuraman's (1994) DP representation:  $G_x = \sum_{i=1}^{\infty} p_i(x) \delta_{\theta_i(x)}$ . De Iorio et al. (2004) proposed a fixed- $p$  DDP, while Griffin and Steel (2006) allowed the weights to depend on predictors. Dunson, Pillai and Park (2007) instead used predictor-dependent convex combinations of DP components.

There is also a rich literature on using mixture priors in hierarchical latent variable models. Bush and MacEachern (1996) and Kleinman and Ibrahim (1998) proposed DP mixtures on the distributions of random effects. Fokoue and Titterington (2003) and Fokoue (2005) proposed mixtures of factor analyzers (MFA) corresponding to a finite mixture of multivariate normal kernels with a factor-analytic decomposition of the component-specific covariances. Dunson (2006) used dynamic mixtures of DPs to allow a latent variable distri-

bution to change nonparametrically across groups. More recently, Chen et al. (2009) and Carvalho et al. (2008) proposed nonparametric Bayes MFA allowing an uncertain number of factors. Lee, Lu, and Song (2008) placed a truncated DP on the distribution of the latent variables within a structural equation model (SEM), while Yang and Dunson (2010) proposed a centering approach to ensure identifiability of the latent factor distributions.

The above approaches have relied on discrete mixture models, which have a number of well known complications motivating alternative methods for modeling unknown densities, such as Polya trees (Mauldin et al, 1992; Lavine, 1992, 1994) and logistic Gaussian processes (LGP) (Lenk 1988, 1991; Tokdar 2007). Polya trees have appealing properties in terms of denseness, conjugacy and posterior consistency but have disadvantages in terms of favoring overly spiky densities. LGP has sound theoretical properties and smoothness of the densities can be controlled through the covariance kernel in the GP. However, posterior computation is a major hurdle. Recently, Jara and Hanson (2010) proposed dependent tail-free processes where they modeled the tail-free probabilities with LGP dependent on covariates. Their approach is shown to approximate the Polya tree marginally at each predictor value. An alternative was suggested by Tokdar, Zhu and Ghosh (2010) relying on LGP for density regression with dimensionality reduction.

In this article, we focus on a new approach for nonparametric density estimation and regression that induces a prior on the unknown density through placing a flexible prior on a nonlinear regression function  $\theta$  in a latent factor model. By using GP priors for  $\theta$ , we obtain substantial control over the smoothness of the induced densities in a very different manner than that achieved by LGP-based models. By using the same latent factor within models for the response and predictors, we induce a flexible single factor density regression model for

addressing the curse of dimensionality implicit in density regression approaches that attempt to allow the conditional density  $f(y|x)$  to vary completely flexibly with predictors. Unlike LGP-based models, the proposed model has appealing conjugacy properties facilitating posterior computation. In addition, the method has appealing theoretical properties in terms of large support and posterior consistency. The proposed class of models is related to Gaussian process latent variable models (GP-LVM) proposed in the machine learning literature (Lawrence, 2005; Silva and Gramacy, 2010), but our modeling details are different and the focus of this literature has been on nonlinear dimensionality reduction with no consideration of density regression or theoretical properties.

Section 2 deals with the density estimation model and explores some theoretical properties of the proposed prior. Section 3 extends our model to a class of single factor density regression models and latent factor regression models. Posterior computation is developed in section 4 with simulation studies in section 5. In section 6, we illustrate our method in a real data example. We conclude this paper with a brief discussion in section 7. Proofs are in the appendix.

## 2. DENSITY ESTIMATION

### 2.1. Model Specification

In this section, we define our model and study properties in univariate density estimation. For now, we seek to model the density of i.i.d. continuous variables, where the density is unknown and belongs to  $\mathcal{F}$ , the set of densities on  $\mathfrak{R}$  with respect to Lebesgue measure. We propose to induce a prior on the unknown density through the following nonlinear latent

variable model,

$$\begin{aligned} y_i &= \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim \Gamma_\sigma, \\ \mu &\sim \Pi^*, \quad \sigma \sim \nu, \quad x_i \sim \text{Uniform}(0, 1), \end{aligned} \tag{2}$$

where  $\mu \in \Theta$  is an unknown  $[0, 1] \rightarrow \Re$  function,  $x_i$  is a uniformly distributed latent variable, and the error distribution  $\Gamma_\sigma$  is centered at 0 and has scale parameter  $\sigma$ . Hence, in the special case in which  $\mu(x) = \mu$ , so that the regression function is a constant, and  $\Gamma_\sigma$  is normal, we have  $f(y; \mu, \sigma^2) = N(y; \mu, \sigma^2)$  so we obtain a normal density. The density of  $y$  conditionally on the unknown regression function  $\mu$  and  $\sigma$  is obtained on marginalizing out the latent variable as

$$f(y; \mu, \sigma) = f_{\mu, \sigma}(y) = \int_0^1 \Gamma_\sigma(y - \mu(x)) dx. \tag{3}$$

To complete the specification, we let  $\mu \sim \Pi^*$ ,  $\sigma \sim \nu$  and obtain the marginal density

$$f(y) = \int_0^\infty \int_\Theta \int_0^1 \Gamma_\sigma(y - \mu(x)) dx \Pi^*(d\mu) \nu(d\sigma). \tag{4}$$

Hence, a prior  $f \sim \Pi$  is induced through assigning independent priors to  $\mu$  and  $\sigma$  in expression (3). The prior  $\Pi$  is over the space of densities  $\mathcal{F}$ , so that realizations from this prior correspond to different continuous densities. When the prior on  $\mu$  is a Gaussian process and the error distribution is  $N(0, \sigma^2)$  (denoted as  $\Gamma_\sigma = \phi_\sigma$ ), we refer to  $f_{\mu, \sigma}$  as the Gaussian process transformation (GPT) and the induced prior  $f \sim \Pi$  as a Gaussian process transformation prior. Although  $\mu$  is not formally a transformation function without making a monotonicity restriction and letting  $\sigma \rightarrow 0$ , this terminology is nonetheless expressive of the form of the proposed model.

The GPT prior does not have the kernel mixture form (1). There will be no clustering of subjects or label switching issues. Instead, the prior  $f \sim \Pi$  is induced through adding a Gaussian residual to a Gaussian process regression model in a uniform latent variable. This is a simple structure aiding computation and interpretability. One can control the smoothness of the density through the covariance in the GP prior for the regression function  $\mu$  and the size of the scale parameter  $\sigma$ . In limiting cases, one can obtain realizations of  $\mu$  concentrated close to a flat line, leading to a normal density as a special case. In addition, by making  $\sigma$  small and choosing the GP covariance to generate a very bumpy  $\mu$ , one can obtain arbitrarily bumpy densities. In practice, by choosing hyperpriors for key covariance parameters, we obtain a data adaptive approach that often outperforms discrete kernel mixtures. The performance of discrete kernel mixtures relies on the ability to accurately approximate the density with few components, and DP mixtures tend to heavily favor a small number of dominate kernels. This tendency can sometimes lead to relatively poor estimation, as illustrated in section 5.

## 2.2. Theoretical Properties

To further justify the proposed prior, we show large support and posterior consistency properties. Large support is an important property in that it ensures that our prior can generate densities that are arbitrarily close to any true density  $f_0$  in a large class, a defining property for a nonparametric Bayesian procedure and a necessary condition to allow the posterior to concentrate in small neighborhoods of the truth. Instead of focusing narrowly on GPT priors, we provide broad theoretical results for priors in the general class of expression (2).

Before proceeding, it is necessary to define some notation and concepts. We denote the Kullback-Leibler (KL) divergence of  $f_{\mu,\sigma}$  from  $f_0$  as  $KL(f_{\mu,\sigma}, f_0)$  and an  $\epsilon$ -sized KL

neighborhood around  $f_0$  as  $KL_\epsilon(f_0)$ . The sup-norm distance is denoted by  $\|\cdot\|_\infty$ . Our development will rely on the fact that any density  $f_0$  can be calculated from a “true function”  $\mu_0 = F_0^{-1}$ , with  $F_0$  denoting the cumulative distribution function. To generate  $y_i \sim f_0$ , one can equivalently draw  $x_i \sim \text{Uniform}(0, 1)$  and let  $y_i = \mu_0(x_i)$ . This corresponds to the limiting case as  $\sigma \rightarrow 0$  in model (2) with  $\mu = \mu_0$ . Hence, any true density can be represented as the limiting case

$$f_0(y) = \lim_{\sigma \rightarrow 0} \int_0^1 \Gamma_\sigma(y - \mu_0(x)) dx, \quad (5)$$

assuming  $\Gamma_\sigma$  is chosen so that such a limit exists. As reasonable regularity conditions, we assume that  $f_0$  is strictly positive and finite,  $\sup_x |\mu_0(x)| < \infty$  and  $0 < \Gamma_\sigma(u) < \infty$  for finite  $u$ . A uniformly bounded  $\mu_0$  along with the condition on  $\Gamma_\sigma$  implies that for  $\mu$  belonging to a ball of finite radius around  $\mu_0$ ,  $f_{\mu, \sigma}$  is strictly positive and finite for all  $\sigma \in \mathfrak{R}^+$ , which ensures a finite KL divergence for a suitable subset of  $\mu$  values.

**Theorem 1.** *Let  $\sup_x |\mu_0(x)| < \infty$ ,  $\Gamma_\sigma$  be normal, Laplace or Cauchy with scale parameter  $\sigma$  and  $f_0$  be the corresponding density in  $\mathcal{F}$  defined as in equation (5). If  $\mu_0$  is in the sup-norm support of  $\Pi^*$  and  $\nu\{\sigma : \sigma \in (0, \eta)\} > 0$  for all  $\eta > 0$ , then  $\Pi(KL_\epsilon(f_0)) > 0$  for all  $\epsilon > 0$ .*

Theorem 1 allows us to verify that the induced prior on the density  $f$  assigns positive probability to KL neighborhoods of any strictly positive and finite true density  $f_0$ . From Schwartz (1965), if the true density  $f_0$  is in the KL support of the prior for  $f$ , the posterior distribution for  $f$  will concentrate asymptotically in arbitrarily small weak neighborhoods of  $f_0$ . Theorem 1 requires the prior  $\mu \sim \Pi^*$  to place positive probability in sup-norm neighborhoods of the inverse cdf  $F_0^{-1}$ . Although one can verify this condition for certain choices of  $\Pi^*$ , such as appropriately chosen Gaussian process priors, it is nonetheless somewhat strin-

gent. We show in Theorem 2 that this condition can be relaxed to only require that the prior  $\mu \sim \Pi^*$  assigns positive probability to L-1 neighborhoods of any element  $\mu_0$  of  $\Theta$ . It is well known that positive sup-norm support automatically guarantees positive L-1 support but the converse is not true.

