

Nonparametric Bayes Conditional Distribution Modeling with Variable Selection

Yeonseung Chung¹ and David B. Dunson¹

¹*Biostatistics Branch*

MD A3-03, National Institute of Environmental Health Sciences

P.O. Box 12233, Research Triangle Park, NC 27709

E-mail: chungy@email.unc.edu

Summary. This article considers 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 with local regions and globally. We first introduce the probit stick-breaking process (PSBP) as a prior for an uncountable collection of predictor-dependent random probability measures 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; Kernel stick-breaking process; Mixture of experts; Hypothesis testing; Stochastic search variable selection.

1. Introduction

This article focuses on flexible modeling of the conditional density of a response variable Y given multiple predictors $\mathbf{X} = (X_1, \dots, X_p)'$. 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 et al. (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 et al. (1999) and Hyndman and Yao (2002) among others. Müller et al. (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 et al. (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 et al. (2007b), and Dunson and Park (2008).

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 (2007) 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) and Wang and Dunson (2007) developed methods for estimating and testing of stochastically ordered distributions.

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 probability measures. 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 non-linear mean and variance regression (Chan et al., 2006; Leslie, Kohn and Nott, 2007) to settings in 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 DP components. Dahl and Newton (2007) and Dunson et al. (2007a) independently developed methods that use a DP to cluster predictor effects. Kim et al.

(2006) proposed to use a DPM model for selecting classifying variables in a multivariate response while clustering subjects based on the selected variables. Cai and Dunson (2007) used a weighted convolution of DPs as a prior for a random effects distribution, while allowing selection of random effects. 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 considers basic properties. Section 3 discusses the PSBPM for the conditional distribution modeling with variable selection. Section 4 develops an MCMC sampling SSVS algorithm for the PSBPM. Section 5 and 6 include a simulation study and an epidemiological application, respectively. Section 7 concludes with discussion.

2. The Probit Stick-Breaking Process

2.1 Formulation

Consider an uncountable collection of predictor-dependent random probability measures, $\mathcal{P}_{\mathcal{X}} = \{P_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, where \mathcal{X} is the sample space for the predictors $\mathbf{x} = (x_1, \dots, x_p)'$. The random measures $P_{\mathbf{x}}$ are defined on $(\Omega, \mathcal{B}(\Omega))$ where Ω is a complete and separable metric space and $\mathcal{B}(\mathcal{A})$ denotes a Borel σ -algebra of subsets of \mathcal{A} . Let \mathcal{Q} be a probability measure on $(\mathcal{M}, \mathcal{N})$ where \mathcal{M} is the space of $\mathcal{P}_{\mathcal{X}}$ and \mathcal{N} is a corresponding σ -algebra of subsets of \mathcal{M} . We propose a new choice of \mathcal{Q} deemed the probit stick-breaking process (PSBP).

To induce \mathcal{Q} , we start with a stick-breaking formulation for each $P_{\mathbf{x}}$ as:

$$P_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{\boldsymbol{\theta}_h}, \quad \forall \mathbf{x} \in \mathcal{X}, \quad (1)$$

where $\pi_h(\mathbf{x})$ is a probability weight on the h th component and $\delta_{\boldsymbol{\theta}}$ is a probability measure with all its mass at $\boldsymbol{\theta}$. We assume $\boldsymbol{\theta}_h \sim P_0$ where P_0 is a probability measure on $(\Omega, \mathcal{B}(\Omega))$ which $P_{\mathbf{x}}$ is defined on. In order to induce a prior for $\pi_h(\mathbf{x})$, independently from $\boldsymbol{\theta}_h$, we introduce the following countable sequences of mutually independent random components:

$$\alpha_h \sim N(\mu, 1), \quad \boldsymbol{\psi}_h = \{\psi_{hj}\}_{j=1}^p \sim G, \quad \boldsymbol{\Gamma}_h = \{\Gamma_{hj}\}_{j=1}^p \sim H, \quad (2)$$

where G and H are distributions over a measurable Polish spaces $(\mathcal{L}_\psi, \mathcal{B}(\mathcal{L}_\psi))$ and $(\mathcal{L}_\Gamma, \mathcal{B}(\mathcal{L}_\Gamma))$, respectively. Using $\alpha_h, \boldsymbol{\psi}_h$, and $\boldsymbol{\Gamma}_h$, we form the probability weights $\pi_h(\mathbf{x})$ as:

$$\begin{aligned} \pi_h(\mathbf{x}) &= \Phi(\eta_h(\mathbf{x})) \prod_{l < h} \left\{ 1 - \Phi(\eta_l(\mathbf{x})) \right\} \\ \text{with } \eta_h(\mathbf{x}) &= \alpha_h - \sum_{j=1}^p \psi_{hj} |x_j - \Gamma_{hj}|, \quad \forall \mathbf{x} \in \mathcal{X} \end{aligned} \quad (3)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, $N(0, 1)$. Then, we obtain the following lemma. Proof is in Appendix.

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

By Lemma 1, $P_{\mathbf{x}}$ in (1) is a well defined probability measure on (Ω, \mathcal{B}) for all $\mathbf{x} \in \mathcal{X}$ and the formulation from (1) through (3) defines a prior \mathcal{Q} for $\mathcal{P}_{\mathcal{X}}$ deemed the probit stick-breaking process (PSBP). The shorthand notation $\mathcal{P}_{\mathcal{X}} \sim PSBP(\mu, G, H, P_0)$ is used to denote that $\mathcal{P}_{\mathcal{X}}$ follows the PSBP with hyperparameters, μ, G, H, P_0 .

In order to motivate the formulation, we first discuss a special case where $G = \delta_{\mathbf{0}_p}$ and $\mathbf{0}_p$ is $p \times 1$ vector of zeros. In this case, $\eta_h(\mathbf{x}) = \alpha_h$ and $\pi_h(\mathbf{x}) = \Phi(\alpha_h) \prod_{l < h} (1 - \Phi(\alpha_l))$ for all $\mathbf{x} \in \mathcal{X}$. Because $\pi_h(\mathbf{x})$ does not depend on \mathbf{x} , we obtain

$$P_{\mathbf{x}} = P = \sum_{h=1}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h} \quad \text{with} \quad \pi_h = \Phi(\alpha_h) \prod_{l < h} (1 - \Phi(\alpha_l)), \quad \forall \mathbf{x} \in \mathcal{X} \quad (4)$$

Note that P in (4) is quite similar to the stick-breaking representation of the DP(λP_0) (Sethuraman, 1994) where $\pi_h = V_h \prod_{l < h} (1 - V_l)$ with $V_h \sim \text{Beta}(1, \lambda)$. As $\lambda > 0$ controls the precision in the DP with small values favoring allocating most of the probability to the

first few components, $\mu \in \mathfrak{R}$ in the PSBP controls precision with large values assigning high probability to the first few components.

