NONPARAMETRIC BAYESIAN INFERENCES ON PREDICTOR-DEPENDENT RESPONSE DISTRIBUTIONS

by
Yeonseung Chung

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2008

Approved by:

Dr. David Dunson, Advisor
Dr. Young Truong, Reader
Dr. Amy Herring, Reader
Dr. Donglin Zeng, Reader
Dr. Jiu-Chiuan Chen, Reader

ABSTRACT

YEONSEUNG CHUNG: Nonparametric Bayesian Inferences on Predictor-Dependent Response Distributions.
(Under the direction of Dr. David Dunson.)

A common statistical problem in biomedical research is to characterize the relationship between a response and predictors. The heterogeneity among subjects causes the response distribution to change across the predictor space in distributional characteristics such as skewness, quantiles and residual variation. In such settings, it would be appealing to model the conditional response distributions as flexibly changing across the predictors while conducting variable selection to identify important predictors both locally (within some local regions) and globally (across the entire range of the predictor space) for the response distribution change.

Nonparametric Bayes methods have been very useful for flexible modeling where nonparametric distributions are assumed unknown and assigned priors such as the Dirichlet process (DP). In recent years, there has been a growing interest in extending the DP to a prior model for predictor-dependent unknown distributions, so that the extended priors are applied to flexible conditional distribution modeling. However, for most of the proposed extensions, construction is not simple and computation can be quite difficult. In addition, literature has been focused on estimation and few attempts have been made to address related hypothesis testing problems such as variable selection.

Paper 1 proposes a local Dirichlet process (lDP) as a generalization of the Dirichlet process to provide a prior distribution for a collection of random probability measures indexed by predictors. The lDP involves a simple construction, results in a marginal Dirichlet process prior for the random measure at any specific predictor value, and leads to a straightforward posterior computation. In paper 2, we propose a more general approach not only estimating the conditional response distribution but also identifying important predictors for the response distribution change both with local regions and globally. This is achieved through the probit stick-breaking process mixture (PSBPM) of normal linear regressions where the PSBP is a newly proposed prior for dependent probability measures and particularly convenient to incorporate variable selection structure. In paper 3, we extend the paper 2 method for longitudinal outcomes which are correlated within subject. The PSBPM of linear mixed effect (LME) model is considered allowing for individual variability within a mixture component.

ACKNOWLEDGMENTS

Though my name solely appears on the cover of this dissertation, many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible, making my graduate experience one that I will cherish forever.

I would like to express my deepest appreciation to my advisor, Dr. David Dunson, without whom this dissertation would not have been possible. David continually and convincingly conveyed a spirit of adventure towards research, even as I struggled to face many unexpected hurdles. I will never forget his patient and endless support that helped me overcome many crises and his willingness to walk with me throughout my journey. His words of encouragement, quiet urgings and careful editing of my writing will never be forgotten. Besides David, I would also like to thank my committee members, Dr. Amy Herring, Dr. Donglin Zeng, Dr. Young Thruong, and Dr. Jiu-Chiuan Chen for their insightful comments and constructive criticisms.

In addition to my academic mentors, many friends have helped me remain strong through these difficult years. Their care and support have helped me overcome setbacks and stay focused on my graduate study. I am especially grateful for the support of one of my best friends, Jayeon Jeong, and to my faith family at First Korean Baptist Church.

I also appreciate the financial support from the National Institute of Environmental Health and Sciences that funded my doctoral research.

Most importantly, this dissertation would not have been possible without the love and patience of my family. My family, to whom this dissertation is dedicated, has been a constant source of love, concern, support and strength. I cannot express enough my heart-felt gratitude for my family.

Finally, I would like to thank God for giving me the wisdom and strength to complete my doctoral degree. I am grateful for God's provision of joys, challenges, and grace for growth through the five years of my doctoral training.

CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In biomedical research, one wishes to study the relationship between a response and predictors. A common interest may be to characterize how the mean response changes as the predictors change. However, there often exists heterogeneity among subjects in the impact of predictors on the response. Such heterogeneity causes the response distribution to change as the predictors change, not only in mean but also in other characteristics such as skewness, quantiles and residual variation. In addition, important predictors for the response distribution may change unexpectedly across the predictor space. Hence, it is appealing if one can flexibly estimate the conditional distribution of a response addressing the distributional changes across the predictor space and can perform hypothesis testing to detect the changes in distribution or to identify important predictors for the distribution change both within some local regions and globally. In particular, 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.

In such settings, nonparametric Bayes methods are very useful where unknown quantities are assigned nonparametric probability measures that are also assumed unknown and assigned priors such as the Dirichlet process (DP) (Ferguson, 1973, 1974). In particular, the Dirichlet process mixture (DPM) (Lo, 1984; Escobar, 1994; Escobar and West, 1995) model has popularly been used to smooth any distributional shape as an infinite mixture model. For the problem of flexible characterization of predictor-dependent response distributions, one may fit DPM models separately for different predictor levels, which results in smooth estimation of predictor-specific response distributions. However, the approach of fitting several DPM models at different predictor levels is disadvantageous in that it neither models trends nor borrows information across the predictor space, which is particularly important in applications having a modest number of subjects. In addition, the approach requires some arbitrary categorization for continuous predictors, which can discard valuable information. Furthermore, as the number of predictor categories increases, estimation and testing efficiency may decrease.

In recent years, there has been a growing interest in extending the DP to a prior model for predictor-dependent unknown probability measures. Most of this literature has relied on extending the stick-breaking representation of the DP (Sethuraman, 1994) and has been stimulated by the dependent Dirichlet process (DDP) framework proposed by MacEachern (1999, 2000, 2001), which replaces the atoms in the stick-breaking representation of the DP with stochastic processes. The DDP structure has been adopted to develop ANOVA-type dependence for random probability measures (De Iorio et al., 2004), for flexible spatial modeling (Gelfand et al., 2004), and for inferences on stochastic ordering (Dunson and Peddada, 2008). The specification of the DDP used in applications incorporates dependence only through the atoms while assuming fixed weights. In other recent work, Griffin and Steel (2006; 2008) proposed an order-based DDP (πDDP) which allows varying weights, while Duan et al. (2005) developed a multivariate stick-breaking process for spatial data. In addition, Teh et al. (2004) proposed the hierarchical Dirichlet process (hDP) which generates group-specific random probability measures having group-dependent weights but sharing atoms in their stick-breaking forms.

Alternatively, convex combinations of independent DPs have been used for modeling collections of dependent random measures. Müller et al. (2004) used this idea to allow dependence across experiments combining a global component and experiment-specific components drawn from DPs. Dunson (2006) proposed an alternative dynamic mixture of DPs (DMDP), which is appropriate for modeling of changes with a categorical predictor or discrete time index. A related idea was used by Pennell and Dunson (2006) to develop dynamic frailty models for event time data. In addition, Rodriguez et al. (2008) used DP-type combination of DPs called nested DP (nDP) which was motivated by the idea of clustering groups and subjects within a group simultaneously. Recently, the idea has been extended to continuous covariate cases by Dunson et al. (2007) and Dunson and Park (2008).

However, for most of the DP-extended priors discussed so far, they are limited either to the cases of categorical predictors or, for continuous predictor cases, to complicated computation causing the methods to be unaccessible in many applications. In addition, this literature has been focused on estimation and few attempts have been made to address related hypothesis testing problems such as variable selection or detecting distributional changes both globally (across the entire predictor space) and locally (within some local predictor regions). In fact, there has been limited focus on hypothesis testing and model selection in Bayesian nonparametric literature. Basu and Chib (2003) proposed an approach for calculating marginal likelihoods and Bayes factors for comparing DPM models. But this approach is not directly applicable to our local variable selection problem. Pennell and Dunson (2008) proposed a method for testing distributional changes in response across an ordinal predictor while Dunson and Pedadda (2008) tested equalities in group specific response distributions against a stochastic ordering. Both approaches deal with a categorical predictor whereas we seek for a methodology that can incorporate a mix of continuous predictors as well as categorical predictors.

Motivated by this, paper 1 proposes a local Dirichlet process (lDP) as a generalization of the Dirichlet process to provide a prior distribution for a collection of random probability measures indexed by predictors. The lDP should be useful to other alternative prior models for dependent random probability measures in that it involves a simple construction, results in a marginal Dirichlet process prior for the random measure at any specific predictor value, and leads to a straightforward posterior computation. Theoretical properties are considered and a blocked Gibbs sampler is proposed for posterior computation in lDP mixture models. The methods are illustrated in a conditional distribution modeling setting using simulated examples and an epidemiologic application.

In paper 2, we propose a more general approach for conditional distribution modeling where we not only estimate the conditional response distribution addressing the distributional changes across the predictor space, but also 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 linear regressions. A global variable selection structure is incorporated to drop unimportant predictors out from the model using the 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.

In paper 3, we extend the method proposed in paper 2 to a more general setting where outcomes are measured multiple times per subject (e.g. longitudinal data analysis) and correlated within subject. We consider a probit stick-breaking process mixture (PSBPM) of linear mixed effects (LME) model. The PSBPM of LME model characterizes the conditional response distribution as predictor-dependent mixture of LME model which accounts for individual variability within each mixture component and induces nonlinear effects of predictors on the response distribution characteristics such as mean or quantiles. In addition, the model is formulated for conducting formal hypothesis testing of variable selection and goodness-of-fit for a LME model. The methods are illustrated through a simulation study and application to a German study of childhood growth.

CHAPTER 2

LITERATURE REVIEW

This chapter consists of literature review for: (1) the Dirichlet process (DP) as a prior model for a random probability measure, (2) various extensions of the DP as a prior model for a collection of predictor-dependent probability measures, (3) nonparametric Bayes estimation for predictor-dependent response distributions, (4) nonparametric Bayes hypothesis testing in predictor-dependent response distributions.

2.1 The Dirichlet Process (DP)

Bayesian inference involves placing distributions over variables in a statistical model. More flexibly, one can place a prior distribution over the space of distributions. The Dirichlet process (DP) is a popularly used prior model for an unknown distribution. In this section, the literature about the DP and its important properties are reviewed.

2.1.1 Definition

The Dirichlet distribution forms the first step toward understanding the Dirichlet process (DP). The Dirichlet distribution is a multi-parameter generalization of the Beta distribution. Consider a k-dimensional random vector $\mathbf{p} = \{p_1, \ldots, p_k\}$. The Dirichlet distribution on $\mathbf{p}$ is given by

$$P(\mathbf{p}|\alpha, M) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{k} \Gamma(\alpha m_i)} \prod_{i=1}^{k} p_i^{\alpha m_i - 1}, \tag{2.1}$$

where $M = \{m_1, \ldots, m_k\}$ is the mean value of $\mathbf{p}$ and α is a precision parameter that says how concentrated the distribution is around M. Both M and $\mathbf{p}$ sum to unity. α can be regarded as the number of pseudo-measurements observed to obtain M. The greater the number of pseudo-measurements is, the more our confidence in M is, and hence, the more the distribution is concentrated around M.

The Dirichlet distribution defines a distribution over a space of discrete distributions. Let $\mathbf{p} = \{p_1, \ldots, p_k\}$ be a probability distribution on the discrete space $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_k\}$ such that $P(X = \mathcal{X}_i) = p_i$, where X is a random variable in the space $\mathcal{X}$. Sampling a Dirichlet from (2.1) results in a distribution $\mathbf{p}$ on the discrete space $\mathcal{X}$. The Dirichlet distribution defined on a space of discrete probability measures on $\mathcal{X}$ can be noted as:

$$\mathbf{p}(\mathcal{X}_1), \ldots, \mathbf{p}(\mathcal{X}_k) \sim \text{Dirichlet}(\alpha m_1, \ldots, \alpha m_k) \tag{2.2}$$

If we consider a continuous sample space Θ and its disjoint partition such as $\Theta = \cup_{i=1}^{k} B_i$, it is apparent that a Dirichlet distribution exists on every disjoint partition of a continuous space Θ because the partition $\{B_1, \ldots, B_k\}$ is itself a discrete space.

Now consider a probability space $(\Theta, \mathcal{B}, G)$, where $\Theta \subset \Re^d$, $\mathcal{B}$ corresponds to the Borel σ-algebra of subsets of $\Re^d$, and G is a probability measure on $(\Theta, \mathcal{B})$. Also, consider another probability space $(\mathcal{G}, \mathcal{C}, \mathcal{P})$, where $\mathcal{G}$ is the space of probability measures G defined on $(\Theta, \mathcal{B})$ and $\mathcal{C}$ is the corresponding σ-algebra. Then, the DP with base measure G_0 and precision α, denoted as DP(αG_0), is a measure $\mathcal{P}$ defined on $(\mathcal{G}, \mathcal{C})$ such that $G(B_1), \ldots, G(B_k) \sim$ Dirichlet$(\alpha G_0(B_1), \ldots, \alpha G_0(B_k))$ for any disjoint partition $\{B_1, \ldots, B_k\}$ of Θ (Ferguson, 1973, 1974).

2.1.2 Polya urn scheme

The formal definition of the DP described in the previous section does not lend itself to an understanding of its distributional property. One way of understanding the DP more intuitively is connecting it to the Polya urn scheme (Blackwell and MacQueen, 1973; Escobar, 1994).

Consider an urn with α balls, of which initially αm_j are of color j, $1 \leq j \leq k$ (assuming for now that all the αm_j are integers). We draw balls at random from the urn, replacing each ball by two balls of the same color. Let $X_i = j$ if the ith ball is of color j. Then,

$$\begin{aligned} p(X_1 = j) &= \frac{\alpha m_j}{\alpha} \\ p(X_2 = j | X_1) &= \frac{\alpha m_j + 1(X_1 = j)}{\alpha + 1} \\ &\vdots \\ p(X_n = j | X_1, \ldots, X_{n-1}) &= \frac{\alpha m_j + \sum_{k=1}^{n-1} 1(X_k = j)}{\alpha + n - 1} \end{aligned} \tag{2.3}$$

We call this sequence $X_1, \ldots, X_n$ as a Polya urn sequence.

Let ϕ_i be ith sample from G with $G \sim \mathrm{DP}(\alpha G_0)$. Then, it was shown that marginalizing over G, ϕ_i is generated according to the following sequence:

$$\begin{aligned} \phi_1 &\sim G_0 \\ \phi_2 | \phi_1 &\sim \frac{\alpha G_0 + \delta_{\phi_1}}{\alpha + 1} \\ &\vdots \\ \phi_n | \phi_1, \ldots, \phi_{n-1} &\sim \frac{\alpha G_0 + \sum_{k=1}^{n-1} \delta_{\phi_k}}{\alpha + n - 1}, \end{aligned} \tag{2.4}$$

where δ_{ϕ_i} is a degenerate measure concentrated at ϕ_i. The sequence in (2.4) can be viewed as a Polya urn sequence by considering the limit as the number of colors in the Polya urn tends to a continuum. We call (2.4) the Polya urn scheme for the marginal distribution of a sample ϕ_i from a random probability measure G following a $\mathrm{DP}(\alpha G_0)$.

The Polya urn scheme of the DP results in a clustering structure amongst $\phi_1, \ldots, \phi_n$ with the following conditional distribution of each ϕ_i, given $\boldsymbol{\phi}^{(i)} = \{\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n\}$ (MacEachern, 1994; West et al., 1994).

$$\phi_i | \boldsymbol{\phi}^{(i)} \sim \left(\frac{\alpha}{\alpha + n - 1} \right) G_0 + \left(\frac{1}{\alpha + n - 1} \right) \sum_{j=1}^{k^{(i)}} n_j^{(i)} \delta_{\theta_j^{(i)}}, \tag{2.5}$$

where $\boldsymbol{\phi}^{(i)}$ takes on $k^{(i)}$ distinct values that are $\theta_j^{(i)}$ for $j = 1, \ldots, k^{(i)}$, and $n_j^{(i)}$ is the number of samples taking $\theta_j^{(i)}$ in $\boldsymbol{\phi}^{(i)}$. Similarly, the predictive distribution of a future ϕ_i for $i = n+1$ given $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_n\}$ follows:

$$\phi_{n+1} | \boldsymbol{\phi} \sim \left(\frac{\alpha}{\alpha + n} \right) G_0 + \left(\frac{1}{\alpha + n} \right) \sum_{j=1}^{k} n_j \delta_{\theta_j}, \tag{2.6}$$

where ϕ takes on k distinct values that are θ_j for $j = 1, \ldots, k$, and n_j is the number of samples taking θ_j in $\boldsymbol{\phi}$.

2.1.3 Stick-breaking representation

An important representation of the DP is the stick-breaking representation constructed by Sethuraman (1994). The random probability measure G sampled from a DP(αG_0) is represented as:

$$G = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}, \tag{2.7}$$

where $p_h = V_h \prod_{l=1}^{h-1} (1 - V_l)$ with $V_h \overset{iid}{\sim} Beta(1, \alpha)$ and $\theta_h \overset{iid}{\sim} G_0$ for $h = 1, \ldots, \infty$. The p_h are called stick-breaking random weights and θ_h called random atoms. It was shown that $\sum_{h=1}^{\infty} p_h \approx 1$ almost surely to ensure G is an appropriate probability measure. This stick-breading representation makes clear the fact that a realization G from a DP(αG_0) is a discrete distribution with infinitely many atoms. This representation forms the basis for developing efficient Gibbs sampling algorithm that doesn't require marginalization over G (discussed in section 2.1.5) and for extending of the DP to prior models for correlated random probability measures (reviewed in section 2.2.1).

2.1.4 Dirichlet process mixture (DPM)

The DP generates a random distribution that is almost surely discrete, which is problematic in modeling a continuous distribution. A simple solution to this problem is to use a Dirichlet process mixture (DPM) (Lo, 1984; Escobar, 1994; Escobar and West, 1995). Let y_i be ith subject's response following a continuous distribution F with unknown parameter ϕ_i, where ϕ_i follows an unknown probability measure G. By placing a DP prior for G, we model the marginal distribution F as a DPM model ensuring that y_i has a continuous distribution while still relaxing the distributional assumptions. The hierarchical structure of the DPM model is expressed as:

$$\begin{aligned} y_i|\phi_i &\sim F(\cdot;\phi_i) \\ \phi_i|G &\sim G(\cdot) \\ G|\alpha,\boldsymbol{\psi} &\sim DP(\alpha G_0(\cdot;\boldsymbol{\psi})), \end{aligned} \tag{2.8}$$

where $\boldsymbol{\psi}$ are the parameters of the parametric distribution G_0.

In recent years, with the availability of simple and efficient methods for posterior computation (see section 2.1.5), the DPM model has been widely used in many applications, which include finance (Kacperczyk et al., 2003), econometrics (Chib and Hamilton, 2002), epidemiology (Dunson, 2005), genomics (Xing et al., 2004; Kim et al., 2007), medicine (Kottas et al., 2002; Bigelow and Dunson, 2008), and machine learning (Beal et al., 2002 and Blei et al., 2004).

2.1.5 Posterior computation for DPMs

Analytic derivation of the posterior distributions for random quantities of interest is prohibited for the DPM models. Much of the DPM literature has been devoted to develop the Markov chain Monte Carlo (MCMC) techniques which allow sampling-based posterior inference in the DPM models.

There are two possible approaches in the MCMC techniques for the DPM models. The first one, called the marginal method, was introduced by Escobar (1994) and Escobar and West (1995), and has been substantially improved in MacEachern (1994), Müller et al. (1996) and particularly in MacEachern and Müller (1998) and Neal (2000). The marginal approach integrates out the random distribution G over the DP prior and uses convenient Polya urn representation within a Gibbs sampler to obtain posterior samples. Although simple to implement, this marginal method has a main drawback that a single-component updating Gibbs sampler is used to sample from a multivariate distribution of dependent variables, which results in problems in moving clusters around the posterior space, and therefore the mixing can be very slow even for moderately sized data sets. To improve the slow mixing problem, several MCMC samplers have been recently proposed based on the split-merge moves. Green and Richardson (2001) proposed one based on the reversible-jump procedure in which numerous ways to propose the split move are possible. Jain and Neal (2004) introduced a Metropolis-Hastings technique with split-merge proposals for conjugate DPM models and the idea was extended to a non-conjugate cases (Jain and Neal, 2007). Dahl (2003) suggested a sequentially-allocated split-merge sampler (SAMS) as an alternative to Jain and Neal (2004) approach.

Another MCMC tool, called the conditional approach or blocked Gibbs sampler, has been advocated by Ishwaran and Zarepour (2000) and Ishwaran and James (2001), who found that it can lead to considerably more robust convergence properties than the marginal approach. The conditional method, instead of integrating it out, augments the random distribution G and updates it as part of the MCMC algorithm. By doing so, the variables to be updated are partitioned in a small number of blocks, where the variables within each block are conditionally independent given the variables in other blocks, which leads to efficient updating of the variables as a block. This is advantageous over the marginal approach, which destroys conditional dependence structure. Moreover, we can directly obtain realizations of random distribution G, which allows for the inference on the underlying distribution G. However, because the Dirichlet process cannot be finitely represented, Ishwaran and Zarepour (2000) suggest to approximate it using a truncation of the random distribution for practical implementation, which produces error depending on the truncation accuracy. Avoiding such approximation, Papaspiliopoulos and Roberts (2004) designed an MCMC algorithm which samples from the exact posterior distribution of all quantities of interest based on the technique of retrospective sampling. More recently, Walker (2007) proposed slice sampling idea which avoids both marginalization and truncation.