**Theorem 2.** *Let  $\sup_x |\mu_0(x)| < \infty$ ,  $\Gamma_\sigma = \phi_\sigma$  and  $f_0$  be the corresponding density in  $\mathcal{F}$  defined in equation (5). If  $\mu_0$  is in the L-1 support of  $\Pi^*$  and  $\nu\{\sigma : \sigma \in (0, \eta)\} > 0$  for all  $\eta > 0$ , then  $\Pi(KL_\epsilon(f_0)) > 0$  for all  $\epsilon > 0$ .*

As the prior  $f \sim \Pi$  is specified indirectly through priors  $\mu \sim \Pi^*$  and  $\sigma \sim \nu$ , it is desirable for elicitation purposes to verify that, for sufficiently small  $\sigma$ ,  $\mu \approx \mu_0$  implies that  $f_{\mu, \sigma} \approx f_0$ . Theorem 3 provides such a verification assuming Gaussian errors. This implies one can potentially center the prior for the density  $f$  on an initial parametric guess  $\tilde{f}$  by centering  $\mu \sim \Pi^*$  on the inverse cdf  $\tilde{F}^{-1}$  while choosing the prior for  $\sigma$  to have mode near zero. The data will then inform about the degree to which  $\mu$  deviates from  $\tilde{F}^{-1}$  and  $\sigma$  deviates from 0.

**Theorem 3.** *For  $\mu_0 \in \Theta$  and  $\Gamma_\sigma = \phi_\sigma$ , let  $f_0$  be the density resulting from equation (5). Then for  $\mu \in N_{\epsilon_1}(\mu_0)$ , with  $N_{\epsilon_1}(\mu_0)$  an  $\epsilon_1$ -sized L-1 neighborhood around  $\mu_0$ , and  $\sigma \in (\epsilon_2, \epsilon_2^*)$ , we have  $f_{\mu, \sigma} \in N_{\frac{\epsilon_1}{\epsilon_2}}(f_0)$  for arbitrarily small  $\epsilon_1, \epsilon_2, \epsilon_2^*$  such that  $0 < \epsilon_1 < \epsilon_2 < \epsilon_2^*$ .*

Although Theorems 1-2 lead to weak posterior consistency, small weak neighborhoods around  $f_0$  are topologically too large and may include densities that are quite different from  $f_0$  in shape and other characteristics. Hence, it is appealing to establish a strong posterior consistency result in which the posterior probability allocated to arbitrarily small L-1 neighborhoods of  $f_0$  increases towards one exponentially fast with increasing sample size. Focusing on the GPT prior described above, we show in Theorem 4 that strong posterior

consistency holds under some conditions on the prior. Notably, for an appropriately chosen GPT prior, we obtain L-1 posterior consistency for all strictly positive and finite true densities  $f_0 \in \mathcal{F}$  without assuming further regularity conditions on  $f_0$ . In contrast, for DP and LGP priors, one typically requires additional regularity conditions. For example, Ghosal et al. (1999) showed strong consistency of DP mixture priors under some tail assumptions, whereas Tokdar et al. (2007) considered densities on fixed bounded intervals when showing strong consistency for logistic Gaussian process priors.

**Theorem 4.** *Suppose  $\mu \sim GP(m, c)$  and define  $f_0$  as in equation (5) where  $\Gamma_\sigma = \phi_\sigma$ . Let  $U = \{f_{\mu, \sigma} : \int |f_{\mu, \sigma} - f_0| dy < \epsilon, \mu \in \Theta, \sigma \in \mathbb{R}^+\}$ . Suppose  $\sigma$  has compact support, the mean function  $m(\cdot)$  is continuously differentiable with  $\sup_{x \in (0,1)} m(x) < \infty$ , and that the covariance function  $c(\cdot, \cdot)$  has continuous fourth derivatives. Then,  $f_0$  is in the KL support of  $\Pi$  implies that the posterior is strongly consistent at  $f_0$ .*

### 3. SINGLE FACTOR DENSITY REGRESSION

Having seen how the GPT yields a flexible approach for nonparametric density estimation with sound theoretical properties, let us now make explicit its application in density regression. We consider the scenario in which both the response and the predictors are continuous and induce a regression model for the conditional response density given predictors

$\mathbf{z}_i=(z_{i1}, \dots, z_{ip})$ :

$$\begin{aligned}
y_i &= \mu^Y(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_Y^2), \\
z_{ik} &= \mu^{Z_k}(x_i) + \epsilon_{ik}^*, \quad \epsilon_{ik}^* \sim N(0, \sigma_{Z_k}^2), k = 1, 2, \dots, p, \\
x_i &\sim \text{Uniform}(0, 1), \\
\mu^Y &\sim \Pi^Y, \quad \mu^{Z_k} \sim \Pi^{Z_k}, k = 1, 2, \dots, p, \\
\sigma^Y &\sim \nu, \quad \sigma_{Z_k} \sim \nu,
\end{aligned} \tag{6}$$

where  $\mu^Y, \mu^{Z_k} \in \Theta$  are unknown  $[0, 1] \rightarrow \mathfrak{R}$  functions,  $\epsilon$ 's are independent errors and  $x_i$  is the latent variable. For simplicity, we assume the same prior  $\nu$  on the precision of the measurement errors in each component model, though this assumption is trivial to relax. Expression (6) is a multivariate generalization of the univariate density estimation model (2). Marginally each of the variables is assigned exactly the prior in (2) and to allow dependence we incorporate the same latent factor  $x_i$  in each of the models.

Our goal in defining a joint model is to induce a flexible but parsimonious model for the conditional density of  $y_i$  given the predictors  $\mathbf{z}_i$ . In estimating conditional densities for multiple predictors, one encounters a daunting dimensionality problem in that one is attempting to estimate a density nonparametrically while allowing arbitrary changes in this density across a multivariate predictor space. Clearly, as  $p$  increases even for large samples there will be many regions of the predictor space that have sparse observations. As a compromise between flexibility and parsimony in addressing the curse of dimensionality, we propose to use a single factor model in which the marginals for each variable are fully flexible but restrictions come in through assuming dependence on a single  $x_i$ . Extensions to the multiple factor case are straightforward.

To obtain insight into basic properties of model (6), we focus initially on the case in which  $p = 1$  for simplicity, though the results can easily be generalized to multiple predictors. On marginalizing out the latent variable  $x_i$ , we obtain the joint and conditional densities as

$$\begin{aligned}\psi(y_i, z_i | \mu^Y, \mu^Z, \sigma_Y, \sigma_Z) &= \psi_\beta(y_i, z_i) = \int_0^1 \phi_{\sigma_Y}(y_i - \mu^Y(x)) \phi_{\sigma_Z}(z_i - \mu^Z(x)) dx, \\ \zeta(y_i | z_i, \mu^Y, \mu^Z, \sigma_Y, \sigma_Z) &= \zeta_{z_i, \beta}(y_i) = \frac{\int_0^1 \phi_{\sigma_Y}(y_i - \mu^Y(x)) \phi_{\sigma_Z}(z_i - \mu^Z(x)) dx}{\int_0^1 \phi_{\sigma_Z}(z_i - \mu^Z(x)) dx},\end{aligned}\quad (7)$$

where  $\beta = (\mu^Y, \mu^Z, \sigma_Y, \sigma_Z)$ . The priors on  $(\mu^Y, \mu^Z, \sigma_Y, \sigma_Z)$  induce a prior on the set of densities  $\{\psi_\beta(y, z), \beta \in \Theta \times \Theta \times \mathfrak{R}^+ \times \mathfrak{R}^+\}$  and  $\{\zeta_\beta(y|z), z \in \mathfrak{R}, \beta \in \Theta \times \Theta \times \mathfrak{R}^+ \times \mathfrak{R}^+\}$ . Some algebra and application of Fubini's theorem show us that

$$\begin{aligned}E(Y|Z = z) &= \frac{\int_0^1 \mu^Y(x) \phi_{\sigma_Z}(z - \mu^Z(x)) dx}{\int_0^1 \phi_{\sigma_Z}(z - \mu^Z(x)) dx}, \\ V(Y|Z = z) &= \sigma_Y^2 + \frac{\int_0^1 (\mu^Y(x))^2 \phi_{\sigma_Z}(z - \mu^Z(x)) dx}{\int_0^1 \phi_{\sigma_Z}(z - \mu^Z(x)) dx} - \left( \frac{\int_0^1 \mu^Y(x) \phi_{\sigma_Z}(z - \mu^Z(x)) dx}{\int_0^1 \phi_{\sigma_Z}(z - \mu^Z(x)) dx} \right)^2.\end{aligned}\quad (8)$$

Expression (9) shows that both the conditional mean and variance are allowed to depend on predictors, and in general the same is true for all conditional moments. Denote the joint and conditional densities arising out of arbitrary parameter values  $\beta_0 = (\mu_0^Y, \mu_0^Z, \sigma_{0Y}, \sigma_{0Z})$  as  $\psi_{\beta_0}(y, z)$  and  $\zeta_{\beta_0}(y|z)$  respectively. We make similar regularity assumptions as in section 2, that  $\psi_{\beta_0}$  is strictly positive and finite and  $\mu_0^Y, \mu_0^Z$  are uniformly bounded. The regularity condition on  $\Gamma_\sigma(\cdot)$  is satisfied by the normal kernel. This ensures finite KL divergence for suitable subsets of  $\mu^Y$  and  $\mu^Z$  values.

**Theorem 5.** *Let  $\mu_0^Y, \mu_0^Z$  satisfy the regularity conditions just stated and  $\psi_{\beta_0}$  be the corresponding joint density. Suppose  $\mu_0^Y, \mu_0^Z$  belong to the sup-norm support of  $\Pi^Y$  and  $\Pi^Z$  respectively and  $\sigma_{0Y}, \sigma_{0Z}$  belong to the support of  $\nu$ . Then  $P(KL_\epsilon(\psi_{\beta_0})) > 0$  for all  $\epsilon > 0$ .*

**REMARK 2.** Using similar techniques as in proof of theorem 6, it can be shown that  $\text{KL}(\zeta_{z,\beta_0}, \zeta_{z,\beta}) < \epsilon$  for all  $\epsilon > 0$ , pointwise for each fixed  $z$ .

**REMARK 3.** Theorem 6 also can be extended to the case when the error distribution in model (6) is Laplace or Cauchy using similar techniques as in proof of theorem 1.

The proof proceeds along the same lines as theorem 1 and a sketch is given in the appendix. Also applying Schwartz theorem, we get weak consistency at  $\psi_{\beta_0}$ . Although the above results are general for any priors satisfying the conditions, we will focus on Gaussian process priors in the sequel.

#### 4. POSTERIOR COMPUTATION

For simplicity, we focus on the single predictor density regression case when outlining an MCMC algorithm for posterior computation. Let  $Y_{n \times 1}$  and  $Z_{N \times 1}$  denote the vector of observations and covariates, respectively. We are interested in prediction of  $y_{n+1}, \dots, y_N$  based on  $z_{n+1}, \dots, z_N$ . Let  $\mu_Y^n$  ( $n \times 1$ ) and  $\mu_Z^N$  ( $N \times 1$ ) denote the realizations of the GP  $\mu^Y$  and  $\mu^Z$  at the latent variable values  $\mathbf{x} = (x_1, \dots, x_n, x_{n+1}, x_N)'$ . From the GP prior, we have  $\mu_Y^n \sim N_n(m_Y^n, \mathbf{K}_Y^n)$  and  $\mu_Z^N \sim N_N(m_Z^N, \mathbf{K}_Z^N)$ . The covariance kernels are squared exponential with  $\mathbf{K}_Y(x, x') = \frac{1}{\phi_Y} \exp \left\{ -C_Y(x - x')^2 \right\}$  and  $\mathbf{K}_Z(x, x') = \frac{1}{\phi_Z} \exp \left\{ -C_Z(x - x')^2 \right\}$ . We specify conjugate gamma priors:  $\sigma_Y^{-2} \sim Ga(a_\sigma, b_\sigma)$ ,  $\sigma_Z^{-2} \sim Ga(aa_\sigma, bb_\sigma)$ ,  $\phi_Y \sim Ga(a_\phi, b_\phi)$  and  $\phi_Z \sim Ga(aa_\phi, bb_\phi)$ . For updating the latent variables  $\mathbf{x}$ , we adopt the griddy Gibbs approach using a set of evenly distributed grid points  $g_1^*, g_2^*, \dots, g_G^* \in (0, 1)$ . Let  $\mathbf{D}_Y$  and  $\mathbf{D}_Z$  be diagonal matrices having  $\sigma_Y^2$  and  $\sigma_Z^2$  as their diagonal elements respectively. Let  $\mu_Y^n(-i)$  include all elements of  $\mu_Y^n$  except  $\mu^Y(x_i)$ , and similarly for  $\mu_Z^N(-i)$ . The Gibbs sampling algorithm alternates between the following steps.

