Nonparametric Bayes Conditional Distribution Modeling With Variable Selection

Yeonseung CHUNG and David B. DUNSON

This article considers a methodology for flexibly characterizing the relationship between a response and multiple predictors. Goals are (1) to estimate the conditional response distribution addressing the distributional changes across the predictor space, and (2) to identify important predictors for the response distribution change both within local regions and globally. We first introduce the probit stick-breaking process (PSBP) as a prior for an uncountable collection of predictor-dependent random distributions and propose a PSBP mixture (PSBPM) of normal regressions for modeling the conditional distributions. A global variable selection structure is incorporated to discard unimportant predictors, while allowing estimation of posterior inclusion probabilities. Local variable selection is conducted relying on the conditional distribution estimates at different predictor points. An efficient stochastic search sampling algorithm is proposed for posterior computation. The methods are illustrated through simulation and applied to an epidemiologic study.

KEY WORDS: Conditional distribution estimation; Hypothesis testing; Kernel stick-breaking process; Mixture of experts; Stochastic search variable selection.

1. INTRODUCTION

This article focuses on flexible modeling of the conditional density of a response variable y given multiple predictors $\mathbf{x} = (x_1, \ldots, x_p)'$. We treat $f(y|\mathbf{x})$ as unknown and potentially changing in shape as $\mathbf{x}$ varies. In addition, our emphasis is on selecting the subset of predictors that have any impact on the response distribution change, either within some local regions of the predictor space or globally. Subset selection is of interest in performing inferences on effects of particular predictors and in building sparse predictive models. Sparsity is of paramount importance in modeling of conditional distributions with many candidate predictors due to the curse of dimensionality.

There is a rich literature on frequentist methods for conditional distribution estimation. Fan, Yao, and Tong (1996) proposed a double-kernel local linear approach. Fan and Yim (2004) developed a cross validation approach for bandwidth selection. Related frequentist methods have been considered by Hall, Wolff, and Yao (1999) and Hyndman and Yao (2002) among others. Müller, Erkanli, and West (1996) proposed a Bayesian approach to nonlinear regression, which was conceptually related to the double-kernel approach. In particular, in order to induce a prior on the unknown function, $E(y|\mathbf{x})$, Müller, Erkanli, and West (1996) proposed to model the joint density of $(y, \mathbf{x})$ using a Dirichlet process mixture (DPM) of Gaussians (Lo 1984; Escobar 1994; Escobar and West 1995). Alternative classes of nonparametric priors that can potentially be used for modeling $f(y|\mathbf{x})$ have been proposed by MacEachern (1999), Griffin and Steel (2006, 2007), Dunson and Park (2008), and Chung and Dunson (2009).

The focus in the above literature has been on estimation and, to our knowledge, there has been essentially no consideration of the important problems of variable selection and hypothesis testing in the general setting of conditional distribution modeling with multiple discrete and continuous candidate predictors. The methods that have been recently proposed are limited in scope to particular cases. Pennell and Dunson (2008) developed a method for testing for changes in unknown distributions across levels of an ordinal predictor. Based on dependent Dirichlet processes (DDPs) with fixed weights, Dunson and Peddada (2008) developed methods for estimating and testing of stochastically ordered distributions across groups.

This article proposes a general Bayesian nonparametric approach for variable selection and hypothesis testing in conditional distribution modeling, avoiding the fixed weights assumption that limits flexibility in building sparse models. We first introduce the probit stick-breaking process (PSBP) as a new choice of prior for an uncountable collection of predictor-dependent random distributions. The PSBP has distinct advantages over previous formulations in terms of computational tractability, which is particularly important in variable selection settings as marginal likelihoods need to be calculated. For modeling conditional distributions, we propose a PSBP mixture (PSBPM) of normal linear regressions, resulting in an infinite mixture with mixing weights varying with predictors.

The primary emphasis of this article is on variable selection and we allow predictors to drop out of the model through zeroing of coefficients in the PSBPM specification. This is carefully formulated to allow development of an efficient stochastic search variable selection (SSVS) algorithm, which can be used to simultaneously search the model space, estimate posterior inclusion probabilities for the predictors, and obtain model-averaged conditional density estimates and predictive distributions. In addition, local variable selection is conducted using the total variation distance of the conditional distribution estimates at different predictor points. Our approach generalizes the SSVS algorithms for linear regression (George and McCulloch 1997) and nonlinear mean and variance regression (Chan et al. 2006; Leslie, Kohn, and Nott 2007) to settings in

Yeonseung Chung is Research Fellow, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave. SPH2, 4th Floor, Boston, MA 02115 (E-mail: *ychung@hsph.harvard.edu*). David B. Dunson is Professor, Department of Statistical Science, Duke University, 218 Old Chemistry Building, Box 90251, Durham, NC 27707 (E-mail: *dunson@stat.duke.edu*). This research was supported in part by the Intramural Research Program of the NIH, NIEHS. The authors thank Lynne Wagenknecht and Dora Il'yasova for generously providing the IRAS study data and for their helpful comments on the approach and Hui Zou for generously providing the R code for implementing the ALLR. The authors also thank editors and referees for their helpful comments and suggestions.

Journal of the American Statistical Association
December 2009, Vol. 104, No. 488, Theory and Methods
DOI: 10.1198/jasa.2009.tm08302

which conditional response distributions change nonparametrically with predictors.

There have been a number of recent articles considering variable selection and hypothesis testing in models with Dirichlet process (DP) components. Dahl and Newton (2007) and MacLehose et al. (2007) independently developed methods that use a DP to cluster predictor effects. Kim, Tadesse, and Vannucci (2006) proposed to use a DPM model for selecting classifying variables in a multivariate response while clustering subjects based on the selected variables. Basu and Chib (2003) proposed a general MCMC algorithm for calculating Bayes factors for comparing DPMs.

None of these methods consider the general problem of selecting predictors to include in a flexible model for the conditional distribution of a response variable. Our proposed approach allows the quantiles of the response distribution to change differentially with predictors, while accommodating local and global variable selection and hypothesis testing. This is useful both when interest focuses on assessing the effects of predictors, and when one wants to build a flexible but parsimonious model for prediction. Section 2 proposes the PSBP and Section 3 discusses the PSBPM for the conditional distribution modeling with variable selection. Section 4 develops an MCMC sampling SSVS algorithm for the PSBPM. Sections 5 and 6 include a simulation study and an epidemiological application, respectively. Section 7 concludes with discussion.

2. THE PROBIT STICK–BREAKING PROCESS

Consider an uncountable collection of predictor-dependent random distributions, $P_{\mathcal{X}} = \{P_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, where $\mathcal{X}$ is the sample space for the predictors $\mathbf{x} = (x_1, \ldots, x_p)'$. In this section, we propose a new choice of prior $\mathcal{P}_{\mathcal{X}}$ for $P_{\mathcal{X}}$ which is deemed the PSBP. Using $P_{\mathbf{x}}$ directly as a conditional response distribution is unappealing, because $P_{\mathbf{x}}$ is almost surely discrete under the PSBP prior, as shown below. We instead use $P_{\mathbf{x}}$ as a predictor-dependent mixture distribution for the parameters in a normal linear regression model. This induces a flexible nonparametric mixture model for the conditional response density $f(y|\mathbf{x})$. This model will be described in detail in Section 3 after introducing the PSBP formulation.

In order to motivate the PSBP formulation, we start with the stick-breaking representation of the DP as a prior for one random distribution P. According to Sethuraman (1994), if $P \sim \mathrm{DP}(\lambda P_0)$,

$$P = \sum_{h=1}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h}, \qquad \boldsymbol{\theta}_h \sim P_0, \tag{1}$$

where $\pi_h = V_h \prod_{l<h}(1 - V_l)$ is a probability weight that is formulated from a stick-breaking process with $V_h \sim \mathrm{Beta}(1, \lambda)$ for $h = 1, \ldots, \infty$, and $\delta_{\boldsymbol{\theta}}$ is a point mass at $\boldsymbol{\theta}$. Starting with a unit length stick, V_1 is the proportion of the stick broken off and assigned to $\boldsymbol{\theta}_1$, V_2 is the proportion of the remaining $(1 - V_1)$ length stick allocated to $\boldsymbol{\theta}_2$, and so on. The induced prior on P will assign probability one to almost surely discrete distributions having an infinite collection of atoms at random locations. Hence, such a prior is typically not appropriate when P is used as the distribution of the data, but is very useful as a prior for a mixture distribution. There is an extremely rich literature on DP mixture models (Lo 1984; Escobar and West 1995).

To construct a prior for $P_{\mathcal{X}}$, a generalization of (1) can be considered as

$$P_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{\boldsymbol{\theta}_h}, \qquad \boldsymbol{\theta}_h \sim P_0 \text{ for all } \mathbf{x} \in \mathcal{X}, \tag{2}$$

where $\pi_h(\mathbf{x})$ is a predictor-dependent probability weight that is constructed from the following predictor-dependent stick-breaking process

$$\pi_h(\mathbf{x}) = V_h(\mathbf{x}) \prod_{l<h} \{1 - V_l(\mathbf{x})\} \quad \text{for all } \mathbf{x} \in \mathcal{X}, \tag{3}$$

where $V_h \sim Q$ with $V_h : \mathcal{X} \to [0, 1]$ a bounded stochastic process over $\mathcal{X}$. Although different choices of Q have been proposed by Griffin and Steel (2006), Reich and Fuentes (2007), Dunson and Park (2008), and Chung and Dunson (2009), such approaches lead to challenging computation. It is appealing to consider a specification that results in a conjugate structure for the random components in $V_h(\mathbf{x})$, and hence simple posterior computation. Such conjugacy is particularly important in developing an efficient algorithm for variable selection and calculation of posterior inclusion probabilities, which will be discussed in detail in Section 4.

With this motivation, we first modify (1) as follows:

$$\begin{aligned} P &= \sum_{h=1}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h}, \qquad \boldsymbol{\theta}_h \sim P_0, \\ \pi_h &= \Phi(\eta_h) \prod_{l<h} \{1 - \Phi(\eta_l)\}, \end{aligned} \tag{4}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, $\mathrm{N}(0, 1)$ and $\eta_h \sim \mathrm{N}(\mu, 1)$ for $h = 1, \ldots, \infty$. The difference between (1) and (4) is that the stick-breaking random variables in (4) arise through a probit model, $\Phi(\eta_h)$, instead of assuming beta distributions $V_h \sim \mathrm{Beta}(1, \lambda)$. In the special case in which $\lambda = 1$ and $\mu = 0$, the two priors for P are identical, because both stick-breaking random variables are assigned $\mathrm{Uniform}(0, 1)$ prior distributions. In general, the priors will not be identical, but will be similar. In the DP prior, λ controls the precision, with values close to zero favoring stick-breaking variables close to one, which leads to high probability allocated to the first few components. In a similar manner, μ controls the precision in prior (4), with large values leading to high probability allocated to the initial components. By placing hyperpriors on λ or μ, one can allow the data to have a stronger influence on the posterior for the weights on the different components.

In order to generalize (4) to obtain a prior $\mathcal{P}_{\mathcal{X}}$ for a collection of predictor-dependent distributions $P_{\mathcal{X}}$, we let

$$\begin{aligned} P_{\mathbf{x}} &= \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{\boldsymbol{\theta}_h}, \qquad \boldsymbol{\theta}_h \sim P_0, \\ \pi_h(\mathbf{x}) &= \Phi(\eta_h(\mathbf{x})) \prod_{l<h} \big\{1 - \Phi(\eta_l(\mathbf{x}))\big\} \quad \text{for all } \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{5}$$

where $\eta_h(\mathbf{x}) = \alpha_h + f_h(\mathbf{x})$ with $\alpha_h \sim \mathrm{N}(\mu, 1)$ and $f_h : \Re^p \to \Re$ being an unknown regression function of $\mathbf{x}$. We refer to the prior $\mathcal{P}_{\mathcal{X}}$ defined in (5) as a PSBP. Letting $\boldsymbol{\phi}_h$ be the finitely-many parameters characterizing $f_h(\mathbf{x})$ with $\boldsymbol{\phi}_h \sim H$, the shorthand notation $P_{\mathcal{X}} \sim \mathrm{PSBP}(\mu, H, P_0)$ is used to denote that $P_{\mathcal{X}}$ follows

the PSBP with hyperparameters, μ, H, P_0. Lemma 1 demonstrates that the prior is coherent, since the probability weights sum to one almost surely for all possible $\mathbf{x}$. The proof is in the Appendix.