Although the PSBP special case in (4) and the DP are very closely related, the PSBP has considerable advantages in generalizations to accommodate predictor-dependence in the stick-breaking weights as in (3). Given \mathbf{x} , each $\pi_h(\mathbf{x})$ is linked through the index h to each location Γ_h . If h th location Γ_h is far from \mathbf{x} , $\eta_h(\mathbf{x})$ is a large negative number, 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 for $\pi_h(\mathbf{x})$, small $\Phi(\eta_h(\mathbf{x}))$ leaves more portion of the stick for other locations to take and $\pi_h(\mathbf{x})$ is small relative to the other $\pi_l(\mathbf{x})$ for $l \neq h$. In addition, by allowing ψ_h to vary with h , we accommodate spatially-adaptive dependence, with more rapid changes occurring in certain regions of \mathcal{X} .

Current generalizations of the DP to incorporate predictor-dependence in $\pi_h(\mathbf{x})$, including the π DDP (Griffin and Steel, 2007) and the KSBP (Dunson and Park, 2008), have more complicated structure than (3) and the updating algorithm for the random components in $\pi_h(\mathbf{x})$ is not straightforward. However, the probit-based weight structure in (3) allows for using a data augmentation approach in order to obtain conjugacy so that the random components $\alpha_h, \psi_h, \Gamma_h$ are more efficiently updated as discussed in section 4. Achieving conjugacy is particularly important in developing an efficient algorithm for variable selection and calculation of posterior model probabilities.

2.2 Moments

We first consider the moments of $P_{\mathbf{x}}$ conditionally on $\alpha_h, \psi_h, \Gamma_h$, but marginalizing out the atoms θ_h over P_0 . For all $B \in \mathcal{B}(\Omega)$, the first and second moments are

$$\begin{aligned} E\{P_{\mathbf{x}}(B)|\alpha_h, \psi_h, \Gamma_h\} &= \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) E\{\delta_{\theta_h}(B)\} = P_0(B) \\ E\{P_{\mathbf{x}}(B)^2|\alpha_h, \psi_h, \Gamma_h\} &= \left[\sum_{h=1}^{\infty} \pi_h(\mathbf{x})^2 E\{\delta_{\theta_h}(B)^2\} \right] \end{aligned}$$

$$\begin{aligned}
& + \left[\sum_{h=1}^{\infty} \sum_{l \neq h} \pi_h(\mathbf{x}) \pi_l(\mathbf{x}) E\{\delta_{\boldsymbol{\theta}_h}(B)\} E\{\delta_{\boldsymbol{\theta}_l}(B)\} \right] \\
& = \sum_{h=1}^{\infty} \pi_h(\mathbf{x})^2 \left[E\{\delta_{\boldsymbol{\theta}_h}(B)^2\} - P_0(B)^2 \right] + P_0(B)^2 \\
& = \|\pi_h(\mathbf{x})\|^2 \{P_0(B) - P_0(B)^2\} + P_0(B)^2 \\
& = \|\pi_h(\mathbf{x})\|^2 P_0(B) + \{1 - \|\pi_h(\mathbf{x})\|^2\} P_0(B)^2 \tag{5}
\end{aligned}$$

Also, the correlation is

$$\begin{aligned}
\text{Corr}\{P_{\mathbf{x}}(B), P_{\mathbf{x}'}(B) | \alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h\} & = \frac{\sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \pi_h(\mathbf{x}')}{\{\sum_{h=1}^{\infty} \pi_h(\mathbf{x})^2\}^{1/2} \{\sum_{h=1}^{\infty} \pi_h(\mathbf{x}')^2\}^{1/2}} \\
& = \frac{\langle \pi_h(\mathbf{x}), \pi_h(\mathbf{x}') \rangle}{\|\pi_h(\mathbf{x})\| \cdot \|\pi_h(\mathbf{x}')\|} \tag{6}
\end{aligned}$$

Note that the correlation is bounded above by 1 from the Cauchy-Schwarz inequality and goes to 1 in the limit as $\mathbf{x} \rightarrow \mathbf{x}'$. Because the correlation is not dependent on B , we obtain a single quantity given \mathbf{x} and \mathbf{x}' . Also, the correlation does not depend on the choice of P_0 .

Next, we consider the moments of $P_{\mathbf{x}}$ marginalizing out $\alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h$ as well as $\boldsymbol{\theta}_h$. Letting $U_h(\mathbf{x}) = \Phi(\eta_h(\mathbf{x}))$, we regard $U_h(\mathbf{x})$ as a random variable following a probability distribution $F_{\mathbf{x}}$. Note that $F_{\mathbf{x}}$ is induced through $N(\mu, 1)$, G , and H although its analytical expression is not straightforward. Because $\alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h$ are iid, we have $U_h(\mathbf{x}) \stackrel{iid}{\sim} F_{\mathbf{x}}$ for $h = 1, \dots, \infty$. Letting $\mu(\mathbf{x}) = E_{F_{\mathbf{x}}}\{U_h(\mathbf{x})\}$, $\mu^{(2)}(\mathbf{x}) = E_{F_{\mathbf{x}}}\{U_h(\mathbf{x})^2\}$, and $\mu(\mathbf{x}, \mathbf{x}') = E_{F_{\mathbf{x}}}\{U_h(\mathbf{x})U_h(\mathbf{x}')\}$, we can show that

$$E\{P_{\mathbf{x}}(B)\} = P_0(B)$$

$$\text{Var}\{P_{\mathbf{x}}(B)\} = \frac{\mu^{(2)}(\mathbf{x})\{P_0(B) - P_0(B)^2\}}{2\mu(\mathbf{x}) - \mu^{(2)}(\mathbf{x})}$$

$$\begin{aligned}
\text{Corr}\{P_{\mathbf{x}}(B), P_{\mathbf{x}'}(B)\} & = \left[\frac{\mu(\mathbf{x}, \mathbf{x}')}{\mu(\mathbf{x}) + \mu(\mathbf{x}') - \mu(\mathbf{x}, \mathbf{x}')} \right] \\
& \times \left[\frac{\{2\mu(\mathbf{x}) - \mu^{(2)}(\mathbf{x})\}\{2\mu(\mathbf{x}') - \mu^{(2)}(\mathbf{x}')\}}{\mu^{(2)}(\mathbf{x})\mu^{(2)}(\mathbf{x}')} \right]^{1/2} \tag{7}
\end{aligned}$$

Similar to the conditional moments, the correlation is not dependent either on B or on P_0 and only depends on the moments of $U_h(\mathbf{x})$. Proofs for (6) and (7) follow similar lines for the moments of the KSBP (Dunson and Park, 2008).