*Step1:* Update  $\sigma_Y^2$  and  $\sigma_Z^2$  using  $\pi(\sigma_Y^{-2}|-) \sim \text{Ga}(a_\sigma + n/2, b_\sigma + \frac{1}{2} \sum_{i=1}^n (y_i - \mu^Y(x_i))^2)$  and  $\pi(\sigma_Z^{-2}|-) \sim \text{Ga}(a_\sigma + N/2, b_\sigma + \frac{1}{2} \sum_{i=1}^N (z_i - \mu^Z(x_i))^2)$  respectively.

*Step2:* To sample the latent variables, choose  $x_i = g_k^*$  with probability  $p_{ik}$ , where

$$\begin{aligned} p_{ik} = P[x_i = g_k^*|-] &= \frac{p_{ik}^Y p_{ik}^Z N(\mu^Y(x_i = g_k^*)|\mu_Y^n(-i)) N(\mu^Z(x_i = g_k^*)|\mu_Z^N(-i))}{\sum_{l=1}^G p_{il}^Y p_{il}^Z N(\mu^Y(x_i = g_l^*)|\mu_Y^n(-i)) N(\mu^Z(x_i = g_l^*)|\mu_Z^N(-i))}, \text{ if } i \leq n \\ &= \frac{p_{ik}^Z N(\mu^Z(x_i = g_k^*)|\mu_Z^N(-i))}{\sum_{l=1}^G p_{il}^Z N(\mu^Z(x_i = g_l^*)|\mu_Z^N(-i))}, \text{ if } n < i \leq N, \end{aligned}$$

where  $p_{ik}^Y = N(y_i; \mu^Y(x_i = g_k^*), \sigma_Y^2)$ ,  $p_{ik}^Z = N(z_i; \mu^Z(x_i = g_k^*), \sigma_Z^2)$  and  $k=1, 2, \dots, G$ .

*Step3:* Update  $\mu_Y^n$  and  $\mu_Z^N$  using  $\pi(\mu_Y^n|-) = N_n\left((\mathbf{D}_Y^{-1} + (\mathbf{K}_Y^n)^{-1})^{-1}(\mathbf{D}_Y^{-1}Y + (\mathbf{K}_Y^n)^{-1}m_Y^n), (\mathbf{D}_Y^{-1} + (\mathbf{K}_Y^n)^{-1})^{-1}\right)$  and  $\pi(\mu_Z^N|-) = N_N\left((\mathbf{D}_Z^{-1} + (\mathbf{K}_Z^N)^{-1})^{-1}(\mathbf{D}_Z^{-1}Z + (\mathbf{K}_Z^N)^{-1}m_Z^N), (\mathbf{D}_Z^{-1} + (\mathbf{K}_Z^N)^{-1})^{-1}\right)$  respectively.

*Step4:* Update  $\mu_Y^{*G} = \{\mu^Y(g_1^*), \dots, \mu^Y(g_G^*)\}$  and  $\mu_Z^{*G} = \{\mu^Z(g_1^*), \dots, \mu^Z(g_G^*)\}$  using the conditional normal distributions  $N(\mu_Y^{*G}|\mu_Y^n)$  and  $N(\mu_Z^{*G}|\mu_Z^N)$  respectively.

*Step5:* Update  $\phi_Y$  and  $\phi_Z$  using  $\pi(\phi_Y|-) \sim \text{Ga}(a_\phi + \frac{n}{2}, b_\phi + \frac{1}{2}(\mu_Y^n - m_Y^n)'(\mathbf{K}_Y^n)^{-1}(\mu_Y^n - m_Y^n))$  and  $\pi(\phi_Z|-) \sim \text{Ga}(a_\phi + \frac{N}{2}, b_\phi + \frac{1}{2}(\mu_Z^N - m_Z^N)'(\mathbf{K}_Z^N)^{-1}(\mu_Z^N - m_Z^N))$  respectively.

*Step6:* Update  $C_Y$  and  $C_Z$  using Metropolis random walk for  $\log(C_Y)$  and  $\log(C_Z)$ .

For prediction of  $y_k$  based on  $z_k$ ,  $k = n+1, \dots, N$ , we use  $\pi(y_k|-) = N(y_k; \mu^Y(x_k), \sigma_Y^2)$ , while

the conditional density estimate is calculated as  $\hat{f}(y|z) = \frac{\frac{1}{G} \sum_{k=1}^G \phi_{\sigma_Y}(y - \mu^Y(g_k^*)) \phi_{\sigma_Z}(z - \mu^Z(g_k^*))}{\frac{1}{G} \sum_{k=1}^G \phi_{\sigma_Z}(z - \mu^Z(g_k^*))}$ .

## 5. SIMULATION STUDY

To assess the performance of the GPT approach in density estimation as well as density regression, we conducted several simulation studies. We chose the mean function for the GP as  $m(x) = 2\sin(x) + \cos(x)$  and utilized the squared exponential covariance kernel. For computational purposes, we worked with the standardized data and then transformed it back in the final step. The hyperparameters for the gamma priors were chosen to be one

throughout. Although we used 75 grid points for the griddy Gibbs approach, the number of points could be as low as 60. The number of iterations used was 10000 with a burn in of 1000. The convergence for the main quantities such as  $\mu$  was rapid with good mixing. All results are reported over 5 replicates.

### 5.1. Univariate Density Estimation

To see how well the GPT does in practice for density estimation, we looked at a variety of scenarios, where the truth was generated from the densities considered in Marron and Wand (1992), which are essentially finite mixtures of Gaussians. We present the results from four of those cases which we thought to be interesting deviations from normality and could be potentially encountered in applications. These are the 2nd, 6th, 8th and 9th Marron-Wand densities. The sample size used was 100. For comparison, we looked at DP mixture of Gaussians (Escobar and West, 1995), mixtures of Polya trees (Hanson, 2006) and frequentist kernel estimates using a Gaussian kernel (and the bandwidth selection method of Sheather and Jones, 1991). More specifically, for both DP mixtures and mixtures of Polya trees, we used the DP package in R and the standard hyperparameter values therein. We used algorithm 8 of Neal (2000) with  $m=1$  for DP mixtures of Gaussians. For frequentist kernel, we used the function “density” in R with Gaussian kernel. Overall, we found that varying the hyperparameter values within a reasonable range does not significantly alter the density estimation results for a sample size of 100, for any of the competitors. Table 1 presents the L-1 distance between true and estimated densities while figure 1 depicts the density plots.

**Table 1: Marron-Wand Curves: L-1 distance between true and estimated densities**

Method	L-1 Distance			
	MW 2	MW 6	MW 8	MW 9
GPT	0.031	0.035	0.031	0.028
DPM	0.035	0.036	0.03	0.038
Polya tree mixture	0.065	0.036	0.045	0.042
Frequentist Kernel	0.145	0.031	0.033	0.028

From table 1, we see that even when the truth is generated from a finite mixture of Gaussians, the GPT tends to do better or at least as well as the DP mixture of Gaussians. Mixtures of Polya trees have somewhat worse performance and result in overly spiky looking estimates.

## 5.2. Density Regression

For density regression, we generated a univariate response by allowing the conditional mean as well as the residual error distribution to vary with the covariate. We compared the out of sample predictive performance of GPT with other competitors such as DP mixture of bivariate normals (Müller, Erkanli and West, 1996), Bayesian additive regression trees (BART) (Chipman, George and McCulloch, 2010), GP mean regression (O’Hagan and Kingman, 1978) and treed GP (Gramacy and Lee, 2008), based on standard packages in R. We used the DP package for DP mixtures of Gaussians and the Bayestree package for the other three methods, and the hyperparameter values therein. The density regression results did not change significantly on varying the hyperparameter values within a reasonable range, for all the competitors. We used the following scheme for simulations:

$$Z \sim F_Z, \quad y_i = \lambda \exp\left(-\frac{e^{z_i}}{1 + e^{z_i}}\right) + \frac{e^{z_i}}{1 + e^{z_i}} \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

where  $F_Z$  is the distribution of the predictors which was chosen to be a trimodal density (9th Marron-Wand curve). We chose  $\lambda = 3$  and split the total sample size of 100 into training set of 50 and test set of 50. The above data generating model allows the shape of the conditional density to change with predictors, hence making prediction non-trivial. Table 2 shows the performance of the GPT along with a few competitors. We computed the mean square error (MSE), 95% coverage for the mean (COV), as well as the L-1 distance between true and estimated densities at 25th, 50th and 75th percentiles of the predictor distribution.