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

The probit stick-breaking weight structure has several advantages compared with other formulations proposed for $\pi_h(\mathbf{x})$. First, using data augmentation, it facilitates easy posterior computation for α_h and $\boldsymbol{\phi}_h$ with carefully chosen functional form $f_h(\mathbf{x})$ and conjugate priors for $\boldsymbol{\phi}_h$ as discussed in Section 4. Second, predictor-dependence can easily be accommodated in $\pi_h(\mathbf{x})$ through $f_h(\mathbf{x})$. Although a variety of choices for $f_h(\mathbf{x})$ are possible (e.g., using splines or Gaussian processes), we focus on the following simple form throughout the paper

$$f_h(\mathbf{x}) = -\sum_{j=1}^{p} \psi_{hj}|x_j - \Gamma_{hj}|, \tag{6}$$

where $\boldsymbol{\phi}_h = \{\psi_{hj}, \Gamma_{hj}\}_{j=1}^{p} \sim H$ and

$$H \equiv \prod_{j=1}^{p} \left\{ \mathrm{N}_+\left(\psi_{hj}; \mu_{\psi_j}, \tau_{\psi_j}^{-1}\right) \times \sum_{m=1}^{M_j} \frac{1}{M_j} \delta_{\Gamma_{jm}^*}(\Gamma_{hj}) \right\}. \tag{7}$$

Here, N_+ denotes a normal distribution truncated below by zero to ensure that $\psi_{hj} \geq 0$ and Γ_{jm}^* for $m = 1, \ldots, M_j$ are discrete points over a reasonable range of the jth predictor. Letting $\boldsymbol{\Gamma}_h = \{\Gamma_{hj}\}_{j=1}^{p}$, we can think of $\boldsymbol{\Gamma}_h$ as random locations scattered over the predictor space. Hence, if the hth location $\boldsymbol{\Gamma}_h$ is far from $\mathbf{x}$, $f_h(\mathbf{x})$ and $\eta_h(\mathbf{x})$ are large negative numbers, so that $\Phi(\eta_h(\mathbf{x}))$ is a positive number close to zero. Because $\Phi(\eta_h(\mathbf{x}))$ is the portion to be taken from the remainder of the unit length stick and assigned to $\pi_h(\mathbf{x})$, small $\Phi(\eta_h(\mathbf{x}))$ leaves a greater portion of the stick for other locations, so $\pi_h(\mathbf{x})$ is small relative to the other $\pi_l(\mathbf{x})$ for $l \neq h$. In addition, by allowing $\boldsymbol{\psi}_h = \{\psi_{hj}\}_{j=1}^{p}$ to vary with h, we accommodate spatially adaptive dependence, with more rapid changes occurring in certain regions of $\mathcal{X}$.

3. CONDITIONAL DISTRIBUTION MODELING WITH VARIABLE SELECTION

3.1 Model Specification

Let y be a univariate continuous response and $\mathbf{x} = (x_1, \ldots, x_p)'$ be a vector of p continuous predictors. We consider the following PSBPM model for $f(y|\mathbf{x})$:

$$\begin{aligned} f(y|\mathbf{x}) &= \int \mathrm{N}(y; \mathbf{x}_0'\boldsymbol{\beta}, \tau^{-1})\, dP_{\mathbf{x}}(\boldsymbol{\beta}, \tau), \\ P_{\mathcal{X}} &= \{P_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\} \sim \mathrm{PSBP}(\mu, H, P_0), \end{aligned} \tag{8}$$

where $\mathbf{x}_0 = (1, \mathbf{x}')'$ is the predictor vector including an intercept and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ is a vector of regression coefficients. Applying the stick-breaking form in (5) with $\boldsymbol{\theta}_h = (\boldsymbol{\beta}_h^*, \tau_h^*)$ and $\boldsymbol{\beta}_h^* = (\beta_{h0}^*, \ldots, \beta_{hp}^*)'$, we obtain

$$f(y|\mathbf{x}) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \mathrm{N}(y; \mathbf{x}_0'\boldsymbol{\beta}_h^*, \tau_h^{*-1}), \tag{9}$$

which is an infinite mixture of normal linear regressions with mixture weights varying with predictors. The finite mixture of linear regression framework has been considered in the neural computing literature under the name of Hierarchical Mixtures of Experts (HME) (Jordan and Jacobs 1994). Some Bayesian work for the finite HME model includes Peng, Jacobs, and Tanner (1996), Jiang and Tanner (1999), and Geweke and Keane (2007). The infinite HME can be obtained using nonparametric Bayesian approaches proposed by Müller, Erkanli, and West (1996), Griffin and Steel (2006, 2007), and Dunson and Park (2008).

In our experience based on simulation studies, the predictor-dependent mixture structure in (9) tends to produce accurate estimates of $f(y|\mathbf{x})$ in regions of the predictor space for which ample data are available. However, as the number of predictors increase and the observations become increasingly sparse, estimation performance [judged in terms of the Kullback–Leibler (KL) divergence from the true density and/or mean integrated square error] tends to diminish. In addition, it is often of primary interest in many applications to conduct local or global variable selection and hypothesis testing to identify important predictors in conditional distribution modeling. This has not been addressed in the literature.

In order to address the curse of dimensionality in estimation and our interest in testing and variable selection, we incorporate a variable selection structure through H and P_0 in (8). Letting γ_{hj} be an inclusion indicator variable for the jth predictor in the hth mixture component, we induce H and P_0 through the following distributions for $\boldsymbol{\phi}_h$ and $\boldsymbol{\theta}_h$:

$$\begin{aligned} \boldsymbol{\phi}_h &= \{\boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h\} \\ &\sim \prod_{j=1}^{p} \left\{ 1(\gamma_{hj} = 0)\delta_0(\psi_{hj}) + 1(\gamma_{hj} = 1)\mathrm{N}_+\left(\psi_{hj}; \mu_{\psi_j}, \tau_{\psi_j}^{-1}\right) \right\} \\ &\quad \times \prod_{j=1}^{p} \left\{ \sum_{m=1}^{M_j} \frac{1}{M_j} \delta_{\Gamma_{jm}^*}(\Gamma_{hj}) \right\}, \\ \boldsymbol{\theta}_h &= (\boldsymbol{\beta}_h^*, \tau_h^*) \\ &\sim \mathrm{N}_{p_{\gamma_h}+1}\left(\boldsymbol{\beta}_{\gamma_h,h}^*; \mathbf{0}, \boldsymbol{\Sigma}_{\gamma_h,h}\right) \times \delta_{\mathbf{0}}\left(\boldsymbol{\beta}_{\bar{\gamma}_h,h}^*\right) \\ &\quad \times \mathrm{Gamma}(\tau_h^*; a_\tau, b_\tau), \end{aligned} \tag{10}$$

where $\boldsymbol{\beta}_{\gamma_h,h}^*$ is the vector of regression coefficients corresponding to $\gamma_{hj} = 1$ including the intercept, $\boldsymbol{\beta}_{\bar{\gamma}_h,h}^*$ is the coefficient vector with $\gamma_{hj} = 0$, and $p_{\gamma_h} = \sum_{j=1}^{p} \gamma_{hj}$. Note that γ_{hj} controls local inclusion of the jth predictor in model (9), with $\gamma_{hj} = 0$ implying that $\psi_{hj} = 0$ and $\beta_{hj}^* = 0$. A value of $\beta_{hj}^* = 0$ implies the jth predictor is assigned a coefficient of zero in the hth linear regression model in (9), while a value of $\psi_{hj} = 0$ leads to excluding the jth predictor from the hth predictor-dependent stick-breaking weight in the expression for $\pi_h(\mathbf{x})$. Clearly, if $\gamma_{hj} = 0$ for $h = 1, \ldots, \infty$, then the jth predictor will be globally excluded from the model. To allow uncertainty in γ_{hj}, we let

$$\gamma_{hj} \sim \mathrm{Bernoulli}(\gamma_{hj}; \kappa_j), \tag{11}$$

where κ_j is the prior probability of $\gamma_{hj} = 1$ for the jth predictor. To borrow information across mixture components, we use the sparseness-favoring prior of Lucas et al. (2006), with

$$\begin{aligned} \kappa_j &\sim 1(w_j = 0)\delta_0(\kappa_j) \\ &\quad + 1(w_j = 1)\,\mathrm{Beta}\left(\kappa_j; a_{\kappa_j}, b_{\kappa_j}\right) \quad \text{for } j = 1, \ldots, p, \\ w_j &\sim \mathrm{Bernoulli}(w_j; 0.5), \end{aligned} \tag{12}$$

which modifies the typical beta hyper-prior to allow exclusion of a predictor from all the mixture components.

In Bayes variable selection, it is important to choose the prior distributions for the coefficients within each model carefully. In variable selection for normal linear regression, Zellner's g-prior (Zellner 1986) is widely used, with mixtures of g-priors (Liang et al. 2008) providing a clear improvement. These priors can be used directly for the coefficients in each mixture component as follows:

$$\begin{aligned} \boldsymbol{\beta}^*_{\gamma_h,h}|\tau^*_h &\sim \mathrm{N}\big(\boldsymbol{\beta}^*_{\gamma_h,h}; \mathbf{0}, \boldsymbol{\Sigma}_{\gamma_h,h}\big), \\ \boldsymbol{\Sigma}_{\gamma_h,h} &= ng^{-1}\big(\mathbf{X}'_{\gamma_h}\mathbf{X}_{\gamma_h}\big)^{-1}/\tau^*_h \quad \text{with} \\ g &\sim \mathrm{Gamma}(g; a_g, b_g), \end{aligned} \tag{13}$$

where n is the number of subjects and $\mathbf{X}_{\gamma_h}$ is the design matrix corresponding to $\gamma_{hj} = 1$ including the intercept. In addition, it is necessary to choose proper priors for the model-specific parameters, and the scale of these priors relative to the data can have an important impact. The mixture of g-priors used above for the component-specific regression parameters and residual precisions automatically accounts for the measurement scale of the predictors. In parametric regression models, the residual precision can be assigned an improper prior to avoid sensitivity to the measurement scale of y. However, as we allow the residual precision τ^*_h to vary across mixture components, an improper prior can lead to an artifactual tendency to allocate all individuals to a single component with high probability. To bypass this problem, we follow a common practice in Bayesian mixture modeling and normalize y prior to analysis, so that default values of a_τ, b_τ can be recommended without considering the measurement scale of the data. We also normalize the predictors to facilitate selection of a default choice for the prespecified grid $\{\Gamma^*_{jm}\}_{m=1}^{M_j}$ and hyperparameters $\mu_{\psi_j}, \tau_{\psi_j}$. These choices will be described in Section 4.2.

3.2 Hypothesis Formulation

We first consider a global null hypothesis for selecting important predictors. As discussed with the variable selection structure in (10), one can consider a global point null hypothesis for exclusion of the jth predictor as $H_{0j}: \gamma_{hj} = 0$ for $h = 1, \ldots, \infty$. However, considering such H_{0j} seems overly restrictive because the weights $\pi_h(\mathbf{x})$ in (9) tend to decrease towards zero rapidly as h increases, suggesting that the mixture components of higher order than some moderate number N may not be practically important for modeling $f(y|\mathbf{x})$. In addition, the infiniteness in H_{0j} makes the calculation of prior and posterior probabilities for the null hypotheses infeasible. If one can determine a finite number N such that $\sum_{N+1}^{\infty} \pi_h(\mathbf{x}) \approx 0$, one may focus on the mixture components of lower order than N for inference.

One possible strategy is to base hypothesis testing only on the subset of components that are occupied by subjects in the sample, and hence have posterior distributions that differ from their priors. This results in an empirical Bayes-type approach in which the data inform about the complexity of the null hypothesis. In particular, we formalize the null hypothesis of no effect of the jth predictor as follows:

$$H^N_{0j}: \gamma_{hj} = 0 \quad \text{for } h = 1, \ldots, N, \tag{14}$$

where N is a finite number large enough so that the posterior distributions of $\gamma_{hj}|\kappa_j$ for $h > N$ are not different from the prior distributions of $\gamma_{hj}|\kappa_j$. In order to find such an N, we note that the PSBPM in (8) implies the following hierarchical model for the data

$$\begin{aligned} y_i|S_i, P_{\mathcal{X}} &\sim \mathrm{N}\big(y_i; \mathbf{x}'_{i0}\boldsymbol{\beta}^*_{S_i}, \tau^{*-1}_{S_i}\big), \\ S_i|P_{\mathcal{X}} &\sim \sum_{h=1}^{\infty} \pi_h(\mathbf{x}_i)\delta_h(S_i), \\ P_{\mathcal{X}} &= \{P_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\} \sim \mathrm{PSBP}(\mu, H, P_0), \end{aligned} \tag{15}$$

where y_i is the ith subject's response and S_i is a latent variable such that $S_i = h$ denotes that the ith subject is assigned to the hth mixture component. Given (y_i, S_i) for $i = 1, \ldots, n$, we obtain $N = \max_{i=1}^n(S_i)$ for which the following theorem holds. The proof is in the Appendix.