3. Conditional Distribution Modeling With Variable Selection

3.1 Model Specification

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

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

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

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

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 include Peng et al. (1996), Jiang and Tanner (1999) and Geweke and Keane (2007). The infinite HME can be obtained using nonparametric Bayesian approaches proposed by Müller et al. (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, which 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 G and P_0 in (8). Letting γ_{hj} be an inclusion indicator variable for the j th predictor in the h th mixture component, we induce G and P_0 through the following distributions for $\boldsymbol{\psi}_h$ and $\boldsymbol{\theta}_h$.

$$\begin{aligned}\boldsymbol{\psi}_h = \{\psi_{hj}\}_{j=1}^p &\sim \prod_{j=1}^p \left\{ 1(\gamma_{hj} = 0)\delta_0(\psi_{hj}) + 1(\gamma_{hj} = 1)N_+(\psi_{hj}; \mu_{\psi_j}, \tau_{\psi_j}^{-1}) \right\} \\ \boldsymbol{\theta}_h = (\boldsymbol{\beta}_h^*, \tau_h^*) &\sim N_{p_{\gamma_h}+1}(\boldsymbol{\beta}_{\gamma_h, h}^*; \mathbf{0}, \boldsymbol{\Sigma}_{\gamma_h, h}) \times \delta_0(\boldsymbol{\beta}_{\gamma_h, h}^*) \times \text{Gamma}(\tau_h^*; a_\tau, b_\tau),\end{aligned}\quad (10)$$

where N_+ denotes a truncated normal distribution bounded below by zero, $\boldsymbol{\beta}_{\gamma_h, h}^*$ is the vector of regression coefficients corresponding to $\gamma_{hj} = 1$ including intercept, $\boldsymbol{\beta}_{\gamma_h, h}^*$ 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 j th predictor, with $\gamma_{hj} = 0$ implying that $\psi_{hj} = 0$ and $\beta_{hj}^* = 0$. A value of $\beta_{hj}^* = 0$ leads to the j th predictor assigned a coefficient of zero in the h th linear regression model in (9), while a value of $\psi_{hj} = 0$ leads to excluding the j th predictor from the h th predictor-dependent stick-breaking weight in the expression for $\pi_h(\mathbf{x})$. Clearly, if $\gamma_{hj} = 0$ for $h = 1, \dots, \infty$, then the j th predictor will be globally excluded from the model. To allow uncertainty in γ_{hj} , we let

$$\gamma_{hj} \sim \text{Bernoulli}(\gamma_{hj}; \kappa_j), \quad (11)$$

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

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

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.

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

where n is the number of subjects and \mathbf{X}_{γ_h} is the design matrix corresponding to $\gamma_{hj} = 1$ including intercept.

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 j th predictor as $H_{0j} : \gamma_{hj} = 0$ for $h = 1, \dots, \infty$. 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 the 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 j th predictor as follows:

$$H_{0j}^N : \gamma_{hj} = 0 \quad \text{for } h = 1, \dots, N, \quad (14)$$

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 examine the following hierarchical structure of the PSBPM in (8).

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

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

Theorem 1. Suppose $y_i|\mathbf{x}_i \sim f(y|\mathbf{x})$ and $f(y|\mathbf{x})$ is assumed to be a PSBPM as in (8) with G and P_0 chosen as in (10) and (11). Let $l(\mathbf{y}, \mathbf{S}|H_{0j})$ and $l(\mathbf{y}, \mathbf{S}|H_{0j}^N)$ be the marginal likelihoods for (\mathbf{y}, \mathbf{S}) under H_{0j} and H_{0j}^N where $\mathbf{y} = (y_1, \dots, y_n)'$ and $\mathbf{S} = (S_1, \dots, S_n)'$. Then, the ratio $R = \frac{l(\mathbf{y}, \mathbf{S}|H_{0j})}{l(\mathbf{y}, \mathbf{S}|H_{0j}^N)}$ 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_{0j}^N , 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_{0j}^N 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_{0j}^N . 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 j th 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 $\mathbf{x}_{(j)}^* = (x_1^*, \dots, x_{j-1}^*, x_{j+1}^*, \dots, x_p^*)'$. Letting $d(x_j, x'_j | \mathbf{x}_{(j)}^*) = \sup_{y \in \mathcal{R}} |F(y | x_j, \mathbf{x}_{(j)}^*) - F(y | x'_j, \mathbf{x}_{(j)}^*)|$, we propose a local interval null hypothesis as:

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

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 j th predictor space as:

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

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 j th 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 G and P_0 chosen as in (10) with (11), (12) and (13). For H , we consider $\Gamma_h = \{\Gamma_{hj}\}_{j=1}^p \sim \prod_{j=1}^p \sum_{m=1}^{M_j} \delta_{\Gamma_{mj}^*}(\Gamma_{hj})$ where Γ_{mj}^* for $m = 1, \dots, M_j$ are pre-specified grid values for j th predic-

tor. In addition, we assume $\mu \sim N(\mu; \mu_\mu, \tau_\mu^{-1})$. In order to sample finite number of random components for $P_{\mathbf{x}}$, we rely on a modification of the blocked Gibbs sampler (Ishwaran and James, 2001) with the 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 $l = 1, \dots, S_i - 1$ and $Z_{il} = 1$ for $l = S_i$ where

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

For $S_i = T$, we introduce Z_{il}^* only for $l = 1, \dots, T - 1$ 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 marginal likelihoods for (\mathbf{y}, \mathbf{S}) (Step 11). Note that this step generalizes the SSVS step for linear regression (George and McCulloch, 1997).

4.2 Default Choices for Hyperparameters

Prior to analysis, we standardize the response and predictors. For the standardized data, we propose the following default choices for the hyperparameters. For G , $\mu_{\psi_j} = 0, \tau_{\psi_j} = 1$ for $j = 1, \dots, p$. For P_0 , $a_g = b_g = 0.5$ and $a_\tau = b_\tau = 0.5$. For H , we choose 50 equally spaced grid points for Γ_{mj}^* in $(-2.5, 2.5)$ for all j . For others, $a_{\kappa_j} = b_{\kappa_j} = 0.5$ for all j and $\mu_\mu = 0, \tau_\mu = 1$. For local null hypotheses, we let $\epsilon = 0.05$. 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. We first generate $x_{ij} \stackrel{iid}{\sim} \text{Uniform}(x_{ij}; -2, 2)$ for $i = 1, \dots, n$ and $j = 1, \dots, p$. The response is generated for a null case (1) and two alternative cases (2) and (3).