Alternatively to the MCMC methods, other techniques for posterior inference in the DPM models have been proposed. This literature include sequential importance sampling (MacEachern et al., 1999) and variational methods (Blei and Jordan, 2006).

2.2 DP-Extended Priors for Dependent Probability Measures

The DP is a prior model for an unknown probability measure. However, modeling the relationship between predictors and the unknown probability measures cannot be achieved directly using the DP. In this section, the work for developing prior models for predictor-dependent unknown probability measures is reviewed.

2.2.1 Dependent Dirichlet process (DDP) and its variations

MacEachern (1999, 2000, 2001) proposed the dependent Dirichlet process (DDP) as a prior model for dependent distributions. Consider a collection of predictor-dependent random measures $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, where $\mathcal{X}$ is a predictor space. The DDP defines a probability distribution $G_{\mathbf{x}}$ for each $\mathbf{x}$ as $G_{\mathbf{x}} = \sum p_h \delta_{\theta_{\mathbf{x}h}}$, where p_h are stick-breaking random weights as in a DP and $\theta_{\mathbf{x}h}$ are predictor-dependent random atoms. In the "single-p DDP", a special case of the DDP, the weights p_h are common to all $G_{\mathbf{x}}$ while the dependence is induced across $\mathbf{x}$ by assuming that $\boldsymbol{\theta}_h = \{\theta_{\mathbf{x}h} : \mathbf{x} \in \mathcal{X}\}$ are iid realizations of a stochastic process in $\mathbf{x}$ (e.g. Gaussian process). Independence across h, together with the stick-breaking weights p_h, guarantees that $G_{\mathbf{x}}$ marginally follows a DP. Dependence in the sample path of the stochastic process $\boldsymbol{\theta}_h$

introduces the desired dependence across $\mathbf{x}$.

De Iorio et al. (2004) used the DDP structure to develop an ANOVA-like probability model over a collection of random distributions. They considered a categorical predictor $\mathbf{x}$, which, for simplicity of explanation, is bivariate as $\mathbf{x} = (v, w)'$ with $v \in \{1, \ldots, V\}$ and $w \in \{1, \ldots, W\}$. Using the DDP framework to induce an ANOVA-type dependence among $G_{\mathbf{x}}$, they defined $G_{\mathbf{x}} = \sum p_h \delta_{\theta_{\mathbf{x}h}}$, where $\theta_{\mathbf{x}h} = m_h + A_{vh} + B_{wh}$ with $m_h \overset{iid}{\sim} G^0_m$, $A_{vh} \overset{iid}{\sim} G^0_{Av}$, and $B_{wh} \overset{iid}{\sim} G^0_{Bw}$ for $v = 1, \ldots, V$ and $w = 1, \ldots, W$. They referred to this probability model as ANOVA-DDP(α, G^0) where $G^0 = (G^0_m, \{G^0_{Av}\}^V_{v=1}, \{G^0_{Bw}\}^W_{w=1})$.

Gelfand et al. (2005) applied the DDP framework to develop a spatial DP process (SDP) model for spatial data. They considered a point-referenced spatial data assumed to arise as a sample from a realization of a random field (random process) $Y_D = \{Y(s) : s \in D\}$ with $D \subset \Re^d$. Simply using the DDP idea, Y_D was modeled as a random spatial process denoted by $\sum_{h=1}^{\infty} p_h \delta_{\boldsymbol{\theta}_{h,D}}$, where p_h are the stick-breaking weights as in DP and $\boldsymbol{\theta}_{h,D} = \{\theta_h(s) : s \in D\}$ is a stochastic process G_0 over the D. Letting $s^{(n)} = (s_1, \ldots, s_n)$ be a specific distinct locations in D where the observations are collected, the full data set consists of the collection of vectors $Y = \{Y(s_1), \ldots, Y(s_n)\}$. Then, the SDP induces a random probability measure $G^{(n)}$ on the space of distribution functions for Y as $G^{(n)} \sim DP(\alpha G_0^{(n)})$, where $G_0^{(n)}$ is an n-variate distribution for Y induced by G_0. With the joint distribution using the same set of stick-breaking weights for any group of n locations, the SDP results in the common surface selection for all locations in the group.

The DDP structure also has been used by Dunson and Peddada (2008) to propose restricted dependent Dirichlet process (rDDP) which has a full support in the space of stochastically ordered random distributions. They considered a collection of probability measures $\{P_1, \ldots, P_k\}$ that belongs to the following convex subset of $\mathcal{P}^K$:

$$C_E = \{(P_1, \ldots, P_K) \subset \mathcal{P}^K : P_i \preceq P_j \quad \forall (i, j) \in E\}, \tag{2.9}$$

where $\mathcal{P}^K$ is the set of $K \times 1$ collections of probability measures on $(\mathcal{X}, \mathcal{B})$ and $E \subset (1, \ldots, K)^2$

is a partial ordering. Here, $P_i \preceq P_j$ denotes that P_j is stochastically larger than P_i such as $P_i(x,\infty) \leq P_j(x,\infty)$ for all x. As a prior for $(P_1,\ldots,P_K) \in C_E$, the proposed rDDP defines P_k as:

$$P_k = \sum_{h=1}^{\infty} p_h \delta_{\theta_{hk}}, \quad p_h = V_h \prod_{l<h}(1-V_l), \quad k = 1,\ldots,K, \tag{2.10}$$

where p_h are stick-breaking weights with $V_h \overset{iid}{\sim} \text{Beta}(1,\alpha)$ and $\boldsymbol{\theta}_h = (\theta_{h1},\ldots,\theta_{hK})' \overset{iid}{\sim} P_0$ are random atoms. Here, P_0 is a Borel probability measure defined on S_E, where $S_E = \{(s_1,\ldots,s_K) \in \mathcal{X}^K : s_i \leq s_j \quad \forall (i,j) \in E\}$.

Relaxing the fixed weight assumption in the DDP framework used so far, Griffin and Steel (2006) proposed an ordered DDP (πDDP) which results in predictor-dependent weights as well as predictor-dependent atoms. For a collection of predictor-dependent random distributions, $\mathcal{G}_{\mathcal{X}}$, the πDDP defines $G_{\mathbf{x}}$ as:

$$G_{\mathbf{x}} \;=\; \sum_{l=1}^{n(\mathbf{x})} p_l(\mathbf{x}) \delta_{\theta_{\pi_l(\mathbf{x})}}, \quad p_l(\mathbf{x}) = V_{\pi_l(\mathbf{x})} \prod_{j<l}(1 - V_{\pi_j(\mathbf{x})}), \quad \forall \mathbf{x} \in \mathcal{X} \tag{2.11}$$

where $\theta_h \overset{iid}{\sim} G_0$, $V_h \overset{iid}{\sim} \text{Beta}(1,\alpha)$, for $h = 1,\ldots,\infty$, and $\boldsymbol{\pi}(\mathbf{x}) = \{\pi_1(\mathbf{x}),\ldots,\pi_{n(\mathbf{x})}(\mathbf{x})\}$ is an ordering for the stick-breaking construction of $G_{\mathbf{x}}$ at the predictor point $\mathbf{x}$, which is assigned an ordering process $\{\boldsymbol{\pi}(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$.

Duan et al. (2006) also relaxed the fixed weight assumption in the Gelfand et al. (2005) SDP and proposed a generalized SDP (GSDP) as a multivariate generalization of the stick-breaking prior, which enables different surfaces to be assumed at different locations. The GSDP generates a spatial process for Y_D such that, for any set of locations $s^{(n)} \in D$,

$$Y(s_1),\ldots,Y(s_n) \sim \sum_{i_1=1}^{\infty} \ldots \sum_{i_n=1}^{\infty} p_{i_1,\ldots,i_n} \delta_{\theta^*_{i_1}(s_1)},\ldots,\delta_{\theta^*_{i_n}(s_n)}, \tag{2.12}$$

where the θ^*_j are independent and identically distributed as G_0, i_j is an abbreviation for $i(s_j)$, $j = 1,2,\ldots,n$, and the weights $\{p_{i_1,\ldots,i_n}\}$, conditionally on the locations, have a distribution defined on the infinite dimensional simplex:

$$\mathbf{P} = \{p_{i_1,\ldots,i_n} \geq 0 : \sum_{i_1=1}^{\infty} \ldots \sum_{i_n=1}^{\infty} p_{i_1,\ldots,i_n} = 1\}. \tag{2.13}$$

Another relevant extension of the DP is the hierarchical Dirichlet process (hDP) proposed by Teh et al. (2004). The hDP generates group-specific random distributions G_j from a DP(αG_0), where G_0 is drawn from another DP(γH). The fact that G_0 is discrete ensures that G_j are dependent across different groups through sharing the atoms. The stick-breaking forms of G_0 and G_j are more informative of this dependence structure as:

$$G_0 = \sum_{h=1}^{\infty} \beta_h \delta_{\theta_h}, \quad G_j = \sum_{h=1}^{\infty} p_{jh} \delta_{\theta_h}, \tag{2.14}$$

where β_h and θ_h are random stick-breaking weights and atoms from DP(γH) and p_{jh} are group-dependent random weights constructed based on the random weights β_h of DP(γH) and p_h of DP(α, G_0). This can be viewed as the case of group-dependent weights with fixed atoms in the DDP framework.

2.2.2 Convex combinations of DPs

A different approach that has been used for developing prior models inducing predictor-dependence is using linear combinations of realizations of Dirichlet process(es). Müller et al. (2004) proposed a prior model for k experiment-dependent distributions as a linear combination of one global component G_0^* and k experiment-specific innovation measures $G_1^*, \ldots, G_k^*$, with all G_h^* assigned independent $DP(\alpha_h, G_{0h})$, for $h = 0, \ldots, k$. The hth experiment-specific distribution is expressed as $G_h = \pi G_0^* + (1 - \pi) G_h^*$.

Dunson (2006) proposed a related approach, which incorporates information on ordering in an ordinal predictor and avoids the over-specification problem of Müller et al. (2004) approach, that is, $k + 1$ random measures are incorporated to characterize k unknown distributions. The proposed dynamic mixture of Dirichelt process (DMDP) defines the predictor-specific random distributions as:

$$\begin{aligned} G_1 &\sim DP(\alpha_0 G_0) \\ G_2 &= (1-\pi_1)G_1 + \pi_1 G_1^*, \\ &\vdots \\ G_h &= (1-\pi_{h-1})G_{h-1} + \pi_{h-1}G_{h-1}^* \\ &= w_{h1}G_1 + \sum_{l=1}^{h-1} w_{h,l+1}G_l^* \end{aligned} \tag{2.15}$$

where $G_1^* \sim DP(\alpha_1 G_{01})$, $G_l^* \overset{ind}{\sim} \mathrm{DP}(\alpha_l G_{0l})$ for $l = 1, \ldots, h-1$, $0 \leq \pi_l \leq 1$, $w_{hl} = \pi_{l-1}\prod_{m=1}^{h-1}(1-\pi_m)$ for $l = 1, \ldots, h-1$ and $w_{hl} = \pi_{h-1}$ for $l = h$ with $\pi_0 = 1$, G_l^* are innovation distributions that characterize changes in the distribution caused by increasing the predictor level from l to $l+1$, and G_{0l} are known distributions. As we move from predictor level from l to $l+1$, we decrease the weights assigned to $G_1, G_1^*, \ldots, G_{l-1}^*$ and introduce a new unknown distribution to our mixture, G_l^*. This evolution in G_h is reasonable in a situation where as the predictor level increases, more changes in the response distribution are expected.

Rodriguez et al. (2008) proposed another type of linear combination of DP realizations, called nested Dirichlet process (nDP), which defines hth group-dependent random distribution as $G_h = \sum_{l=1}^{\infty} \pi_l^* G_l^*$,where $\pi_l^* = v_l^* \prod_{j=1}^{l-1}(1 - v_j^*)$ with $v_j^* \overset{iid}{\sim} Beta(1, \alpha)$, and $G_l^* \overset{iid}{\sim} DP(\beta G_0)$. The i.i.d. realizations of a $DP(\beta G_0)$, G_l^*, are linearly combined with another DP stick-breaking weights π_l^*. When the nDP is considered as a prior for group-dependent mixture distributions in a mixture model, it induces clustering predictor groups while identifying clusters of subjects within each predictor group.

While the above three prior models were for a categorical predictor case, similar idea was used for a continuous predictor case. Dunson et al. (2006) proposed a weighted mixture of DPs (WMDP) which defines a random distribution at a particular predictor point $\mathbf{x}$ as:

$$G_{\mathbf{x}} = \sum_{j=1}^{n} b_j(\mathbf{x}) G_{\mathbf{x}_j}^*, \quad G_{\mathbf{x}_j}^* \overset{iid}{\sim} \mathrm{DP}(\alpha G_0), \quad \text{for} \quad j = 1, \ldots, n, \quad \forall \mathbf{x} \in \mathcal{X}, \tag{2.16}$$

where $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), \ldots, b_n(\mathbf{x}))'$ is a predictor-dependent weight function mapping from $\mathcal{X} \to \mathcal{P}_n$, with $\mathcal{P}_n$ being the n-dimensional probability simplex, so that $b_j(\mathbf{x}) \geq 0$, $j = 1, \ldots, n$, and $\mathbf{b}(\mathbf{x})'\mathbf{1}_n = 1$, for all $\mathbf{x} \in \mathcal{X}$. The collection $\mathcal{G}^*_{\mathbf{X}} = \{G^*_{\mathbf{x}_i}; i = 1, \ldots, n\}$ consists of i.i.d. realizations from a DP(αG_0) indexed by sampled predictor values $\mathbf{x}_i$. Hence, the WMDP for $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}}; \mathbf{x} \in \mathcal{X}\}$ is defined by placing a DP-distributed *basis* random measures at each of the sample predictor values, and then mixing across these *basis* measures to construct an uncountable collection of random probability measures for all possible predictor values $\mathbf{x} \in \mathcal{X}$. For the weight function $\mathbf{b}(\mathbf{x})$, they used a kernel-based weighting scheme such as:

$$b_j(\mathbf{x}) = \frac{\gamma_j K(\mathbf{x}, \mathbf{x}_j)}{\sum_{l=1}^n \gamma_l K(\mathbf{x}, \mathbf{x}_l)}, \quad j = 1, \ldots, n, \quad \forall \mathbf{x} \in \mathcal{X}, \tag{2.17}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)'$ represent weights on the different *basis* locations, and $K : \mathcal{X} \times \mathcal{X} \to \Re^+$ is a kernel function, such as $K(\mathbf{x}, \mathbf{x}') = \exp(-\psi||\mathbf{x} - \mathbf{x}'||^2)$.

Avoiding the sample dependence of the WMDP, which is problematic from a Bayesian perspective and results in some unappealing properties, Dunson and Park (2008) proposed a class of kernel stick-breaking process (KSBP) to be used as a sample-free prior for $\mathcal{G}_{\mathcal{X}}$, which induces a predictor-dependent prediction rule upon marginalization. The KSBP defines $G_{\mathbf{x}}$ as:

$$\begin{aligned} G_{\mathbf{x}} &= \sum_{h=1}^{\infty} P_h(\mathbf{x}) G^*_h, \\ P_h(\mathbf{x}) &= W(\mathbf{x}; V_h, \Gamma_h) \prod_{l<h} \{1 - W(\mathbf{x}; V_l, \Gamma_l)\} \\ W(\mathbf{x}; V_h, \Gamma_h) &= V_h K(\mathbf{x}, \Gamma_h), \quad \forall \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{2.18}$$

where $\Gamma_h \overset{iid}{\sim} H$ are random locations, $V_h \overset{ind}{\sim} \text{beta}(a_h, b_h)$ are probability weights, and $G^*_h \overset{iid}{\sim} \mathcal{Q}$ are probability measures, for $h = 1, \ldots, \infty$. Here, H is a probability measure defined on $\mathcal{X}'$ which may or may not correspond to $\mathcal{X}$, $\mathcal{Q}$ is a probability measure on a space of probability measures such as a DP, and $K : \Re^p \times \Re^p \to [0, 1]$ is a bounded kernel function, which is initially assumed to be known. Note that (2.18) formulates $G_{\mathbf{x}}$ as a predictor-dependent mixture over an infinite sequence of *basis* probability measures G^*_h that are located at Γ_h for $h = 1, \ldots, \infty$.

Bases located close to $\mathbf{x}$ and having a smaller index, h, will tend to receive higher probability weight. In this manner, the KSBP accommodates dependence between $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$.

2.3 Nonparametric Bayes Estimation for Predictor-Dependent Response Distributions

Recall that our goal is to flexibly characterize the relationship between a response $y \in \mathcal{Y}$ and predictors $\mathbf{x} = (x_1, \ldots, x_p)' \in \mathcal{X}$. The challenge is that a parametric form for the conditional distribution of y given $\mathbf{x}$ is unknown and the changes in the distributional shape needs to be addressed across the predictors $\mathbf{x}$. In this section, nonparametric Bayes methods for flexible conditional distribution modeling are reviewed.

2.3.1 DPM of regressions and its predictor-dependent extension

It is well known that a mixture of a sufficiently-large number of normal densities can approximate any smooth density. For a flexible characterization of the conditional density of a response given predictors, one can consider the following mixture of regression models:

$$f(y_i|\mathbf{x}_i) = \int f(y_i|\mathbf{x}_i, \phi_i) dG_{\mathbf{x}_i}(\phi_i), \tag{2.19}$$

where $f(y_i|\mathbf{x}_i, \phi_i) = N(y_i; \mathbf{x}_i'\boldsymbol{\beta}_i, \sigma_i^2)$, with $\phi_i = (\boldsymbol{\beta}_i, \sigma_i^2)$. Depending on the choice of $G_{\mathbf{x}_i}$, the model (2.19) encompasses a wide variety of flexible regression models as special cases including normal linear regression, linear regression with the residual density modeled as a finite/infinite mixture of normals, and a finite/infinite mixture of regressions.

In nonparametric Bayes methods, one choice of $G_{\mathbf{x}_i}$ is such that $G_{\mathbf{x}_i} \equiv G$ is assumed unknown and assigned a DP(αG_0) prior, under which the model (2.19) becomes an infinite mixture of regressions as:

$$f(y_i|\mathbf{x}_i) = \sum_{h=1}^{\infty} p_h N(y_i; \mathbf{x}_i'\boldsymbol{\beta}_h, \sigma_h^2), \tag{2.20}$$

where p_h are an infinite sequence of random stick-breaking weights and $(\boldsymbol{\beta}_h, \sigma_h^2)$ are random atoms i.i.d. sampled from G_0. The DPM of regressions in (2.20) is relatively flexible in that it incorporates an infinite number of normal linear regression components with a few components having most of the weights depending on the precision prior α, particularly when the number of mixture components is uncertain. However, assuming that the weights p_h are constant across the predictors still restricts the conditional density in that the shape of residual variation is the same across the predictors and the mean regression structure is linear as:

$$E(y_i|\mathbf{x}_i) = \sum_{h=1}^{\infty} p_h \mathbf{x}_i'\boldsymbol{\beta}_h = \sum_{h=1}^{\infty} p_h \sum_{j=1}^{p} x_{ij}\beta_{hj} = \sum_{j=1}^{p} x_{ij} \sum_{h=1}^{\infty} p_h \beta_{hj} = \mathbf{x}_i'\bar{\boldsymbol{\beta}}, \tag{2.21}$$

where $\bar{\boldsymbol{\beta}} = \sum_{h=1}^{\infty} p_h \beta_{hj}$.

Applying the prior models for a collection of predictor-dependent random distributions, the assumption $G_{\mathbf{x}_i} \equiv G$ in the DPM of regression model can be relaxed. As reviewed in section 2.2, any prior model $\mathcal{P}$ for $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}_i}\ \mathbf{x}_i \in \mathcal{X}\}$ can be incorporated in the mixture model specified in (2.19) as:

$$\begin{aligned} f(y_i|\mathbf{x}_i) &= \int f(y_i|\mathbf{x}_i, \phi_i) dG_{\mathbf{x}_i}(\phi_i), \\ \mathcal{G}_{\mathcal{X}} &= \{G_{\mathbf{x}_i}\ \mathbf{x}_i \in \mathcal{X}\}, \quad \mathcal{G}_{\mathcal{X}} \sim \mathcal{P}, \end{aligned} \tag{2.22}$$

where $f(y_i|\mathbf{x}_i, \phi_i) = N(y_i; \mathbf{x}_i'\boldsymbol{\beta}_i, \sigma_i^2)$, with $\phi_i = (\boldsymbol{\beta}_i, \sigma_i^2)$. Dunson et al. (2007), Griffin and Steel, (2006), and Dunson and Park (2008) applied their prior models to the model (2.22) and showed good performances in characterizing predictor-dependent response distributions.