Theorem 1. Suppose $y_i|\mathbf{x}_i \sim f(y|\mathbf{x})$ and $f(y|\mathbf{x})$ follows the PSBPM model in (8) with H and P_0 chosen as in (10) and (11). Let $l(\mathbf{y}, \mathbf{S}|H_{0j})$ and $l(\mathbf{y}, \mathbf{S}|H^N_{0j})$ be the marginal likelihoods for $(\mathbf{y}, \mathbf{S})$ under H_{0j} and H^N_{0j} where $\mathbf{y} = (y_1, \ldots, y_n)'$ and $\mathbf{S} = (S_1, \ldots, S_n)'$. Then, the ratio $R = \frac{l(\mathbf{y},\mathbf{S}|H_{0j})}{l(\mathbf{y},\mathbf{S}|H^N_{0j})}$ does not depend on $(\mathbf{y}, \mathbf{S})$.

Theorem 1 implies that the complete data $(\mathbf{y}, \mathbf{S})$ contain no information to distinguish between H_{0j} and H^N_{0j}, so the prior and posterior distributions for $\gamma_{hj}|\kappa_j$ for $h > N$ become the same. Hence, inferences based on higher-order null hypotheses than H^N_{0j} may be unreliable being overly sensitive to the choice of prior. This sensitivity to the prior may result in lack of consistency in hypothesis testing, and other unappealing properties. Basing hypothesis tests in nonparametric models on finitely many parameters is also appealing from a practical perspective, since calculation of posterior probabilities and Bayes factors becomes feasible.

Next, we consider local hypothesis testing for the predictors identified as important by testing H^N_{0j}. Because it is not straightforward how to use γ_{hj} for local null hypothesis formulation, we rely on the model-averaged conditional distribution estimates at different predictor points. For the jth predictor, one may consider testing if the conditional distributions are different between x_j and x'_j adjusted for the other predictors at fixed values $\mathbf{x}^*_{(j)} = (x^*_1, \ldots, x^*_{j-1}, x^*_{j+1}, \ldots, x^*_p)'$. Letting $d(x_j, x'_j)|_{\mathbf{x}^*_{(j)}} = \sup_{y\in\Re} |F(y|x_j, \mathbf{x}^*_{(j)}) - F(y|x'_j, \mathbf{x}^*_{(j)})|$, we propose a local interval null hypothesis as

$$H_{0j}\big(x_j, x'_j|\mathbf{x}^*_{(j)}\big) : d(x_j, x'_j)|_{\mathbf{x}^*_{(j)}} < \epsilon, \tag{16}$$

where ϵ is a small positive constant. This null implies the total variation distance between the conditional distributions at x_j and x'_j adjusted for the other predictors is negligible. Prior or posterior probabilities can be calculated by specifying a fine grid of values for y wide enough to cover the minimum and maximum of y_i. Using (16), we can further consider the local null hypothesis of equality of the conditional distributions across a region $A_j \subset \mathcal{X}_j$ with $\mathcal{X}_j$ the sample space for the jth predictor, as

$$H_{0j}\big(A_j|\mathbf{x}^*_{(j)}\big) : \sup_{x_j, x'_j \in A_j} \big\{d(x_j, x'_j)|_{\mathbf{x}^*_{(j)}}\big\} < \epsilon. \tag{17}$$

This implies that the total variation distance between the conditional distributions at any two points in A_j adjusted for the other predictors is negligible. Considering that the PSBPM characterizes the conditional distributions very flexibly, hypothesis testing for (16) and (17) would be sensitive to the choice of $\mathbf{x}^*_{(j)}$, in particular, when the jth predictor interacts with any of the other predictors. Given the flexibility of the model, inferences on the interactions among predictors are not trivial and can be further research topics.

4. POSTERIOR COMPUTATION

4.1 Model and MCMC Algorithm

We develop an MCMC algorithm for the PSBPM following the specification in (8) with H and P_0 chosen as in (10) through (13). Additionally, we assume $\mu \sim \mathrm{N}(\mu; \mu_\mu, \tau_\mu^{-1})$. In order to sample a finite number of random components for $P_{\mathbf{x}}$, we rely on a modification of the blocked Gibbs sampler (Ishwaran and James 2001) with a truncation level T.

The updating steps are in the Appendix. Note that all full conditionals are very straightforward. In step 1, S_i is sampled from a multinomial. For updating the weight components, $\alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h$, we use a data augmentation approach. For $S_i = h$, we introduce $Z_{il} = 0$ for $l = 1, \ldots, S_i - 1$ and $Z_{il} = 1$ for $l = S_i$ where

$$\begin{aligned} Z_{il} &= 1(Z^*_{il} > 0), \\ Z^*_{il} &\sim \mathrm{N}\left(Z^*_{il}; \alpha_h - \sum_{j=1}^{p} \psi_{hj}|x_{ij} - \Gamma_{hj}|, 1\right). \end{aligned} \tag{18}$$

For $S_i = T$, we introduce Z^*_{il} only for $l = 1, \ldots, T-1$ because we let $\Phi(\eta_T(\mathbf{x})) = 1$ so that $\sum_{h=1}^{T} \pi_h(\mathbf{x}) = 1$. Given Z^*_{il}, we update $\alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h$ from their conjugate full conditionals (steps 2–4). The atoms, $\boldsymbol{\beta}^*_h, \tau^*_h$, and other hyperparameters are also updated from their conjugate full conditionals (steps 5–10). Finally, we update γ_{hj} based on the likelihoods for $(\mathbf{y}, \mathbf{S})$ [equivalently for $(\mathbf{y}, \mathbf{Z}^*)$ where $\mathbf{Z}^* = \{Z^*_{il}, l = 1, \ldots, S_i\}_{i=1}^n$] with ψ_{hj} and β^*_{hj} marginalized out (step 11). Such marginalization is facilitated in particular by the conjugacy obtained for ψ_{hj} by the probit weight structure of the PSBP. Note that this step generalizes the SSVS step for linear regression (George and McCulloch 1997).

4.2 Default Choices for Hyperparameters

As motivated in Section 3.1, we recommend standardizing the data prior to analysis. Assuming standardization, we recommend the following default choices for the hyperparameters. For H, $\mu_{\psi_j} = 0$, $\tau_{\psi_j} = 100$, and Γ^*_{jm} are 50 equally spaced grid points in $(-3.5, 3.5)$ for all j. For P_0, $a_g = b_g = 0.5$ and $a_\tau = b_\tau = 0.5$. For others, $a_{\kappa_j} = b_{\kappa_j} = 0.5$ for all j and $\mu_\mu = 0$, $\tau_\mu = 1$. We let $\epsilon = 0.1$ in defining local null hypotheses as this implies negligible local changes in the conditional densities under the null in simulations. For truncation, we let $T = 20$ which tends to be large enough in most applications to obtain good performance in variable selection and conditional distribution estimation, though for large datasets and complex problems a higher truncation level may be needed to obtain an accurate approximation to the true infinite-dimensional PSBPM model. We have found good performance for these choices of hyperparameter values in a wide variety of simulation studies, a subset of which will be presented in the next section. It is important to acknowledge that results are not entirely robust to hyperparameter choice in that high variance priors can lead one to overly favor the null hypothesis corresponding to exclusion of all the candidate predictors. This is a well-known issue in Bayesian methods for model and variable selection, and is by no means unique to the nonparametric mixture models considered here. Refer, for example, to Liang et al. (2008) for a recent review of default priors for parametric variable selection.

5. SIMULATION STUDY

In order to illustrate the proposed method and to assess the performance, we conduct a simulation study. Predictors are generated as $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})' \overset{\text{iid}}{\sim} \mathrm{N}_p(\mathbf{0}, \mathbf{C})$ for $i = 1, \ldots, n$ where $\mathbf{C}$ follows a covariance structure with homogeneous marginal variance $\sigma^2_{\mathbf{x}}$ and the pairwise correlation $\mathrm{corr}(x_{ij}, x_{ij'}) = \rho^{|j-j'|}$. We consider $n = 1000$, $p = 10$, $\sigma^2_{\mathbf{x}} = 1$, and $\rho = 0.1, 0.5, 0.9$. The response is generated for a null case (1) and two alternative cases (2) and (3).

$$\begin{aligned} &(1) \quad y_i \overset{\text{iid}}{\sim} 0.5\mathrm{N}(y_i; 1, 1) + 0.5\mathrm{N}(y_i; -1, 0.5^2), \\ &(2) \quad y_i \overset{\text{iid}}{\sim} \mathrm{N}(y_i; \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2), \\ &(3) \quad y_i \overset{\text{iid}}{\sim} 0.5\mathrm{N}\big(y_i; 10, \big(5 + 20|\min(x_{i1}, 0)|\big)^2\big) \\ &\qquad\qquad + 0.5\mathrm{N}(y_i; -10 + 6x_{i3}, 5^2). \end{aligned}$$

Case (1) is a mixture of two normals with no change in $f(y|\mathbf{x})$ across $\mathbf{x}$. Case (2) is a standard normal linear regression where $f(y|\mathbf{x})$ changes only in mean as $\mathbf{x}$ changes. In particular, we consider $\boldsymbol{\beta} = (1, -1, 0, 1, 0, 0, 0, 0, 0, 0)'$ with 3 different values for σ such that the signal-to-noise ratio (SNR) $= \frac{\boldsymbol{\beta}'\mathbf{C}\boldsymbol{\beta}}{\sigma^2}$ is about 0.3333, 1, 3 (equivalently, the theoretical model $R^2 = \frac{\boldsymbol{\beta}'\mathbf{C}\boldsymbol{\beta}}{\boldsymbol{\beta}'\mathbf{C}\boldsymbol{\beta}+\sigma^2}$ is about 0.25, 0.5, 0.75). Case (3) is a mixture of two normals where the variance for the 1st mixture component decreases monotonically as x_1 increases and the location for the 2nd component shifts to the right as x_3 increases. In particular, x_1 has a local impact only when $x_1 < 0$ having no effect on $E(y|\mathbf{x})$ while x_3 has a global impact on $E(y|\mathbf{x})$.

5.1 Simple Application of PSBPM

After standardizing y and $\mathbf{x}$, we applied the PSBPM with the priors and hyperparameters discussed in Sections 3 and 4 to all cases. The MCMC algorithm described in Section 4.1 was run for 10,000 iterations, with the first 5000 iterations discarded as a burn-in. The MCMC chain appeared to converge rapidly and to mix efficiently based on the trace plots.

In case (1), the estimated marginal inclusion probability for the jth predictor ($\widehat{P}_j$), with $P_j = 1 - \Pr(H^N_{0j}|\text{data})$, was ≤ 0.05 for all j, suggesting that none of the predictors are important. In case (2), $\widehat{P}_j = 1$ for $j = 1, 2, 4$ and ≤ 0.07 for the other j, implying that PSBPM correctly selects important predictors in a simple normal linear regression case. The true conditional response density $f(y|\mathbf{x}^*)$, with $\mathbf{x}^*$ various predictor points, was almost the same as the predictive density $\hat{f}(y|\mathbf{x}^*)$, with its 95% credible intervals very narrow and enclosing the true density in both cases (figures not shown).

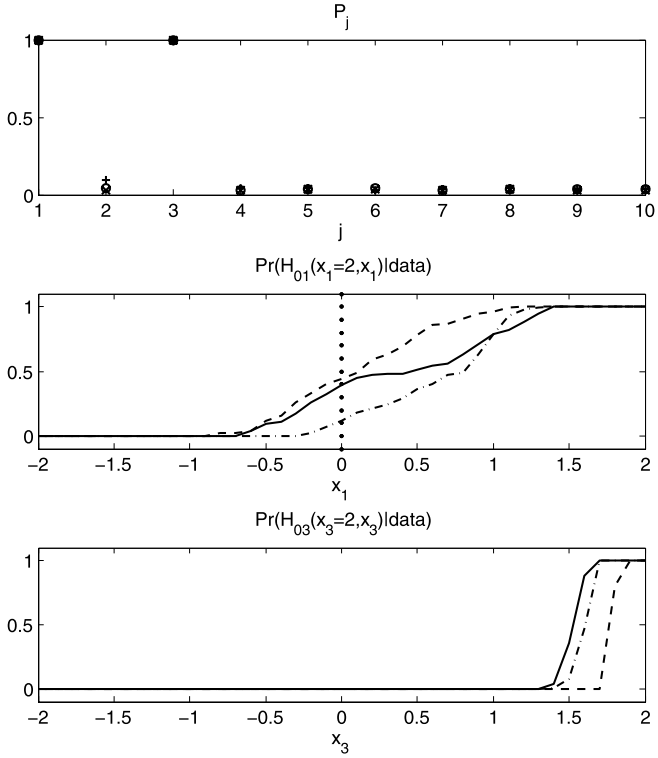

Figure 1. Case (3) with $\rho = 0.1$ ('x' or solid), 0.5 ('o' or dash-dotted), 0.9 ('+' or dashed): Top—$\widehat{P}_j$ for $j = 1, \ldots, 10$; middle—$\widehat{\Pr}(H_{01}(x_1 = 2, x_1)|\text{data})$ with x_1 varying across 40 grid points; bottom—$\widehat{\Pr}(H_{03}(x_3 = 2, x_3)|\text{data})$ with x_3 varying across 40 grid points.