$$\begin{aligned}
 (1) \quad y_i &\stackrel{iid}{\sim} 0.5N(y_i; 1, 1) + 0.5N(y_i; -1, 0.5^2) \\
 (2) \quad y_i &\stackrel{iid}{\sim} N(y_i; 2x_{i1} - 3x_{i2} + x_{i4} - x_{i5}, 1) \\
 (3) \quad y_i &\stackrel{iid}{\sim} 0.5N(y_i; 10, (2 + 4e^{-\min(x_{i1}, 0)})^2) + 0.5N(y_i; -10 + 5x_{i2}, 5^2) \quad (19)
 \end{aligned}$$

The case (1) is a mixture of two normals with no change in $f(y|\mathbf{x})$ across \mathbf{x} . The case (2) is a standard normal linear regression where $f(y|\mathbf{x})$ changes only in mean as \mathbf{x} changes. The 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_2 increases. In particular, x_1 has a local impact only when $x_1 < 0$ having no effect on $E(y|\mathbf{x})$ while x_2 has a global impact on $E(y|\mathbf{x})$.

5.1 Simple Application of PSBPM

We begin with $p = 10$ and $n = 1000$. After standardizing y , we applied the PSBPM with the priors and hyperparameters discussed in sections 3 and 4. The MCMC algorithm described in section 4.1 was run for 10,000 iterations, with the first 5,000 iterations discarded as burn-ins. The MCMC chain appeared to converge rapidly and to mix efficiently based on the trace plots.

In case (1), $\Pr(H_{0j}^N | \text{Data})$ was above 0.9 for all j , suggesting that none of the predictors are important. 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. In case (2), $\Pr(H_{0j}^N|\text{Data}) = 0$ for $j = 1, 2, 4, 5$ and above 0.8 for the other j , implying that the PSBPM correctly selects important predictors in a simple normal linear regression case. The true and predictive response densities were almost the same at various predictor points \mathbf{x}^* .

In case (3), Figure 1 shows that $\Pr(H_{0j}^N|\text{Data})$ are 0 for $j = 1, 2$ and above 0.8 for $j \geq 3$. The PSBPM correctly identified x_1 and x_2 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 $\Pr(H_{01}(\max(x_1), x_1)|\text{Data})$ is 0 for $x_1 < 0$, increases towards 1 for $x_1 > 0$ reflecting the local impact of x_1 . Meanwhile, $\Pr(H_{02}(\max(x_2), x_2)|\text{Data})$ is 0 across x_2 because x_2 is globally important. Figure 2 shows that the predictive density (dashed) with its 95% credible intervals (dash-dotted) closely follows the true one (solid) reflecting the shape change across x_1 and x_2 .

In order to evaluate scalability to larger numbers of candidate predictors, we applied the PSBPM for all 3 cases in (19) with $p = 10, 15, 20$ and $n = 800, 1000, 1200$. The results were similar to $p = 10$ and $n = 1000$ implying that the PSBPM is robust to moderate sample sizes and can handle reasonably many predictors. In addition, we conducted a sensitivity analysis for different choices of hyperparameters within a reasonable range and found similar results regardless of the choice. Finally, we applied the method to 100 replicates of each simulation case and found that the results were consistent among the replicates. (Results not shown.)

5.2 Comparison with a simple and a competing methods

In order to illustrate the potential of the PSBPM, we compare it with a simple method and a competing method for the simulation cases in (19) with $p = 10$ and $n = 1000$. For a simple one, we consider a standard linear regression with SSVS (George and McCulloch, 1997) (LR-SSVS) where prior structure for regression coefficients is consistent to (10) with (11), (12),

(13). As a competitor, we consider Bayesian Additive Regression Trees (BART) (Chipman et al., 2006). Although BART focuses on mean response, we chose BART because there is no competing method which performs variable selection in the general setting of conditional distribution modeling and the BART is a recently proposed flexible mean regression model shown to be comparable with its competitors while allowing for variable selection based on the partial dependence plot (PDP) (Chipman et al., 2006). Implementing BART using the R statistical software, we consider a default setting for priors and hyperparameters.

In case (1), we obtained $\Pr(\beta_j = 0|\text{Data}) \approx 1$ for all j with LR-SSVS and none of the predictors appeared to have an impact on the mean response with BART based on the PDP. In case (2), both LR-SSVS and BART correctly identified x_j for $j = 1, 2, 4, 5$ as important. Predictive performance for $E(y|\mathbf{x})$ was good for both methods in both cases. This implies that the PSBPM, LR-SSVS and BART are comparable in a null case or a simple linear regression case with respect to variable selection and mean prediction. However, for a non-normal response data such as case (1), the LR-SSVS and BART would not be comparable with PSBPM for distribution prediction because of their normality assumption. Although there is a recent extension of BART that allows nonparametric modeling of the residual distribution using DP mixtures, our approach is still dramatically more flexible in allowing the residual distribution to change flexibly over the predictor space. In addition, the PDP of BART is not a formal approach for variable selection not being comparable to a posterior exclusion probability of Bayes Factor provided by the LR-SSVS or PSBPM.

In case (3), LR-SSVS detected only x_2 as important with $\Pr(\beta_2 = 0|\text{Data})=0$. $\Pr(\beta_1 = 0|\text{Data})=0.87$ and $\Pr(\beta_j = 0|\text{Data})$ was above 0.9 for $j \geq 3$. Meanwhile, BART showed a strong evidence that x_1 has an impact but not so much for the other predictors. This suggests that the PSBPM identifies important predictors correctly while LR-SSVS and BART fail to do so, in particular, when predictors have impacts not only on the mean but also on the shape or tails of the response distribution substantially. This is not an unusual scenario

in applications, since such behavior is a natural consequence when the predictors are not related to the typical response but instead to risk of extremes. For example, these extremes may correspond to adverse health responses or unusual financial or meteorological events. In addition, we compared the three methods with respect to mean prediction for 200 in-sample predictor points. Figure 3 shows the scatter plot for predictive mean $\hat{E}(y|\mathbf{x})$ and true mean $E(y|\mathbf{x})$ along with the observed response y versus x_2 . PSBPM (top) and LR-SSVS (middle) were comparable in that $\hat{E}(y|\mathbf{x})$ is almost indistinguishable from $E(y|\mathbf{x})$ while BART (bottom) performed poorer with $\hat{E}(y|\mathbf{x})$ scattering around $E(y|\mathbf{x})$.