**Table 2: Mean square error and L-1 distance between true and estimated densities**

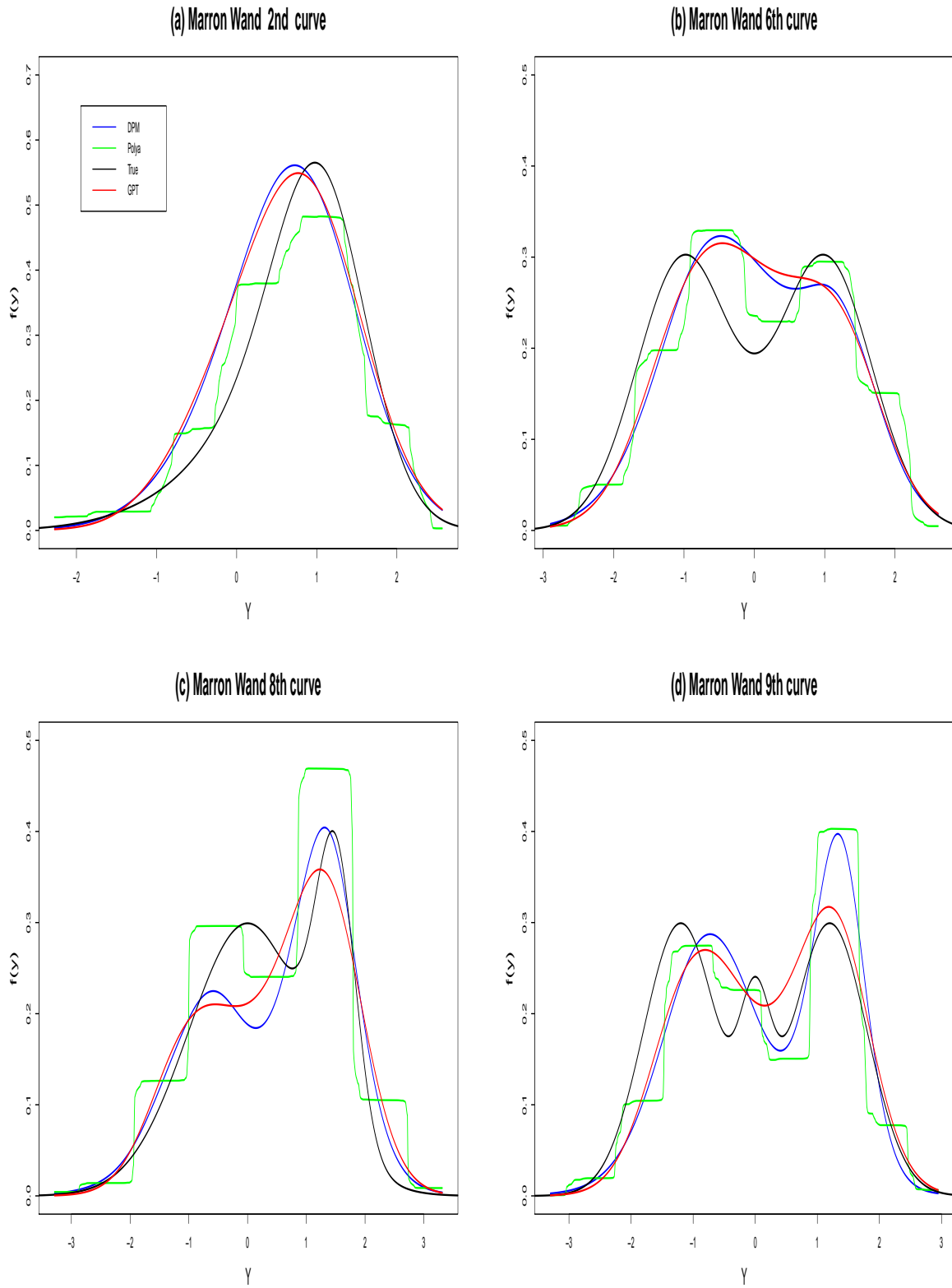


Figure 1: Marron-Wand curves - density estimates for GPT, DPM and Polya tree mixtures.

Method	MSE	COV(%)	L-1 Distance		
			25th	50th	75th
GPT	1.26	94	0.08	0.04	0.06
DPM	1.53	42	0.03	0.04	0.06
BART	1.59	46	0.13	0.026	0.10
GP reg	1.52	76	0.14	0.03	0.10
treed GP	1.6	72	0.07	0.09	0.04

The results in table 2 are consistent with our experience in simulations- when the predictor distribution is multimodal and the shape of the conditional density is allowed to change with predictors, then the GPT tends to do as well or better than DP mixture of Gaussians. For the above study, the average number of components in the conditional distribution obtained from DP mixtures was around 15 which is quite high for a sample size of 50. As illustrated in table 2, BART, treed GP and the GP mean regression methods are primarily mean regression methods and so cannot possibly do well in terms of characterizing the entire conditional of response given predictors. They might perhaps estimate the mean surface reasonably well, but eventually fail in capturing multimodality or tail behavior, the latter often being an important focus of inferences.

## 6. EPIDEMIOLOGY APPLICATION

### 6.1 Study Background

DDT is a cheap and popular alternative for reducing the transmission of malaria, but has been shown to have negative effects on public health. In order to study the association between the DDT metabolite DDE and preterm delivery, Longnecker et al. (2001) measured DDE in mother's serum in the third trimester of pregnancy and also recorded gestational age at delivery (GAD) as well as age. In order to assess the dose-response relationship, they did logistic regression with response as dichotomized GAD (preterm or normal depending on a cut-off of 37 weeks of completed gestation) and explanatory variables as categorized DDE

based on empirical quantiles. Their results showed a significant dose-response relationship which had important public health implications. In order to get a more complete picture of the effect of DDE on preterm birth, Dunson et al. (2008) analysed the data using kernel stick breaking processes using 2313 observations and looked at the conditional distribution of GAD over different quantiles of DDE as well as the dose response relationship between preterm birth and DDE. Their analysis showed an increasing bump in the left tail with increasing DDE.

## 6.2 Analysis and Results

We used the GPT to analyze the dose response relationship in a subset of 182 women of advanced maternal age ( $\geq 35$  yrs) in the above dataset. We examined the conditional distribution of GAD at 10th, 60th, 90th and 99th percentile of DDE. Further, we looked at the dose response relationship between preterm birth and DDE, by examining the left tail of GAD over varying doses of DDE. We used normalized data for analysis and converted it back in the final step. Using the principles of theorem 3, we were able to incorporate prior information on the response density within our approach, which is not straightforward in other methods. To start with, we obtained simple frequentist kernel estimates (as in section 5.1) of the univariate response density using an earlier data on gestational age at delivery. This was then converted into an inverse cdf (using a linear approximation) to be used as a mean function for the response component. By choosing the prior on the normal residual precision as  $\text{Ga}(25,1)$  for the response component (thus allowing the residual variance to have mode near 0), we let the data influence the deviation of the posterior from the prior guess. Given the limited sample size and the complexity of the data we are trying to model, we adjusted the hyperparameter settings to reflect our prior belief about the data. The starting

value for the length-scale parameter in the covariance kernel in the Metropolis random walk was chosen to be 25, so as to have smooth Gaussian process prior. Instead of working with DDE, we used  $\log(\text{DDE})$  which resembled a Gaussian distribution, and let the mean function be 0 for the predictor component, with a  $\text{Ga}(1,1)$  prior for the corresponding residual precision.

Figure 2 shows the conditional distribution curves for GPT along with 90% credible intervals. Although we focused on a small subsample of 182 women of advanced maternal age, the GPT results for the conditional density are remarkably similar to the ones reported in Dunson et al. (2008), which suggests that there is no systematic difference for women of advanced maternal age. The conditional densities show an increasing bump in the left tail with increasing DDE, suggesting increased risk of preterm birth at higher doses. This is further supported by dose-response curves for  $P(\text{GAD} < T)$  in figure 4, with different choices for cut-off  $T$ . Although the dose-response curve is mostly flat for  $T=33$  weeks, the relationship becomes more significant as cut-off increases, with the dose-response tapering off at  $T=40$  weeks. This suggests that increased risk of preterm birth at higher DDE dosage is attributable to premature deliveries between 33 and 37 weeks. Trace plots of  $f(y|z)$  for different DDE percentiles (not shown) exhibit excellent rates of convergence and mixing. For comparison, figure 3 shows the density estimates from the DP mixture of Gaussians which has a tendency to overly favor multimodal densities, which is as expected given our simulation study results. These results were obtained using DP package in R (and the hyperparameter values therein), which utilizes algorithm 8 of Neal (2000) with  $m=1$ .

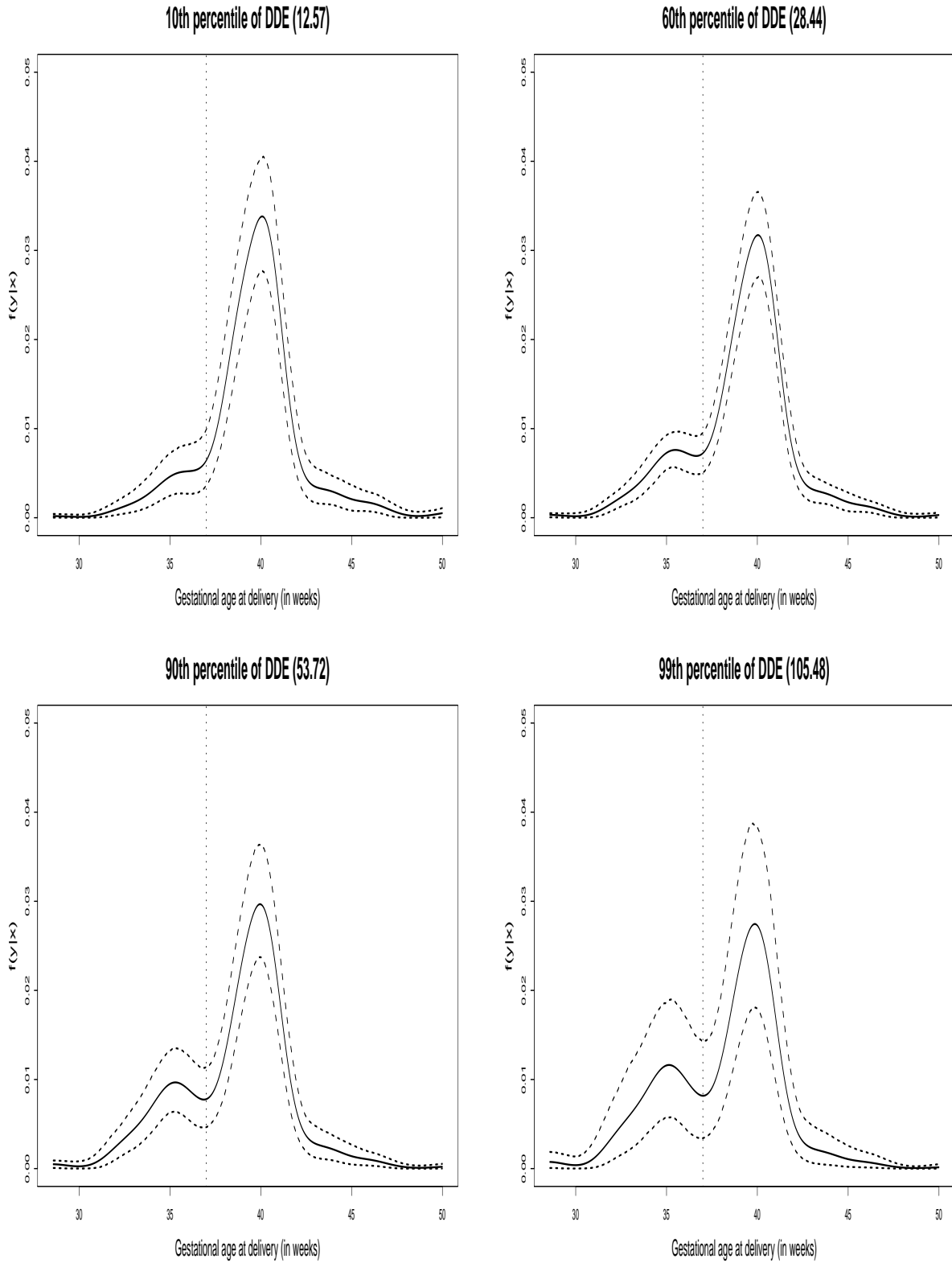


Figure 2: GPT conditional density estimates and 90% credible intervals for 10th, 60th, 90th, 99th DDE quantiles. Vertical dashed line for cut-off at 37 weeks.