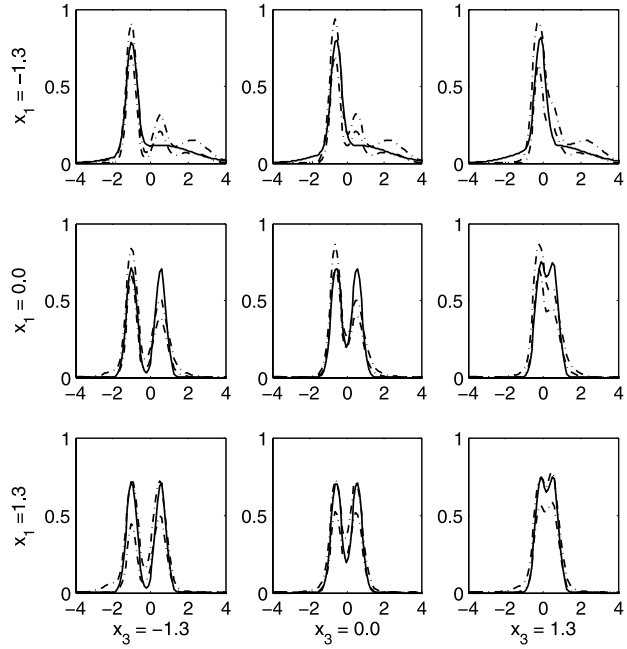

Figure 2. Case (3) with $\rho = 0.5$: Predictive (dashed) conditional response density $\hat{f}(y|\mathbf{x}^*)$ with 95% credible intervals (dash-dotted) at $\mathbf{x}^* = (x_1, \bar{x}_2, x_3, \ldots, \bar{x}_{10})$ with x_1 and x_3 varying among 5th, 50th, 95th empirical percentiles; true density $f(y|\mathbf{x}^*)$ (solid).

In case (3), Figure 1 shows that $\widehat{P}_j = 1$ for $j = 1, 3$ and ≤ 0.05 for the other j. The PSBPM correctly identified x_1 and x_3 as important for the change in $f(y|\mathbf{x})$ although x_1 is only locally important having no impact on $E(y|\mathbf{x})$. Figure 1 also shows that $\widehat{\Pr}(H_{01}(x_1 = 2, x_1)|\text{data})$ is 0 for $x_1 < 0$ and increases towards 1 for $x_1 > 0$ reflecting the local impact of x_1. In addition, $\widehat{\Pr}(H_{03}(x_3 = 2, x_3)|\text{data})$ is 0 across x_3 because x_3 is globally important. Figure 2 shows that $\hat{f}(y|\mathbf{x}^*)$ (dashed) with its 95% credible intervals (dash-dotted) closely follows $f(y|\mathbf{x}^*)$ (solid) reflecting the shape change across x_1 and x_3. Figure 2 is for $\rho = 0.5$ and the results were almost the same for $\rho = 0.1, 0.9$.

Letting $\widehat{P}_j \geq 0.5$ correspond to significance evidence for rejecting H_{0j}^N against the alternative, we obtained $\geq$98% power and $\leq$6% Type I error rate based on the 50 replicated datasets in all cases except that the powers for x_1 and x_2 were 94% and 92% in case (2) with SNR $= 1/3$ and $\rho = 0.9$ (refer to Tables 2 and 3 in Section 5.2), suggesting that the approach seems to have good frequentist operating characteristics. Using another significance cutoff $\widehat{P}_j \geq 0.9$ did not change the results as expected with enough sample size $n = 1000$ except for the case (2) with SNR $= 1/3$ and $\rho = 0.9$. In addition, we conducted a sensitivity analysis for different hyperparameter choices in PSBP(μ, H, P_0). (i) μ-related: $\mu_\mu = 0, 3$, $\tau_\mu = 1, 10$; (ii) H-related: $\mu_{\psi_j} = 0, 0.1$, $\tau_{\psi_j} = 1, 10, 100$; (iii) P_0-related: $a_g = b_g = 0.5, 1$, $a_\tau = b_\tau = 0.5, 1$. We obtained almost identical results with these choices to the default choice of hyperparameters in all cases.

In order to evaluate scalability to larger numbers of candidate predictors, we applied PSBPM for all the cases with $p = 10, 15, 20$ and $n = 800, 1000, 1200$. The results were similar to $n = 1000$ and $p = 10$ in all cases, implying that PSBPM is robust to moderate shifts from $(n, p) = (1000, 10)$ combination. Finally, we implemented PSBPM for a relatively small sample size $n = 200$ with $p = 10$. In case (1), we obtained almost identical results to $n = 1000$. In case (2), results were also similar to $n = 1000$ except that the power loss for x_1 and x_2 was greater than in $n = 1000$ in case (2) with SNR $= 1/3$ and $\rho = 0.9$ (refer to Tables 5 and 6 in Section 5.2). In case (3), the powers for x_1 were reduced to 78%, 70%, 56% for $\rho = 0.1, 0.5, 0.9$, respectively, with the significance cutoff $\widehat{P}_j \geq 0.5$ while the power for x_3 and Type I error rate remained similar (refer to Table 6).

5.2 Comparison With Other Methods

We compared PSBPM with a simple method and three competing methods for all cases. As a simple approach, we considered a standard linear regression with SSVS (George and McCulloch 1997) (LR-SSVS) with the prior for the regression coefficients consistent with (10)–(13). As competitors, we considered Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch 2008), Adaptive Lasso Linear Regression (ALLR) (Zou 2006), and the Bayesian Treed Gaussian process (BTGP) (Gramacy and Lee 2008). BART is a recently proposed flexible mean regression model which allows for informal variable selection through a partial dependence plot. ALLR is a linear regression with a regularization technique where L1 penalties are weighted differently for each coefficient and variable selection is done through shrinkage towards zero.

We considered two variants of ALLR: (1) ALLR1—the approach proposed in Zou (2006); (2) ALLR2—a two-stage procedure in which the squared residuals from the ALLR1 fit are used as the response in a second application of ALLR1, with

the selected predictors consisting of the union of the predictors selected in the two stages. The first stage is designed to select predictors of the mean in a homoscedastic linear model, while the second stage is designed to select predictors of the variance. The BART and ALLR models assume homoscedasticity, while BTGP accommodates heteroscedasticity, but is not designed for variable selection.

Each of these competitors to PSBPM is focused on flexible mean regression, and such approaches are not designed to detect changes in skewness, modality or shape of the conditional response distribution with predictors. However, we were unable to find a more flexible competitor. Implementing BART and BTGP using the R statistical package, we used the default setting for priors and hyperparameters. For ALLR, we used 5-fold cross-validation following the guidelines in Zou (2006) for the choice of tuning parameters.

To assess mean prediction performance, we generate a dataset with a sample size n and calculate root mean squared error (RMSE) $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{E}(y_i|\mathbf{x}_i) - E(y_i|\mathbf{x}_i))^2}$ for an additional n test samples. Then, we repeated this for 50 replicate datasets and obtain mean RMSE (mRMSE) and standard error (s.e.). We applied the 2nd step in ALLR2 to the fit of ALLR1 so we report RMSE only once from ALLR1. For variable selection, we calculated the percentage of replicates for which the jth predictor is selected as significant (PS_j). We used both $\widehat{P}_j \geq 0.5$ and $\widehat{P}_j \geq 0.9$ as cutoffs for significance in LR-SSVS and PSBPM, while significance in ALLR corresponds to nonzero regression coefficient estimates. Because the partial dependence plot from BART is subjective to interpret and does not provide an automated approach to variable selection and BTGP does not do variable selection, we do not report PS_j for BART and BTGP. Results for $n = 1000$ and $n = 200$ are summarized in Tables 1–3 and Tables 4–6, respectively.

For mean prediction, refer to Tables 1 and 4. In case (1), PSBPM and ALLR performed the best leading to the smallest mRMSE. In cases (2) and (3), PSBPM, LR-SSVS, and ALLR performed comparably except that PSBPM was slightly poorer for the case (2) with SNR $= 1/3$ and $\rho = 0.9$. In all cases, BART and BTGP yielded relatively large mRMSE. Even though the PSBPM method allows flexible estimation of not only $\text{E}(y|\mathbf{x})$ but also the entire conditional distribution $f(y|\mathbf{x})$, it still has comparable overall performance in out-of-sample prediction of the mean to the best of the mean regression methods we considered.

For variable selection, refer to Tables 2 and 3 for $n = 1000$ and Tables 5 and 6 for $n = 200$. Firstly with $n = 1000$ and $\widehat{P}_j \geq 0.5$, PS_j was 0% for all j with all four methods reflecting the truth in case (1) (not reported). In case (2), PS_j was 100% for $j = 1, 2, 4$ and $\leq$6% for all other j with LR-SSVS and PSBPM except that $\text{PS}_1 = 94\%$, $\text{PS}_2 = 92\%$ for PSBPM in case (2) with SNR $= 1/3$ and $\rho = 0.9$, while ALLR1 and ALLR2 yielded $\text{PS}_j = 100\%$ for $j = 1, 2, 4$, and 10%–30% for all other j. In case (3), PSBPM had $\text{PS}_j \geq 98\%$ for $j = 1, 3$ and $\leq$6% for all other j while ALLR2 had $\text{PS}_j \geq 94\%$ for $j = 1, 3$ and 6%–34% for all other j. LR-SSVS and ALLR1 had $\text{PS}_j \geq 80\%$ for $j = 3$ and $\text{PS}_j \leq 24\%$ for all other j, failing to detect x_1 as an important predictor. Using $\widehat{P}_j \geq 0.9$ as a cutoff did not change PS_j for

Table 1. Summary of out-of-sample mean prediction performance for simulated data with $n = 1000$ in terms of mRMSE (s.e. in parentheses)

			mRMSE (s.e.)		
Case	SNR	Method	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
(1)	NA	PSBPM	0.03 (0.018)	0.02 (0.019)	0.03 (0.019)
		LR-SSVS	0.04 (0.021)	0.03 (0.022)	0.04 (0.022)
		ALLR	0.03 (0.020)	0.02 (0.020)	0.03 (0.021)
		BART	0.17 (0.022)	0.17 (0.019)	0.15 (0.018)
		BTGP	0.10 (0.023)	0.10 (0.026)	0.11 (0.042)
(2)	1/3	PSBPM	0.07 (0.028)	0.08 (0.033)	0.11 (0.054)
		LR-SSVS	0.06 (0.022)	0.06 (0.017)	0.06 (0.023)
		ALLR	0.06 (0.023)	0.06 (0.020)	0.07 (0.021)
		BART	0.20 (0.019)	0.20 (0.016)	0.20 (0.019)
		BTGP	0.13 (0.020)	0.15 (0.027)	0.17 (0.033)
	1	PSBPM	0.06 (0.024)	0.06 (0.025)	0.08 (0.038)
		LR-SSVS	0.06 (0.023)	0.06 (0.022)	0.06 (0.020)
		ALLR	0.06 (0.024)	0.06 (0.023)	0.06 (0.023)
		BART	0.19 (0.016)	0.20 (0.016)	0.21 (0.017)
		BTGP	0.13 (0.016)	0.15 (0.030)	0.16 (0.033)
	3	PSBPM	0.05 (0.026)	0.06 (0.024)	0.06 (0.025)
		LR-SSVS	0.05 (0.026)	0.06 (0.024)	0.05 (0.019)
		ALLR	0.06 (0.025)	0.06 (0.023)	0.05 (0.018)
		BART	0.17 (0.018)	0.18 (0.013)	0.19 (0.012)
		BTGP	0.15 (0.024)	0.19 (0.031)	0.22 (0.032)
(3)	NA	PSBPM	0.09 (0.037)	0.08 (0.031)	0.09 (0.040)
		LR-SSVS	0.05 (0.023)	0.05 (0.026)	0.07 (0.034)
		ALLR	0.05 (0.024)	0.05 (0.025)	0.07 (0.030)
		BART	0.28 (0.043)	0.29 (0.058)	0.31 (0.045)
		BTGP	0.22 (0.062)	0.28 (0.188)	0.37 (0.245)

Table 2. Summary of variable selection performance for simulated data with $n = 1000$ in terms of PS_j (PS_j = % of simulated studies where jth predictor was judged significant)

			PS_j (%)									
Case	ρ	Method	$j = 1^*$	2^*	3	4^*	5	6	7	8	9	10
(2)	0.1	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
SNR = 1/3		LR-SSVS	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	2 [0]	4 [0]
		ALLR1	100	100	20	100	20	8	22	12	16	22
		ALLR2	100	100	20	100	20	8	22	12	16	22
	0.5	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	100 [100]	100 [100]	2 [0]	100 [100]	2 [0]	0 [0]	6 [0]	2 [0]	0 [0]	0 [0]
		ALLR1	100	100	24	100	30	16	16	18	14	20
		ALLR2	100	100	24	100	30	16	16	18	14	20
	0.9	PSBPM	94 [82]	92 [70]	2 [2]	100 [100]	2 [2]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	100 [100]	100 [100]	4 [0]	100 [100]	0 [0]	2 [0]	2 [0]	2 [0]	2 [0]	0 [0]
		ALLR1	100	100	24	100	22	22	28	14	24	18
		ALLR2	100	100	24	100	22	22	28	14	24	18
			$j = 1^*$	2^*	3	4^*	5	6	7	8	9	10
(2)	0.1	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
SNR = 1		LR-SSVS	100 [100]	100 [100]	2 [0]	100 [100]	2 [0]	0 [0]	2 [0]	0 [0]	0 [0]	0 [0]
		ALLR1	100	100	20	100	18	10	20	22	16	20
		ALLR2	100	100	20	100	18	10	20	22	16	20
	0.5	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	100 [100]	100 [100]	4 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		ALLR1	100	100	12	100	28	14	18	22	16	18
		ALLR2	100	100	12	100	28	14	18	22	16	18
	0.9	PSBPM	100 [100]	100 [100]	6 [4]	100 [100]	4 [2]	0 [0]	0 [0]	2 [2]	0 [0]	0 [0]
		LR-SSVS	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	2 [0]
		ALLR1	100	100	12	100	24	20	10	14	30	20
		ALLR2	100	100	12	100	24	20	10	14	30	20

NOTE: For PSBPM and LR-SSVS, numbers not in [ ] are PS_j from using $\widehat{P}_j > 0.5$ and numbers in [ ] are from using $\widehat{P}_j > 0.9$. Predictors marked with * are the ones that are significant in truth.