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), $[140, 200]$ (pre-diabetes), 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 multi-ethnic cohort. Figure 5

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 non-linear 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 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 4. For $i = 1, \dots, 868$, $y_i = 2\text{-hour glucose level (mg/dl)}$, $x_{i1} = \text{IS}$, $x_{i2} = \text{age}$, $x_{i3} = \text{WTH}$, $x_{i4} = \text{BMI}$, $x_{i5} = \text{DBP}$, and $x_{i6} = \text{SBP}$. Prior to the analysis, we standardized both response and predictors. Firstly, we applied the simple LR-SSVS and obtained $\Pr(\beta_j = 0|\text{Data}) = 0.00, 0.65, 0.00, 0.94, 0.14, 0.01$, for $j = 1, \dots, 6$. In order to better meet the normality assumption, we fit the LR-SSVS for log-transformed response and obtained $\Pr(\beta_j = 0|\text{Data}) = 0.00, 0.14, 0.00, 0.71, 0.00, 0.23$, for $j = 1, \dots, 6$. IS, WTH, DBP, and SBP were found to be important and age was added with log-transformation. Secondly, we applied the BART and found strong evidence for the effect of IS and some evidence for the other predictors with/without log transformation based on the partial-dependence plots. However, the residual plots showed that the constant normal residual assumption is strongly violated so the results may not be reliable.

Next, we applied the PSBPM and obtained $\Pr(H_j^N|\text{Data}) = 0.00, 0.00, 0.87, 0.97, 0.97, 0.78$, indicating that only IS and age are important predictors. The results for IS and age

are consistent with LR-SVSS and BART applied to log-transformed glucose level while inconsistent results were shown for the other predictors. We suspect that such inconsistency may result from the restrictive assumption of LR-SSVS and BART for the residual 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 $\mathbf{x}^* = (x_1, x_2, \bar{x}_3, \dots, \bar{x}_{10})$ with x_1 and x_2 varying among 5th, 50th, 95th empirical percentiles. Figure 5 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.

7. Discussion

We propose a nonparametric Bayesian approach for conditional distribution modeling with variable selection. We first introduce the probit stick-breaking process (PSBP) as a new choice of prior for an uncountable collection of predictor-dependent random probability measures and consider a PSBP mixture (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 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.

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.

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:

$$l(\mathbf{y}, \mathbf{S} | \boldsymbol{\Lambda}) = \int \prod_{i=1}^n (y_i | \mathbf{x}_i, \boldsymbol{\theta}_{S_i}) \prod_{h=1}^{\infty} (\boldsymbol{\theta}_h | \boldsymbol{\gamma}_h) d\boldsymbol{\Theta} \times \int \prod_{i=1}^n (S_i | \mathbf{x}_i, \boldsymbol{\Xi}) \prod_{h=1}^{\infty} (\boldsymbol{\xi}_h | \boldsymbol{\gamma}_h) d\boldsymbol{\Xi} \quad (20)$$

Because $S_i \leq N$, we reexpress (20) as:

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

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

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

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

MCMC algorithm

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

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

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

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

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

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

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

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

4. Update Γ_{hj} for $j = 1, \dots, p$ and $h = 1, \dots, T - 1$: If $\gamma_{hj} = 0$, don't update. If $\gamma_{hj} = 1$,

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

5. Update β_h^* for $h = 1, \dots, T$: With $\beta_h^* = (\beta_{\gamma_h, h}^*, \beta_{\gamma_h, h}^*)$, $\beta_{\gamma_h, h}^* = \mathbf{0}$.

$$\beta_{\gamma_h, h}^* \sim N(\beta_{\gamma_h, h}^*; [\tau_h^* \mathbf{X}'_{\gamma_h, h} \mathbf{X}_{\gamma_h, h} + \Sigma_{\gamma_h, h}^{-1}]^{-1} [\tau_h^* \mathbf{X}'_{\gamma_h, h} \mathbf{y}_h], [\tau_h^* \mathbf{X}'_{\gamma_h, h} \mathbf{X}_{\gamma_h, h} + \Sigma_{\gamma_h, h}^{-1}]^{-1}),$$

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 $h = 1, \dots, T$: With $k_h = \sum_{i=1}^n 1(S_i = h)$ and $p_{\gamma_h} = \sum_{j=1}^p \gamma_{hj}$,

$$\begin{aligned} \tau_h^* &\sim \text{Gamma}(\tau_h^*; a_\tau + \frac{k_h}{2} + \frac{p_{\gamma_h} + 1}{2}), \\ b_\tau &+ \frac{1}{2} (\mathbf{y}_h - \mathbf{X}_{\gamma_h, h} \beta_{\gamma_h, h}^*)' (\mathbf{y}_h - \mathbf{X}_{\gamma_h, h} \beta_{\gamma_h, h}^*) + \frac{g}{2n} \beta_{\gamma_h, h}^{*'} (\mathbf{X}'_{\gamma_h} \mathbf{X}_{\gamma_h}) \beta_{\gamma_h, h}^* \end{aligned}$$

7. Update g :

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

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

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

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

$$\Pr(w_j = 1) = \frac{\frac{\Gamma(b_{\kappa_j} + T) \Gamma(a_{\kappa_j} + b_{\kappa_j})}{\Gamma(b_{\kappa_j}) \Gamma(a_{\kappa_j} + b_{\kappa_j} + T)}}{1 + \frac{\Gamma(b_{\kappa_j} + T) \Gamma(a_{\kappa_j} + b_{\kappa_j})}{\Gamma(b_{\kappa_j}) \Gamma(a_{\kappa_j} + b_{\kappa_j} + T)}}$$

10. Update μ :

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

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

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

$$\begin{aligned} a_{hj} &= \kappa_j \times \int \prod_{i:S_i \geq h, S_i \neq T} N(Z_{ih}^*; \alpha_h - \sum_{j=1}^p \psi_{hj} |x_{ij} - \Gamma_{hj}|, 1) N_+(\psi_{hj}; \mu_{\psi_j}, \tau_{\psi_j}^{-1}) d\psi_{hj} \\ &\quad \times \int \prod_{S_i=h} N(y_i; \mathbf{x}'_{i0} \boldsymbol{\beta}_h^*, \tau_h^{*-1}) N(\beta_{hj}^*; \mu_{\beta_j}, \tau_{\beta_j}^{-1}) d\beta_{hj}^* \\ b_{hj} &= (1 - \kappa_j) \times \prod_{i:S_i \geq h, S_i \neq T} N(Z_{ih}^*; \alpha_h - \sum_{k=1, k \neq j}^p \psi_{hk} |x_{ik} - \Gamma_{hk}|, 1) \\ &\quad \times \prod_{S_i=h} N(y_i; \mathbf{x}'_{(-j)i0} \boldsymbol{\beta}_{(-j)h}^*, \tau_h^{*-1}), \end{aligned}$$

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 $N_{p\gamma_h}(\boldsymbol{\beta}_{\gamma_h, h}^*; \mathbf{0}, ng^{-1}(\mathbf{X}'_{\gamma_h} \mathbf{X}_{\gamma_h})^{-1}/\tau_h^*)$.