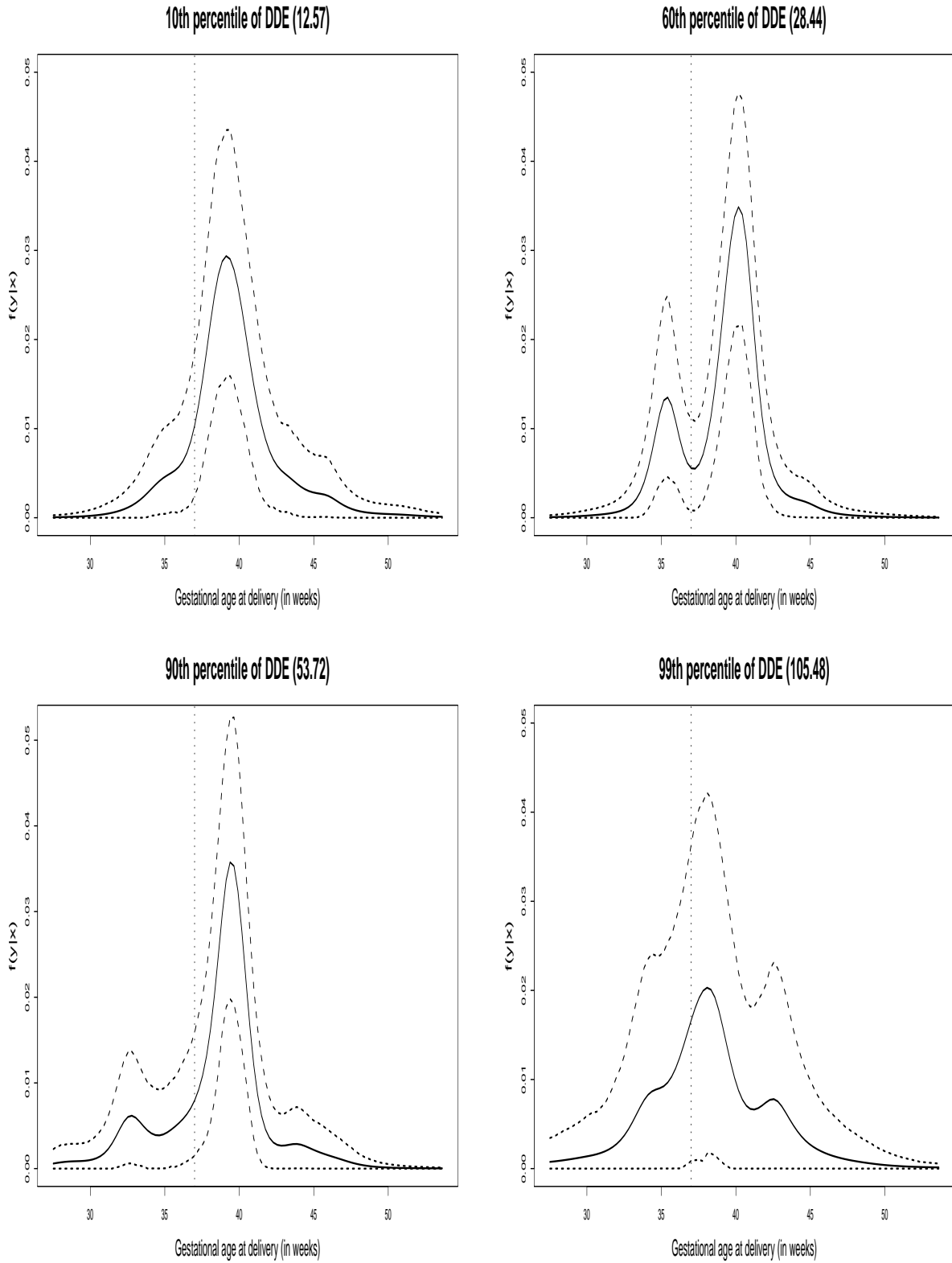


Figure 3: DPM conditional density estimates and 90% credible intervals for 10th, 60th, 90th, 99th DDE quantiles. Vertical dashed line for cut-off at 37 weeks.

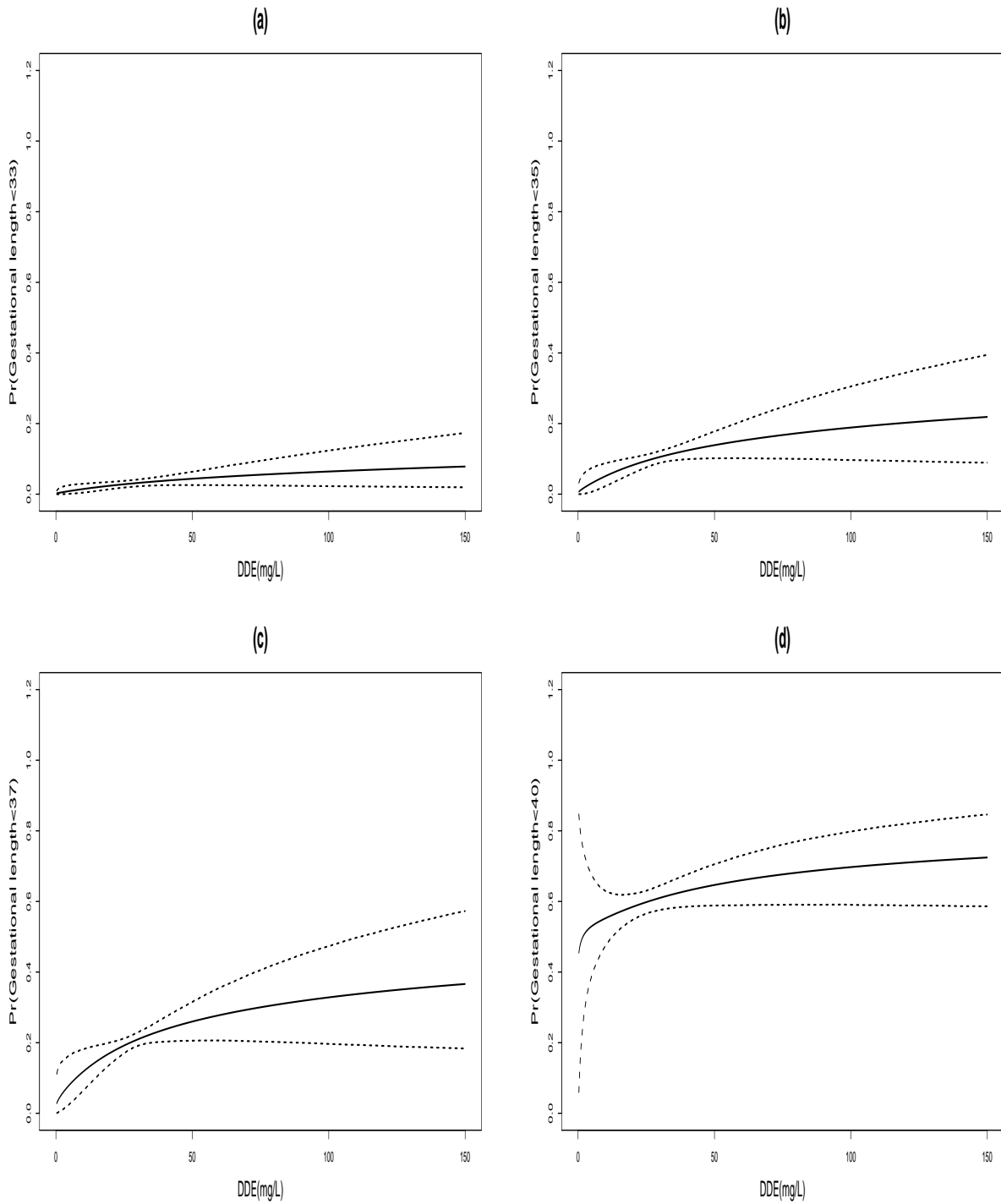


Figure 4: Estimated probability that gestational age at delivery is less than  $T$  weeks versus DDE dose, for (a)  $T = 33$ , (b)  $T = 35$ , (c)  $T = 37$ , (d)  $T = 40$ . Solid lines are posterior means and dashed lines are pointwise 90% credible intervals.

## 7. Discussion

In this paper, we discuss single factor transformation priors for density regression. This novel method provides us with a flexible non-discrete mixture alternative to be used in a variety of situations including density estimation, density regression, hierarchical latent variable models and even mixed models. This method combines theoretical flexibility and inherent transparency, thus providing an attractive alternative tool to practitioners.

### APPENDIX: PROOF OF THEOREMS

#### Proof of Theorem 1:

That  $f_{\mu,\sigma}$  exists is seen from equation (3). The limiting case as  $\sigma \rightarrow 0$  can be explained as the case when  $Y$  has the distribution function  $F = \mu^{-1}$  and marginalizing out the latent variable  $x \in [0, 1]$ . Thus  $\lim_{\sigma \rightarrow 0} \int_0^1 \Gamma_\sigma(y - \mu(x)) dx$  exists. Under regularity conditions, we can write the KL divergence between  $f_{\mu,\sigma}$  and  $f_0$  as

$$KL(f_{\mu,\sigma}, f_0) = \int f_0 \log \frac{f_0}{f_{\mu,\sigma}} = \int f_0(y) \left[ \frac{\lim_{\sigma \rightarrow 0} \int_0^1 \Gamma_\sigma(y - \mu_0(x)) dx}{\int_0^1 \Gamma_\sigma(y - \mu(x)) dx} \right] dy. \quad (9)$$

Note,  $\Gamma_\sigma(y - \mu(x)) \geq \frac{\Gamma_\sigma(y - \mu_0(x))}{h_\sigma(y, \mu(x) - \mu_0(x))}$ , where  $h_\sigma(y, \mu - \mu_0) = e^{\frac{1}{2\sigma^2}(\mu - \mu_0)^2 - \frac{1}{\sigma^2}(y - \mu_0)(\mu - \mu_0)}$  for normal error, while for Laplace error,  $h_\sigma(y, \mu - \mu_0) = e^{\frac{1}{\sigma}(|\mu_0 - \mu|)}$ . Further  $\sup_x |\log(h_\sigma(y, \mu(x) - \mu_0(x)))| \rightarrow 0$  for all  $y \in \mathfrak{R}$  as  $\|\mu - \mu_0\|_\infty, \sigma^2$  go to 0 with  $\|\mu - \mu_0\|_\infty / \sigma^2 \rightarrow 0$ . Under regularity conditions,  $\frac{f_0}{f_{\mu,\sigma}}$  has a finite upper bound, hence we can use dominated convergence theorem subsequently. Observe, for a fixed  $\sigma$ ,  $\int_0^1 \Gamma_\sigma(y - \mu(x)) dx \geq \frac{1}{\sup_{x \in (0,1)} h_\sigma(y, \mu(x) - \mu_0(x))} \int_0^1 \Gamma_\sigma(y - \mu_0(x)) dx$ . As  $\|\mu - \mu_0\|_\infty, \sigma^2$  go to 0 with  $\|\mu - \mu_0\|_\infty / \sigma^2 \rightarrow 0$ , and applying dominated convergence theorem,  $0 < \int_{\mathfrak{R}} f_0(y) \log \frac{f_0(y)}{f_{\mu,\sigma}(y)} dy \leq \lim_{\sigma \rightarrow 0} \lim_{\|\mu - \mu_0\|_\infty \rightarrow 0} \log(\sup_x h_\sigma(y, \mu(x) - \mu_0(x))) \rightarrow 0$ . For Cauchy errors,  $f_0(y) = \lim_{\sigma \rightarrow 0} \int_0^1 \frac{1}{\pi\sigma} \frac{1}{\left(1 + \frac{(y - \mu_0(x))^2}{\sigma^2}\right)} dx$  and

$$f_{\mu,\sigma}(y) = \int_0^1 \frac{1}{\pi\sigma} \frac{1}{\left(1 + \frac{(y-\mu(x))^2}{\sigma^2}\right)} dx = \int_0^1 \frac{1}{\pi\sigma} \frac{1}{\left(1 + \frac{1}{\sigma^2} [(y-\mu_0(x))^2 - 2(y-\mu_0(x))(\mu(x)-\mu_0(x)) + (\mu(x)-\mu_0(x))^2]\right)} dx.$$

As  $\|\mu - \mu_0\|_\infty, \sigma^2$  go to 0 with  $\|\mu - \mu_0\|_\infty/\sigma^2 \rightarrow 0$ , and applying dominated convergence theorem,  $\int_{\mathfrak{R}} f_0(y) \log \frac{f_0(y)}{f_{\mu,\sigma}(y)} dy \rightarrow 0$ . Thus taking appropriate limits,  $KL(f_{\mu,\sigma}, f_0)$  goes to 0 under Gaussian, Laplace and Cauchy errors. Hence we can choose a suitably small  $\epsilon_1$  and  $\epsilon_2$  with  $0 < \epsilon_1 < \epsilon_2$  such that  $\left\{ \|\mu - \mu_0\|_\infty \leq \epsilon_1, 0 < \sigma \leq \epsilon_2 \right\} \Rightarrow KL(f_{\mu,\sigma}, f_0) \leq \epsilon$ . Using the assumptions on the support of priors  $\Pi^*$  and  $\nu$ , we have  $\Pi(KL_\epsilon(f_0)) > 0$ .