2.3.2 DPM for the joint distribution of response and predictors

Alternatively, instead of assuming that the predictors to be included are known, Müller et al. (1996) proposed a different approach for flexible characterization of the conditional density of response given predictors. The joint density of response and predictors was modeled as a DP normal mixture model, which leads to the conditional density as an infinite mixture of regression models, with the mixture weights varying with predictors. Let y_i be ith subject's response and $\mathbf{x}_i$ be ith subject's predictors. For $\mathbf{z}_i = (y_i, \mathbf{x}_i')'$, the DP normal mixture model is expressed as:

$$\begin{aligned} f(\mathbf{z}_i) &= \int N(\mathbf{z}_i; \boldsymbol{\mu}_{zi}, \boldsymbol{\Sigma}_{zi}) dG(\boldsymbol{\mu}_{zi}, \boldsymbol{\Sigma}_{zi}) \\ G &\sim DP(\alpha G_0) \end{aligned} \tag{2.23}$$

This leads to $f(\mathbf{z}_i) = \sum_{h=1}^{\infty} p_h N(\mathbf{z}_i; \boldsymbol{\mu}_{zh}^*, \boldsymbol{\Sigma}_{zh}^*)$, which also leads to the conditional density $f(y_i|\mathbf{x}_i)$ as:

$$\begin{aligned} f(y_i|\mathbf{x}_i) &= \frac{f(\mathbf{z}_i)}{f(\mathbf{x}_i)} = \frac{\sum_{h=1}^{\infty} p_h N(\mathbf{z}_i; \boldsymbol{\mu}_{zh}^*, \boldsymbol{\Sigma}_{zh}^*)}{\sum_{h=1}^{\infty} p_h N(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}h}^*, \boldsymbol{\Sigma}_{\mathbf{x}h}^*)} \\ &= \frac{\sum_{h=1}^{\infty} p_h N(y_i; \mathbf{x}_i'\boldsymbol{\beta}_h^*, \sigma_h^{*2}) N(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}h}^*, \boldsymbol{\Sigma}_{\mathbf{x}h}^*)}{\sum_{h=1}^{\infty} p_h N(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}h}^*, \boldsymbol{\Sigma}_{\mathbf{x}h}^*)} \\ &= \sum_{h=1}^{\infty} p_h(\mathbf{x}_i) N(y_i; \mathbf{x}_i'\boldsymbol{\beta}_h^*, \sigma_h^{*2}), \end{aligned} \tag{2.24}$$

which is an infinite mixture of regression models, with the mixture weights $p_h(\mathbf{x}_i)$ varying with predictors, where $p_h(\mathbf{x}_i) = \frac{p_h N(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}h}^*, \boldsymbol{\Sigma}_{\mathbf{x}h}^*)}{\sum_{h=1}^{\infty} p_h N(\mathbf{x}_i; \boldsymbol{\mu}_{\mathbf{x}h}^*, \boldsymbol{\Sigma}_{\mathbf{x}h}^*)}$.

2.4 Nonparametric Bayes Hypothesis Testing in Predictor-Dependent Response Distributions

A flexible characterization of the relationship between a response and predictors involves not only the estimation of the predictor-specific response distribution but also the hypothesis testing of the distributional changes across the predictors or for model selection through the identification of significantly associated predictors both globally and locally. In this section, little work addressing the related problems is reviewed.

2.4.1 Model selection

Basu and Chib (2003) proposed an approach comparing semi-parametric models, constructed under the DPM framework, with alternative semi-parametric Bayesian models. They considered a model space $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$, where one (or more) of the models is a DPM model. Given a data $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, they suggested a formal Bayesian approach comparing any two models $\mathcal{M}_r$ and $\mathcal{M}_s$ using the Bayes factor which is the ratio of marginal likelihoods as:

$$\mathbf{B}_{rs} = \frac{m(\mathbf{y}|\mathcal{M}_r)}{m(\mathbf{y}|\mathcal{M}_s)} \tag{2.25}$$

For the DPM models, the calculation of marginal likelihoods required an integration over the space of the infinite dimensional parameter G, and an additional integration over the prior distribution of the hyper-parameters. Since the direct calculation is impossible, they based the marginal likelihood estimation on the approach of Chib (1995), which uses the representation of the marginal likelihood that is amenable to calculation by MCMC methods. The Chib (1995) approach required the estimation of both likelihood and posterior ordinates of the DPM model at a single high-density point. The posterior ordinate computation was simply based on the output produced by the MCMC algorithms while the computation for the likelihood ordinate was devised via collapsed sequentially importance sampling (SIS), which is a variant of the SIS method introduced by Kong et al. (1994). Although the method is innovative for the comparison of DPM models including covariates and hierarchical prior structure, it is not directly applicable to the local variable selection problem.

2.4.2 Testing for distributional changes

Pennell and Dunson (2007) proposed a method for testing for distribution changes across an ordinal predictor. They considered the predictor-specific response distribution as predictor-dependent mixture model using the DMDP prior (Dunson, 2006) for the mixture distributions. The model is expressed as:

$$\begin{aligned} f_h(\mathbf{y}_{hi}) &= \int N(\mathbf{y}_{hi}; \boldsymbol{\mu}_{hi}, \boldsymbol{\Sigma}_{hi}) dG_h(\boldsymbol{\mu}_{hi}, \boldsymbol{\Sigma}_{hi}) \\ \{G_1, \ldots, G_K\} &\sim DMDP(\pi_l, \alpha_l, G_{0l}, l = 1, \ldots, K-1) \end{aligned} \tag{2.26}$$

Here, it is immediately apparent that differences in mixture distributions, G_h and G_{h+1}, imply differences in the response distributions, f_h and f_{h+1}. Hence, the local null hypothesis was defined with respect to the total variation (tv) distances between two mixture distributions of adjacent predictor levels as follows:

$$H_{0h} : ||G_{h+1} - G_h||_{TV} \leq \epsilon, \quad \text{for} \quad h = 1, \ldots, K-1, \tag{2.27}$$

where ϵ is some small constant such that when H_{0h} holds, there is no appreciable difference in the mixture distributions across groups h and $h+1$. It was shown that the tv distance between G_{h+1} and G_h is totally controlled by π_h, so the local null can be equivalently stated as:

$$H_{0h} : \pi_h \leq \epsilon^*, \quad \text{for} \quad h = 1, \ldots, K-1, \tag{2.28}$$

Placing a prior $\pi_h \sim p_{0h}\text{Uniform}(0, \epsilon^*) + (1 - p_{0h})\text{Uniform}(\epsilon^*, 1)$ for π_h, for $h = 1, \ldots, K-1$, allowed for calculating the posterior probability for H_{0h}. The global null of no changes in response distribution across groups (H_0) corresponds to the intersection of theses local nulls.

Furthermore, Dunson and Peddada (2008) proposed a method for testing equalities of distributions against stochastically ordered alternatives in the rDDP mixture model framework (Refer to the summary in section 2.2.2). For a two group problem where $P_1 \preceq P_2$, they considered rDDP with P_0 chosen as:

$$f(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f_1(\boldsymbol{\theta}_1)\{\pi_0\delta_0(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) + (1-\pi_0)f_2(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)\}, \tag{2.29}$$

where $\boldsymbol{\mathcal{X}} = \Re$, $f_1(\cdot)$ is a density on $\Re$ (e.g. Gaussian), $0 \leq \pi_0 \leq 1$ is the prior probability of $\boldsymbol{\theta}_1 = \boldsymbol{\theta} = 2$, and $f_2(\cdot)$ is a density with support on $\Re^+$ (e.g. truncated Gaussian). To formally assess how close P_1 and P_2, they defined a distance metric as:

$$d_{12} = \max_{x \in \mathcal{X}} \left| P_2(x, \infty) - P_1(x, \infty) \right| \tag{2.30}$$

and formulate the null hypothesis that P_1 and P_2 are close in some sense to an alternative of stochastic ordering as:

$$H_0 : d_{12} \leq \epsilon \quad \text{and} \quad d_{12} > \epsilon, \tag{2.31}$$

where $\epsilon > 0$ is a small positive constant. The posterior probability for the null hypotheses were calculated based on the fact that $d_{12} = \sum_{h=1}^{\infty} p_h 1(\beta_h > 0) \sim \text{Beta}(\alpha(1-\pi_0), \alpha\pi_0)$. The idea was generalized to multiple group cases and the global null hypothesis was defined as:

$$\begin{aligned} H_0 &: \cup_{k=1}^{K-1} H_{0k} \\ \text{where} \quad H_{0k} &: d_{k,k+1} \leq \epsilon \quad \text{and} \quad H_{1k} d_{k,k+1} > \epsilon \end{aligned} \tag{2.32}$$

CHAPTER 3

THE LOCAL DIRICHLET PROCESS (LDP)

3.1 Introduction

In recent years, there has been a dramatic increase in applications of nonparametric Bayes methods, motivated largely by the availability of simple and efficient methods for posterior computation in Dirichlet process mixture (DPM) models (Lo, 1984; Escobar, 1994; Escobar and West, 1995). The DPM models incorporate Dirichlet process (DP) priors (Ferguson, 1973, 1974) for components in Bayesian hierarchical models, resulting in an extremely flexible class of models. Due to the flexibility and ease in implementation, DPM models are now routinely implemented in a wide variety of applications, ranging from machine learning (Beal et al., 2002 and Blei et al., 2004) to genomics (Xing et al., 2004 and Kim et al., 2006).

In many settings, it is natural to consider generalizations of the DP and DPM-based models to accommodate dependence. For example, one may be interested in studying changes in a density with predictors. Following Lo (1984), one can use a DPM for Bayes inference on a single density as follows:

$$f(y) = \int_{\Omega} k(y, u)\, G(du), \tag{3.1}$$

where $k(y, u)$ is a non-negative valued kernel defined on $(\mathcal{D} \times \Omega, \mathcal{F} \times \mathcal{B})$ such that for each $u \in \Omega$, $\int_{\mathcal{D}} k(y, u) dy = 1$ and for each $y \in \mathcal{D}$, $\int_{\Omega} k(y, u) G(du) < \infty$ with $\mathcal{D}, \Omega$ Borel subsets of Euclidean spaces and $\mathcal{F}, \mathcal{B}$ the corresponding σ-fields, and G is a finite random probability measure on $(\Omega, \mathcal{B})$ following a DP. A natural extension for modeling of a conditional density $f(y|\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}$, with $\mathcal{X}$ a Lebesgue measurable subset of $\Re^p$, is as follows:

$$f(y|\mathbf{x}) = \int_{\Omega} k(y, u)\, G_{\mathbf{x}}(du), \tag{3.2}$$

where the mixing measure $G_{\mathbf{x}}$ is now indexed by the predictor value. We are then faced with modeling a collection of random mixing measures denoted as $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$.

Recent work on defining priors for collections of random probability measures has primarily relied on extending the stick-breaking representation of the DP (Sethuraman, 1994). This literature was stimulated by the dependent Dirichlet process (DDP) framework proposed by MacEachern (1999, 2000, 2001), which replaces the atoms in the Sethuraman (1994) representation with stochastic processes. The DDP framework has been adopted to develop ANOVA-type models for random probability measures (De Iorio et al., 2004), for flexible spatial modeling (Gelfand et al., 2004), in time series applications (Caron et al., 2006), and for inferences on stochastic ordering (Dunson and Peddada, 2008). The specification of the DDP used in applications incorporates dependence only through the atoms while assuming fixed weights. In other recent work, Griffin and Steel (2006) proposed an order-based DDP (πDDP) which allows varying weights, while Duan et al. (2005) developed a multivariate stick-breaking process for spatial data.

Alternatively, convex combinations of independent DPs can be used for modeling collections of dependent random measures. Müller et al.(2004) proposed this idea to allow dependence across experiments and discrete dynamic settings were considered by Pennell and Dunson (2006) and Dunson (2006). Recently, the idea has been extended to continuous covariate cases by Dunson et al. (2007) and Dunson and Park (2008).

Some desirable properties of a prior for a collection, $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\}$, of predictor-dependent probability measures are: (1) increasing dependence in $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$ with decreasing distance between $\mathbf{x}$ and $\mathbf{x}'$; (2) simple and interpretable expressions for the expectation and variance of each $G_{\mathbf{x}}$ as well as the correlation between $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$; (3) $G_{\mathbf{x}}$ has a marginal DP prior for all $\mathbf{x} \in \mathcal{X}$; (4) posterior computation can proceed efficiently through a straightforward MCMC algorithm in a broad variety of applications. Although the DDP, πDDP and the prior proposed by Duan et al. (2005) achieve (1), πDDP and Duan et al. (2005) approaches are not straightforward to implement in general applications. The fixed stick-breaking weights version of the DDP tends to be easy to implement, but has the disadvantage of not allowing locally adaptive mixture weights. The kernel mixture approaches of Dunson et al. (2007) and Dunson and Park (2008) lack the marginal DP property (3). Property (3) is appealing in that there is rich theoretical literature on DPs, showing posterior consistency (Ghosal et al., 1999 and Lijoi et al., 2005) and rates of convergence (Ghosal and Van der Vaart, 2007).

This article proposes a simple extension of the DP, which provides an alternative to the fixed weights DDP in order to allow local adaptivity, while also achieving properties (1)-(4). The prior is constructed by first assigning stick-breaking weights and atoms to random locations in a predictor space. Each predictor-dependent random probability measure is formulated using the random weights and atoms located in a neighborhood about that predictor value. Dependence is induced by local sharing of random components. We call this prior the local Dirichlet process (lDP).

Section 2 describes stick-breaking priors for collections of predictor-dependent random probability measures. Section 3 introduces the lDP and discusses properties. Computation is described in section 4. Sections 5 and 6 include simulation studies and an epidemiologic application. Section 7 concludes with a discussion. Proofs are included in appendices.

3.2 Predictor-Dependent Stick-Breaking Priors

3.2.1 Stick-Breaking Priors

Ishwaran and James (2001) proposed a general class of stick-breaking priors for random probability measures. This class provides a useful starting point in considering extensions to allow predictor dependence.

Definition 1. A random probability measure, G, has a *stick-breaking prior* (SBP) if

$$G = \sum_{h=1}^{N} p_h \delta_{\theta_h}, \quad 0 \leq p_h \leq 1, \quad \sum_{h=1}^{N} p_h = 1 \quad a.s., \tag{3.3}$$

where δ_θ is a discrete measure concentrated at θ, $p_h = V_h \prod_{l<h}(1 - V_l)$ are random weights with $V_h \overset{ind}{\sim} Beta(a_h, b_h)$ independently from $\theta_h \overset{iid}{\sim} G_0$ with G_0 a non-atomic base probability measure. For $N = \infty$, the condition $\sum_{h=1}^{N} p_h = 1$ a.s. is satisfied by Lemma 1 in Ishwaran and James (2001). For finite N, the condition is satisfied by letting $V_N = 1$.

There are many processes that fall into this class of SBP. The DP corresponds to the special case in which $N = \infty$, $a_h = 1$ and $b_h = \alpha$ as established in Sethuraman (1994). The two-parameter Poisson-Dirichlet process corresponds to the case where $N = \infty$, $a_h = 1 - a$, and $b_h = b + ha$ with $0 \leq a < 1$ and $b > -a$ (Pitman 1995, 1996). Additional special cases are listed in Ishwaran and James (2001).

3.2.2 Predictor-Dependent Stick-Breaking Priors

Consider an uncountable collection of predictor-dependent random probability measures, $\mathcal{G}_\mathcal{X} = \{G_\mathbf{x} : \mathbf{x} \in \mathcal{X}\}$. The predictor space $\mathcal{X}$ is a Lebesgue measurable subset of Euclidian space and the random measures $G_\mathbf{x}$ are defined on $(\Omega, \mathcal{B})$ where Ω is a complete and separable metric space and $\mathcal{B}$ is a corresponding Borel σ-algebra. Let $\mathcal{P}$ be a probability measure on $(\mathcal{M}, \mathcal{N})$ where $\mathcal{M}$ is the space of uncountable collections of random probability measures $G_\mathbf{x}$ and $\mathcal{N}$ is the corresponding Borel σ-algebra. Then, $\mathcal{G}_\mathcal{X} \sim \mathcal{P}$ denotes that $\mathcal{P}$ is a prior on the random collection $\mathcal{G}_\mathcal{X}$.

We call $\mathcal{P}$ a *predictor-dependent stick-breaking prior* (SBP$_{\mathcal{X}}$) if $G_{\mathbf{x}} \in \mathcal{G}_{\mathcal{X}} \sim \mathcal{P}$ can be represented as:

$$\begin{aligned} G_{\mathbf{x}} &= \sum_{h=1}^{N(\mathbf{x})} p_h(\mathbf{x}) \delta_{\theta_h(\mathbf{x})} \\ \text{with} \quad 0 \leq p_h(\mathbf{x}) \leq 1 \qquad &\text{and} \qquad \sum_{h=1}^{N(\mathbf{x})} p_h(\mathbf{x}) = 1 \quad a.s., \quad \forall \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{3.4}$$

where the random weights $p_h(\mathbf{x})$ have a stick-breaking form, $p_h(\mathbf{x})$ and $\theta_h(\mathbf{x})$ are predictor-dependent, and $N(\mathbf{x})$ is also indexed by the predictor value $\mathbf{x}$. Depending on how we form $p_h(\mathbf{x})$, $\theta_h(\mathbf{x})$ and $N(\mathbf{x})$, different dependencies among $G_{\mathbf{x}}$ are induced. Several interesting priors, such as the DDP, πDDP and the prior proposed by Duan et al. (2005) fall into the SBP$_{\mathcal{X}}$ class. In the next section, we propose a new choice of SBP$_{\mathcal{X}}$ deemed the local Dirichlet process (lDP).

3.3 Local Dirichlet Process

3.3.1 Formulation

Formulating the local Dirichlet process (lDP) starts with obtaining the following three sequences of mutually independent global random components:

$$\mathbf{\Gamma} = \{\Gamma_h, h = 1, \ldots, \infty\}, \, \mathbf{V} = \{V_h, h = 1, \ldots, \infty\}, \, \mathbf{\Theta} = \{\theta_h, h = 1, \ldots, \infty\}, \tag{3.5}$$

where $\Gamma_h \overset{iid}{\sim} H$ are locations, $V_h \overset{iid}{\sim} Beta(1, \alpha)$ are probability weights, and $\theta_h \overset{iid}{\sim} G_0$ are atoms. G_0 is a probability measure on $(\Omega, \mathcal{B})$ on which $G_{\mathbf{x}}$ will be defined and H is a probability measure on $(\mathcal{X}', \mathcal{A})$ where $\mathcal{A}$ is a Borel σ-algebra of subsets of $\mathcal{X}'$ and $\mathcal{X}'$ is a Lebesgue measurable subset of Euclidian space that may or may not correspond to the predictor space $\mathcal{X}$. For a given predictor space $\mathcal{X}$, we introduce the probability space $(\mathcal{X}', \mathcal{A}, H)$ such that it satisfies the following regularity condition from which one can deduce $\mathcal{X} \subset \mathcal{X}'$:

Condition 1. For all $\mathbf{x} \in \mathcal{X}$ and $\psi > 0$, $H(\eta^{\psi}_{\mathbf{x}}) > 0$, where $\eta^{\psi}_{\mathbf{x}} = \{\mathbf{x}' : d(\mathbf{x}, \mathbf{x}') < \psi,\ \mathbf{x}' \in \mathcal{X}'\}$ is defined as a ψ-neighborhood around a point $\mathbf{x} \in \mathcal{X}$ with $d : \mathcal{X} \times \mathcal{X}' \to \Re^{+}$ being some distance measure.

Next, focusing on a local predictor point $\mathbf{x} \in \mathcal{X}$, we define sets of local random components for $\mathbf{x}$ as:

$$\boldsymbol{\Gamma}(\mathbf{x}) = \{\Gamma_h, h \in \mathcal{L}_{\mathbf{x}}\},\ \mathbf{V}(\mathbf{x}) = \{V_h, h \in \mathcal{L}_{\mathbf{x}}\},\ \boldsymbol{\Theta}(\mathbf{x}) = \{\theta_h, h \in \mathcal{L}_{\mathbf{x}}\}, \tag{3.6}$$

where $\mathcal{L}_{\mathbf{x}} = \{h : d(\mathbf{x}, \Gamma_h) < \psi,\ h = 1, \ldots, \infty\}$ is a predictor-dependent set indexing the locations belonging to the ψ-neighborhood of $\mathbf{x}$, $\eta^{\psi}_{\mathbf{x}}$, which is defined on $\mathcal{X}'$ by ψ and $d(\cdot,\cdot)$. Hence, the sets $\mathbf{V}(\mathbf{x})$ and $\boldsymbol{\Theta}(\mathbf{x})$ contain the random weights and atoms that are assigned to the locations $\boldsymbol{\Gamma}(\mathbf{x})$ in $\eta^{\psi}_{\mathbf{x}}$. Here, ψ controls the neighborhood size. For simplicity, we treat ψ as fixed throughout the paper, though one can obtain a more flexible class of priors by assuming a hyper prior for ψ.