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.
- Cai, B. and Dunson, D.B. (2007). Bayesian variable selection in nonparametric random effects models. *ISDS Discussion Paper*, **05-16**, Duke University, Durham, NC, USA.
- 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.
- 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., Herring, A.H. and Mulherin-Engel, S.M. (2007a). Bayesian selection and clustering of polymorphisms in functionally-related genes. *Journal of the American Statistical Association*, in press.
- Dunson, D.B. and Park, J-H. (2008). Kernel stick-breaking process. *Biometrika*, in press.
- Dunson, D.B. and Peddada, S.D. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrika*, in press.
- Dunson, D.B., Pillai, N., and Park, J-H. (2007b). Bayesian density regression. *Journal of the Royal Statistical Society, Series B*, **69**, 163-183.
- 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., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83**, 189-206.
- Fan, J.Q. and Yim, T.H. (2004). A cross validation method for estimating conditional densities. *Biometrika*, **91**, 819-834.
- 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.
- Griffin, J.E. and Steel, M.F.J. (2006) Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, **101**, 179-194.

- Griffin, J.E. and Steel, M.F.J. (2007) Bayesian nonparametric modelling with the Dirichlet process regression smoother. *CRiSM Working Paper*, 07-05.
- 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.
- Jacobs, R.A., Peng, F.C. and Tanner, M.A. (1997). A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks*, **10**, 231-241.
- 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.
- 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.
- Leslie, D.S., Kohn, R. and Nott, D.J. (2007). A general approach to heteroscedastic linear regression. *Statistics and Computing*, **17**, 131-146.
- 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, *Bayesian Inference for Gene Expression and Proteomics*, 155-176, Cambridge University Press.
- MacEachern, S.N. (1999), Dependent Nonparametric Processes, *Proceedings of the Bayesian Section of the American Statistical Association*, 50-55.
- Müller, P., Erkanli, A. and West, M. (1996). Bayesian Curve Fitting using Multivariate Normal Mixtures. *Biometrika*, **83**, 67-79.
- Pennell, M.L. and Dunson, D.B. (2007). Nonparametric Bayes testing of changes in a response distribution with an ordinal predictor. *Biometrics*, revision submitted.
- Sethuraman, J. (1994), A Constructive Definition of the Dirichlet Process Prior, *Statistica Sinica*, **2**, 639-650.
- Wang, L. and Dunson, D.B. (2007). Bayesian isotonic density regression. *ISDS Discussion Paper*, **07-11**, Duke University, Durham, NC.
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions, *Bayesian Inference and Decision Techniques: Essay in Honor of Bruno de Finetti*, 233-243, North-Holland/Elsevier.

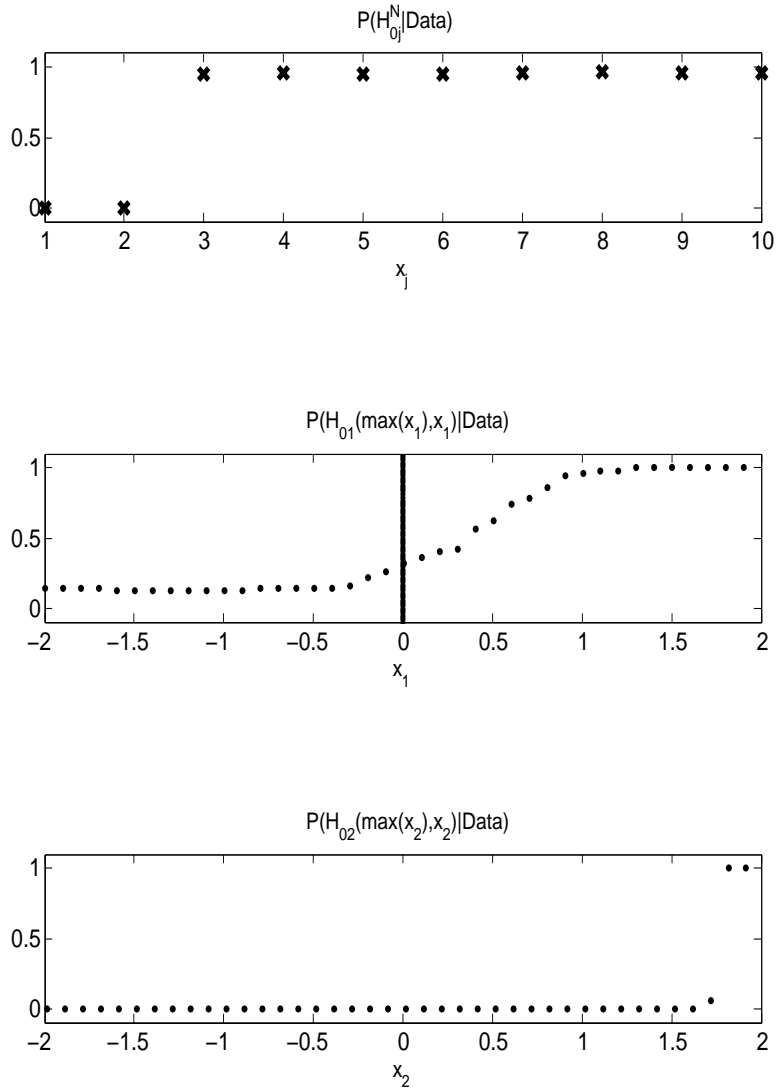


Figure 1: Top - Posterior probabilities for H_{0j}^N for $j = 1, \dots, 10$; Middle - Posterior probabilities for $H_{01}(\max(x_1), x_1)$ with x_1 varying across 40 grid points; Bottom - Posterior probabilities for $H_{02}(\max(x_2), x_2)$ with x_2 varying across 40 grid points