### Proof of Theorem 2:

Using the regularity conditions and Taylor's series expansion, we have for fixed  $y, \mu, \sigma$ ,

$$\log \frac{f_0(y)}{f_{\mu,\sigma}(y)} = \sum_{k=1}^{n_0} (-1)^k \frac{\{(f_0(y) - 1)^k - (f_{\mu,\sigma}(y) - 1)^k\}}{k} + \delta_1^y(n_0) - \delta_2^y(n_0), \text{ for a fixed } n_0,$$

where  $\delta_1^y(n_0) - \delta_2^y(n_0)$  is uniformly bounded in  $y$ . Using the identity  $a^n - b^n = (a-b)(\sum_{k=1}^n a^{n-k} b^{k-1})$ ,

and denoting  $g_0 = f_0 - 1$  and  $g_{\mu,\sigma} = f_{\mu,\sigma} - 1$ , we have,

$$\begin{aligned} \int \left| \sum_{k=1}^{n_0} (-1)^k \frac{\{(f_0 - 1)^k - (f_{\mu,\sigma} - 1)^k\}}{k} \right| dy &\leq \sum_{k=1}^{n_0} \int \left| \frac{(-1)^k}{k} (f_0 - f_{\mu,\sigma}) \left( \sum_{l=1}^k g_{\mu,\sigma}^{k-l} g_0^{l-1} \right) \right| dy \\ &\leq \sum_{k=1}^{n_0} \sup_y \left| \frac{(-1)^k}{k} \left( \sum_{l=1}^k g_{\mu,\sigma}^{k-l} g_0^{l-1} \right) \right| \int |f_0(y) - f_{\mu,\sigma}(y)| dy = K(n_0) \int |f_0(y) - f_{\mu,\sigma}(y)| dy, \end{aligned}$$

where  $K(n_0) = \sum_{k=1}^{n_0} \sup_y \left| \frac{(-1)^k}{k} \left( \sum_{l=1}^k g_{\mu,\sigma}^{k-l} g_0^{l-1} \right) \right|$  is a finite constant depending on  $n_0$ , for  $\mu$

belonging to a finite  $L_1$  ball around  $\mu_0$ , using the regularity conditions. Further, using

similar methods as in the proof of theorem 3, we can show that for  $\epsilon_1 < \epsilon_2 < \epsilon_2^*$ ,

$$\{\mu \in N_{\epsilon_1}(\mu_0), \sigma \in (\epsilon_2, \epsilon_2^*)\} \Rightarrow \int |f_0(y) - f_{\mu,\sigma}(y)| dy < \frac{\epsilon_1}{\epsilon_2}. \quad (10)$$

Using inequality (10), we have for  $\mu \in N_{\epsilon_1}(\mu_0)$  and  $\sigma \in (\epsilon_2, \epsilon_2^*)$ ,

$$\begin{aligned} \int \left| \sum_{k=1}^{n_0} (-1)^k \frac{\{(f_0 - 1)^k - (f_{\mu,\sigma} - 1)^k\}}{k} \right| dy &\leq K(n_0) \frac{\epsilon_1}{\epsilon_2}. \text{ Also note, } KL(f_0, f_{\mu,\sigma}) \leq \int f_0 \left| \log \frac{f_0}{f_{\mu,\sigma}} \right| \\ &= \int f_0 \left| \sum_{k=1}^{n_0} (-1)^k \frac{\{(f_0 - 1)^k - (f_{\mu,\sigma} - 1)^k\}}{k} \right| + \delta_1^y(n_0) - \delta_2^y(n_0) \leq K(n_0) \frac{\epsilon_1}{\epsilon_2} + \Delta(n_0) = \epsilon, \end{aligned}$$

for a finite  $K(n_0)$  and suitably small  $\Delta(n_0) = \sup_y |\delta_1^y(n_0) - \delta_2^y(n_0)|$  with  $\epsilon_1, \epsilon_2$  depending on  $\epsilon$ . Under positive L-1 support by  $\Pi^*$ , the rest follows by similar arguments as in theorem 1.

### Proof of Theorem 3:

Note that,  $|f_{\mu,\sigma}(y) - f_0(y)| \leq |\sup_{x \in (0,1)} \phi_\sigma(y - \mu(x)) - f_0|$ , which is integrable  $\forall(\mu, \sigma)$ .

Then using dominated convergence theorem,  $\lim_{\sigma \rightarrow 0} \int |f_{\mu,\sigma} - f_0| = \int \lim_{\sigma \rightarrow 0} |f_{\mu,\sigma} - f_0|$ . Now applying Fatou's Lemma and Fubini's Theorem successively, we have,

$$\begin{aligned}
\int \lim_{\sigma \rightarrow 0} |f_{\mu,\sigma} - f_0| dy &\leq \int \lim_{\sigma \rightarrow 0} \int_0^1 |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \mu_0(x))| dx dy \\
&\leq \liminf_{\sigma \rightarrow 0} \int \int_0^1 |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \mu_0(x))| dx dy \quad (\text{Fatou's lemma}) \\
&= \liminf_{\sigma \rightarrow 0} \int_0^1 \int |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \mu_0(x))| dy dx \quad (\text{Fubini's Theorem}) \\
&= \liminf_{\sigma \rightarrow 0} \left\{ \int_{x|\mu_0 > \mu} \int |\phi_\sigma(y - \mu(x)) - \phi_\sigma(y - \mu_0(x))| dy dx \right. \\
&\quad \left. + \int_{x|\mu_0 < \mu} \int |\phi_\sigma(y - \mu_0(x)) - \phi_\sigma(y - \mu(x))| dy dx \right\}.
\end{aligned}$$

In the proof of lemma 1 of Ghosal, Ghosh and Ramamoorthy (1999), it was shown that for fixed  $\theta_1 < \theta_2$ ,  $\|\phi_\sigma(y - \theta_1) - \phi_\sigma(y - \theta_2)\| < \frac{\theta_2 - \theta_1}{\sigma}$ , which would imply

$$\begin{aligned}
\int \lim_{\sigma \rightarrow 0} |f_{\mu,\sigma} - f_0| dy &\leq \liminf_{\sigma \rightarrow 0} \left\{ \int_{x|\mu_0 > \mu} \frac{\mu_0(x) - \mu(x)}{\sigma} dx + \int_{x|\mu_0 < \mu} \frac{\mu(x) - \mu_0(x)}{\sigma} dx \right\} \\
&= \liminf_{\sigma \rightarrow 0} \int_0^1 \frac{|\mu(x) - \mu_0(x)|}{\sigma} dx \leq \lim_{\sigma \rightarrow 0} \int_0^1 \frac{|\mu(x) - \mu_0(x)|}{\sigma} dx. \quad (11)
\end{aligned}$$

As  $\|\mu - \mu_0\|_\infty, \sigma^2$  go to 0 with  $\|\mu - \mu_0\|_\infty / \sigma^2 \rightarrow 0$ , the above limit exists and goes to 0. Given that the limit exists and goes to 0, we can now choose sufficiently small  $\epsilon_1, \epsilon_2, \epsilon_2^*$  with  $0 < \epsilon_1 < \epsilon_2 < \epsilon_2^*$  such that for  $\int_0^1 |\mu_0(x) - \mu(x)| dx < \epsilon_1 \equiv \mu \in N_{\epsilon_1}(\mu_0)$  and  $\sigma \in (\epsilon_2, \epsilon_2^*)$ , we would have  $\int |f_{\mu,\sigma} - f_0| dy < \frac{\epsilon_1}{\epsilon_2}$ , using (11).

### Proof of Theorem 4:

Our proof is based on theorem 2 of Ghosal, Ghosh and Ramamoorthi (1999) who gave a set of alternate sufficient conditions for almost sure convergence of the posterior of strong neighborhoods. Their result involves conditions on the size of the parameter space in terms of L-1 metric entropy. Before proceeding, let us review L-1 metric entropy and theorem 2 of Ghosal, Ghosh and Ramamoorthi (1999).

**DEFINITION 1.** For  $\mathcal{G} \subset \mathcal{F}$  and  $\delta > 0$ , L-1 metric entropy  $J(\delta, \mathcal{G})$  is defined as the logarithm of minimum of all  $k$ , such that there exist  $f_1, f_2, \dots, f_k$  in  $\mathcal{F}$  with the property  $\mathcal{G} \subset \cup_{i=1}^k \{f : \int |f - f_i| dy < \delta\}$ .

**Theorem 6.** (Ghosal, Ghosh and Ramamoorthi) Let  $\Pi$  be a prior on  $\mathcal{F}$ . Suppose  $f_0 \in \mathcal{F}$  is in the Kullback-Leibler support of  $\Pi$  and let  $U = \{f : \int |f - f_0| dy < \epsilon\}$ . If there is a  $\delta < \epsilon/4$ ,  $c_1, c_2 > 0$ ,  $\beta < \epsilon^2/8$  and  $\mathcal{F}_n \subset \mathcal{F}$  such that for all large  $n$ :

- (1)  $\Pi(\mathcal{F}_n^c) < c_1 \exp(-nc_2)$ , and,
- (2) The L-1 metric entropy,  $J(\delta, \mathcal{F}_n) < n\beta$ ,

then  $\Pi(U|Y_1, Y_2, \dots, Y_n) \rightarrow 1$  a.s.  $P_{f_0}$ .

The constants  $\delta, c_1, c_2, \beta$  and  $\mathcal{F}_n$  are allowed to depend on fixed neighborhoods, or equivalently on  $\epsilon$ .