PSBPM except for the power loss in case (2) with SNR = 1/3 and $\rho = 0.9$. For LR-SSVS, it increased Type I error rate and reduced power particularly in case (3). Hence, PSBPM maintains a high power and low Type I error rate in almost all cases, whereas LR-SSVS does not detect x_1 in case (3). The ALLR procedures tend to produce an unacceptably high Type I error rate except when none of the predictors is important [case (1)]. The over-selecting tendency of ALLR with tuning parameters chosen via five-fold cross validation is known (Wang, Li, and Tsai 2007; Zou, Hastie, and Tibshirani 2007), and more recent recommendations are to use BIC for tuning parameter selection. The overall conclusions were similar with $n = 200$ in cases (1) and (2) as shown in Tables 5 and 6. In case (3), powers were reduced in all methods and refer to the details in Table 6.

6. EPIDEMIOLOGICAL APPLICATION

6.1 Motivation and Background

In epidemiological studies for diabetes, interest can be on characterizing the relationship between glucose tolerance (GT) and insulin sensitivity (IS) and other diabetes risk factors. GT is measured by 2-hour plasma glucose level (mg/dl) in the oral glucose tolerance test and indicates how fast glucose is cleared from the blood. GT is also used to diagnose type 2 diabetes using <140 (normal), [140, 200] (prediabetes), and >200 (diabetes). IS provides an indicator of how well the body responds to insulin, a hormone regulating movement of glucose from the blood to body cells. Although it is well known that low IS is related to poor GT (high 2-hour plasma glucose level), previous studies have either categorized IS and GT prior to analysis or focused on linear associations. These approaches discard information and can yield misleading inferences. Biologically, one anticipates changes in the shape of the 2-hour glucose distribution with changes in IS and other risk factors for diabetes, such as age, blood pressures, or obesity measures.

Data were obtained from the Insulin Resistance Atherosclerosis Study (IRAS) (Wagenknecht et al. 1995), which was a prospective study designed to assess the relationships among IS and cardiovascular disease risk factors in a large multiethnic cohort. Figure 3 plots 2-hour plasma glucose level against IS, age, waist-to-hip ratio (WTH), body mass index (BMI), diastolic blood pressure (DBP), and systolic blood pressure (SBP). Examining the data, one notes a large right skew in the glucose distribution, with the distributional shape changing with IS. The changes of the glucose distribution with BMI may be local, while the other predictors may have negligible impact on the glucose distribution. As linear or nonlinear mean or median regression models are not supported for these data, our goal is to apply the proposed method that allows the distribution of 2-hour glucose to change flexibly with the different risk factors

Table 3. Summary of variable selection performance for simulated data with $n = 1000$ in terms of PS_j (PS_j = % of simulated studies where jth predictor was judged significant)

			PS_j (%)									
Case	ρ	Method	$j = 1^*$	2^*	3	4^*	5	6	7	8	9	10
(2)	0.1	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
SNR = 3		LR-SSVS	100 [100]	100 [100]	2 [0]	100 [100]	4 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		ALLR1	100	100	14	100	16	6	16	18	18	16
		ALLR2	100	100	14	100	16	6	16	18	18	16
	0.5	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	100 [100]	100 [100]	4 [0]	100 [100]	0 [0]	0 [0]	0 [0]	2 [0]	0 [0]	2 [0]
		ALLR1	100	100	20	100	14	24	14	24	32	26
		ALLR2	100	100	20	100	14	24	14	24	32	26
	0.9	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	100 [100]	100 [100]	0 [0]	100 [100]	2 [0]	0 [0]	2 [0]	2 [0]	0 [0]	0 [0]
		ALLR1	100	100	6	100	22	18	30	18	16	16
		ALLR2	100	100	6	100	22	18	30	18	16	16
			$j = 1^*$	2	3^*	4	5	6	7	8	9	10
(3)	0.1	PSBPM	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	2 [2]	0 [0]	0 [0]	0 [0]
		LR-SSVS	4 [0]	0 [0]	100 [98]	2 [0]	2 [0]	0 [0]	0 [0]	2 [0]	0 [0]	0 [0]
		ALLR1	24	8	100	12	10	10	8	16	6	6
		ALLR2	100	12	100	16	12	10	10	20	6	8
	0.5	PSBPM	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	4 [0]	4 [0]	98 [96]	4 [0]	6 [0]	0 [0]	2 [0]	2 [0]	0 [0]	2 [0]
		ALLR1	20	10	100	10	10	14	12	20	12	12
		ALLR2	100	14	100	14	12	18	18	20	14	18
	0.9	PSBPM	98 [98]	6 [6]	100 [100]	0 [0]	0 [0]	0 [0]	2 [2]	0 [0]	0 [0]	0 [0]
		LR-SSVS	8 [2]	6 [0]	80 [38]	14 [2]	4 [0]	8 [0]	6 [0]	4 [2]	4 [2]	4 [2]
		ALLR1	20	16	88	22	16	24	20	14	20	20
		ALLR2	100	26	94	28	20	34	24	24	26	28

NOTE: For PSBPM and LR-SSVS, numbers not in [ ] are PS_j from using $\widehat{P}_j > 0.5$ and numbers in [ ] are from using $\widehat{P}_j > 0.9$. Predictors marked with * are the ones that are significant in truth.

under study, while also allowing risk factors to drop out of the model and to have effects that are local to particular regions of the predictor space.

6.2 Analysis

We analyzed the IRAS study data focusing on the relationship between 2-hour glucose level and 6 predictors shown in Figure 3. For $i = 1, \ldots, 868$, y_i = 2-hour glucose level (mg/dl), x_{i1} = IS, x_{i2} = age, x_{i3} = WTH, x_{i4} = BMI, x_{i5} = DBP, and x_{i6} = SBP. Prior to the analysis, natural log transformation was used for IS [$\log(\text{IS} + 1)$] (Laws et al. 1997) and we standardized both response and predictors. Firstly, we applied the PSBPM with the default choice of hyperparameters and obtained $\widehat{P}_j = 1, 1, 0.03, 0.02, 0.03, 0.03$ implying that IS and age are important factors for the change of glucose distribution. In order to examine how IS and age affect the 2-hour glucose distribution, we obtained predictive density $\hat{f}(y|\mathbf{x}^*)$ at $\mathbf{x}^* = (x_1, x_2, \bar{x}_3, \ldots, \bar{x}_{10})$ with x_1 and x_2 varying among 5th, 50th, 95th empirical percentiles. Figure 4 shows that the glucose density has a very heavy right tail for low IS (x_1) but, as IS increases, the right tail disappears making the mode become higher. In fact, the right tail seems to characterize the group of people whose 2-hour glucose level is above 200 mg/dl (reference line is 0.2 with standardization). This implies that there may be underlying genetic factors or unadjusted risk factors that can explain such heavy right tail shape of 2-hour glucose level for the people with low IS other than the predictors included in the current model. In addition, the right tail becomes heavier as age (x_2) increases especially for those subjects with low IS, meaning that aging is also related to poor GT. Local hypothesis testing for IS and age adjusting for the other predictors showed that both IS and age globally affects the glucose distribution with no interaction between IS and aging. All of these results were not sensitive to different hyperparameter choices mentioned in simulations.

In order to compare the results with LR-SSVS and ALLR, we created 20 independent train/test splits by randomly selecting 4/5 of the data as a training set and the remaining 1/5 as a test set. Based on each training set, each method was used to predict the corresponding test set and we calculated mRMSE and mean correlation (mCORR) between y_i and $\hat{E}(y_i|\mathbf{x}_i)$ with their s.e. For variable selection, PS_j ($\widehat{P}_j \geq 0.9$ for significance) was obtained based on the fit of training set. Table 7 shows that mRMSE and mCORR are similar among the three methods whereas PS_j were very inconsistent. However, the residual plots from LR-SSVS and ALLR showed that the constant normal residual assumption is strongly violated so we suspect that their results may not be reliable.

Based on a train/test split, we compared the three methods based on the estimated 95% predictive intervals for the test

Table 4. Summary of out-of-sample mean prediction performance for simulated data with $n = 200$ in terms of mRMSE (s.e. in parentheses)

			mRMSE (s.e.)		
Case	SNR	Method	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
(1)	NA	PSBPM	0.06 (0.063)	0.07 (0.066)	0.06 (0.042)
		LR-SSVS	0.08 (0.054)	0.11 (0.069)	0.14 (0.066)
		ALLR	0.05 (0.047)	0.07 (0.062)	0.05 (0.041)
		BART	0.22 (0.039)	0.22 (0.042)	0.18 (0.036)
		BTGP	0.21 (0.053)	0.24 (0.126)	0.49 (0.284)
(2)	1/3	PSBPM	0.15 (0.065)	0.16 (0.095)	0.20 (0.065)
		LR-SSVS	0.15 (0.049)	0.15 (0.059)	0.18 (0.060)
		ALLR	0.15 (0.051)	0.15 (0.059)	0.17 (0.059)
		BART	0.29 (0.033)	0.30 (0.041)	0.29 (0.038)
		BTGP	0.27 (0.037)	0.26 (0.050)	0.35 (0.125)
	1	PSBPM	0.13 (0.050)	0.14 (0.053)	0.14 (0.059)
		LR-SSVS	0.14 (0.050)	0.14 (0.047)	0.14 (0.052)
		ALLR	0.13 (0.048)	0.14 (0.050)	0.15 (0.051)
		BART	0.30 (0.041)	0.33 (0.044)	0.32 (0.037)
		BTGP	0.32 (0.050)	0.31 (0.060)	0.37 (0.111)
	3	PSBPM	0.13 (0.049)	0.13 (0.058)	0.14 (0.060)
		LR-SSVS	0.13 (0.048)	0.13 (0.056)	0.13 (0.058)
		ALLR	0.13 (0.047)	0.13 (0.056)	0.13 (0.057)
		BART	0.28 (0.046)	0.30 (0.045)	0.32 (0.042)
		BTGP	0.36 (0.065)	0.35 (0.066)	0.38 (0.089)
(3)	NA	PSBPM	0.10 (0.039)	0.10 (0.052)	0.11 (0.048)
		LR-SSVS	0.15 (0.060)	0.15 (0.051)	0.17 (0.047)
		ALLR	0.12 (0.060)	0.13 (0.057)	0.14 (0.055)
		BART	0.33 (0.043)	0.35 (0.066)	0.33 (0.073)
		BTGP	0.39 (0.344)	0.42 (0.196)	1.19 (1.111)

set. Figure 5 shows that the predictive intervals from PSBPM become larger as the observed y increases while the intervals from LR-SSVS or ALLR are relatively larger across the observed y. The percentage of the test data points falling inside the 95% predictive interval was 96%, 96%, 95% while the average width of the predictive intervals was 2.75, 3.29, 3.14 for PSBPM, LR-SSVS, ALLR, respectively, suggesting that the PSBPM results in narrower predictive intervals while maintaining the same coverage with LR-SSVS and ALLR.