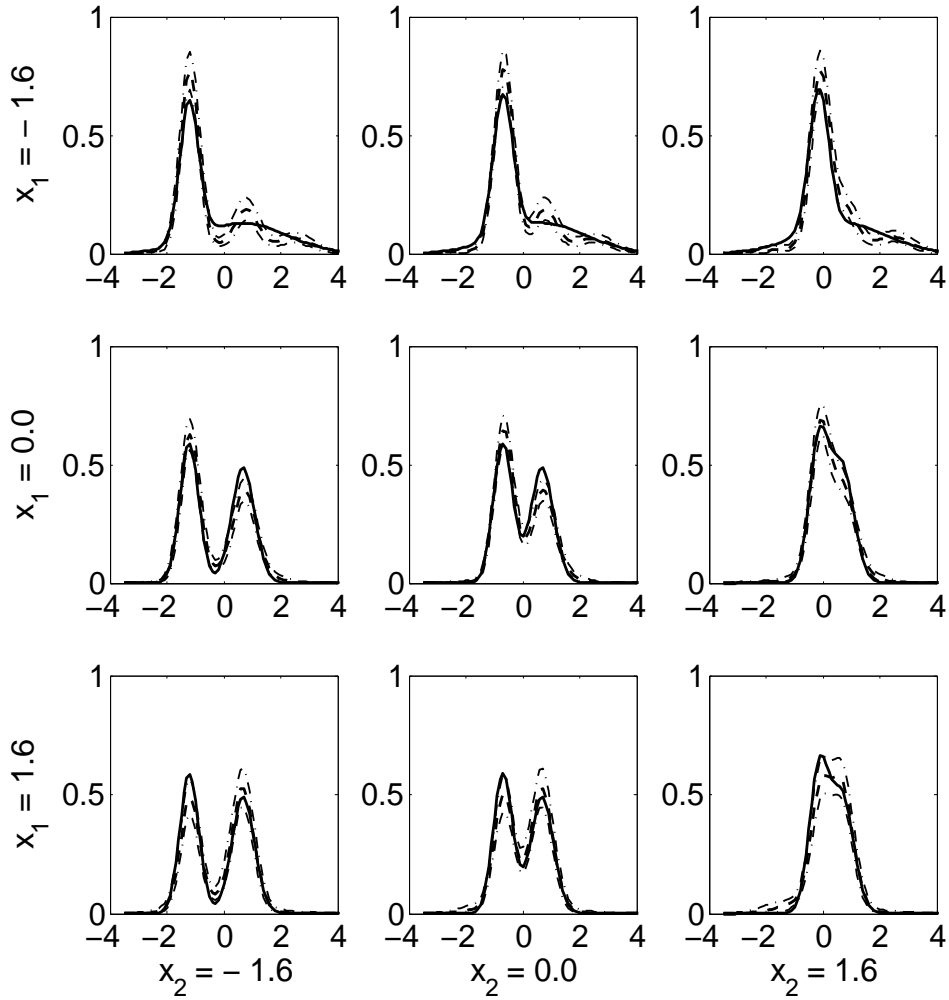


Figure 2: True (solid), Predictive (dashed) conditional response density $\hat{f}(y|\mathbf{x}^*)$ with 95% credible intervals (dash-dotted) at $\mathbf{x}^* = (x_1, x_2, \bar{x}_3, \dots, \bar{x}_{10})$ with x_1 and x_2 varying among 5th, 50th, 95th empirical percentiles

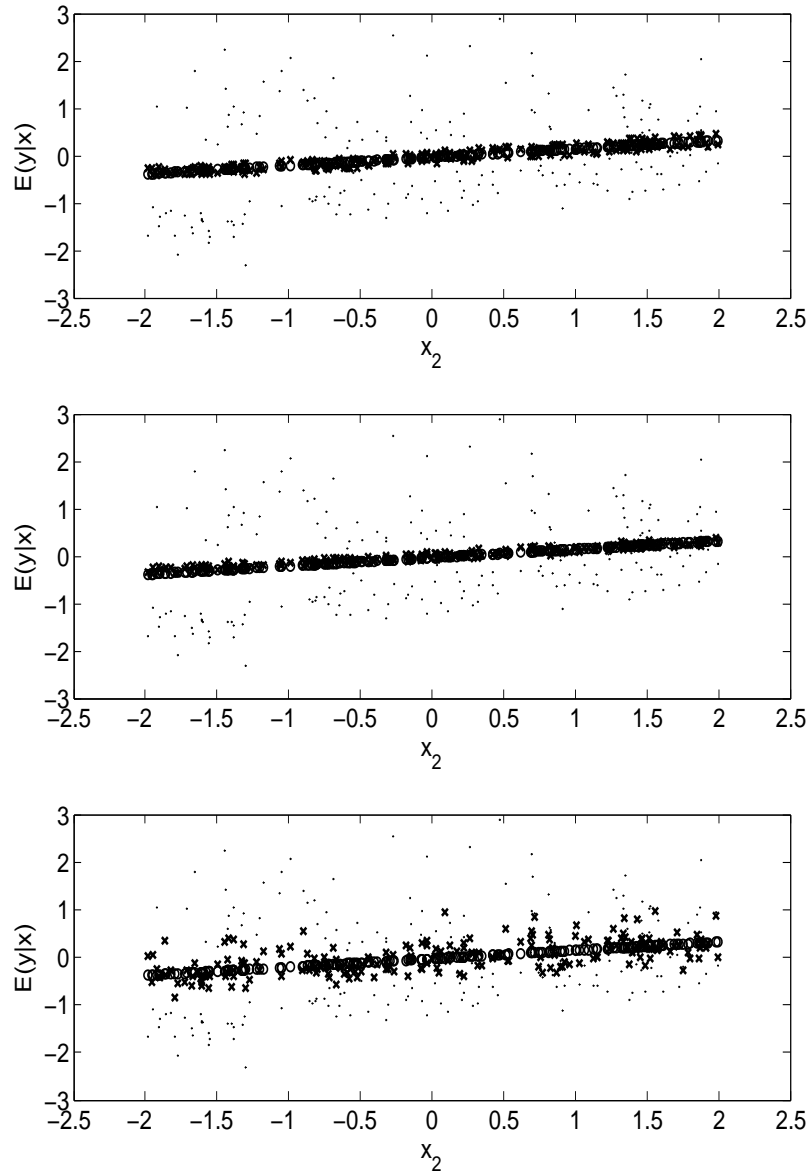


Figure 3: True mean $E(y|x)$ ('o'), Predictive mean $\hat{E}(y|x)$ ('x'), observed data y ('*') across x_2 : Top - PSBPM; Middle - LR-SSVS; Bottom - BART