We start off by assuming that  $\nu\left(0 < \sigma < M\right) = 1$  with  $M < \infty$  and  $\sup_{x \in (0,1)} m(x) < \infty$ . Let  $A_n = [-a_n, a_n]$ ,  $B_n = [-b_n, b_n]$  and let us denote  $\tilde{\mu} = \sup_{x \in (0,1)} |\mu(x) - m(x)| > 0$ . Let us consider the sieve  $\mathcal{F}_n := \left\{f_{\mu, \sigma}(y) : y \in A_n = [-a_n, a_n], \text{ with } a_n = b_n + c_n M \text{ and } 0 < \mu(\tilde{x}) \leq b_n\right\}$ , where  $a_n, b_n, c_n$  are  $O(n^{1/2})$ , strictly positive and increase to  $\infty$ . Clearly  $\mathcal{F}_n \subset \mathcal{F}$  and  $\mathcal{F}_n$

increases to  $\mathcal{F}$  as  $n \rightarrow \infty$ .

$$\begin{aligned}
\text{Now } \Pi(\mathcal{F}_n^c) &= \Pi\left(f_{\mu,\sigma} : y \in A_n^c, \tilde{\mu} > b_n\right) + \Pi\left(f_{\mu,\sigma} : y \in A_n, \tilde{\mu} > b_n\right) + \Pi\left(f_{\mu,\sigma} : y \in A_n^c, \tilde{\mu} \leq b_n\right) \\
&= \Pi^*\left(\mu : \tilde{\mu} > b_n\right) + \Pi\left(f_{\mu,\sigma} : y \in A_n^c, \tilde{\mu} \leq b_n\right) \\
&\leq \Pi^*\left(\mu : \tilde{\mu} > b_n\right) + \Pi\left(f_{\mu,\sigma} : y \in A_n^c, \mu(x) \in B_n, x \in (0, 1)\right). \tag{12}
\end{aligned}$$

$$\begin{aligned}
\text{For } c_0 = \text{variance of the GP, the second term is } &\Pi\left(f_{\mu,\sigma} : y \in A_n^c, \mu(x) \in B_n, x \in (0, 1)\right) \\
&= \int_0^M \int_{B_n} \int_0^1 \left\{ P\left(y < -a_n | \mu(\cdot), \sigma, x\right) + P\left(y > a_n | \mu(\cdot), \sigma, x\right) \right\} \phi_{c_0}(\mu(x) - m(x)) dx d\mu \nu(d\sigma) \\
&\leq \int_0^M \int_{B_n} \int_0^1 \left\{ P\left(\frac{y + b_n}{M} < \frac{-a_n + b_n}{M}\right) + P\left(\frac{y - b_n}{M} > \frac{a_n - b_n}{M}\right) \right\} \phi_{c_0}(\mu(x) - m(x)) dx d\mu \nu(d\sigma) \\
&\leq \int_0^M \int_{-\infty}^{\infty} \int_0^1 \left\{ \Phi(-c_n) + \left(1 - \Phi(c_n)\right) \right\} \phi_{c_0}(\mu(x) - m(x)) dx d\mu \nu(d\sigma) \\
&= \frac{2}{\sqrt{2\pi}} \int_{-\infty}^{-c_n} \exp^{-\frac{1}{2}u^2} du \leq \frac{2}{\sqrt{2\pi}} \exp^{-\frac{1}{4}c_n^2} \int_{-\infty}^{-c_n} \exp^{-\frac{1}{4}u^2} du \leq L \exp^{-\frac{1}{4}c_n^2}, \tag{13}
\end{aligned}$$

since  $c_n$  is strictly positive. The inequality in the second step is obtained by bounding  $P\left(y < -a_n | \mu, \sigma\right)$  by the left tail of a  $N(-b_n, M^2)$  distribution and bounding  $P\left(y > a_n | \mu, \sigma\right)$  with the right tail of a  $N(b_n, M^2)$  distribution. Also Choi and Schervish (2004) showed in Lemma 5, that for 0 mean Gaussian process  $\eta$  on  $[0,1]$  having a continuously differentiable mean function and a covariance kernel having fourth partial derivatives,  $P\left(\sup_{x \in (0,1)} |\eta(x)| > V\right) \leq A \exp\left(-dV^2\right)$ , where  $A$  and  $d$  are positive constants. Substituting  $\mu(\tilde{x})$  for  $\eta(x)$  in the above inequality we obtain,

$$\Pi^*\left(\mu : \tilde{\mu} > b_n\right) \leq A \exp\left(-db_n^2\right), \text{ for some positive constants } A \text{ and } d. \tag{14}$$

Using inequalities (13) and (14) in equation (12), we have

$$\Pi(\mathcal{F}_n^c) \leq A \exp\left(-db_n^2\right) + L \exp\left(-\frac{1}{4}c_n^2\right) \leq B \exp(-p\gamma_n),$$

where  $\gamma_n = O(n)$ , and  $B$  and  $p$  are positive constants. Thus we have shown that condition 1 of Theorem 6 holds. Before showing that the second sufficient condition holds, let us prove the following lemma.

**Lemma 1.** For  $y \in A_n$  and  $\mu \in B_n$ , let  $f_{\mu,\sigma}$  be defined as in eqn (3). Then  $\sup_y |f_{\mu,\sigma}(y)| \leq \gamma$  and  $\sup_{y \in A_n} \left| \frac{d}{dy} f_{\mu,\sigma}(y) \right| \leq \beta$ , where both  $\gamma, \beta$  are finite and  $d_1, d_2 > 0$ .

**Proof:** From the definition of  $f_{\mu,\sigma}$  in equation (3), it is easy to see that

$\sup_{y \in A_n} f_{\mu,\sigma}(y) \leq \sup_{y \in A_n} \sup_{x \in (0,1)} \phi_\sigma(y - \mu(x)) \leq \gamma < \infty$  (using the form of normal pdf). Again, using interchange of integral and differentiation, we have

$$\begin{aligned} \sup_{y \in A_n} \left| \frac{d}{dy} f_{\mu,\sigma}(y) \right| &= \sup_{y \in A_n} \int_0^1 \left| \frac{d}{dy} \phi_\sigma(y - \mu(x)) \right| dx \\ &= \sup_{y \in A_n} \int_0^1 \left| -\frac{1}{\sqrt{2\pi\sigma^2}} \left( \frac{y - \mu(x)}{\sigma^2} \right) \exp \left\{ -\frac{1}{2\sigma^2} (y - \mu(x))^2 \right\} \right| dx \\ &\leq \sup_{y \in A_n} \sup_{x \in (0,1)} \left| -\frac{1}{\sqrt{2\pi\sigma^2}} \left( \frac{y - \mu(x)}{\sigma^2} \right) \exp \left\{ -\frac{1}{2\sigma^2} (y - \mu(x))^2 \right\} \right| = \beta < \infty. \end{aligned}$$

Thus our lemma is proved. Now theorem 2.7.1 of van der Vaart and Wellner (1996) shows that if  $\mathcal{X}$  be a bounded convex subset of  $\mathfrak{R}$  with non-empty interior, and  $C_1^1(\mathcal{X}) :=$  set of all continuous functions  $g: \mathcal{X} \rightarrow \mathfrak{R}$ , with  $\|g\|_1 = \sup_x |g(x)| + \sup_{x,y} \frac{|g(x) - g(y)|}{|x - y|} \leq 1$ , then there exist a constant  $k$  such that  $\log N(\epsilon, C_1^1(\mathcal{X}), \|\cdot\|_\infty) \leq k\lambda(\mathcal{X})/\epsilon$ , where  $\|\cdot\|_\infty$  denotes the sup-norm distance,  $N$  is the  $\epsilon$ -covering number in sup-norm and  $\lambda$  is the Lesbesgue measure of the set  $\left\{ x : |x - \mathcal{X}| < 1 \right\}$ . We adapt this theorem slightly for our setup, where we denote  $\mathcal{X}_n = A_n = [-a_n, a_n]$  with  $\lambda(\mathcal{X}_n) = \lambda \left\{ y : |y - \mathcal{X}_n| < 1 \right\} \propto a_n$ ,  $C_1^1(\mathcal{X}_n) := \left\{ f_{\mu,\sigma} : \sup_{y \in A_n} |f_{\mu,\sigma}(y)| < \gamma, \sup_{y \in A_n} \left| \frac{d}{dy} f_{\mu,\sigma}(y) \right| < \beta \right\}$ . Using similar reasoning as in Lemma 4 of Choi and Schervish (2007), we replace  $g$  by  $f_{\mu,\sigma}/2\max(\beta, \gamma)$ , and get the following for  $f_{\mu,\sigma} \in C_1^1(\mathcal{X}_n)$ :

$$\begin{aligned} \|g\|_1 &= \sup_y |g(y)| + \sup_{x,y} \frac{|g(x) - g(y)|}{|x - y|} = \sup_y \left| \frac{f_{\mu,\sigma}(y)}{2\max(\beta, \gamma)} \right| + \sup_{x,y} \frac{|f_{\mu,\sigma}(x) - f_{\mu,\sigma}(y)|}{2\max(\beta, \gamma)|x - y|} \\ &\leq \sup_y \left| \frac{f_{\mu,\sigma}(y)}{2\max(\beta, \gamma)} \right| + \frac{\sup_y \left| \frac{d}{dy} f_{\mu,\sigma}(y) \right|}{2\max(\beta, \gamma)} < 1, \text{ (using mean value theorem).} \end{aligned}$$

Thus using theorem 2.7.1 of Van der Vaart and Wellner (1996), we obtain,  $\log N(\epsilon, C_1^1(\mathcal{X}_n), \|\cdot\|_\infty) \leq K' a_n \max(\beta, \gamma)/\epsilon = K'' a_n/\epsilon$ . It can also be easily observed from lemma 1 that  $\mathcal{F}_n \subset C_1^1(\mathcal{X}_n)$ , which would mean,  $\log N(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) \leq \log N(\epsilon, C_1^1(\mathcal{X}_n), \|\cdot\|_\infty) \leq K'' a_n/\epsilon$ . Also, if we replace  $\epsilon$  by  $\epsilon_n = \frac{\delta}{2a_n}$  for each  $\mathcal{F}_n$  above, then  $\log N(\frac{\delta}{2a_n}, \mathcal{F}_n, \|\cdot\|_\infty) \leq 2K'' a_n^2/\delta = O(n)$ . Now note that

$$\left\{ f_{\mu, \sigma}(y) : \sup_{y \in A_n} |f_{\mu, \sigma}(y) - f^i(y)| < \frac{\delta}{2a_n} \right\} \subseteq \left\{ f_{\mu, \sigma}(y) : \int_{y \in A_n} |f_{\mu, \sigma}(y) - f^i(y)| dy < \delta \right\}, \forall i.$$

The above implies that for a fixed  $\mathcal{F}_n$ , we can replace each sup-norm ball of radius  $\frac{\delta}{2a_n}$  by a L-1 ball of radius  $\delta$  having the same center, so as to obtain  $\log N(\frac{\delta}{2a_n}, \mathcal{F}_n, \|\cdot\|_\infty) = J(\delta, \mathcal{F}_n)$ . Choosing  $\delta < \epsilon/4$ , we can choose sufficiently small  $K_1$  in  $a_n^2 = K_1 n$  to ensure that  $J(\delta, \mathcal{F}_n) \leq n\alpha$ , with  $\alpha < \epsilon^2/8$ . Hence the second sufficient condition is satisfied. Thus using theorem 2 of Ghosal, Ghosh and Ramamoorthi (1999), we obtain our result.