Using the local random components in (3.6), we consider the following form for $G_{\mathbf{x}}$:

$$G_{\mathbf{x}} = \sum_{l=1}^{N(\mathbf{x})} p_l(\mathbf{x}) \delta_{\theta_{\pi_l(\mathbf{x})}} \quad \text{with} \quad p_l(\mathbf{x}) = V_{\pi_l(\mathbf{x})} \prod_{j<l} (1 - V_{\pi_j(\mathbf{x})}), \tag{3.7}$$

where $N(\mathbf{x})$ is the cardinality of $\mathcal{L}_{\mathbf{x}}$ and $\pi_l(\mathbf{x})$ is the lth ordered index in $\mathcal{L}_{\mathbf{x}}$. Then, condition 1 ensures that the following lemma holds (refer to the proof of lemma 1 in the appendix).

Lemma 1. For all $\mathbf{x} \in \mathcal{X}$, $N(\mathbf{x}) = \infty$ and $\sum_{l=1}^{N(\mathbf{x})} p_l(\mathbf{x}) = 1$ almost surely.

By Lemma 1, it is apparent that $G_{\mathbf{x}}$ formed as in (3.7) is a well-defined stick-breaking random probability measure for $\mathbf{x}$. It is also straightforward that we can define $G_{\mathbf{x}}$ for all $\mathbf{x} \in \mathcal{X}$ by (3.6) and (3.7) using the global components in (3.5). Therefore, given α, G_0, H, ψ with a choice of $d(\cdot,\cdot)$, the steps from (3.5) through (3.7) defines a new choice of predictor-dependent stick-breaking prior (SBP$_{\mathcal{X}}$) for $\mathcal{G}_{\mathcal{X}}$, deemed the local Dirichlet process (lDP). We use the shorthand notation $\mathcal{G}_{\mathcal{X}} = \{G_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\} \sim \text{lDP}(\alpha, G_0, H, \psi)$ to denote that $\mathcal{G}_{\mathcal{X}}$ is assigned a lDP with hyperparameters α, G_0, H, ψ.

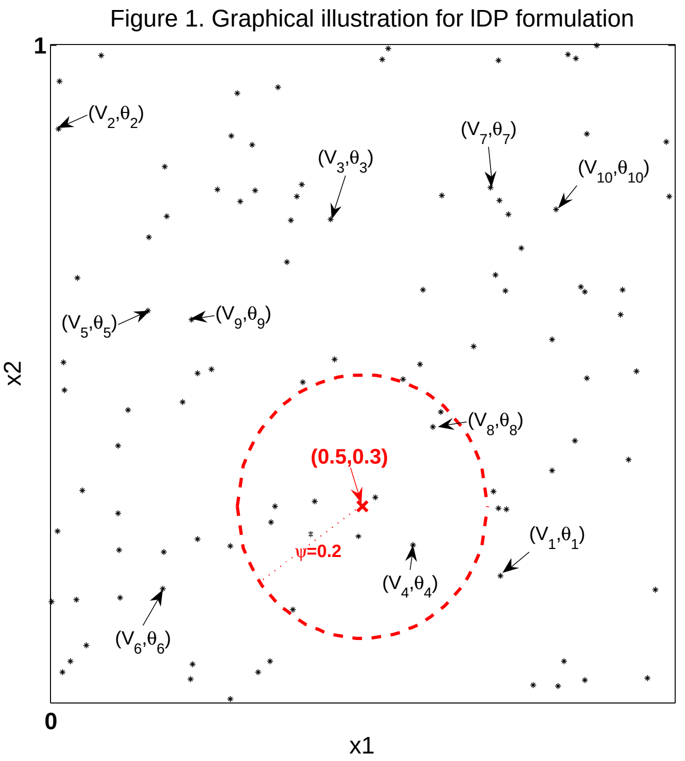

FIGURE 3.1: Black asterisks are the first 100 random locations generated on $\mathcal{X}' = [0,1]^2$ from H=Uniform([0,1]^2). Red dashed circle indicates the neighborhood of the red crossed predictor point $\mathbf{x} = (0.5, 0.3)'$ determined by Euclidian distance d($\cdot$) and $\psi = 0.2$. (V_h, θ_h) for $h = 1, \ldots, 10$ are the the first 10 random pairs of weight and atom assigned to the first 10 random locations Γ_h for $h = 1, \ldots, 10$.

Figure 3.1 illustrates the lDP formulation graphically for a case where $\mathcal{X} = [0,1]^2$ and $\mathcal{G}_{\mathcal{X}} \sim \text{lDP}(\alpha, G_0, H, \psi)$ with H=Uniform($[0,1]^2$) leading to $\mathcal{X} = \mathcal{X}'$ and $\psi = 0.2$. For a simple illustration, we consider Euclidian distance for $d(\cdot,\cdot)$ for bivariate predictors. Random locations in $[0,1]^2$ are generated from a uniform distribution, with the first 100 locations plotted as '$*$' in Figure 3.1. The random pair of weight and atom (V_h, θ_h) is placed at location Γ_h, with the first 10 pairs labeled in Figure 3.1. For a predictor value $\mathbf{x} = (0.5, 0.3)'$, the red dashed circle indicates the neighborhood of $\mathbf{x}$, $\eta^{\psi}_{\mathbf{x}}$. Then, $G_{\mathbf{x}}$ at $\mathbf{x} = (0.5, 0.3)'$ is constructed using the weights and atoms within the dashed circle in the order of the index to formulate the stick-breaking representation. For all other $\mathbf{x} \in \mathcal{X}$, $G_{\mathbf{x}}$ are formed following the same steps.

From Figure 3.1, it is apparent that the dependence between $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$ increases as the distance between $\mathbf{x}$ and $\mathbf{x}'$ decreases. For closer $\mathbf{x}$ and $\mathbf{x}'$, their neighborhoods overlap more so that similar components are used for constructing $G_{\mathbf{x}}$ and $G_{\mathbf{x}'}$, while if $\mathbf{x}$ and $\mathbf{x}'$ are far apart, there will be at most a small area of intersection so that few to none of the random components are shared. In the non-overlapping case, $G_{\mathbf{x}}$ and $G'_{\mathbf{x}}$ are assigned independent DP priors, as is clear from Theorem 1 and the subsequent development.

Theorem 1. If $\mathcal{G}_{\mathcal{X}} \sim lDP(\alpha, G_0, H, \psi)$, for any $\mathbf{x} \in \mathcal{X}$, $G_{\mathbf{x}} \sim DP(\alpha G_0)$.

The marginal DP property shown in Theorem 1 is appealing in allowing one to rely directly on the rich literature on properties of the DP to obtain insight into the prior for the random probability measure at any particular predictor value. However, unlike the DP, the lDP allows the probability measure to vary with predictors, while borrowing information across local regions of the predictor space. This is accomplished through incorporating shared random components. Due to the sharing and to the almost sure discreteness property of each $G_{\mathbf{x}}$, the lDP will induce local clustering of subjects according to their predictor values. Theorem 2 illustrates this local clustering property more clearly.

Theorem 2. Suppose $\mathcal{G}_{\mathcal{X}} \sim lDP(\alpha, G_0, H, \psi)$ and $\phi_i | G_{\mathbf{x}_i} \overset{ind}{\sim} G_{\mathbf{x}_i}$, for $i = 1, \ldots, n$, with $\mathbf{x}_i$ denoting the predictor value for subject i. Then,

$$\kappa_{\mathbf{x}_i,\mathbf{x}_j} = \Pr(\phi_i = \phi_j \,|\, \mathbf{x}_i, \mathbf{x}_j, \alpha, \psi) = \frac{2P_{\mathbf{x_i},\mathbf{x_j}}}{(1 + P_{\mathbf{x_i},\mathbf{x_j}})\alpha + 2}, \text{ for any } \mathbf{x}_i, \mathbf{x}_j \in \mathcal{X},$$

where $P_{\mathbf{x}_i,\mathbf{x}_j} = \frac{H(\eta^{\psi}_{\mathbf{x}_i} \bigcap \eta^{\psi}_{\mathbf{x}_j})}{H(\eta^{\psi}_{\mathbf{x}_i} \bigcup \eta^{\psi}_{\mathbf{x}_j})}$ is the conditional probability of Γ_h falling within the intersection region $\eta^{\psi}_{\mathbf{x}_i} \cap \eta^{\psi}_{\mathbf{x}_j}$ given $\Gamma_h \in \eta^{\psi}_{\mathbf{x}_i} \cup \eta^{\psi}_{\mathbf{x}_j}$.

The clustering probability $\kappa_{\mathbf{x}_i,\mathbf{x}_j}$ increases from 0 when $\eta^{\psi}_{\mathbf{x}_i} \bigcap \eta^{\psi}_{\mathbf{x}_j} = \emptyset$ to $1/(\alpha+1)$ when $\mathbf{x}_i = \mathbf{x}_j$ which is the case of $P_{\mathbf{x}_i,\mathbf{x}_j} = 1$. This implies that, for fixed α, the clustering probability under $\mathcal{G}_{\mathcal{X}} \sim \text{lDP}(\alpha, G_0, H, \psi)$ is bounded above by the clustering probability under the global DP, which takes $G_{\mathbf{x}} \equiv G \sim DP(\alpha G_0)$, leading to $\Pr(\phi_i = \phi_j \,|\, \alpha) = 1/(\alpha+1)$. Also, note that small values of the precision parameter α will induce V_h values that are close to one. This in turn causes a small number of atoms in each neighborhood to dominate, inducing few local clusters. However, when ψ is small and hence neighborhood sizes are small, there will still be many clusters across $\mathcal{X}$.

It is interesting to consider relationships between the lDP and other priors proposed in the literature in limiting special cases. First, note that the lDP converges to the DP as $\psi \rightarrow \infty$, so that all the neighborhoods around each of the predictor values encompass the entire predictor space. Also, the $\text{lDP}(\alpha, G_0, H, \psi)$ corresponds to a limiting case of the kernel stick-breaking process (KSBP) (Dunson and Park, 2008), in which the kernel is defined as $K(\mathbf{x}, \Gamma) = 1\big(d(\mathbf{x}, \Gamma) < \psi\big)$ and the DP placed at each location have precision parameters $\rightarrow 0$.

3.3.2 Moments and Correlation

From Theorem 1 and properties of the DP, $\mathcal{G}_{\mathcal{X}} \sim lDP(\alpha, G_0, H, \psi)$ implies, for any $\mathbf{x} \in \mathcal{X}$,

$$E\{G_{\mathbf{x}}(B)\} = G_0(B) \quad \text{and} \quad Var\{G_{\mathbf{x}}(B)\} = \frac{G_0(B)(1 - G_0(B))}{1+\alpha}, \quad \forall B \in \mathcal{B} \tag{3.8}$$

Next, let us consider the correlation between $G_{\mathbf{x}_1}$ and $G_{\mathbf{x}_2}$, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$. First, we show the correlation conditionally on the locations $\mathbf{\Gamma}$ but marginalizing out the weights $\mathbf{V}$ and atoms $\mathbf{\Theta}$. As discussed in section 3.1, if $\mathbf{\Gamma}$ is given, the lDP can be regarded as a special case of the πDDP. Hence, following Theorem 1 in Griffin and Steel (2006), for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,

$$\begin{aligned} \rho_{\mathbf{x}_1,\mathbf{x}_2}(\mathbf{\Gamma}) &= Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B) | \mathbf{\Gamma}\} \\ &= \frac{2}{\alpha+2} \sum_{h \in \mathcal{L}_{\mathbf{x}_1} \cap \mathcal{L}_{\mathbf{x}_2}} \left(\frac{\alpha}{\alpha+2}\right)^{\#\mathcal{S}_h} \left(\frac{\alpha}{\alpha+1}\right)^{\#\mathcal{S}'_h}, \quad \forall B \in \mathcal{B}, \end{aligned} \tag{3.9}$$

where $\#\mathcal{S}$ is the cardinality of the set $\mathcal{S}$, $\mathcal{S}_h = \mathcal{A}_{1h} \cap \mathcal{A}_{2h}$, $\mathcal{S}'_h = \mathcal{A}_{1h} \cup \mathcal{A}_{2h} - \mathcal{S}_h$, and $\mathcal{A}_{kh} = \{\pi_j(\mathbf{x}_k) : j < l, \pi_l(\mathbf{x}_k) = h\}$ for $h \in \mathcal{L}_{\mathbf{x}_1} \cap \mathcal{L}_{\mathbf{x}_2}$. In other words, $\#\mathcal{S}_h$ is the number of indices on the locations $\mathbf{\Gamma}$ that are below h and are shared in the neighborhoods of $\mathbf{x}_1$ and $\mathbf{x}_2$, while $\#\mathcal{S}'_h$ is the number of indices that are below h and belong to the neighborhoods of either $\mathbf{x}_1$ or $\mathbf{x}_2$ but not both. For a given h, reducing $\#S_h$ by one induces adding two elements to S'_h, thus reducing the correlation, as expected. From expression (3.9), it is clear that the neighborhoods around $\mathbf{x}_1$ and $\mathbf{x}_2$ are increasingly overlapping and the correlation between $G_{\mathbf{x}_1}$ and $G_{\mathbf{x}_2}$ increases as $\mathbf{x}_1 \to \mathbf{x}_2$. Expression (3.9) is particularly useful in being free of dependence on B.

Marginalizing the correlation in (3.9) over the prior for the random locations $\mathbf{\Gamma}$ is equivalent to marginalizing out the $\#\mathcal{S}_h$ and $\#\mathcal{S}'_h$ for $h \in \mathcal{L}_{\mathbf{x}_1} \cap \mathcal{L}_{\mathbf{x}_2}$. In considering the correlation between $G_{\mathbf{x_1}}$ and $G_{\mathbf{x_2}}$, we can ignore the Γ_h for $h \in \{1, \ldots, \infty\} \setminus \mathcal{L}_{\mathbf{x}_1} \cup \mathcal{L}_{\mathbf{x}_2}$ and focus on the Γ_h only for $h \in \mathcal{L}_{\mathbf{x}_1} \cup \mathcal{L}_{\mathbf{x}_2}$. Let γ_j be the jth ordered component of $\mathcal{L}_{\mathbf{x}_1} \cup \mathcal{L}_{\mathbf{x}_2}$. For example, if $\mathcal{L}_{\mathbf{x}_1} \cup \mathcal{L}_{\mathbf{x}_2} = \{1, 3, 5, 6, 7, 8, \ldots\}$, $\gamma_1 = 1$, $\gamma_2 = 3, \gamma_3 = 5, \gamma_4 = 6, \cdots$. Let $Z_{\gamma_j} = 1(\gamma_j \in \mathcal{L}_{\mathbf{x}_1} \cap \mathcal{L}_{\mathbf{x}_2})$ be an indicator for whether Γ_{γ_j} are shared by the neighborhoods of $\mathbf{x}_1$ and $\mathbf{x}_2$ or not. Then, the formula in (3.9) can be reexpressed with respect to Z_{γ_j} as follows:

$$\begin{aligned}
\rho_{\mathbf{x}_1,\mathbf{x}_2}(\mathbf{\Gamma}) &= Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B)|\mathbf{\Gamma}\} \\
&= \frac{2}{\alpha+2}\sum_{j=1}^{\infty} Z_{\gamma_j}\left(\frac{\alpha}{\alpha+2}\right)^{\sum_{k=1}^{j-1} Z_{\gamma_k}}\left(\frac{\alpha}{\alpha+1}\right)^{j-1-\sum_{k=1}^{j-1} Z_{\gamma_k}} \qquad (3.10)
\end{aligned}$$

Note that it is straightforward to show that $Z_{\gamma_j} \overset{iid}{\sim} \text{Bernoulli}(P_{\mathbf{x}_1,\mathbf{x}_2})$, for $j = 1, \ldots, \infty$, with $P_{\mathbf{x}_1,\mathbf{x}_2} = \frac{H(\eta^{\psi}_{\mathbf{x}_1} \bigcap \eta^{\psi}_{\mathbf{x}_2})}{H(\eta^{\psi}_{\mathbf{x}_1} \bigcup \eta^{\psi}_{\mathbf{x}_2})}$ the conditional probability of Γ_h falling within the intersection region $\eta^{\psi}_{\mathbf{x}_1} \cap \eta^{\psi}_{\mathbf{x}_2}$ given $\Gamma_h \in \eta^{\psi}_{\mathbf{x}_1} \cup \eta^{\psi}_{\mathbf{x}_2}$. Finally, marginalizing out $\{Z_{\gamma_j}\}_{j=1}^{\infty}$ results in the following Theorem.

Theorem 3. If $\mathcal{G}_{\mathcal{X}} \sim lDP(\alpha, G_0, H, \psi)$, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$,

$$\rho_{\mathbf{x}_1,\mathbf{x}_2} = Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B)\} = \frac{2P_{\mathbf{x}_1,\mathbf{x}_2}(\alpha+1)}{(1+P_{\mathbf{x}_1,\mathbf{x}_2})\alpha+2}, \quad \forall B \in \mathcal{B}$$

The correlation is expressed only in terms of $P_{\mathbf{x}_1,\mathbf{x}_2}$ and α. Regardless of α, the correlation is

1 if $\mathbf{x}_1 = \mathbf{x}_2$ which implies the neighborhoods around $\mathbf{x}_1, \mathbf{x}_2$ are identical and $P_{\mathbf{x}_1,\mathbf{x}_2}$=1. Also, the correlation is 0 when the neighborhoods are non-overlapping with $P_{\mathbf{x}_1,\mathbf{x}_2}$=0. In addition, $P_{\mathbf{x}_1,\mathbf{x}_2} \leq \rho_{\mathbf{x}_1,\mathbf{x}_2} \leq 1$ and $\rho_{\mathbf{x}_1,\mathbf{x}_2}$ increases as α increases for fixed $P_{\mathbf{x}_1,\mathbf{x}_2}$. When $\alpha \to 0$, the correlation converges to $P_{\mathbf{x}_1,\mathbf{x}_2}$. Meanwhile, when $\alpha \to \infty$, the correlation converges to $\frac{2P_{\mathbf{x}_1,\mathbf{x}_2}}{1+P_{\mathbf{x}_1,\mathbf{x}_2}}$.

Note that $P_{\mathbf{x}_1,\mathbf{x}_2}$ depends on H, ψ, and the locations $\mathbf{x}_1$ and $\mathbf{x}_2$ given a choice of $d(\cdot,\cdot)$. When $\mathcal{X}'$ for H is chosen to satisfy Condition 2, some appealing properties result.

Condition 2. For all $\mathbf{x} \in \mathcal{X}$ with $\mathcal{X}$ being p-dimensional, $\{\mathbf{x}^*; d(\mathbf{x}^*, \mathbf{x}) < \psi, \mathbf{x}^* \in \Re^p\} \subset \mathcal{X}'$.

From condition 2, one can deduce that $\mathcal{X}'$ contains all the points in $\Re^p$ within the distance of ψ from $\mathbf{x}$ for any $\mathbf{x} \in \mathcal{X}$. Under condition 2, with H chosen to be a uniform probability measure on a bounded space $\mathcal{X}'$, $P_{\mathbf{x}_1,\mathbf{x}_2}$ depends only on ψ and $d(\mathbf{x}_1, \mathbf{x}_2)$ which is the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$, but not on the exact locations of $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathcal{X}$. Hence, upon examination of Theorem 3, it is apparent that condition 2 implies an isotropic correlation structure, which is an appealing default in the absence of prior knowledge of changes in the correlation structure according to the locations in $\mathcal{X}$. Figure 3.2 shows how the correlation $\rho_{\mathbf{x}_1,\mathbf{x}_2}$ changes as a function of $d(\mathbf{x}_1, \mathbf{x}_2)$ in the case where $\mathbf{x} \in \mathcal{X} = [0, 1]$ and H is Uniform($[-\psi, 1 + \psi]$) so that $\mathcal{X}' = [-\psi, 1 + \psi]$ and condition 2 holds for different ψ with $d(\cdot,\cdot)$ corresponding to the Euclidian distance. The $\rho_{\mathbf{x}_1,\mathbf{x}_2}$ decays from 1 to 0 as $d(\mathbf{x}_1, \mathbf{x}_2)$ increases and the decay is faster for smaller ψ. As $\psi \to \infty$, the decay line gets closer to a horizontal line at $\rho_{\mathbf{x}_1,\mathbf{x}_2} = 1$, which is the case of lDP=DP. Also, for a given ψ and $d(\mathbf{x}_1, \mathbf{x}_2)$, the $\rho_{\mathbf{x}_1,\mathbf{x}_2}$ is higher as $\alpha \to \infty$. Although the choice of $d(\cdot,\cdot)$ being Euclidian makes the curves in Figure 3.2 close to linear, the curvature can easily be changed by choosing a different distance measure $d(\cdot,\cdot)$.

3.3.3 Truncation Approximation

Finite approximations to infinite stick-breaking priors form the basis for commonly-used computational algorithms (Ishwaran and James, 2001). In this subsection, we discuss a finite dimensional approximation to the lDP.