7. DISCUSSION

We propose a nonparametric Bayesian approach for conditional distribution modeling with variable selection. We first introduce the PSBP as a new choice of prior for an uncountable collection of predictor-dependent random distributions and consider a PSBPM of normal linear regressions, resulting in an infinite mixture with mixing weights varying with predictors. Incorporating variable selection structure in both regression coefficients and mixing weights, we allow predictors to drop out of the model or to be included in the model such that local or global effects for the conditional distribution change can be assessed.

The proposed method is innovative in that it deals with variable selection and local and global hypothesis testing problems in the general setting of conditional distribution modeling. The method should be useful in many applications where interest is not only on the conditional mean response but also on the overall shape or tails of the conditional response distribution, in particular, when the response distribution changes in shape not following standard parametric assumptions across the predictor space. In the present paper, we only illustrated continuous predictor cases but we note that the method can easily be generalized to incorporate categorical predictors (results not shown).

Although the PSBPM performed well in various simulation studies, there is much room to improve because of the model complexity. First, it would not be feasible to implement the method if too many candidate predictors are considered or to obtain reliable results if only small samples are available. In addition, there is a need for the development of efficient approaches for formal hypothesis testing of interactions and for identifying local regions of high-dimensional predictor spaces across which response distributions change. Finally, although the finite approximation for infinite mixture using a truncation level $T = 20$ performed well for the cases considered in this paper, one may avoid such approximation using the exact blocked Gibbs sampler of Papaspiliopoulos and Roberts (2008).

APPENDIX

Proof of Lemma 1

Following the proof of Lemma 1 for the KSBP (Dunson and Park 2008), $\sum_{h=1}^{\infty} \pi_h(\mathbf{x}) = 1$ a.s. iff $\sum_{h=1}^{\infty} \log\{1 - \Phi(\eta_h(\mathbf{x}))\} = -\infty$ a.s. Also, $\sum_{h=1}^{\infty} \log\{1 - \Phi(\eta_h(\mathbf{x}))\} = -\infty$ iff $\sum_{h=1}^{\infty} E[\log\{1 - \Phi(\eta_h(\mathbf{x}))\}] = -\infty$. Because $\log\{1 - \Phi(\eta_h(\mathbf{x}))\} \leq 0$, the condition is satisfied.

Table 5. Summary of variable selection performance for simulated data with $n = 200$ in terms of PS_j (PS_j = % of simulated studies where jth predictor was judged significant)

			PS_j (%)									
Case	ρ	Method	$j = 1^*$	2^*	3	4^*	5	6	7	8	9	10
(2)	0.1	PSBPM	100 [90]	96 [82]	0 [0]	98 [96]	0 [0]	0 [0]	0 [0]	0 [0]	4 [0]	0 [0]
SNR = 1/3		LR-SSVS	100 [100]	100 [94]	4 [0]	100 [98]	4 [0]	10 [0]	0 [0]	4 [0]	12 [2]	4 [0]
		ALLR1	100	100	12	100	18	22	8	14	30	20
		ALLR2	100	100	12	100	18	22	8	14	30	20
	0.5	PSBPM	98 [88]	100 [92]	2 [0]	98 [96]	2 [0]	0 [0]	2 [0]	0 [0]	2 [0]	0 [0]
		LR-SSVS	100 [98]	100 [98]	2 [0]	100 [98]	8 [0]	2 [0]	2 [0]	2 [0]	6 [0]	6 [0]
		ALLR1	100	100	20	100	18	18	30	28	20	16
		ALLR2	100	100	20	100	18	18	30	28	20	16
	0.9	PSBPM	60 [44]	46 [28]	4 [2]	86 [66]	4 [2]	6 [0]	0 [0]	2 [0]	0 [0]	0 [0]
		LR-SSVS	82 [52]	76 [36]	4 [2]	90 [64]	8 [2]	6 [2]	4 [0]	4 [2]	6 [0]	6 [0]
		ALLR1	98	94	30	92	30	38	22	36	44	34
		ALLR2	98	94	30	92	30	38	22	36	44	34
			$j = 1^*$	2^*	3	4^*	5	6	7	8	9	10
(2)	0.1	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	2 [0]	0 [0]	0 [0]	2 [0]	0 [0]	2 [0]
SNR = 1		LR-SSVS	100 [100]	100 [100]	8 [0]	100 [100]	6 [2]	8 [0]	4 [0]	6 [0]	4 [0]	8 [0]
		ALLR1	100	100	18	100	12	24	20	24	14	16
		ALLR2	100	100	18	100	12	24	20	24	14	16
	0.5	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	100 [100]	100 [100]	4 [0]	100 [100]	0 [0]	0 [0]	10 [2]	4 [0]	2 [0]	2 [0]
		ALLR1	100	100	18	100	16	20	28	20	16	22
		ALLR2	100	100	18	100	16	20	28	20	16	22
	0.9	PSBPM	100 [98]	100 [96]	2 [0]	100 [100]	0 [0]	2 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	100 [98]	100 [98]	4 [0]	100 [100]	2 [0]	4 [0]	0 [0]	0 [0]	2 [0]	2 [0]
		ALLR1	100	100	30	100	32	22	16	20	26	34
		ALLR2	100	100	30	100	32	22	16	20	26	34

NOTE: For PSBPM and LR-SSVS, numbers not in [ ] are PS_j from using $\widehat{P}_j > 0.5$ and numbers in [ ] are from using $\widehat{P}_j > 0.9$. Predictors marked with * are the ones that are significant in truth.

Proof of Theorem 1

Let $\boldsymbol{\theta}_h = (\boldsymbol{\beta}_h^*, \tau_h^*)$, $\boldsymbol{\xi}_h = (\alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h)$, and $\boldsymbol{\gamma}_h = \{\gamma_{hj}\}_{j=1}^p$. Also, let $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_h\}_{h=1}^\infty$, $\boldsymbol{\Xi} = \{\boldsymbol{\xi}_h\}_{h=1}^\infty$, and $\boldsymbol{\Lambda} = \{\boldsymbol{\gamma}_h\}_{h=1}^\infty$. Given $\boldsymbol{\Lambda}$, the marginal likelihood for $(\mathbf{y},\mathbf{S})$ is

$$l(\mathbf{y}, \mathbf{S}|\boldsymbol{\Lambda}) = \int \prod_{i=1}^n (y_i|\mathbf{x}_i, \boldsymbol{\theta}_{S_i}) \prod_{h=1}^\infty (\boldsymbol{\theta}_h|\boldsymbol{\gamma}_h)\, d\boldsymbol{\Theta} \times \int \prod_{i=1}^n (S_i|\mathbf{x}_i, \boldsymbol{\Xi}) \prod_{h=1}^\infty (\boldsymbol{\xi}_h|\boldsymbol{\gamma}_h)\, d\boldsymbol{\Xi}. \quad (19)$$

Because $S_i \le N$, we reexpress (19) as

$$\begin{aligned} l(\mathbf{y}, \mathbf{S}|\boldsymbol{\Lambda}) &= \int \prod_{i=1}^n (y_i|\mathbf{x}_i, \boldsymbol{\theta}_{S_i}) \prod_{h=1}^N (\boldsymbol{\theta}_h|\boldsymbol{\gamma}_h)\, d\boldsymbol{\Theta}^N \int \prod_{h>N} (\boldsymbol{\theta}_h|\boldsymbol{\gamma}_h)\, d\boldsymbol{\Theta}_+^N \\ &\quad \times \int \prod_{i=1}^n (S_i|\mathbf{x}_i, \boldsymbol{\Xi}^N) \prod_{h=1}^N (\boldsymbol{\xi}_h|\boldsymbol{\gamma}_h)\, d\boldsymbol{\Xi}^N \\ &\quad \times \int \prod_{h>N} (\boldsymbol{\xi}_h|\boldsymbol{\gamma}_h)\, d\boldsymbol{\Xi}_+^N \\ &= \int \prod_{i=1}^n (y_i|\mathbf{x}_i, \boldsymbol{\theta}_{S_i}) \prod_{h=1}^N (\boldsymbol{\theta}_h|\boldsymbol{\gamma}_h)\, d\boldsymbol{\Theta}^N \\ &\quad \times \int \prod_{i=1}^n (S_i|\mathbf{x}_i, \boldsymbol{\Xi}^N) \prod_{h=1}^N (\boldsymbol{\xi}_h|\boldsymbol{\gamma}_h)\, d\boldsymbol{\Xi}^N \\ &= l(\mathbf{y}, \mathbf{S}|\boldsymbol{\Lambda}^N), \end{aligned}$$

where $\boldsymbol{\Theta}^N = \{\boldsymbol{\theta}_h\}_{h=1}^N$, $\boldsymbol{\Theta}_+^N = \{\boldsymbol{\theta}_h\}_{h>N}$, $\boldsymbol{\Xi}^N = \{\boldsymbol{\xi}_h\}_{h=1}^N$, $\boldsymbol{\Xi}_+^N = \{\boldsymbol{\xi}_h\}_{h>N}$, $\boldsymbol{\Lambda}^N = \{\boldsymbol{\gamma}_h\}_{h=1}^N$, and $\boldsymbol{\Lambda}_+^N = \{\boldsymbol{\gamma}_h\}_{h>N}$. Then,

$$\begin{aligned} R = \frac{l(\mathbf{y}, \mathbf{S}|H_{0j})}{l(\mathbf{y}, \mathbf{S}|H_{0j}^N)} &= \frac{\int l(\mathbf{y}, \mathbf{S}|\boldsymbol{\Lambda})(\boldsymbol{\Lambda}|H_{0j})\, d\boldsymbol{\Lambda}}{\int l(\mathbf{y}, \mathbf{S}|\boldsymbol{\Lambda})(\boldsymbol{\Lambda}|H_{0j}^N)\, d\boldsymbol{\Lambda}} \\ &= \frac{\int l(\mathbf{y}, \mathbf{S}|\boldsymbol{\Lambda}^N)(\boldsymbol{\Lambda}^N|H_{0j})\, d\boldsymbol{\Lambda}^N \times \int (\boldsymbol{\Lambda}_+^N|H_{0j})\, d\boldsymbol{\Lambda}_+^N}{\int l(\mathbf{y}, \mathbf{S}|\boldsymbol{\Lambda}^N)(\boldsymbol{\Lambda}^N|H_{0j}^N)\, d\boldsymbol{\Lambda}^N \times \int (\boldsymbol{\Lambda}_+^N|H_{0j}^N)\, d\boldsymbol{\Lambda}_+^N} \\ &= \frac{\int l(\mathbf{y}, \mathbf{S}|\boldsymbol{\Lambda}^N)(\boldsymbol{\Lambda}^N|H_{0j})\, d\boldsymbol{\Lambda}^N}{\int l(\mathbf{y}, \mathbf{S}|\boldsymbol{\Lambda}^N)(\boldsymbol{\Lambda}^N|H_{0j}^N)\, d\boldsymbol{\Lambda}^N} \\ &= 1, \end{aligned}$$

because $(\boldsymbol{\Lambda}^N|H_{0j}) = (\boldsymbol{\Lambda}^N|H_{0j}^N)$. The ratio R does not depend on $(\mathbf{y}, \mathbf{S})$.