### Proof of Theorem 5:

Let us denote  $\beta = (\mu^Y, \mu^Z, \sigma_Y, \sigma_Z)$ ,  $\beta_0 = (\mu_0^Y, \mu_0^Z, \sigma_{0Y}, \sigma_{0Z})$ ,  $\tilde{\mu} = (\mu^Y, \mu^Z)$ ,  $\tilde{\mu}_0 = (\mu_0^Y, \mu_0^Z)$ ,  $\tilde{\sigma} = (\sigma_Y, \sigma_Z)$  and  $\tilde{\sigma}_0 = (\sigma_{0Y}, \sigma_{0Z})$ . The KL divergence between  $\psi(y, z|\beta) = \psi_\beta$  and  $\psi_{\beta_0}$  is

$$\begin{aligned} KL(\psi_\beta, \psi_{\beta_0}) &= \int \int \psi_{\beta_0} \log \left( \frac{\psi_{\beta_0}}{\psi_\beta} \right) dy dz = \int \int \psi_{\beta_0} \log \left( \frac{\psi_{\beta_0}}{\psi_{\tilde{\mu}, \tilde{\sigma}}} \right) dy dz \\ &= \int \int \psi_{\beta_0} \log \left( \frac{\psi_{\beta_0}}{\psi_{\tilde{\mu}_0, \tilde{\sigma}_0}} \right) dy dz + \int \int \psi_{\beta_0} \log \left( \frac{\psi_{\tilde{\mu}_0, \tilde{\sigma}_0}}{\psi_{\tilde{\mu}, \tilde{\sigma}}} \right) dy dz. \end{aligned} \quad (15)$$

Looking at the second term on the right hand side in equation (15), we have

$$\begin{aligned} \log \left( \frac{\psi_{\tilde{\mu}_0, \tilde{\sigma}_0}}{\psi_{\tilde{\mu}, \tilde{\sigma}}} \right) &\leq \log \left( \sup_x \exp \left\{ \frac{1}{2\sigma_Y^2} (\mu^Y(x) - \mu_0^Y(x))^2 - \frac{1}{\sigma_Y^2} (y - \mu_0^Y(x)) (\mu^Y(x) - \mu_0^Y(x)) \right\} \right) \\ &\quad + \log \left( \sup_x \exp \left\{ \frac{1}{2\sigma_Z^2} (\mu^Z(x) - \mu_0^Z(x))^2 - \frac{1}{\sigma_Z^2} (z - \mu_0^Z(x)) (\mu^Z(x) - \mu_0^Z(x)) \right\} \right). \end{aligned} \quad (16)$$

Letting  $\mu^Y \rightarrow \mu_0^Y$  in sup-norm and  $\mu^Z \rightarrow \mu_0^Z$  in sup-norm, right hand side of (16) goes to 0.

Under regularity conditions, a finite upper bound exists for  $\frac{\psi_{\tilde{\mu}_0, \tilde{\sigma}_0}}{\psi_{\tilde{\mu}, \tilde{\sigma}}}$ , hence applying dominated convergence theorem, the second term on the right hand side in equation (15) goes to 0.

Similarly, a finite upper bound exists for  $\frac{\psi_{\beta_0}}{\psi_{\hat{\mu}_0, \hat{\sigma}}}$ , hence letting  $\sigma_Y \rightarrow \sigma_{0Y}$  and  $\sigma_Z \rightarrow \sigma_{0Z}$  and applying dominated convergence theorem, the first term on the right hand side in equation (15) also goes to 0. Under positive sup-norm support of  $\mu_0^Y$  and  $\mu_0^Z$  and  $\sigma_{0Y}$ ,  $\sigma_{0Z}$  belonging to the support of  $\nu$ , the rest follows by similar arguments as in other proofs.

To prove remark 2, note that for a fixed  $z$

$$KL(\zeta_{z,\beta}, \zeta_{z,\beta_0}) = \int \zeta_{z,\beta_0}(y) \log \left( \frac{\zeta_{z,\beta_0}(y)}{\zeta_{z,\beta}(y)} \right) dy = \int \zeta_{z,\beta_0}(y) \log \left( \frac{\psi_{\beta_0}(y, z)}{\psi_{\beta}(y, z)} \right) dy - \log \left( \frac{f(z|\mu_0^Z, \sigma_{0Z})}{f(z|\mu^Z, \sigma_Z)} \right),$$

where  $f(z|\mu^Z, \sigma_Z) = \int_0^1 \phi_{\sigma_Z}(z - \mu^Z(x)) dx$  is the marginal density. Using similar steps as above, it can be shown that both the terms on the right hand side goes to 0 as  $\mu^Z$  goes to  $\mu_0^Z$  in sup-norm and  $\sigma_Z \rightarrow \sigma_{0Z}$ .

## References

- [1] Brown, E.R., and Ibrahim, J.G. (2003), “A Bayesian semiparametric joint hierarchical model for longitudinal and survival data”, *Biometrics*, 59, 221 - 228.
- [2] Bush, C.A., and MacEachern, S.N. (1996), “A semiparametric Bayesian model for randomised block designs”, *Biometrika*, 83, 275 - 285.
- [3] Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q., and West, M. (2008), “High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics”, *Journal of the American Statistical Association*, 103, 1438 - 1456.
- [4] Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D.B., and Carin, L. (2010), “Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: algorithm and performance bounds”, *IEEE. Transactions on Signal Processing*, 58, 6140 - 6155.

- [5] Chipman, H., George, E., and McCulloch, R. (2010), “BART: Bayesian Additive Regression Trees”, *The Annals of Applied Statistics*, 4, 266 - 298.
- [6] Choi, T., and Schervish, M. (2004), “Posterior Consistency in Nonparametric Regression Problems under Gaussian Process Priors”, Technical Report.
- [7] De Iorio, M., Muller, P., Rosner, G.L., and MacEachern, S. (2004), “An ANOVA Model for Dependent Random Measures”, *Journal of the American Statistical Association*, 99, 205-215.
- [8] Dunson, D.B. (2006), “Bayesian dynamic modeling of latent trait distributions”, *Biostatistics*, 7, 551 - 568.
- [9] Dunson, D. B., Pillai, N., and Park, J. H. (2007), “Bayesian density regression”, *Journal of the Royal Statistical Society, Series B*, 69, 163 - 183.
- [10] Dunson, D.B., and Park, J. H. (2008), “Kernel stick breaking processes”, *Biometrika*, 95, 307 - 323.
- [11] Escobar, M.D., and West, M. (1995), “Bayesian Density Estimation and Inference Using Mixtures”, *Journal of the American Statistical Association*, 90, 577 - 588.
- [12] Ferguson, T. S. (1973), “A Bayesian analysis of some nonparametric problems”, *Annals of Statistics*, 1, 209 - 230
- [13] Ferguson, T. S. (1974), “Prior distributions on spaces of probability measures”, *Annals of Statistics*, 2, 615 - 629.
- [14] Fokoue, E., and Titterton, D.M. (2003), “Mixtures of factor analysers: Bayesian estimation and inference by stochastic simulation”, *Machine Learning*, 50, 73 - 94.

- [15] Fokoue, E. (2005), "Mixtures of factor analyzers: an extension with covariates", *Journal of Multivariate Analysis*, 95, 370 - 384.
- [16] Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), "Posterior consistency of Dirichlet mixtures in density estimation", *Annals of Statistics*, 27, 143 - 158.
- [17] Ghosal, S., and Roy, S. (2006), "Posterior consistency of Gaussian process prior for nonparametric binary regression", *Annals of Statistics*, 34, 2413 - 2429.
- [18] Gramacy R.B., and Lee, H. K. H, (2008), "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling", *Journal of the American Statistical Association*, 103, 1119 - 1130.
- [19] Griffin, J. E. and Steel, M. F. J. (2006), "Order-based dependent Dirichlet processes", *Journal of the American Statistical Association*, 101, 179 - 194.
- [20] Hanson, T. (2006), "Inference for Mixtures of Finite Polya Trees", *Journal of the American Statistical Association*, 101, 1548 - 1565.
- [21] Jara, A., and Hanson, T. (2010), "A class of mixtures of dependent tail-free processes", *Biometrika*, accepted.
- [22] Kleinman, K.P., and Ibrahim, J.G. (1998), "A semiparametric Bayesian approach to the random effects model", *Biometrics*, 54, 921 - 938.
- [23] Lavine, M. (1992), "Some Aspects of Polya Tree Distributions for Statistical Modelling", *Annals of Statistics*, 20, 1222 - 1235.
- [24] Lavine, M. (1994), "More Aspects of Polya Tree Distributions for Statistical Modelling", *Annals of Statistics*, 22, 1161 - 1176.

- [25] Lawrence, N. (2005), “Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models”, *Journal of Machine Learning Research*, 6, 1783 - 1816.
- [26] Lee, S.Y., Lu, B., and Song, X.Y. (2008), “Semiparametric Bayesian analysis of structural equation models with fixed covariates”, *Statistics in Medicine*, 27, 2341 - 2360.
- [27] Lenk, P. J. (1988), “The Logistic Normal Distribution for Bayesian, Nonparametric, Predictive Densities”, *Journal of the American Statistical Association*, 83, 509 - 516.
- [28] Lenk, P. J. (1991), “Towards a Practicable Bayesian Nonparametric Density Estimator”, *Biometrika*, 78, 531 - 543.
- [29] 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 - 4.
- [30] MacEachern, S. N. (1999), “Dependent nonparametric processes”, In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association, pp. 50-55.
- [31] Marron, J. S., and Wand, M. P. (1992), “Exact mean integrated squared error”, *Annals of Statistics*, 20, 712 - 736.
- [32] Mauldin, R.D., Sudderth, W.D., and Williams, S.C. (1992), “Polya Trees and Random Distributions”, *Annals of Statistics*, 20, 1203 - 1221.
- [33] Müller, P., Erkanli, A., and West, M. (1996), “Bayesian curve fitting using multivariate normal mixtures”, *Biometrika*, 83, 67 - 79.
- [34] Neal, R.M. (2000), “Markov Chain sampling methods for Dirichlet process mixture models”, *Journal of Computational and Graphical Statistics*, 9, 249 - 265.

- [35] O’Hagan, A., and Kingman, J. F. C. (1978), “Curve Fitting and Optimal Design for Prediction”, *Journal of the Royal Statistical Society B*, 40, 1 - 42.
- [36] Sethuraman, J. (1994), “A constructive definition of Dirichlet priors”, *Statistica Sinica*, 4, 639 - 650.
- [37] Sheather, S. J., and Jones M. C. (1991), “A reliable data-based bandwidth selection method for kernel density estimation”, *Journal of the Royal Statistical Society B*, 53, 683 - 690.
- [38] Silva, R., and Gramacy, R. (2010), “Gaussian process structural equation models with latent variables”, *Proceedings of the 26th Conference on Uncertainty on Artificial Intelligence, UAI*.
- [39] Schwartz, L. (1965), “On Bayes procedures”, *Z. Wahrsch. Verw. Gebiete*, 4, 10 - 26.
- [40] Tokdar, S. T., and Ghosh, J. K. (2007), “Posterior consistency of logistic Gaussian process priors in density estimation”, *Journal of Statistical Planning and Inference*, 137, 34 - 42.
- [41] Tokdar, S. T., Zhu, Y. M., and Ghosh, J. K. (2010), “Bayesian Density Regression with Logistic Gaussian Process and Subspace Projection Bayesian Analysis”, *Bayesian Analysis*, 5, 319 - 344.
- [42] Van der Vaart, A. W., and Wellner, J. A. (1996), “Weak Convergence and Empirical Processes”, Springer-Verlag, New York.
- [43] Yang, M., and Dunson, D.B. (2010), “Bayesian semiparametric structural equation models with latent variables”, *Psychometrika*, 75, 675-693.