Since the lDP has the marginal DP property, let us recall the finite dimensional DP. Ishwaran

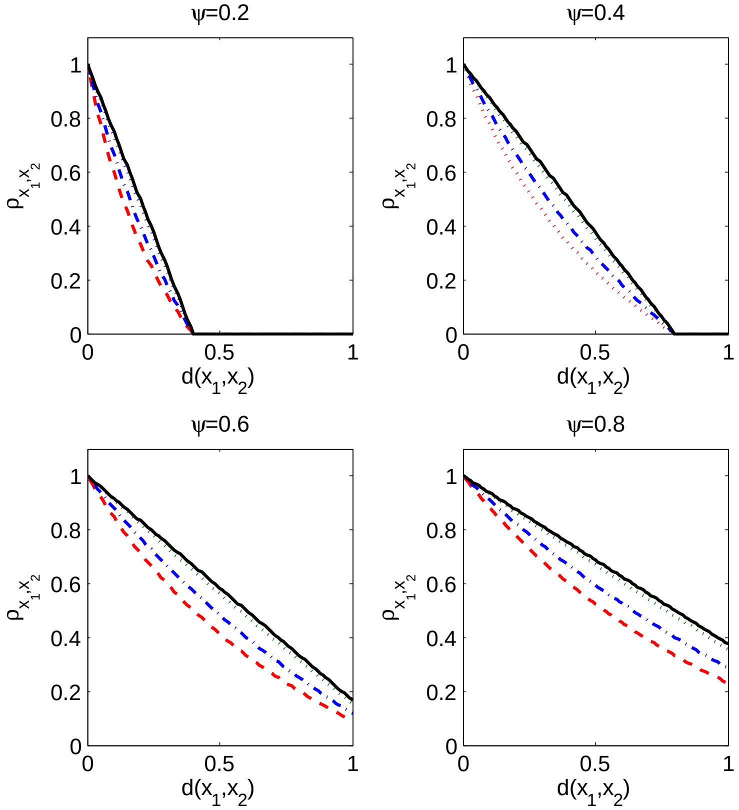

FIGURE 3.2: Change in correlation $\rho_{\mathbf{x}_1,\mathbf{x}_2}$ over the change in distance $d(\mathbf{x}_1,\mathbf{x}_2)$ for different α and ψ: $\alpha = 0.0001$ (red dashed), $\alpha = 1$ (blue dot-dashed), $\alpha = 10$ (green dotted), $\alpha = 10000$ (black solid).

and James (2001) defines an N-truncation of the DP (DP^N) by discarding the $N+1, N+2, \ldots, \infty$ terms and replacing p_N with $1 - \sum_{h=1}^{N-1} p_h$ in the DP stick-breaking form in (3.3). They show that the DP^N approximates the DP well in terms of the total variation (tv) norm of the marginal densities of the data obtained from the corresponding DPM models. According to their Theorem

2,

$$||\mu_N - \mu_\infty|| \leq 4\left[1 - E\left\{\left(\sum_{h=1}^{N-1} p_h\right)^n\right\}\right] \approx 4n \times exp\{-(N-1)/\alpha\}, \tag{3.11}$$

where $||\cdot||$ is tv norm, μ_N and μ_∞ are the marginal probability measures for the data from the DPM^N and DPM models, and n is the sample size. Note that the sample size has a modest effect on the bound for a reasonably large value of N and the bound decreases exponentially with N increasing, implying that even for a fairly large sample size, the DPM^N approximates the DP well with moderate N.

Following a similar route, let us define an N-truncation of the lDP (lDP^N) as follows:

Definition 3. For a finite N, let $\mathbf{\Gamma}^N = \{\Gamma_h, h = 1, \ldots, N\}$, $\mathbf{V}^N = \{V_h, h = 1, \ldots, N\}$, and $\mathbf{\Theta}^N = \{\theta_h, h = 1, \ldots, N\}$ be the sets of global random locations, weights, and atoms, respectively. Distributional assumptions for Γ_h, V_h, and θ_h are the same as in (3.5) and the corresponding local sets are defined as in (3.6). Then, $\mathcal{G}_\mathcal{X} \sim lDP^N(\alpha, G_0, H, \psi)$ if

$$\begin{aligned} G_\mathbf{x} &= \sum_{l=1}^{N(\mathbf{x})-1} p_l(\mathbf{x})\delta_{\theta_{\pi_l(\mathbf{x})}} + \left(1 - \sum_{l=1}^{N(\mathbf{x})-1} p_l(\mathbf{x})\right)\delta_{\theta_{\pi_{N(\mathbf{x})}(\mathbf{x})}} \\ \text{with} \quad p_l(\mathbf{x}) &= V_{\pi_l(\mathbf{x})}\prod_{j<l}(1 - V_{\pi_j(\mathbf{x})}) \quad \text{for} \quad l = 1, \ldots, N(\mathbf{x}) - 1 \end{aligned}$$

The $G_\mathbf{x}$ in Definition 3 has a similar form to $G = \sum_{h=1}^{N} p_h\delta_{\theta_h}$ obtained from the DP^N except that N in G is replaced by $N(\mathbf{x})$ in $G_\mathbf{x}$ and N in DP^N is fixed while $N(\mathbf{x})$ in lDP^N is random. Focusing on a particular predictor value $\mathbf{x}$, it is easy to show that $N(\mathbf{x}) \sim$ Binomial(N,$P_\mathbf{x}$), where N is the total number of global locations in lDP^N and $P_\mathbf{x} = H(\eta_\mathbf{x}^\psi)$ is the probability that a location belongs to the neighborhood around $\mathbf{x}$, $\eta_\mathbf{x}^\psi$. Then, marginalizing out $N(\mathbf{x})$ in the bound on the tv distance between the marginal densities of an observation obtained at a particular predictor value $\mathbf{x}$ from the lDPM and lDPM^N models results in Theorem 4.

Theorem 4. Define a model (3.2) with $\mathcal{G}_\mathcal{X} \sim$ lDP(α, G_0, H, ψ) as local Dirichlet process mixture (lDPM) model. lDPM^N corresponds to (3.2) with $\mathcal{G}_\mathcal{X} \sim$ lDP$^N(\alpha, G_0, H, \psi)$. Suppose an observation is obtained from lDPM^N and lDPM models at $\mathbf{x}$. Then,

$$||\mu_N(\mathbf{x}) - \mu_\infty(\mathbf{x})|| \;\leq\; 4\bigg(\frac{\alpha+1}{\alpha}\bigg)\bigg\{1 - \bigg(\frac{1}{\alpha+1}\bigg)P_{\mathbf{x}}\bigg\}^N,$$

where $\mu_N(\mathbf{x})$ and $\mu_\infty(\mathbf{x})$ are the marginal probability measures for the observation. Notice that the bound decreases exponentially with N increasing, suggesting that we can obtain a good approximation to the lDP using a moderate N, as long as α is small and the neighborhood size is not too small. In particular, we require a large N for a given level of accuracy as $\psi \to 0$, since $P_{\mathbf{x}}$ decreases as the size of $\eta_{\mathbf{x}}^{\psi}$ decreases.

3.4 Posterior Computation

We develop an MCMC algorithm based on the blocked Gibbs sampler (Ishwaran and James, 2001) for an lDPM^N model. For simplicity in exposition, we describe a Gibbs sampling algorithm for a particular hierarchical model, though the approach can be easily adapted for computation in a broad variety of other settings. We let

$$\begin{aligned} f(y_i \,|\, \mathbf{x}_i, \tau) &= \int f(y_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}_i, \tau)\, dG_{\mathbf{x}_i}(\boldsymbol{\beta}_i) \quad \text{for } i = 1, \ldots, n \\ \mathcal{G}_{\mathcal{X}} &\sim lDP^N(\alpha, G_0, H, \psi), \end{aligned} \tag{3.12}$$

where $f(y_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}_i, \tau) = N(y_i; {\mathbf{x}_i}'\boldsymbol{\beta}_i, \tau^{-1})$ with $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})'$. For simplicity, we consider a univariate predictor case where $p = 2$ and $\mathbf{x}_i' = (1, x_i)$ with $d(\cdot,\cdot)$ Euclidian distance but the generalization to multiple predictors or to using different distance metric is straightforward. G_0 is assumed to be $N_p(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, H is assumed to be Uniform(a_Γ, b_Γ) and additional conjugate priors are assigned for τ, α, $\boldsymbol{\mu}_\beta$ and $\boldsymbol{\Sigma}_\beta$.

Let K_i be an indicator variable denoting that $K_i = h$ implies ith subject is assigned to the hth mixture component. Then, the hierarchical structure of the model (3.12) with respect to the random variables is recast as follows.

$$
\begin{aligned}
(y_i|\mathbf{x}_i, \boldsymbol{\beta}^*, \tau, \mathbf{K}) &\sim N(\mathbf{x}_i{}'\boldsymbol{\beta}^*_{K_i}, \tau^{-1}), \quad i = 1, \ldots, n \\
(K_i|\mathbf{V}, \boldsymbol{\Gamma}) &\sim \sum_{l=1}^{N(\mathbf{x}_i)} p_l(\mathbf{x}_i)\delta_{\pi_l(\mathbf{x}_i)}(\cdot), \quad i = 1, \ldots, n \\
(V_h|\alpha) &\sim \text{Beta}(1, \alpha), \quad h = 1, \ldots, N \\
(\Gamma_h) &\sim \text{Uniform}(a_\Gamma, b_\Gamma), \quad h = 1, \ldots, N \\
(\boldsymbol{\beta}^*_h|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) &\sim N_p(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), \quad h = 1, \ldots, N \\
\boldsymbol{\mu}_\beta &\sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_\mu) \\
\boldsymbol{\Sigma}_\beta^{-1} &\sim \text{Wishart}(\{\nu_0\boldsymbol{\Sigma}_0\}^{-1}, \nu_0) \\
\tau &\sim \text{Gamma}(\nu_1, \nu_2) \\
\alpha &\sim \text{Gamma}(\eta_1, \eta_2), \qquad (3.13)
\end{aligned}
$$

where $\boldsymbol{\beta}^* = \{\boldsymbol{\beta}^*_h, h = 1, \ldots, N\}$, $\mathbf{K} = \{K_i, i = 1 \ldots, n\}$, $\mathbf{V} = \{V_h, h = 1, \ldots, N\}$, and $\boldsymbol{\Gamma} = \{\Gamma_h, h = 1, \ldots, N\}$. The full conditionals for each of the random components are based on the following joint distribution.

$$
\begin{aligned}
&(\mathbf{y}, \mathbf{K}, \mathbf{V}, \boldsymbol{\Gamma}, \boldsymbol{\beta}^*, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta, \tau, \alpha) \\
&\quad \propto (\mathbf{y}|\boldsymbol{\beta}^*, \tau, \mathbf{K})(\mathbf{K}|\mathbf{V}, \boldsymbol{\Gamma})(\mathbf{V}|\alpha)(\boldsymbol{\Gamma})(\boldsymbol{\beta}^*|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)(\boldsymbol{\mu}_\beta)(\boldsymbol{\Sigma}_\beta)(\tau)(\alpha) \qquad (3.14)
\end{aligned}
$$

Then, the Gibbs sampler proceeds by sampling from the following conditional posterior distributions:

(a) Conditional for $K_i, i = 1, \ldots, n$

$$\begin{aligned}
(K_i|\mathbf{y}, \mathbf{V}, \boldsymbol{\Gamma}, \boldsymbol{\beta}^*, \tau) &\sim \sum_{l=1}^{N(\mathbf{x}_i)} p_l'(\mathbf{x}_i)\delta_{\pi_l(\mathbf{x}_i)}(K_i) \\
p_l'(\mathbf{x}_i) &= \frac{N(y_i; \mathbf{x}_i'\boldsymbol{\beta}^*_{\pi_l(\mathbf{x}_i)}, \tau^{-1})p_l(\mathbf{x}_i)}{\sum_{l=1}^{N(\mathbf{x}_i)} N(y_i; \mathbf{x}_i'\boldsymbol{\beta}^*_{\pi_l(\mathbf{x}_i)}, \tau^{-1})p_l(\mathbf{x}_i)} \\
p_l(\mathbf{x}_i) &= V_{\pi_l(\mathbf{x}_i)} \prod_{j<l}(1 - V_{\pi_j(\mathbf{x}_i)}) \quad \text{for} \quad l < N(\mathbf{x}_i) \\
p_l(\mathbf{x}_i) &= \prod_{j<l}(1 - V_{\pi_j(\mathbf{x}_i)}) \quad \text{for} \quad l = N(\mathbf{x}_i)
\end{aligned}$$

(b) Conditional for $V_h, h = 1, \ldots, N$

$$(V_h|\mathbf{K}, \boldsymbol{\Gamma}, \alpha) \sim \text{Beta}(1 + \sum_{i=1}^{n} 1(K_i = h \text{ and } K_i \neq \pi_{N(\mathbf{x}_i)}(\mathbf{x}_i)), \alpha + \sum_{i=1}^{n} 1(K_i > h))$$

(c) Conditional for $\Gamma_h, h = 1, \ldots, N$

$$(\Gamma_h|\mathbf{K}, \mathbf{V}) \sim \text{Uniform}(\max[\max_{i;K_i=h}(x_i - \psi), a_\Gamma], \min[\min_{i;K_i=h}(x_i + \psi), a_\Gamma])$$

(d) Conditional for $\boldsymbol{\beta}^*_h, h = 1, \ldots, N$

$$\begin{aligned}
(\boldsymbol{\beta}^*_h|\mathbf{y}, \mathbf{K}, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta, \tau) &\sim N_p(\hat{\boldsymbol{\mu}}_{\beta h}, \hat{\boldsymbol{\Sigma}}_{\beta h}) \\
\hat{\boldsymbol{\mu}}_{\beta h} &= \hat{\boldsymbol{\Sigma}}_{\beta h}[\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta + \tau\mathbf{X}_{ih}\mathbf{y}_{ih}] \\
\hat{\boldsymbol{\Sigma}}_{\beta h} &= [{\boldsymbol{\Sigma}_\beta}^{-1} + \tau\mathbf{X}_{ih}\mathbf{X}_{ih}']^{-1},
\end{aligned}$$

where $\mathbf{y}_{ih}$ is $n_h \times 1$ response vector and $\mathbf{X}_{ih}'$ is $n_h \times p$ design matrix for the subjects with $K_i = h$ and n_h is the number of those subjects.

(e) Conditional for $\boldsymbol{\mu}_\beta$

$$\begin{aligned}(\boldsymbol{\mu}_\beta|\boldsymbol{\beta}^*, \boldsymbol{\Sigma}_\beta) &\sim N_p(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_\mu)\\ \hat{\boldsymbol{\mu}}_0 &= \hat{\boldsymbol{\Sigma}}_\mu[\boldsymbol{\Sigma}_\mu^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_\beta^{-1}\sum_{h=1}^{N}\boldsymbol{\beta}_h^*]\\ \hat{\boldsymbol{\Sigma}}_\mu &= [\boldsymbol{\Sigma}_\mu^{-1} + N\boldsymbol{\Sigma}_\beta^{-1}]^{-1}\end{aligned}$$

(f) Conditional for $\boldsymbol{\Sigma}_\beta^{-1}$

$$(\boldsymbol{\Sigma}_\beta^{-1}|\boldsymbol{\beta}^*, \boldsymbol{\mu}_\beta) \sim \text{Wishart}([\sum_{h=1}^{N}(\boldsymbol{\beta}_h^* - \boldsymbol{\mu}_\beta)(\boldsymbol{\beta}_h^* - \boldsymbol{\mu}_\beta)' + \nu_0\boldsymbol{\Sigma}_0]^{-1}, N + \nu_0)$$

(g) Conditional for τ

$$(\tau|\mathbf{y}, \boldsymbol{\beta}^*, \mathbf{K}) \sim \text{Gamma}(\nu_1 + \frac{n}{2}, \nu_2 + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i{}'\boldsymbol{\beta}_{K_i}^*)^2)$$

(h) Conditional for α

$$(\alpha|\mathbf{V}) \sim \text{Gamma}(\eta_1 + N, \eta_2 - \sum_{h=1}^{N} log(1 - V_h))$$

Note that this Gibbs sampling algorithm consists only of simple steps for sampling from standard distributions and is no more complex than blocked Gibbs samplers for DPMs. In addition, we have observed good computational performance, in terms of mixing and convergence rates, in simulated and real data applications.

3.5 Simulation Examples

We obtained data from two simulated examples, where $n = 500$ and a univariate predictor x_i was simulated from Uniform(0,1). Case 1 was a null case where y_i was generated from a normal regression model N($y_i; -1 + 2x_i, 0.01$). Case 2 was a mixture of two normal linear regression models, with the mixture weights depending on the predictor, with the error variance differing, and with a non-linear mean function for the second component:

$$f(y_i \,|\, \mathbf{x}_i) = e^{-2x_i} N(y_i; x_i, 0.01) + (1 - e^{-2x_i}) N(y_i; x_i^4, 0.04) \tag{3.15}$$

We applied the lDPM^N model in (3.12) to the simulated data with N=50. Based on the results, N=50 seems to be chosen to be large enough since the higher clusters having higher indices are not used in any of the subjects or are used in only a small proportion of them. Also, repeating the analysis for twice N, we obtained very similar results, suggesting that the results are robust to the choice of N, as long as N is not chosen to be small.

For the hyperparameters, we let $\nu_1 = \nu_2 = 0.01$, $\eta_1 = \eta_2 = 2$, $\nu_0 = p$, $\mathbf{\Sigma}_0 = I_p$, $\boldsymbol{\mu}_0 = \mathbf{0}$, $\mathbf{\Sigma}_\mu = n(\mathbf{X}'\mathbf{X})^{-1}$, $a_\Gamma = -0.05$, and $b_\Gamma = 1.05$. The neighborhood size $\psi = 0.05$ was chosen such that the average number of subjects belonging to the neighborhoods around each predictor value in the sample is $\approx$ n/10. We analyzed the simulated data using the proposed Gibbs sampling algorithm run for 10,000 iterations with a 5,000 iteration burn-in. The convergence and mixing of the MCMC algorithm were good (trace plots not shown). Also, results tended to be robust to repeating analysis with reasonable alternative hyperparameter values.

For case 1, as shown in Figure 3.3, the predictive mean regression curve (blue dashed, right bottom panel), the true linear regression function (red solid), and the pointwise 95% credible intervals (green dashed) were almost the same. Figure 3.3 also shows the predictive densities (blue dashed) at the 10th, 25th, 50th, 75th, and 90th sample percentiles of x_i, with these densities almost indistinguishable from the true densities (red solid).

For case 2, Figure 3.4 shows an $x - y$ plot (right bottom panel) of the data along with the estimated predictive mean curve (blue dashed), which closely follows the true mean curve (red solid). Figure 3.4 also shows the estimated predictive densities (blue dashed) correspond approximately to the true densities (red solid) in most cases and the 95% credible intervals (green dashed) closely cover the true densities in all cases.

Repeating the analysis for case 2, but with $\boldsymbol{\beta}_i \overset{iid}{\sim} G$ and $G \sim DP(\alpha G_0)$, we obtained poor results (density estimates diverged substantially from true densities, posterior mean curve failed to capture true non-linear function), suggesting that a DPM model is inadequate.

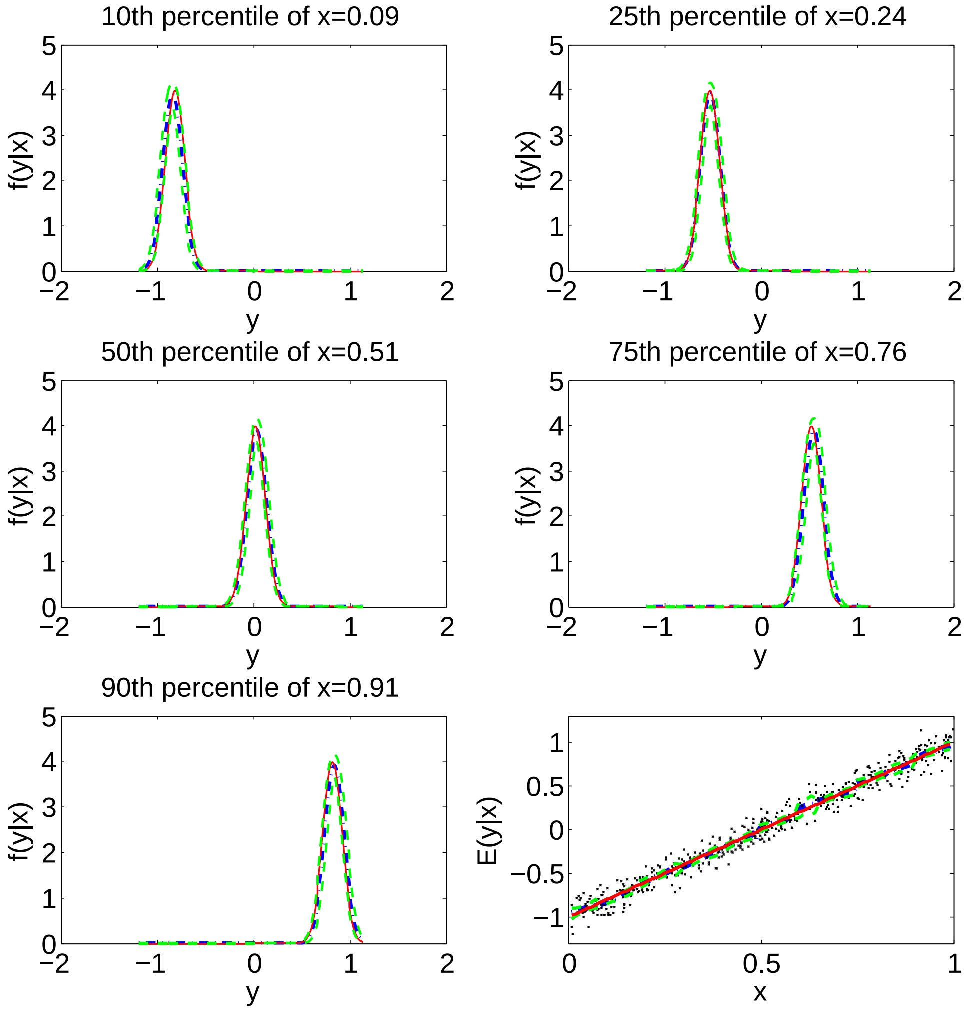

FIGURE 3.3: Results for simulation case 1: True conditional densities of $y|x$ (red solid), predictive conditional densities (blue dot-dashed), and 95% pointwise credible intervals (green dashed). The lower right panel shows the data (black dots), along with true (red solid) and estimated mean (blue dashed) regression curves superimposed with 95% credible line (green dashed).