MCMC Algorithm

1. Update S_i for $i = 1, \ldots, n$: With $\pi_h(\mathbf{x}_i) = \Phi(\eta_h(\mathbf{x}_i)) \prod_{l<h}(1 - \Phi(\eta_l(\mathbf{x}_i)))$,

$$\Pr(S_i = h) = \frac{\pi_h(\mathbf{x}_i)\mathrm{N}(y_i; \mathbf{x}_{i0}'\boldsymbol{\beta}_h^*, \tau_h^{*-1})}{\sum_{h=1}^T \pi_h(\mathbf{x}_i)\mathrm{N}(y_i; \mathbf{x}_{i0}'\boldsymbol{\beta}_h^*, \tau_h^{*-1})}.$$

Table 6. Summary of variable selection performance for simulated data with $n = 200$ in terms of PS_j (PS_j = % of simulated studies where jth predictor was judged significant)

			PS_j (%)									
Case	ρ	Method	$j = 1^*$	2^*	3	4^*	5	6	7	8	9	10
(2)	0.1	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
SNR = 3		LR-SSVS	100 [100]	100 [100]	2 [0]	100 [100]	0 [0]	0 [0]	4 [0]	2 [0]	2 [0]	2 [0]
		ALLR1	100	100	14	100	18	22	14	10	6	22
		ALLR2	100	100	14	100	18	22	14	10	6	22
	0.5	PSBPM	100 [100]	100 [100]	0 [0]	100 [100]	0 [0]	2 [0]	0 [0]	0 [0]	2 [0]	0 [0]
		LR-SSVS	100 [100]	100 [100]	4 [0]	100 [100]	2 [0]	2 [2]	2 [0]	0 [0]	8 [0]	2 [0]
		ALLR1	100	100	14	100	16	20	14	8	22	14
		ALLR2	100	100	14	100	16	20	14	8	22	14
	0.9	PSBPM	100 [100]	100 [100]	2 [0]	100 [100]	0 [0]	2 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	100 [100]	100 [100]	2 [0]	100 [100]	2 [0]	2 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		ALLR1	100	100	18	100	20	16	24	20	36	8
		ALLR2	100	100	18	100	20	16	24	20	36	8
			$j = 1^*$	2	3^*	4	5	6	7	8	9	10
(3)	0.1	PSBPM	78 [38]	0 [0]	100 [100]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	10 [2]	4 [0]	70 [38]	2 [0]	4 [0]	6 [0]	8 [2]	2 [0]	2 [0]	6 [0]
		ALLR1	20	10	82	6	18	22	16	10	6	14
		ALLR2	98	16	84	12	22	26	28	12	14	16
	0.5	PSBPM	70 [40]	2 [0]	100 [98]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]	0 [0]
		LR-SSVS	16 [0]	8 [2]	70 [30]	8 [2]	6 [0]	2 [0]	12 [2]	2 [0]	8 [0]	2 [0]
		ALLR1	22	18	78	16	16	14	22	6	14	8
		ALLR2	98	30	82	20	16	16	24	6	18	10
	0.9	PSBPM	56 [14]	16 [4]	90 [62]	2 [0]	0 [0]	0 [0]	4 [0]	2 [0]	2 [0]	0 [0]
		LR-SSVS	10 [2]	14 [2]	30 [6]	12 [2]	6 [0]	8 [0]	12 [0]	12 [2]	4 [0]	2 [0]
		ALLR1	22	26	36	18	16	10	18	20	22	20
		ALLR2	70	44	50	30	22	12	24	24	24	20

NOTE: For PSBPM and LR-SSVS, numbers not in [ ] are PS_j from using $\widehat{P}_j > 0.5$ and numbers in [ ] are from using $\widehat{P}_j > 0.9$. Predictors marked with * are the ones that are significant in truth.

2. Update α_h for $h = 1, \ldots, T-1$: With $n_h = \sum_{i=1}^n 1(S_i \geq h)$,

$$\alpha_h \sim \mathrm{N}\Big(\alpha_h; [n_h+1]^{-1}\Big[\sum_{i:S_i \geq h} W_{ih}^* + \mu\Big], [n_h+1]^{-1}\Big),$$

where $W_{ih}^* = Z_{ih}^* + \sum_{j=1}^p \psi_{hj}|x_{ij} - \Gamma_{hj}|$.

3. Update ψ_{hj} for $j = 1, \ldots, p$ and $h = 1, \ldots, T-1$: If $\gamma_{hj} = 0$, $\psi_{hj} = 0$. If $\gamma_{hj} = 1$,

$$\psi_{hj} \sim \mathrm{N}_+\Big(\psi_{hj}; \Big[\tau_{\psi_j} + \sum_{i:S_i \geq h} |x_{ij} - \Gamma_{hj}|^2\Big]^{-1} \times \Big[\tau_{\psi_j}\mu_{\psi_j} + \sum_{i:S_i \geq h} |x_{ij} - \Gamma_{hj}|U_{ih}^*\Big], \Big[\tau_{\psi_j} + \sum_{i:S_i \geq h} |x_{ij} - \Gamma_{hj}|^2\Big]^{-1}\Big),$$

where $U_{ih}^* = \alpha_h - Z_{ih}^* - \sum_{k=1,k\neq j}^p \psi_{hk}|x_{ik} - \Gamma_{hk}|$.

4. Update Γ_{hj} for $j = 1, \ldots, p$ and $h = 1, \ldots, T-1$: If $\gamma_{hj} = 0$, do not update. If $\gamma_{hj} = 1$,

$$\Pr(\Gamma_{hj} = \Gamma_{jm}^*) = \Big(\frac{1}{M_j}\prod_{i:S_i \geq h} \mathrm{N}\Big(Z_{ih}^*; \alpha_h - \sum_{k=1,k\neq p} \psi_{hk}|x_{ik} - \Gamma_{hk}| - \psi_{hj}|x_{ij} - \Gamma_{jm}^*|, 1\Big)\Big) \Big/ \Big(\sum_{m=1}^{M_j} \frac{1}{M_j}\prod_{i:S_i \geq h} \mathrm{N}\Big(Z_{ih}^*; \alpha_h - \sum_{k=1,k\neq p} \psi_{hk}|x_{ik} - \Gamma_{hk}| - \psi_{hj}|x_{ij} - \Gamma_{jm}^*|, 1\Big)\Big).$$

5. Update $\boldsymbol{\beta}_h^*$ for $h = 1, \ldots, T$: With $\boldsymbol{\beta}_h^* = (\boldsymbol{\beta}_{\gamma_h,h}^*, \boldsymbol{\beta}_{\bar{\gamma}_h,h}^*)$, $\boldsymbol{\beta}_{\bar{\gamma}_h,h}^* = \mathbf{0}$.

$$\boldsymbol{\beta}_{\gamma_h,h}^* \sim \mathrm{N}\big(\boldsymbol{\beta}_{\gamma_h,h}^*; \big[\tau_h^*\mathbf{X}_{\gamma_h,h}'\mathbf{X}_{\gamma_h,h} + \boldsymbol{\Sigma}_{\gamma_h,h}^{-1}\big]^{-1}\big[\tau_h^*\mathbf{X}_{\gamma_h,h}'\mathbf{y}_h\big], \big[\tau_h^*\mathbf{X}_{\gamma_h,h}'\mathbf{X}_{\gamma_h,h} + \boldsymbol{\Sigma}_{\gamma_h,h}^{-1}\big]^{-1}\big),$$

where $\mathbf{X}_{\gamma_h,h}$ is the design matrix of the predictors corresponding to $\gamma_{hj} = 1$ and $S_i = h$ and $\mathbf{y}_h$ is the response vector corresponding to $S_i = h$.

6. Update τ_h^* for $h = 1, \ldots, T$: With $k_h = \sum_{i=1}^n 1(S_i = h)$ and $p_{\gamma_h} = \sum_{j=1}^p \gamma_{hj}$,

$$\tau_h^* \sim \mathrm{Gamma}\Big(\tau_h^*; a_\tau + \frac{k_h}{2} + \frac{p_{\gamma_h}+1}{2}, b_\tau + \frac{1}{2}\big(\mathbf{y}_h - \mathbf{X}_{\gamma_h,h}\boldsymbol{\beta}_{\gamma_h,h}^*\big)'\big(\mathbf{y}_h - \mathbf{X}_{\gamma_h,h}\boldsymbol{\beta}_{\gamma_h,h}^*\big) + \frac{g}{2n}\boldsymbol{\beta}_{\gamma_h,h}^{*\prime}\big(\mathbf{X}_{\gamma_h}'\mathbf{X}_{\gamma_h}\big)\boldsymbol{\beta}_{\gamma_h,h}^*\Big).$$

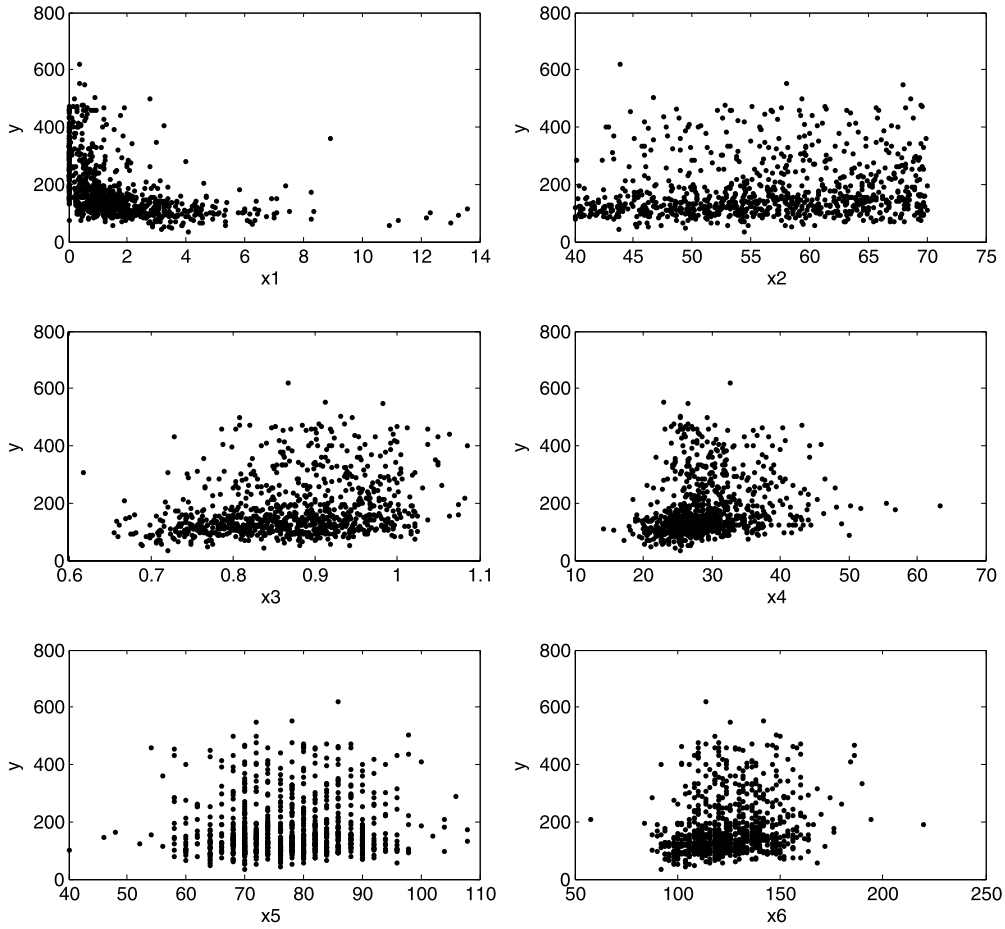

Figure 3. Data from IRAS study: $y =$ 2-hour glucose level (mg/dl); $x_1 =$ insulin sensitivity; $x_2 =$ age; $x_3 =$ waist to hip ratio; $x_4 =$ body mass index; $x_5 =$ diastolic blood pressure; $x_6 =$ systolic blood pressure.

7. Update g:

$$g \sim \text{Gamma}\left(g; a_g + \frac{\sum_{h=1}^{T}(p_{\gamma_h}+1)}{2}, b_g + \sum_{h=1}^{T}\frac{\tau_h^*}{2n}\boldsymbol{\beta}_{\gamma_h,h}^{*\prime}(\mathbf{X}_{\gamma_h}'\mathbf{X}_{\gamma_h})\boldsymbol{\beta}_{\gamma_h,h}^{*}\right).$$

8. Update κ_j for $j = 1, \ldots, p$: If $w_j = 0$, $\kappa_j = 0$. If $w_j = 1$,

$$\kappa_j \sim \text{Beta}\big(a_{\kappa_j} + q_j, b_{\kappa_j} + T - q_j\big) \quad \text{with } q_j = \sum_{h=1}^{T}\gamma_{hj}.$$

9. Update w_j for $j = 1, \ldots, p$: If $\sum_{h=1}^{T}\gamma_{hj} > 0$, $w_j = 1$. If $\sum_{h=1}^{T}\gamma_{hj} = 0$,

$$\Pr(w_j = 1) = \left(\frac{\Gamma(b_{\kappa_j}+T)\Gamma(a_{\kappa_j}+b_{\kappa_j})}{\Gamma(b_{\kappa_j})\Gamma(a_{\kappa_j}+b_{\kappa_j}+T)}\right) \Big/ \left(1 + \frac{\Gamma(b_{\kappa_j}+T)\Gamma(a_{\kappa_j}+b_{\kappa_j})}{\Gamma(b_{\kappa_j})\Gamma(a_{\kappa_j}+b_{\kappa_j}+T)}\right).$$

10. Update μ:

$$\mu \sim \text{N}\left(\mu; [T-1+\tau_\mu]^{-1}\left[\sum_{h=1}^{T-1}\alpha_h + \tau_\mu\mu_\mu\right], [T-1+\tau_\mu]^{-1}\right).$$

11. Update γ_{hj} for $j = 1, \ldots, p$ and $h = 1, \ldots, T$:

$$\Pr(\gamma_{hj} = 1) = \frac{a_{hj}}{a_{hj}+b_{hj}},$$

$$a_{hj} = \kappa_j \times \int \prod_{i:S_i \geq h, S_i \neq T} \text{N}\left(Z_{ih}^*; \alpha_h - \sum_{j=1}^{p}\psi_{hj}|x_{ij} - \Gamma_{hj}|, 1\right) \times \text{N}_+\big(\psi_{hj}; \mu_{\psi_j}, \tau_{\psi_j}^{-1}\big)\, d\psi_{hj} \times \int \prod_{S_i = h} \text{N}(y_i; \mathbf{x}_{i0}'\boldsymbol{\beta}_h^*, \tau_h^{*-1}) \times \text{N}\big(\beta_{hj}^*; \mu_{\beta_j}, \tau_{\beta_j}^{-1}\big)\, d\beta_{hj}^*,$$

$$b_{hj} = (1-\kappa_j) \times \prod_{i:S_i \geq h, S_i \neq T} \text{N}\left(Z_{ih}^*; \alpha_h - \sum_{k=1, k\neq j}^{p}\psi_{hk}|x_{ik} - \Gamma_{hk}|, 1\right) \times \prod_{S_i = h} \text{N}\big(y_i; \mathbf{x}_{(-j)i0}'\boldsymbol{\beta}_{(-j)h}^*, \tau_h^{*-1}\big),$$