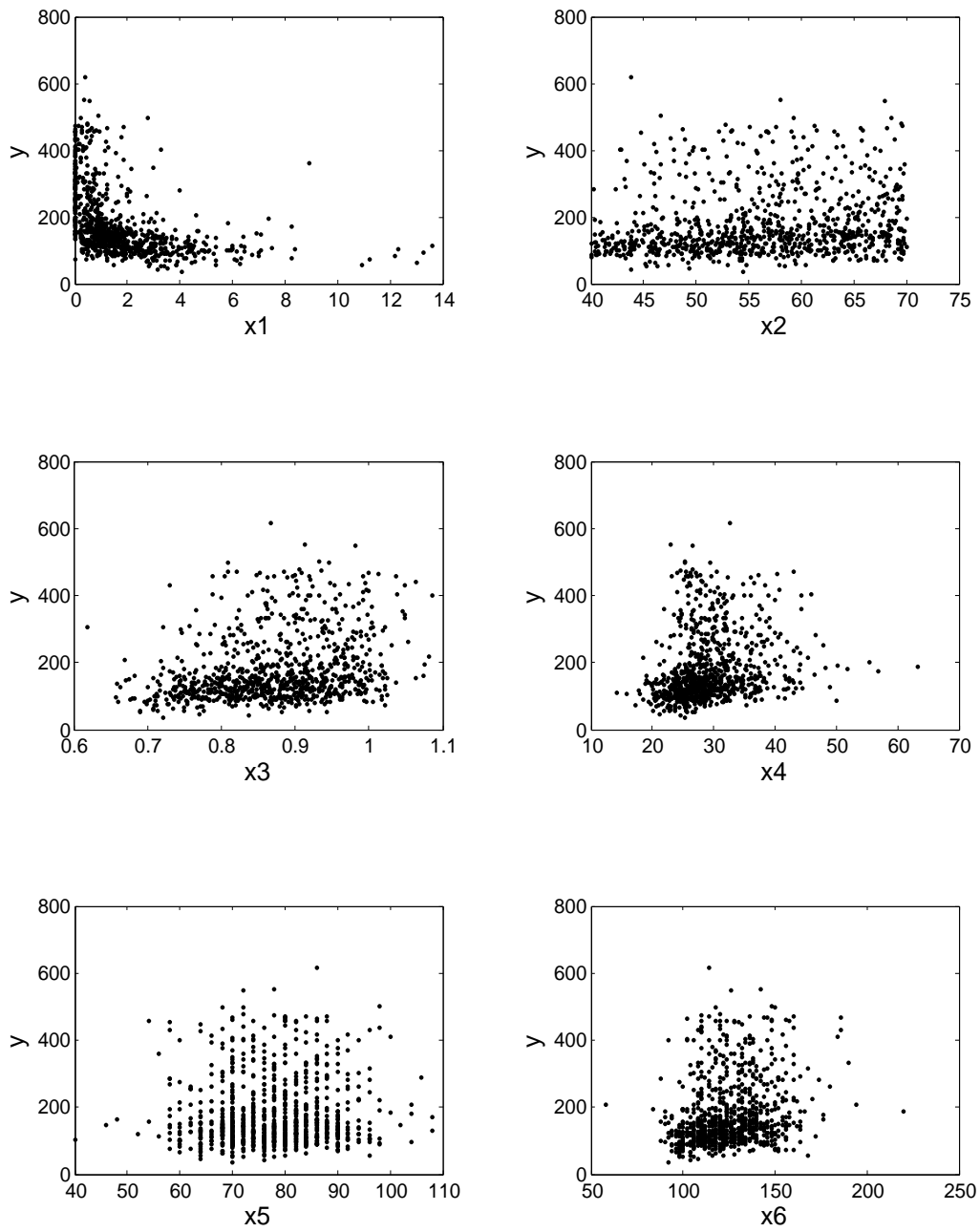


Figure 4: Data from IRAS study : y = 2-hour glucose level (mg/dl); x_1 = insulin sensitivity; x_2 = age; x_3 = waist to hip ratio; x_4 = body mass index; x_5 = diastolic blood pressure; x_6 = systolic blood pressure

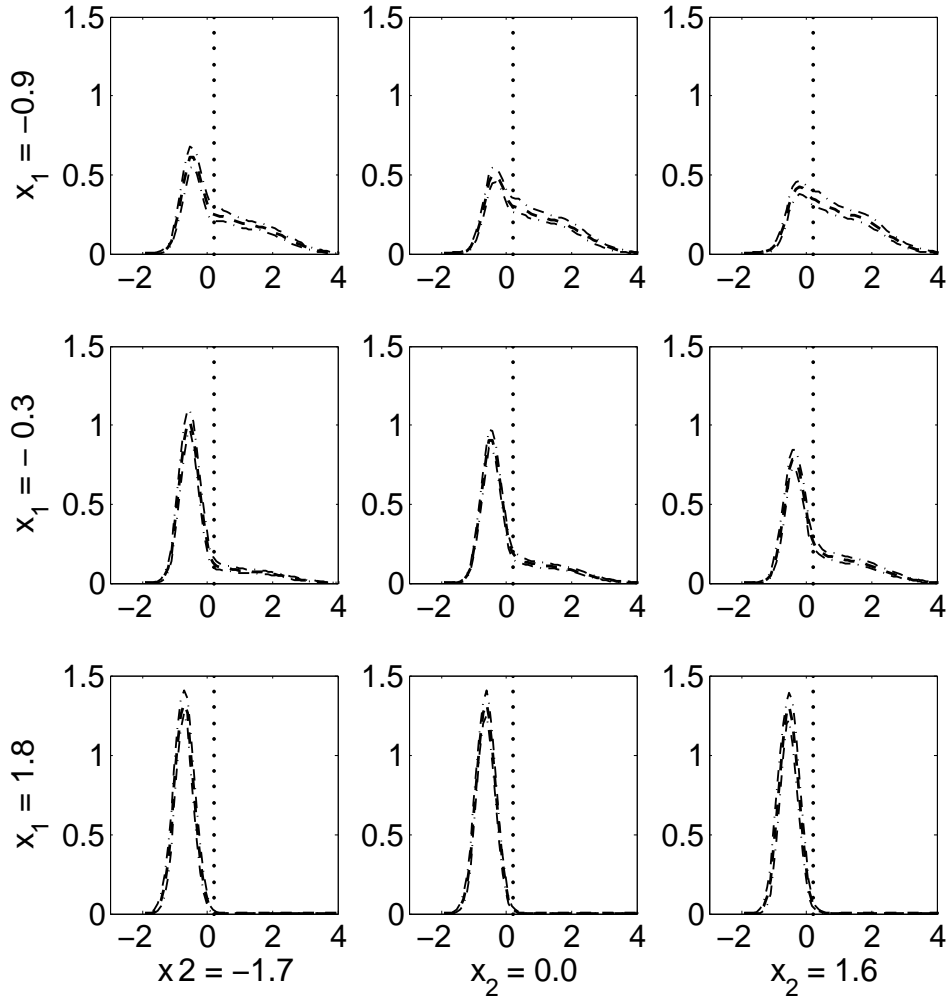


Figure 5: Predictive (dashed) conditional response density $\hat{f}(y|\mathbf{x}^*)$ with 95% credible intervals (dash-dotted) at $\mathbf{x}^* = (x_1, x_2, \bar{x}_3, \dots, \bar{x}_6)$ with x_1 and x_2 varying among 5th, 50th, 95th empirical percentiles