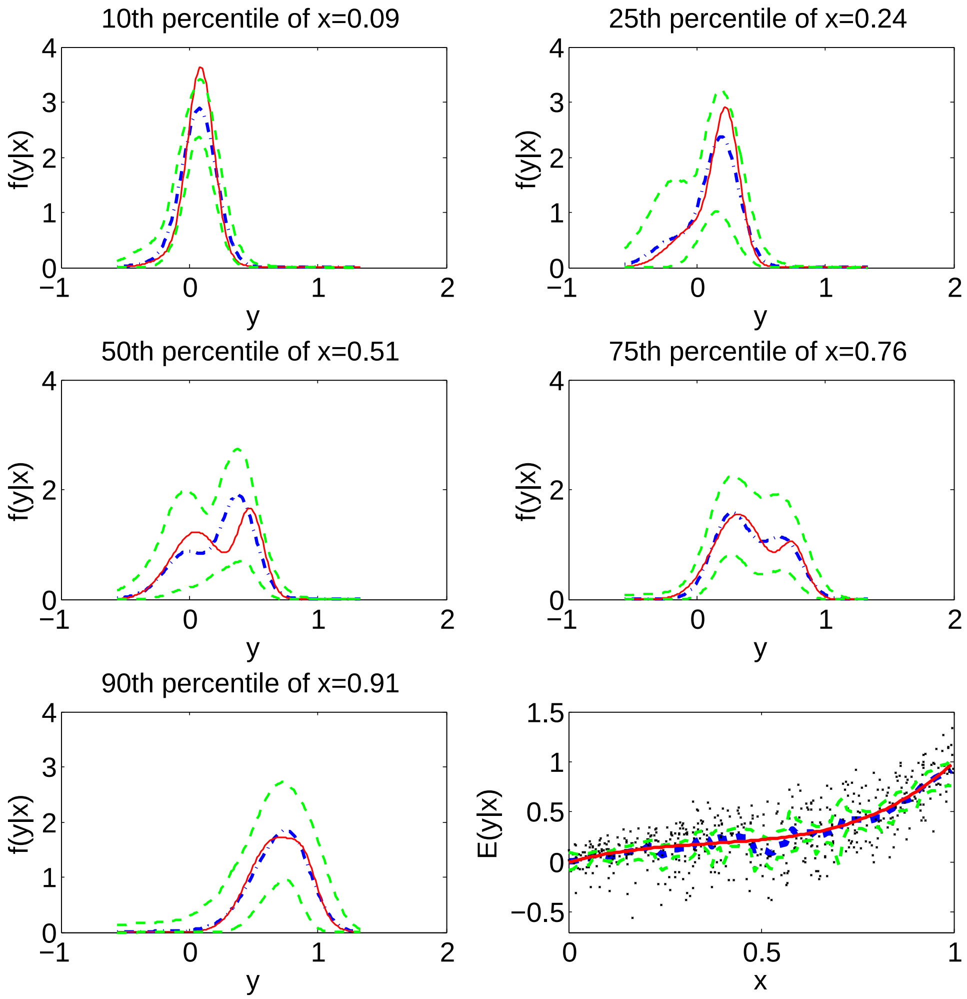

FIGURE 3.4: Results for simulation case 2: True conditional densities of $y|x$ (red solid), predictive conditional densities (blue dot-dashed), and 95% pointwise credible intervals (green dashed). The lower right panel shows the data (black dots), along with true (red solid) and estimated mean (blue dashed) regression curves superimposed with 95% credible line (green dashed).

3.6 Epidemiological Application

3.6.1 Background and Motivation

In diabetic studies, interest often focuses on the relationship between 2-hour serum insulin levels (indicator for insulin sensitivity/resistence) and 2-hour plasma glucose levels (indicator for diabetic risk) that are measured in the oral glucose tolerance test (OGTT). Although most studies examine the mean change of the 2-hour insulin versus 2-hour glucose, it would be more interesting to assess the whole distributional change of the 2-hour insulin level across the range of the 2-hour glucose levels.

We obtained data from a study which followed a sample of Pima Indians from a population near Phoenix, Arizona since 1965. This study was conducted by the National Institute of Diabetes and Digestive and Kidney Disease, with the Pima Indians chosen because of their high risk of diabetes. Using these data, our goal is conducting inferences on changes in the 2-hour serum insulin distribution with changes in 2-hour glucose level without making restrictive assumptions, such as normality or a constant residual variation. Certainly, it is biologically plausible that the insulin distribution is non-normal and should change as the glucose level changes not only in mean but also in other features such as skewness, residual variation, and modality.

3.6.2 Analysis and Results

For woman i ($i = 1, \ldots, 393$), let y_i correspond to the 2-hour serum insulin level measured in μU/ml (micro Units per milliliter) and let x_i denote the 2-hour plasma glucose level measured in mg/dl (milligrams per deciliter). We applied the lDPM^N model described in (3.12), after scaling y and x by dividing by 100. Hyperparameters were set to be the same as in the simulation study except that $\psi = 0.08$ such that n/10 subjects belong to each neighborhood on average and $a_\Gamma = min(x_i) - \psi$, and $b_\Gamma = max(x_i) + \psi$ such that the edge effects are avoided in the inference. We analyzed the data using the proposed Gibbs sampling algorithm run for 10,000 iterations with a 5,000 iteration burn-in. The convergence and mixing of the MCMC algorithm were good (Trace plots not shown) and results were robust with reasonable alternative hyperparameter values.

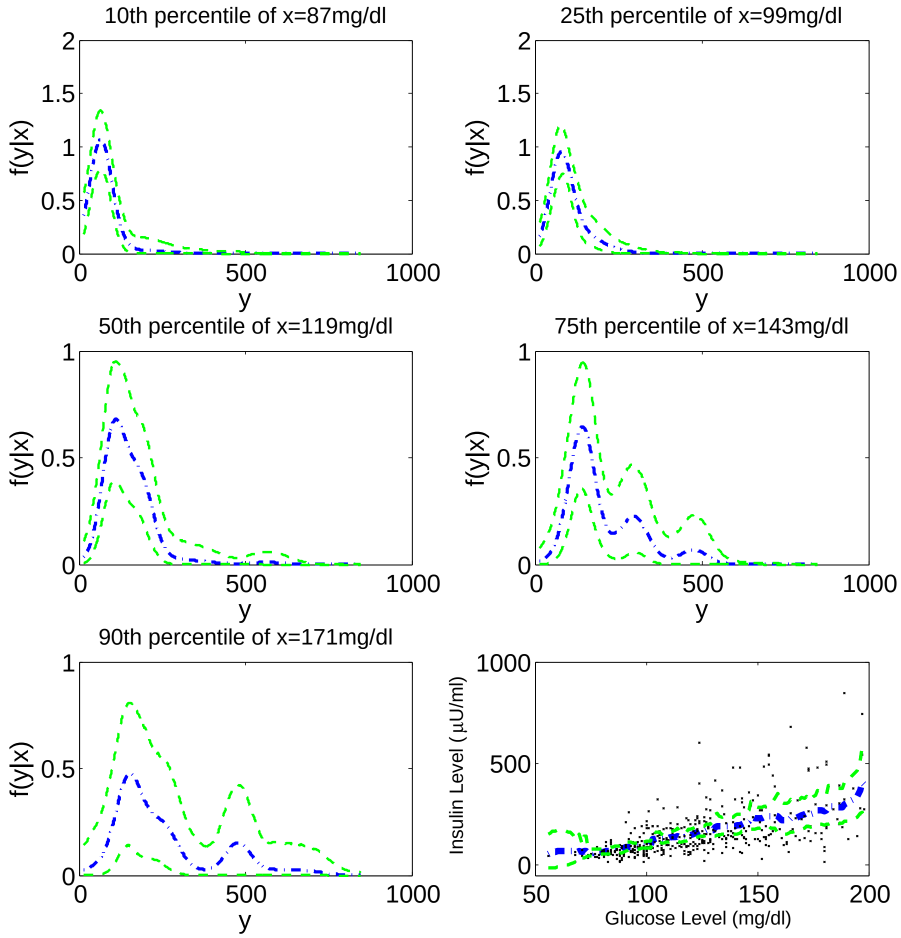

FIGURE 3.5: Results for Pima indian data example: Predictive conditional densities (blue dot-dashed), and 95% pointwise credible intervals (green dashed). The lower right panel shows the data (black dots), along with estimated mean (blue dashed) regression curves superimposed with 95% credible line (green dashed).

Figure 3.5 shows the predictive distributions for the insulin level at various empirical percentiles of the glucose level. As the glucose level increases, there is a slightly nonlinear change in the mean insulin level (right bottom panel) and a dramatic increase in the heaviness of the right tail of the insulin distribution. Also, some multi-modality in the insulin distribution appears as the glucose level falls into the pre-diabetes range (140-200 mg/dl) and closer to the cut point (200mg/dl) for the diagnosis of diabetes. This shift in the shape of the insulin distribution biologically implies that the women with pre-diabetes are expected to have different insulin sensitivities, which may further induce different diabetic risks even for the same glucose level. This may be due to unadjusted covariates or unmeasured risk factors. Such distributional changes in response induced by predictors (e.g. risk factor, exposure, treatment, and etc.) is pervasive in epidemiologic studies, but is not at all well characterized by standard regression models that do not allow the whole distribution to flexibly change with predictors.

3.7 Discussion

This article proposed a new stick-breaking prior for the collection of predictor-dependent random probability measures. The prior, called the lDP, is a useful alternative to recently developed prior models that induce predictor-dependence among distributions. Its marginal DP structure should be useful in considering theoretical properties, such as posterior consistency and rates of convergence. A related formulation was independently developed by Griffin and Steel (2008) although the lDP is appealing in its simplicity for construction and computation. In particular, the construction is intuitive and leads to simple expressions for the dependence in random measures at different locations, while also leading to straightforward posterior computation relying on truncation with a fair amount of accuracy.

Although we have focused on a conditional density estimation application, there are many interesting applications of the lDP to be considered in future work. First, the DP is widely used to induce a prior on a random partition or clustering structure (Quintana, 2006; Kim et al., 2006). In such settings, the DP has the potential disadvantage of requiring an exchangeability assumption, which may be violated when predictors are available that can inform about the clustering. The lDP provides a straightforward mechanism for local, predictor-dependent clustering, which can be used as an alternative to product partition models (Quintana and Iglesias, 2003) and model-based clustering approaches (Fraley and Raftery, 2002). It is of interest to explore the theoretical properties of the induced prior on the random partition. In this respect, it is likely that the hyperparameter ψ plays a key role. Hence, as a more robust data-driven approach one may consider fully Bayes or empirical Bayes methods for allowing uncertainty in ψ.

CHAPTER 4

NONPARAMETRIC BAYES CONDITIONAL DISTRIBUTION MODELING WITH VARIABLE SELECTION

4.1 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 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, $\mathrm{E}(Y \mid \mathbf{X})$, 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. (2007), 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 (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 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 MacLehose et al. (2007) 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. 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.

4.2 The Probit Stick-Breaking Process

4.2.1 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 $\mathbf{x} = (\mathbf{x_1}, \dots, \mathbf{x_p})'$. 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:

$$P_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{\boldsymbol{\theta}_\mathbf{h}}, \quad \forall \mathbf{x} \in \mathcal{X}, \tag{4.1}$$

where $\pi_h(\mathbf{x})$ is a probability weight on the hth component and $\delta_{\boldsymbol{\theta}}$ is a probability measure with all its mass at θ. 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:

$$\alpha_h \sim N(\mu, 1), \quad \boldsymbol{\psi}_h = \{\psi_{hj}\}_{j=1}^{p} \sim G, \quad \boldsymbol{\Gamma}_h = \{\Gamma_{hj}\}_{j=1}^{p} \sim H, \tag{4.2}$$

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:

$$\begin{aligned} \pi_h(\mathbf{x}) &= \Phi(\eta_h(\mathbf{x})) \prod_{\mathbf{l<h}} \Big\{ \mathbf{1} - \boldsymbol{\Phi}(\eta_\mathbf{l}(\mathbf{x})) \Big\} \\ \text{with} \quad \eta_h(\mathbf{x}) &= \alpha_h - \sum_{j=1}^{p} \psi_{hj} |x_j - \Gamma_{hj}|, \quad \forall \mathbf{x} \in \mathcal{X} \end{aligned} \tag{4.3}$$

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

By Lemma 1, $P_{\mathbf{x}}$ in (4.1) is a well defined probability measure on $(\Omega, \mathcal{B})$ for all $\mathbf{x} \in \mathcal{X}$ and the formulation from (4.1) through (4.3) 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, $\eta_h(\mathbf{x}) = \alpha_{\mathbf{h}}$ and $\pi_h(\mathbf{x}) = \mathbf{\Phi}(\alpha_{\mathbf{h}}) \prod_{\mathbf{l}<\mathbf{h}}(\mathbf{1} - \mathbf{\Phi}(\alpha_{\mathbf{h}}))$ for all $\mathbf{x} \in \mathcal{X}$. Because $\pi_h(\mathbf{x})$ does not depend on $\mathbf{x}$, we obtain

$$P_{\mathbf{x}} = P = \sum_{h=1}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h} \quad \text{with} \quad \pi_h = \Phi(\alpha_h) \prod_{l<h}(1 - \Phi(\alpha_l)), \quad \forall \mathbf{x} \in \mathcal{X} \tag{4.4}$$

Note that P in (4.4) 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 $V_h \sim$ 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, $\mu \in \Re$ in the PSBP controls precision with large values assigning high probability to the first few components.

Although the PSBP special case in (4.4) 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 (4.3). Given $\mathbf{x}$, each $\pi_h(\mathbf{x})$ is linked through the index h to each location $\mathbf{\Gamma}_h$. If hth location $\mathbf{\Gamma}_h$ 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 $\boldsymbol{\psi}_h$ 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 πDDP (Griffin and Steel, 2007) and the KSBP (Dunson and Park, 2008), have more complicated structure than (4.3) and the updating algorithm for the random components in $\pi_h(\mathbf{x})$ is not straightforward. However, the probit-based weight structure in (4.3) allows for using a data augmentation approach in order to obtain conjugacy so that the random components $\alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h$ 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.

4.2.2 Moments

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

$$\begin{aligned}
\mathrm{E}\{P_{\mathbf{x}}(B)|\alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h\} &= \sum_{h=1}^{\infty} \pi_h(\mathbf{x})\mathbf{E}\{\delta_{\boldsymbol{\theta}_{\mathbf{h}}}(\mathbf{B})\} = \mathbf{P_0}(\mathbf{B}) \\
\mathrm{E}\{P_{\mathbf{x}}(B)^2|\alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h\} &= \left[\sum_{h=1}^{\infty} \pi_h(\mathbf{x})^{\mathbf{2}}\mathbf{E}\{\delta_{\boldsymbol{\theta}_{\mathbf{h}}}(\mathbf{B})^{\mathbf{2}}\}\right] \\
&\quad + \left[\sum_{h=1}^{\infty}\sum_{l \neq h} \pi_h(\mathbf{x})\pi_{\mathbf{l}}(\mathbf{x})\mathbf{E}\{\delta_{\boldsymbol{\theta}_{\mathbf{h}}}(\mathbf{B})\}\mathbf{E}\{\delta_{\boldsymbol{\theta}_{\mathbf{l}}}(\mathbf{B})\}\right] \\
&= \sum_{h=1}^{\infty} \pi_h(\mathbf{x})^{\mathbf{2}}\left[\mathbf{E}\{\delta_{\boldsymbol{\theta}_{\mathbf{h}}}(\mathbf{B})^{\mathbf{2}}\} - \mathbf{P_0}(\mathbf{B})^{\mathbf{2}}\right] + \mathbf{P_0}(\mathbf{B})^{\mathbf{2}} \\
&= ||\pi_h(\mathbf{x})||^{\mathbf{2}}\{\mathbf{P_0}(\mathbf{B}) - \mathbf{P_0}(\mathbf{B})^{\mathbf{2}}\} + \mathbf{P_0}(\mathbf{B})^{\mathbf{2}} \\
&= ||\pi_h(\mathbf{x})||^{\mathbf{2}}\mathbf{P_0}(\mathbf{B}) + \{\mathbf{1} - ||\pi_{\mathbf{h}}(\mathbf{x})||^{\mathbf{2}}\}\mathbf{P_0}(\mathbf{B})^{\mathbf{2}} \qquad (4.5)
\end{aligned}$$

Also, the correlation is

$$\begin{aligned}
\mathrm{Corr}\{P_{\mathbf{x}}(B), P_{\mathbf{x}'}(B)|\alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h\} &= \frac{\sum_{h=1}^{\infty} \pi_h(\mathbf{x})\pi_{\mathbf{h}}(\mathbf{x}')}{\{\sum_{h=1}^{\infty} \pi_h(\mathbf{x})^{\mathbf{2}}\}^{\mathbf{1/2}}\{\sum_{\mathbf{h=1}}^{\infty} \pi_{\mathbf{h}}(\mathbf{x}')^{\mathbf{2}}\}^{\mathbf{1/2}}} \\
&= \frac{< \pi_h(\mathbf{x}), \pi_{\mathbf{h}}(\mathbf{x}') >}{||\pi_h(\mathbf{x})|| \cdot ||\pi_{\mathbf{h}}(\mathbf{x}')||} \qquad (4.6)
\end{aligned}$$

Note that the correlation is bounded above by 1 from the Cauchy-Schwarz inequality and goes to 1 in the limit as $\mathbf{x} \to \mathbf{x}'$. 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 $U_h(\mathbf{x}) = \boldsymbol{\Phi}(\eta_{\mathbf{h}}(\mathbf{x}))$, 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 $U_h(\mathbf{x}) \overset{\mathbf{iid}}{\sim} \mathbf{F_x}$ for $h = 1, \ldots, \infty$. Letting $\mu(\mathbf{x}) = \mathrm{E}_{\mathbf{F_x}}\{\mathbf{U_h}(\mathbf{x})\}$, $\mu^{(2)}(\mathbf{x}) = \mathrm{E}_{\mathbf{F_x}}\{\mathbf{U_h}(\mathbf{x})^{\mathbf{2}}\}$, and $\mu(\mathbf{x}, \mathbf{x}') = \mathrm{E}_{\mathbf{F_x}}\{\mathbf{U_h}(\mathbf{x})\mathbf{U_h}(\mathbf{x}')\}$, we can show that

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

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

$$\begin{aligned} \mathrm{Corr}\{P_{\mathbf{x}}(B), P_{\mathbf{x}'}(B)\} \quad = \quad & \left[\frac{\mu(\mathbf{x}, \mathbf{x}')}{\mu(\mathbf{x}) + \mu(\mathbf{x}') - \mu(\mathbf{x}, \mathbf{x}')}\right] \\ & \times \left[\frac{\{2\mu(\mathbf{x}) - \mu^{(\mathbf{2})}(\mathbf{x})\}\{\mathbf{2}\mu(\mathbf{x}') - \mu^{(\mathbf{2})}(\mathbf{x}')\}}{\mu^{(2)}(\mathbf{x})\mu^{(\mathbf{2})}(\mathbf{x}')}\right]^{1/2} \end{aligned} \tag{4.7}$$

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 (4.6) and (4.7) follow similar lines for the moments of the KSBP (Dunson and Park, 2008).