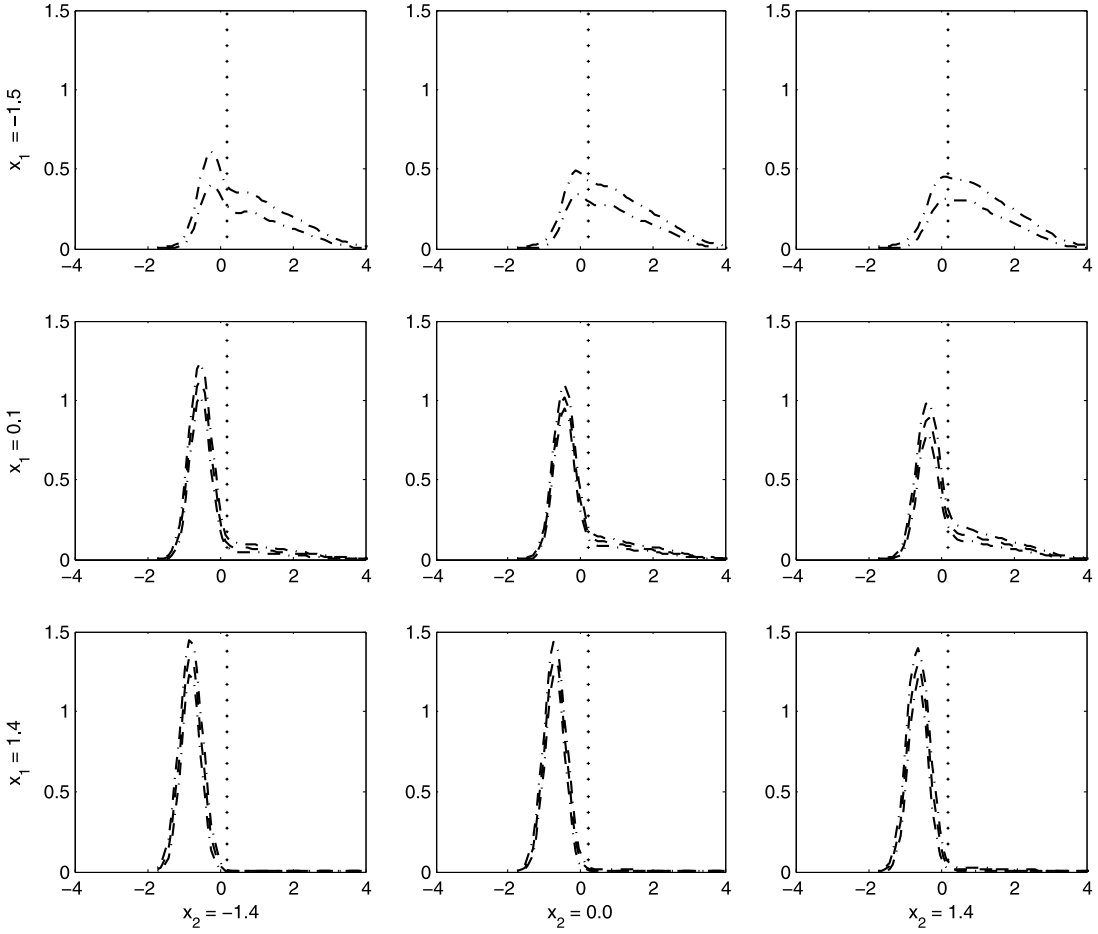

Figure 4. Predictive (dashed) conditional response density $\hat{f}(y|\mathbf{x}^*)$ with 95% credible intervals (dash-dotted) at $\mathbf{x}^* = (x_1, x_2, \bar{x}_3, \ldots, \bar{x}_6)$ with x_1 and x_2 varying among 5th, 50th, 95th empirical percentiles.

where μ_{β_j} and τ_{β_j} in a_{hj} are the conditional mean and precision for β^*_{hj} given $\boldsymbol{\beta}^*_{(-j)\gamma_h,h}$ obtained from $\mathrm{N}_{p_{\gamma_h}}(\boldsymbol{\beta}^*_{\gamma_h,h}; \mathbf{0}, ng^{-1}(\mathbf{X}'_{\gamma_h}\mathbf{X}_{\gamma_h})^{-1}/\tau^*_h)$.

[Received June 2008. Revised June 2009.]

REFERENCES

Basu, S., and Chib, S. (2003), "Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models," *Journal of the American Statistical Association*, 98, 224–235.

Chan, D., Kohn, R., Nott, D., and Kirby, C. (2006), "Locally Adaptive Semiparametric Estimation of the Mean and Variance Functions in Regression Models," *Journal of Computational and Graphical Statistics*, 15, 915–936.

Chipman, H., George, E., and McCulloch, R. (2008), "BART: Bayesian Additive Regression Trees," technical report, University of Chicago.

Chung, Y., and Dunson, D. B. (2009), "The Local Dirichlet Process," *Annals of the Institute of Statistical Mathematics*, to appear.

Dahl, D. B., and Newton, M. A. (2007), "Multiple Hypothesis Testing by Clustering Treatment Effects," *Journal of the American Statistical Association*, 102, 517–526.

Dunson, D. B., and Park, J.-H. (2008), "Kernel Stick-Breaking Process," *Biometrika*, 95, 307–323.

Dunson, D. B., and Peddada, S. D. (2008), "Bayesian Nonparametric Inference on Stochastic Ordering," *Biometrika*, 95, 859–874.

Escobar, M. D. (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Fan, J. Q., and Yim, T. H. (2004), "A Cross Validation Method for Estimating Conditional Densities," *Biometrika*, 91, 819–834.

Table 7. Summary of test-sample mean prediction performance based on mRMSE and mCORR and variable selection performance based on PS_j for IRAS data

			$\mathrm{PS}_j(\%)$					
Method	mRMSE (s.e.)	mCORR (s.e.)	$j=1$	2	3	4	5	6
PSBPM	0.79 (0.060)	0.57 (0.046)	100	70	0	0	0	0
LR-SSVS	0.80 (0.059)	0.56 (0.047)	100	10	10	25	10	0
ALLR1	0.80 (0.059)	0.55 (0.048)	100	60	80	90	50	20
ALLR2	NA	NA	100	65	100	95	50	75

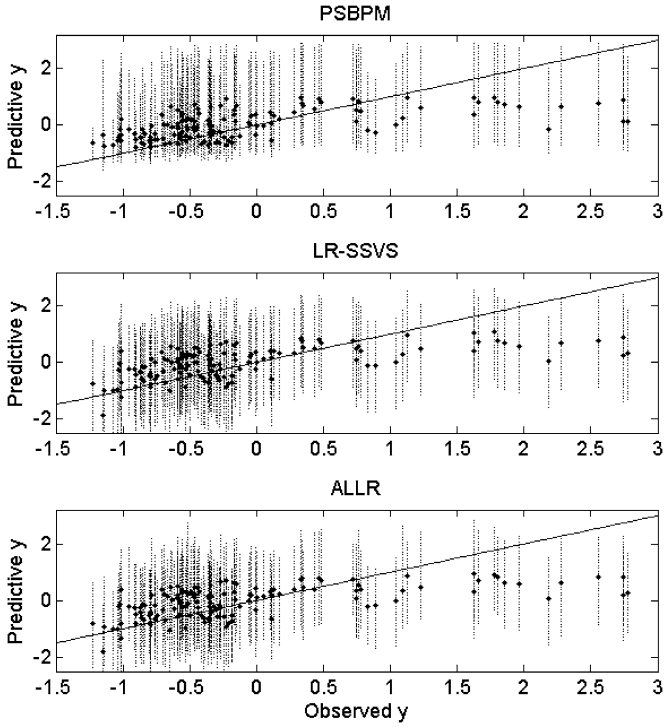

Figure 5. Test dataset predictive performance comparison; y-axis = predictive mean (dots), 95% predictive intervals (vertical lines), reference line (diagonal line); x-axis = observed y; top—PSBPM, middle—LR-SSVS, Bottom—ALLR.

Fan, J. Q., Yao, Q., and Tong, H. (1996), "Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems," *Biometrika*, 83, 189–206.

George, E. I., and Mcculloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Geweke, J., and Keane, M. (2007), "Smoothly Mixing Regressions," *Journal of Econometrics*, 138, 252–290.

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.

Griffin, J. E., and Steel, M. F. J. (2006), "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194.

——— (2007), "Bayesian Nonparametric Modeling With the Dirichlet Process Regression Smoother," Working Paper 07–05, CRiSM.

Hall, P., Wolff, R. C. L., and Yao, Q. W. (1999), "Methods for Estimating a Conditional Distribution Function," *Journal of the American Statistical Association*, 94, 154–163.

Hyndman, R. J., and Yao, Q. (2002), "Nonparametric Estimation and Symmetry Tests for Conditional Density Functions," *Journal of Nonparametric Statistics*, 14, 259–278.

Ishwaran, H., and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Jiang, W., and Tanner, M. A. (1999), "Hierarchical Mixtures of Experts for Exponential Family Regression Models: Approximation and Maximum Likelihood Estimation," *The Annals of Statistics*, 27, 987–1011.

Jordan, M. I., and Jacobs, R. A. (1994), "Hierarchical Mixtures of Experts and the EM Algorithm," *Neural Computation*, 6, 181–214.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006), "Variable Selection in Clustering via Dirichlet Process Mixture Models," *Biometrika*, 93, 877–893.

Laws, A., Hoen, H. M., Selby, J. V., Saad, M. F., Haffner, S. M., and Howard, B. V. (1997), "Differences in Insulin Suppression of Free Fatty Acid Levels by Gender and Glucose Tolerance Status," *Arteriosclerosis, Thrombosis, and Vascular Biology*, 17, 64–71.

Leslie, D. S., Kohn, R., and Nott, D. J. (2007), "A General Approach to Heteroscedastic Linear Regression," *Statistics and Computing*, 17, 131–146.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixture of g Priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423.

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *The Annals of Statistics*, 12, 351–357.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. R., and West, M. (2006), "Sparse Statistical Modeling in Gene Expression Genomics," in *Bayesian Inference for Gene Expression and Proteomics*, Cambridge: Cambridge University Press, pp. 155–176.

MacEachern, S. N. (1999), "Dependent Nonparametric Processes," in *Proceedings of the Bayesian Section*, Alexandria, VA: American Statistical Association, pp. 50–55.

MacLehose, R. F., Dunson, D. B., Herring, A. H., and Hoppin, J. (2007), "Bayesian Methods for Highly Correlated Exposure Data," *Epidemiology*, 18, 199–207.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian Curve Fitting Using Multivariate Normal Mixtures," *Biometrika*, 83, 67–79.

Papaspiliopoulos, O., and Roberts, G. O. (2008), "Retrospective MCMC for Dirichlet Process Hierarchical Models," *Biometrika*, 95, 169–186.

Peng, F., Jacobs, R., and Tanner, M. A. (1996), "Bayesian Inference in Mixtures of Experts and Hierarchical Mixtures of Experts Models With an Application to Speech Recognition," *Journal of the American Statistical Association*, 91, 953–960.

Pennell, M. L., and Dunson, D. B. (2008), "Nonparametric Bayes Testing of Changes in a Response Distribution With an Ordinal Predictor," *Biometrics*, 64, 413–423.

Reich, B. J., and Fuentes, M. (2007), "A Multivariate Semiparametric Bayesian Spatial Modeling Framework for Hurricane Surface Wind Fields," *The Annals of Applied Statistics*, 2, 249–264.

Sethuraman, J. (1994), "A Constructive Definition of the Dirichlet Process Prior," *Statistica Sinica*, 2, 639–650.

Wagenknecht, L. E., Mayer, E. J., Rewers, M. et al. (1995), "The Insulin Resistance Atherosclerosis Study (IRAS) Objectives, Design, and Recruitment Results," *Annals of Epidemiology*, 5, 464–472.

Wang, H., Li, R., and Tsai, C.-L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–568.

Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions," in *Bayesian Inference and Decision Techniques: Essay in Honor of Bruno de Finetti*, Amsterdam: North-Holland, pp. 233–243.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007), "On Degrees of Freedom of the Lasso," *The Annals of Statistics*, 35, 2173–2192.