4.3 Conditional Distribution Modeling With Variable Selection

4.3.1 Model Specification

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

$$\begin{aligned} f(y|\mathbf{x}) &= \int N(y; \mathbf{x_0'}\boldsymbol{\beta}, \tau^{\mathbf{-1}})\mathbf{dP_x}(\boldsymbol{\beta}, \tau) \\ \mathcal{P}_{\mathcal{X}} &= \{P_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\} \sim \text{PSBP}(\mu, \mathbf{G}, \mathbf{H}, \mathbf{P_0}), \end{aligned} \tag{4.8}$$

where $\mathbf{x_0} = (\mathbf{1}, \mathbf{x}')'$ 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 (4.1) with $\boldsymbol{\theta}_h = (\boldsymbol{\beta}_h^*, \tau_h^*)$ and $\boldsymbol{\beta}_h^* = (\beta_{h0}^*, \ldots, \beta_{hp}^*)'$, we obtain

$$f(y|\mathbf{x}) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x})\mathbf{N}(\mathbf{y}; \mathbf{x_0'}\boldsymbol{\beta}^*_{\mathbf{h}}, \tau_{\mathbf{h}}^{*\mathbf{-1}}), \tag{4.9}$$

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 (4.9) 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 (4.8). Letting γ_{hj} be an inclusion indicator variable for the jth predictor in the hth mixture component, we induce G and P_0 through the following distributions for $\boldsymbol{\psi}_h$ and $\boldsymbol{\theta}_h$.

$$\begin{aligned} \boldsymbol{\psi}_h = \{\psi_{hj}\}_{j=1}^p &\sim \prod_{j=1}^p \left\{ 1(\gamma_{hj} = 0)\delta_0(\psi_{hj}) + 1(\gamma_{hj} = 1)N_+(\psi_{hj}; \mu_{\psi_j}, \tau_{\psi_j}^{-1}) \right\} \\ \boldsymbol{\theta}_h = (\boldsymbol{\beta}_h^*, \tau_h^*) &\sim N_{p_{\gamma_h}+1}(\boldsymbol{\beta}^*_{\gamma_h,h}; \mathbf{0}, \boldsymbol{\Sigma}_{\gamma_h,h}) \times \delta_{\mathbf{0}}(\boldsymbol{\beta}^*_{\bar{\gamma}_h,h}) \times \text{Gamma}(\tau_h^*; a_\tau, b_\tau), \end{aligned} \tag{4.10}$$

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, $\boldsymbol{\beta}^*_{\bar{\gamma}_h,h}$ 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, with $\gamma_{hj} = 0$ implying that $\psi_{hj} = 0$ and $\beta^*_{hj} = 0$. A value of $\beta^*_{hj} = 0$ leads to the jth predictor assigned a coefficient of zero in the hth linear regression model in (4.9), 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

$$\gamma_{hj} \sim \text{Bernoulli}(\gamma_{hj}; \kappa_j), \tag{4.11}$$

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

$$\begin{aligned} \kappa_j &\sim 1(w_j = 0)\delta_0(\kappa_j) + 1(w_j = 1)\text{Beta}(\kappa_j; a_{\kappa_j}, b_{\kappa_j}) \quad \text{for} \quad j = 1, \ldots, p \\ w_j &\sim \text{Bernoulli}(w_j; 0.5), \end{aligned} \tag{4.12}$$

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. Theses priors can be used directly for the coefficients in each mixture component as follows.

$$\begin{aligned} \boldsymbol{\beta}^*_{\gamma_h,h}|\tau^*_h &\sim N(\boldsymbol{\beta}^*_{\gamma_h,h}; \mathbf{0}, \boldsymbol{\Sigma}_{\gamma_h,h}) \\ \boldsymbol{\Sigma}_{\gamma_h,h} &= ng^{-1}(\mathbf{X}'_{\gamma_\mathbf{h}}\mathbf{X}_{\gamma_\mathbf{h}})^{-\mathbf{1}}/\tau^*_\mathbf{h} \quad \text{with} \quad \mathbf{g} \sim \text{Gamma}(\mathbf{g};\, \mathbf{a_g}, \mathbf{b_g}), \end{aligned} \tag{4.13}$$

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

4.3.2 Hypothesis Formulation

We first consider a global null hypothesis for selecting important predictors. As discussed with the variable selection structure in (4.10), 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 (4.9) 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 jth predictor as follows:

$$H_{0j}^{N} : \gamma_{hj} = 0 \quad \text{for} \quad h = 1, \ldots, N, \tag{4.14}$$

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 (4.8).

$$\begin{aligned} y_i|S_i, \mathcal{P}_{\mathcal{X}} &\sim N(y_i; \mathbf{x_{i0}'}\boldsymbol{\beta}^*_{\mathbf{S_i}}, \tau^{*\mathbf{-1}}_{\mathbf{S_i}}) \\ S_i|\mathcal{P}_{\mathcal{X}} &\sim \sum_{h=1}^{\infty} \pi_h(\mathbf{x_i})\delta_{\mathbf{h}}(\mathbf{S_i}) \\ \mathcal{P}_{\mathcal{X}} = \{P_{\mathbf{x}} : \mathbf{x} \in \mathcal{X}\} &\sim PSBP(\mu, G, H, P_0), \end{aligned} \tag{4.15}$$

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 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. Proof is in the Appendix.

> **Theorem 1**. Suppose $y_i|\mathbf{x_i} \sim \mathbf{f}(\mathbf{y}|\mathbf{x})$ and $f(y|x)$ is assumed to be a PSBPM as in (4.8) with G and P_0 chosen as in (4.10) and (5.8). Let $l(\mathbf{y}, \mathbf{S}|\mathbf{H_{0j}})$ and $l(\mathbf{y}, \mathbf{S}|\mathbf{H^N_{0j}})$ be the marginal likelihoods for $(\mathbf{y}, \mathbf{S})$ under H_{0j} and H_{0j}^N where $\mathbf{y} = (\mathbf{y_1}, \ldots, \mathbf{y_n})'$ and $\mathbf{S} = (S_1, \ldots, S_n)'$. Then, the ratio $R = \frac{l(\mathbf{y},\mathbf{S}|\mathbf{H_{0j}})}{l(\mathbf{y},\mathbf{S}|\mathbf{H^N_{0j}})}$ 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_{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 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 $\mathbf{x}^*_{\mathbf{(j)}} = (\mathbf{x}^*_\mathbf{1}, \ldots, \mathbf{x}^*_{\mathbf{j-1}}, \mathbf{x}^*_{\mathbf{j+1}}, \ldots, \mathbf{x}^*_\mathbf{p})'$. Letting $d(x_j, x_j')|_{\mathbf{x}^*_{\mathbf{(j)}}} = sup_{y \in \Re} |F(y|x_j, \mathbf{x}^*_{\mathbf{(j)}}) - \mathbf{F}(\mathbf{y}|\mathbf{x}'_\mathbf{j}, \mathbf{x}^*_{\mathbf{(j)}})|$, we propose a local interval null hypothesis as:

$$H_{0j}(x_j, x_j'|\mathbf{x}^*_{\mathbf{(j)}}) : d(x_j, x_j')|_{\mathbf{x}^*_{\mathbf{(j)}}} < \epsilon, \tag{4.16}$$

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 (4.16), we can further consider the local null hypothesis of equality of the conditional distributions across a region $A_j \subset \mathcal{X}_j$ with $\mathcal{X}_j$ jth predictor space as:

$$H_{0j}(A_j|\mathbf{x}^*_{\mathbf{(j)}}) : sup_{x_j, x_j' \in A_j} \{d(x_j, x_j')|_{\mathbf{x}^*_{\mathbf{(j)}}}\} < \epsilon, \tag{4.17}$$

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 (4.16) and (4.17) would be sensitive to the choice of $\mathbf{x}^*_{\mathbf{(j)}}$, in particular, when 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.

4.4 Posterior Computation

4.4.1 Model and MCMC algorithm

We develop an MCMC algorithm for the PSBPM following the specification in (4.8) with G and P_0 chosen as in (4.10) with (5.8), (4.12) and (4.13). For H, we consider $\mathbf{\Gamma}_h = \{\Gamma_{hj}\}_{j=1}^p \sim \prod_{j=1}^p \sum_{m=1}^{M_j} \delta_{\Gamma^*_{mj}}(\Gamma_{hj})$ where Γ^*_{mj} for $m = 1, \ldots, M_j$ are pre-specified grid values for jth predictor. 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, $\alpha_h, \boldsymbol{\psi}_h, \mathbf{\Gamma}_h$, 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

$$\begin{aligned} Z_{il} &= 1(Z^*_{il} > 0) \\ Z^*_{il} &\sim N\Bigg(Z^*_{il}; \alpha_h - \sum_{j=1}^p \psi_{hj}|x_{ij} - \Gamma_{hj}|, 1\Bigg) \end{aligned} \tag{4.18}$$

For $S_i = T$, we introduce Z^*_{il} only for $l = 1, \ldots, T-1$ because we let $\Phi(\eta_T(\mathbf{x})) = \mathbf{1}$ so that $\sum_{h=1}^T \pi_h(\mathbf{x}) = \mathbf{1}$. Given Z^*_{il}, we update $\alpha_h, \boldsymbol{\psi}_h, \mathbf{\Gamma}_h$ 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).

4.4.2 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 $\mu_\mu = 0, \tau_\mu = 1$. We let $\epsilon = 0.05$ in defining local null hypotheses as this implies negligible local changes in the conditional densities under the null in simulations (not shown). For truncation, we let $T = 20$ which was shown to be large enough because N tends to converge to a small number (≤ 10). 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.

4.5 Simulation Study

In order to illustrate the proposed method and to assess the performance, we conduct a simulation study. We first generate $x_{ij} \overset{iid}{\sim} \text{Uniform}(x_{ij}; -2, 2)$ 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).

$$\begin{aligned}
(1) \quad y_i &\overset{iid}{\sim} 0.5N(y_i; 1, 1) + 0.5N(y_i; -1, 0.5^2) \\
(2) \quad y_i &\overset{iid}{\sim} N(y_i; 2x_{i1} - 3x_{i2} + x_{i4} - x_{i5}, 1) \\
(3) \quad y_i &\overset{iid}{\sim} 0.5N(y_i; 10, (2 + 4e^{-min(x_{i1}, 0)})^2) + 0.5N(y_i; -10 + 5x_{i2}, 5^2) \qquad (4.19)
\end{aligned}$$

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 $E(y|\mathbf{x})$.

4.5.1 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 4.1 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 4.1 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 4.2 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 (4.19) 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). Letting $\Pr(H_{0j}|\text{Data}) < 0.05$ as a significant evidence for rejecting H_{0j} against the alternative, we obtained 98% of rejecting rate on average for important predictors and 95% of not-rejecting rate for unimportant predictors out of 100 replicated data sets in each simulation case.

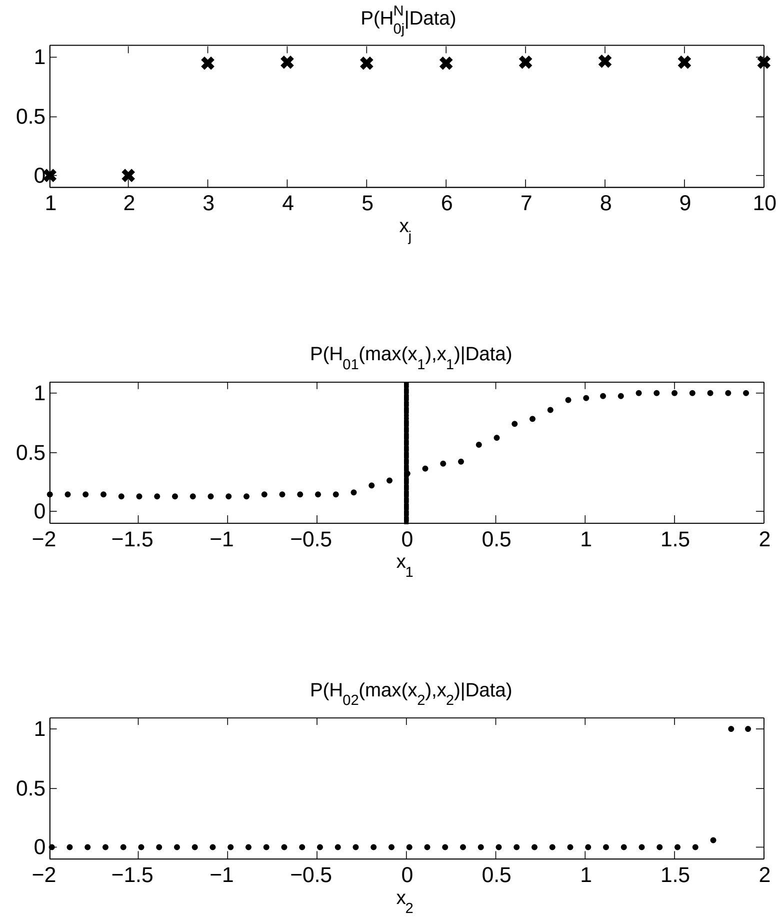

FIGURE 4.1: Top - Posterior probabilities for H_{0j}^N for $j = 1, \ldots, 10$; Middle - Posterior probabilities for $H_{01}(\max(x_1), x_1)$ with x_1 varying across 40 grid points; Bottom - Posterior probabilities for $H_{02}(\max(x_2), x_2)$ with x_2 varying across 40 grid points

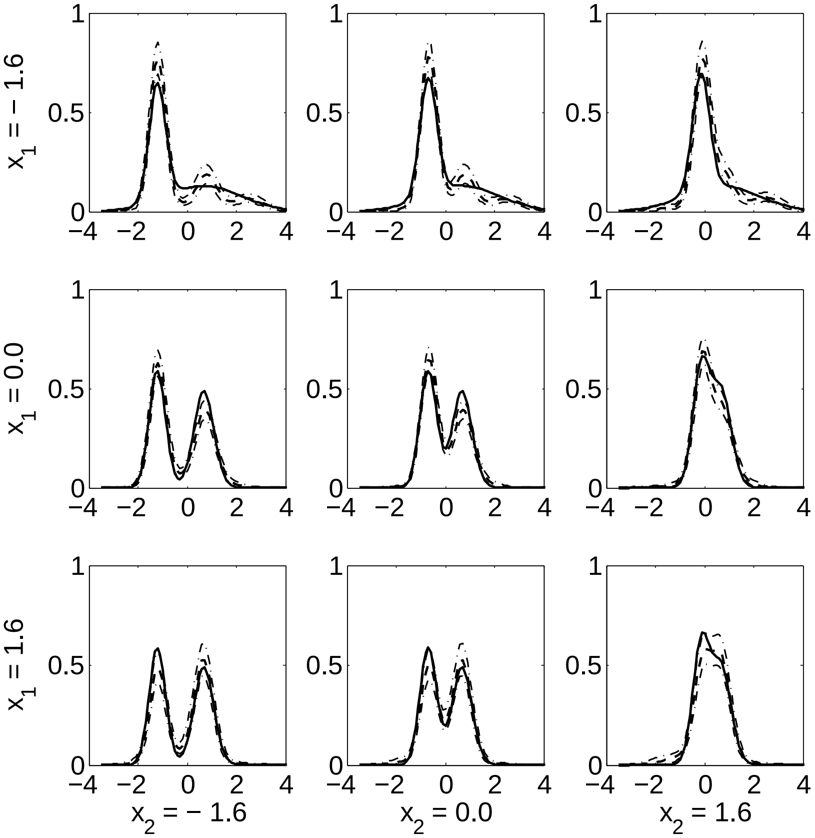

FIGURE 4.2: True (solid), Predictive (dashed) conditional response density $\hat{f}(y|\mathbf{x}^*)$ with 95% credible intervals (dash-dotted) at $\mathbf{x}^* = (\mathbf{x_1}, \mathbf{x_2}, \bar{\mathbf{x}}_\mathbf{3}, \ldots, \bar{\mathbf{x}}_\mathbf{10})$ with x_1 and x_2 varying among 5th, 50th, 95th empirical percentiles

4.5.2 Comparison with a simple and a competing method

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 (4.19) 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 (4.10) with (5.8), (4.12), (4.13). 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($\beta_j = 0|$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 is not a formal approach for variable selection, and is not comparable to the posterior inclusion probabilities and Bayes factors provided by LR-SSVS or PSBM.

In case (3), LR-SSVS detected only x_2 as important with Pr($\beta_2 = 0|$Data)=0. Pr($\beta_1 = 0|$Data)=0.87 and Pr($\beta_j = 0|$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

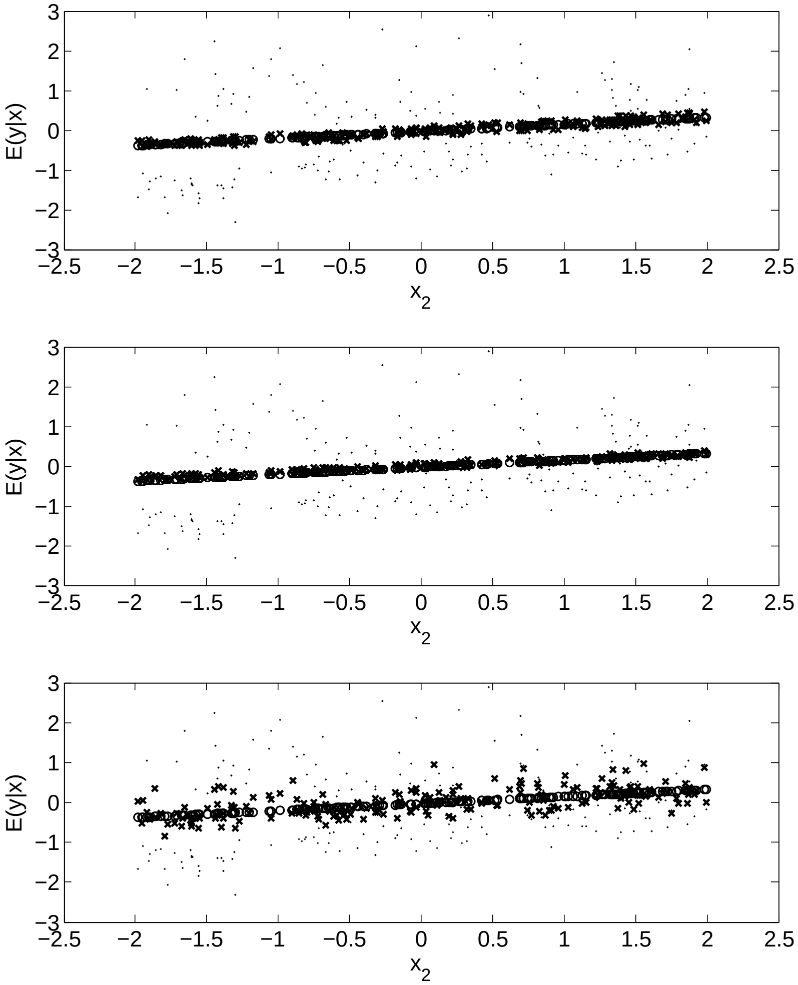

FIGURE 4.3: True mean $E(y|\mathbf{x})$ ('o'), Predictive mean $\hat{E}(y|\mathbf{x})$ ('x'), observed data y ('*') across x_2 : Top - PSBPM; Middle - LR-SSVS; Bottom - BART

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 4.3 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 $E(y|\mathbf{x})$.

4.6 Epidemiological Application

4.6.1 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 4.4 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.

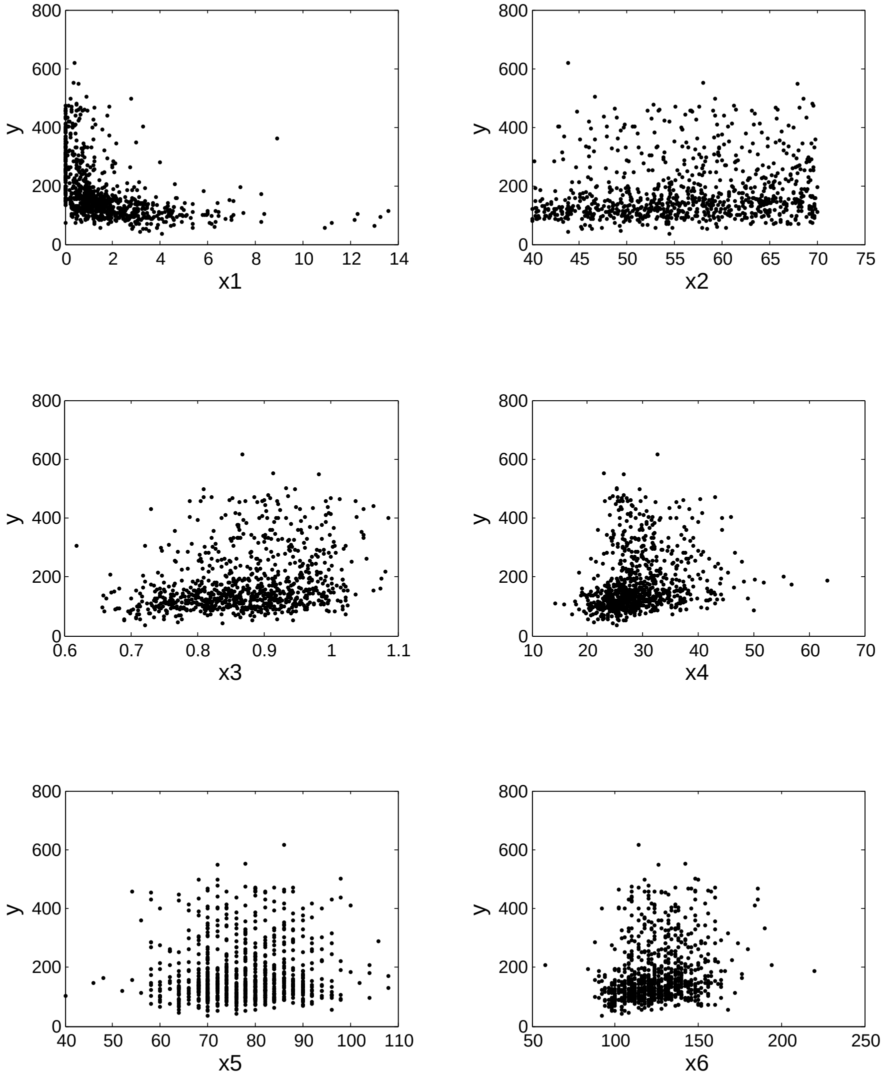

FIGURE 4.4: Data from IRAS study : y = 2-hour glucose level (mg/dl); x_1 = insulin sensitivity; x_2 = age; x_3 = waist to hip ratio; x_4 = body mass index; x_5 = diastolic blood pressure; x_6 = systolic blood pressure

4.6.2 Analysis

We analyzed the IRAS study data focusing on the relationship between 2-hour glucose level and 6 predictors shown in Figure 4.4. 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, 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 $j = 1, \ldots, 6$. 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 $j = 1, \ldots, 6$. 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 $\mathbf{x}^* = (\mathbf{x_1}, \mathbf{x_2}, \bar{\mathbf{x}}_\mathbf{3}, \ldots, \bar{\mathbf{x}}_\mathbf{10})$ with x_1 and x_2 varying among 5th, 50th, 95th empirical percentiles. Figure 4.5 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.

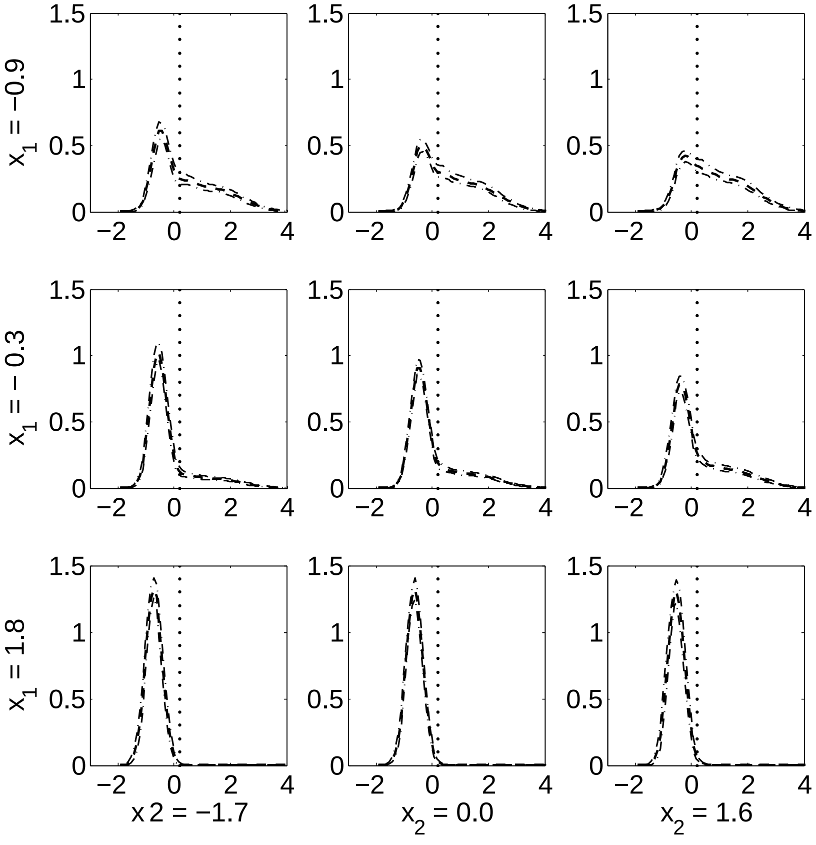

FIGURE 4.5: Predictive (dashed) conditional response density $\hat{f}(y|\mathbf{x}^*)$ with 95% credible intervals (dash-dotted) at $\mathbf{x}^* = (\mathbf{x_1}, \mathbf{x_2}, \bar{\mathbf{x}}_\mathbf{3}, \ldots, \bar{\mathbf{x}}_\mathbf{6})$ with x_1 and x_2 varying among 5th, 50th, 95th empirical percentiles

4.7 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.

CHAPTER 5

BAYES VARIABLE SELECTION IN LATENT CLASS MODELING OF LONGITUDINAL DATA

5.1 Introduction

The enormous increase in the incidence of obesity over the past several decades has led to a great deal of concern among public health researchers, clinicians and the general public. Obesity is a complex health condition, which results from the interplay of genetics, diet and other environmental factors. As weight loss intervention programs for adults are often unsuccessful, there is considerable interest in identifying prenatal and childhood risk factors predictive of the later development of obesity, with the hope that early interventions and behavioral modifications may be more efficacious. Our motivation is drawn from a German study of childhood growth (Fenske et al., 2008), which recorded body mass index (BMI) over time for 3097 children starting at birth and continuing to age 5.

Potentially, one could use a linear mixed effects (LME) model (Laird and Ware, 1982) for data of this type. However, the assumptions of linearity of the growth trajectories and normality of the random effects characterizing variability in the trajectories are clearly questionable. Latent class trajectory (LCT) models (Muthén and Shedden, 1999) provide a flexible alternative, which relies on using a finite mixture of normals for the random effects distribution, while allowing non-linear trajectories through the use of polynomials. In this framework, a polytomous logistic regression model is used to relate predictors to the probability of allocation to each latent class, with data for individuals in a class characterized using an LME model. Nagin (1999) proposed an alternative approach, which instead assumed that individuals within a class had identical random effects, leading to clustering of individuals according to their growth trajectory (Roeder, Lynch and Nagin, 1999). This type of approach can be implemented routine in SAS (Jones, Nagin and Roeder, 2001).

As noted in Bigelow and Dunson (2008), there are some drawbacks to these frequentist finite mixture modeling-based approaches. The first is the need to estimate the number of latent classes, k, and then condition inferences on this estimate. The typical estimation strategy relies on fitting the model for different choices of k, and choosing $\hat{k}$ based on the BIC. The BIC is not theoretically justified in this mixture model setting, and it is appealing to allow uncertainty in estimation of k in performing inferences. In addition, a more biologically realistic model would allow the number of classes represented in the sample to increase with sample size, as there may be occasional introduction of an individual having a rare health condition leading to a very different growth trajectory than observed in previous subjects. To allow uncertainty in estimating k, while allowing the number of classes to grow at a rate proportional to $\alpha \log n$, with n the number of subjects, one can use the following Dirichlet process mixture (DPM) model (Escobar and West, 1995; Bush and MacEachern, 1996; Kleinman and Ibrahim, 1998):

$$\begin{aligned} y_{ij} &= \mathbf{x}_{ij}'\boldsymbol{\beta}_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \tau^{-1}) \\ \boldsymbol{\beta}_i &\sim P, \quad P \sim DP(\alpha P_0), \end{aligned} \tag{5.1}$$

where y_{ij} is the jth observation on individual i, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijq})'$ is a vector of time-dependent predictors, $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iq})'$ are subject-specific random effects, P is an unknown random effects distribution, and $DP(\alpha P_0)$ denotes a Dirichlet process (DP) prior (Ferguson, 1973; 1974) with precision α and base distribution P_0.

As is clear from the Sethuraman (1994) stick-breaking representation of the DP, the semi-parametric Bayes random effects model in (5.1) implies that $\boldsymbol{\beta}_i \sim P = \sum_{h=1}^{\infty} \pi_h \delta_{\boldsymbol{\beta}_h^*}$, with δ_θ denoting a distribution concentrated at θ. Hence, the random effects distribution is discrete and individuals will be allocated to clusters, with each cluster having a distinct random effects vector. As the random effects characterize the growth trajectory, trajectory clusters will be obtained automatically. This property was used by Ray and Mallick (2006) for wavelet-based functional clustering, while Wang, Ray and Mallick (2007) and Bigelow and Dunson (2008) independently extended this type of approach for joint modeling with functional predictors. DPMs have been widely used to allow for unknown random effects distributions in Bayesian hierarchical models, with Ohlssen, Sharples and Spiegelhalter (2007) providing a recent tutorial on the practical implementation.

Unfortunately, DPMs for random effects distributions do not allow us to directly address our interests in identifying predictors of the growth trajectory. In addition, although there is an increasingly-rich literature on methods for generalizing DPMs to allow predictor dependence (Griffin and Steel, 2006; Dunson et al., 2007; Dunson and Park, 2008, among others), such methods do not allow for variable selection, with the exception of a recent approach proposed by Chung and Dunson (2008). The Chung and Dunson (2008) method relied on a probit stick-breaking process (PSBP), which was carefully defined to allow Bayesian variable selection to be implemented via a simple stochastic search variable selection (SSVS) algorithm (George and McCulloch, 1993; 1997). The goal of the current paper is to generalize this approach to the longitudinal data setting, with the applied emphasis being the selection of predictors of trajectories in childhood growth. The proposed approach is highly-flexible in allowing the mean and quantile trajectories to vary flexibly with the selected predictors, allowing one to conduct inferences on risk of overweight or obesity without relying on pre-specified BMI categories. It is important to avoid categorizing BMI to avoid sensitivity to cutoffs and to allow finer-scale inferences. For example, there are considerable clinical differences within the overweight and obesity categories.

Section 2 defines the variable selection problem and nonparametric Bayes approach. Section 3 develops an algorithm for posterior computation. Section 4 considers a simulation study. Section 5 applies the method to the German growth data set, and Section 6 discusses the results.

5.2 Mixture Models for Longitudinal Data with Variable Selection

5.2.1 Predictor-Dependent Mixture Model

For $i = 1, \ldots, n$, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ be the ith subject's longitudinal response vector and $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})'$ be the ith subject's time-varying predictor matrix, where $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijq})'$ denotes the ith subject's predictor vector at time t_{ij}, for $j = 1, \ldots, n_i$. The following normal linear random effects model provides a simple model for characterizing these data:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N_{n_i}(\mathbf{0}, \tau^{-1}\mathbf{I}_{n_i}) \\ \boldsymbol{\beta}_i &\sim N_p(\boldsymbol{\theta}, \Omega), \end{aligned} \tag{5.2}$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iq})'$ are subject-specific random effects, which are assumed independent of the residual $\boldsymbol{\epsilon}_i$. By including non-linear basis functions evaluated at t_{ij} within the $\mathbf{x}_{it}$ vector, one can accommodate flexible non-linear trajectories within the framework of (5.2). For instance, one can consider a cubic spline with two knots as $\mathbf{x}_{ij} = (t_{ij}, t_{ij}^2, t_{ij}^3, (t_{ij} - t_0)_+^3, (t_{ij} - t_1)_+^3)'$, where t_{ij} is the ith subject's jth measurement time and t_0 and t_1 are pre-specified knots.

Extending model (5.2) to a LCT framework, we use (5.2) to characterize the data for subjects within a class, while allowing the random effects distribution parameters and residual precision to vary across classes. In particular, letting $S_i \in \{1, \ldots, N\}$ denote the latent class for subject

i, with N an upper bound on the number of classes occupied by the n subjects, we let

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i | S_i \sim N_{n_i}(\mathbf{0}, \tau_{S_i}^{*-1}\mathbf{I}) \\ \boldsymbol{\beta}_i | S_i &\sim N_q(\boldsymbol{\theta}_{S_i}^*, \Omega_{S_i}^*), \\ S_i | \mathbf{w_i} &\sim \sum_{h=1}^{N} \pi_h(\mathbf{w}_i)\delta_h, \end{aligned} \tag{5.3}$$

where $\{\tau_h^*, \boldsymbol{\theta}_h^*, \Omega_h^*\}$ are parameters specific to class h, for $h = 1, \ldots, N$, $\mathbf{w}_i = (w_{i1}, \ldots, w_{ip})'$ is a vector of predictors for subject i, δ_h is a distribution concentrated at h, and $\pi_h(\mathbf{w}) = \Pr(S_i = h \,|\, \mathbf{w}_i = \mathbf{w})$ is the probability of allocation to class h given predictors $\mathbf{w}$. Note that the allocation probability depends on the predictors which allows for predictor-dependent clustering of subjects into different trajectory classes.

In Bayesian framework, uncertainty can be allowed for the number of latent classes, N, class-specific parameters, $\{\tau_h^*, \boldsymbol{\theta}_h^*, \Omega_h^*\}$, and predictor-dependent allocation probabilities $\pi_h(\mathbf{w})$. Such uncertainty can be obtained by adding the following hierarchy in model (5.2).

$$\begin{aligned} \{\tau_i, \boldsymbol{\theta}_i, \Omega_i\} &\sim P_{\mathbf{w}_i}, \\ \mathcal{P}_{\mathcal{W}} = \{P_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\} &\sim \mathrm{PSBP}(\mu, P_0, G, H), \end{aligned} \tag{5.4}$$

where $P_{\mathbf{w}}$ is random distribution indexed by $\mathbf{w}$, $\mathcal{P}_{\mathcal{W}}$ is an uncountable collection of $P_{\mathbf{w}}$, and PSBP (μ, P_0, G, H) denotes the probit stick-breaking process with hyperparameters of μ, P_0, G, H (Chung and Dunson, 2008) as a prior for $\mathcal{P}_{\mathcal{W}}$. More intuition for the prior structure (5.4) can be obtained by expressing $P_{\mathbf{w}}$ in the stick-breaking representation as follows.

$$\begin{aligned} P_{\mathbf{w}} &= \sum_{h=1}^{\infty} \pi_h(\mathbf{w})\delta_{\boldsymbol{\phi}_\mathbf{h}}, \\ \pi_h(\mathbf{w}) &= \Phi(\eta_h(\mathbf{w})) \prod_{\mathbf{l<h}} \Big\{ \mathbf{1} - \boldsymbol{\Phi}(\eta_\mathbf{l}(\mathbf{w})) \Big\}, \\ \eta_h(\mathbf{w}) &= \alpha_h - \sum_{k=1}^{p} \psi_{hk}|w_k - \Gamma_{hk}|, \quad \forall \mathbf{w} \in \mathcal{W}, \end{aligned} \tag{5.5}$$

where $\pi_h(\mathbf{w})$ are random stick-breaking weights, $\boldsymbol{\phi}_h$ are random atoms corresponding to $\{\tau_h^*, \boldsymbol{\theta}_h^*, \Omega_h^*\}$ in model (5.3), and $\Phi(\cdot)$ is cumulative distribution function of the standard normal, $N(0,1)$. The uncertainty for the weights and atoms is allowed through

$$\boldsymbol{\phi}_h \ \sim \ P_0, \quad \alpha_h \sim N(\alpha_h; \mu, 1), \quad \boldsymbol{\psi}_h = \{\psi_{hk}\}_{k=1}^p \sim G, \quad \boldsymbol{\Gamma}_h = \{\Gamma_{hk}\}_{k=1}^p \sim H, \tag{5.6}$$

where P_0 is a known distribution which class-specific parameters follow, G is a known distribution defined on a positive support, and H is a known distribution from which the random locations are drawn. Although $P_{\mathbf{w}}$ is defined as an infinitely discrete distribution, $\pi_h(\mathbf{w})$ decreases toward zero rapidly as h increases and subject allocation tends to happen mostly among the first N mixture components where N is a finite constant such that $\sum_{h=N+1}^{\infty} \pi_h(\mathbf{w}) \approx \mathbf{0}$. Hence, model (5.2) with (5.4) provides a Bayesian specification of the proposed LCT model specified in (5.3). We call this the PSBP mixture of linear mixed effects models (PSBPM-LME) hereafter.

5.2.2 Variable Selection and Hypothesis Testing

Our emphasis is on identifying the predictors of the longitudinal trajectory. If kth predictor has an impact on classifying subjects into different trajectory classes, the predictor should be included in the allocation probability $\pi_h(\mathbf{w})$. Otherwise, we would drop it from the model. This inclusion or exclusion of a predictor in the model can effectively be done introducing a variable selection structure in G as follows.

$$\boldsymbol{\psi}_h = \{\psi_{hk}\}_{k=1}^p \sim G \equiv \prod_{k=1}^p \{1(\omega_{hk} = 0)\delta_0(\psi_{hk}) + 1(\omega_{hk} \neq 0) N_+(\psi_{hk}; \mu_{\psi_k}, \tau_{\psi_k}^{-1})\}, \tag{5.7}$$

where ω_{hk} is the inclusion indicator for kth predictor in hth mixture weight and N+ denotes a truncated normal distribution bounded below by zero. For uncertainty of predictor inclusion, we let

$$\omega_{hk} \sim \text{Bernoulli}(\omega_{hk}; \kappa_{\omega_k}) \quad \text{for} \quad k = 1, \ldots, p \tag{5.8}$$

Note that ω_{hk} controls local inclusion of the kth predictor in the model, with $\omega_{hk} = 0$ implying that $\psi_{hk} = 0$ which leads to excluding the kth predictor from the hth predictor-dependent allocation probability $\pi_h(\mathbf{w})$. Clearly, if $\omega_{hk} = 0$ for $h = 1, \ldots, \infty$, then the kth predictor will be globally excluded from the model playing no role in subject allocation to trajectory classes.

Based on the structure in (5.7) and (5.8), we proposed the following null hypothesis for excluding kth predictor from the model.

$$H_{0k}^{N} : \omega_{hk} = 0 \quad \text{for} \quad h = 1, \ldots, N \tag{5.9}$$

where $N = \max_{i=1}^{n}\{S_i\}$ and S_i is the class which ith subject belongs to. Conceptually, $N = \infty$ makes more sense because the PSBPM-REM assumes infinite number of latent classes. However, using $N = \infty$ is overly restrictive as $\pi_h(\mathbf{x})$ decreases towards zero rapidly as h increases and $\sum_{N+1}^{\infty} \pi_h(\mathbf{x}) \approx \mathbf{0}$ after a finite number N. In fact, following Theorem 1 of Chung and Dunson (2008), we can show that the ratio of likelihoods under H_{0k}^{N} with $N = \infty$ and $N = \max_{i=1}^{n}\{S_i\}$ for the complete data $(\mathbf{Y}, \mathbf{S})$ with $\mathbf{Y} = \{\mathbf{y_i}\}_{\mathbf{i=1}}^{\mathbf{n}}$ and $\mathbf{S} = \{S_i\}_{i=1}^{n}$ does not depend on $(\mathbf{Y}, \mathbf{S})$. This implies the data has no information to distinguish between $N = \infty$ and $N = \max_{i=1}^{n}\{S_i\}$ in H_{0k}^{N}.

5.3 Posterior Computation

5.3.1 MCMC algorithm

We develop an MCMC algorithm for the PSBPM-REM specified in (5.2) with (5.4) where G is chosen as in (5.7) with (5.8). For P_0 and H, we assume

$$\begin{aligned}
\{\tau_h^*, \boldsymbol{\theta}_h^*, \Omega_h^*\} &\sim P_0 \equiv \text{Gamma}(\tau_h^*; a_\tau, b_\tau) \times \prod_{r=1}^{q} N(\theta_{hr}^*; 0, \lambda_r^{-1}) \times \text{Wishart}(\Omega_h^{*-1}; \nu_0, \Omega_0^{-1}) \\
\boldsymbol{\Gamma}_h = \{\Gamma_{hk}\}_{k=1}^{p} &\sim H \equiv \prod_{k=1}^{p} \sum_{m=1}^{M_k} \delta_{\Gamma_{mk}^*}(\Gamma_{hk}),
\end{aligned}$$

where Γ^*_{mk} for $m = 1, \ldots, M_k$ are pre-specified grid values for kth predictor. In addition, we assume $\lambda_r \sim \text{Gamma}(\lambda_r; a_{\lambda_r}, b_{\lambda_r})$ and $\mu \sim N(\mu; \mu_\mu, \tau_\mu^{-1})$. In order to sample finite number of random components for $P_{\mathbf{w}}$, 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 as in Chung and Dunson (2008). 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

$$\begin{aligned} Z_{il} &= 1(Z^*_{il} > 0) \\ Z^*_{il} &\sim N\left(Z^*_{il}; \alpha_h - \sum_{k=1}^{p} \psi_{hk}|w_{ik} - \Gamma_{hk}|, 1\right) \end{aligned} \tag{5.10}$$

For $S_i = T$, we introduce Z^*_{il} only for $l = 1, \ldots, T-1$ because we let $\Phi(\eta_T(\mathbf{w})) = \mathbf{1}$ so that $\sum_{h=1}^{T} \pi_h(\mathbf{w}) = \mathbf{1}$. Given Z^*_{il}, we update $\alpha_h, \boldsymbol{\psi}_h, \boldsymbol{\Gamma}_h$ from their conjugate full conditionals (Steps 2-4). The atoms, $\boldsymbol{\theta}^*_h, \tau^*_h, \Omega^*_h, \boldsymbol{\beta}_i$, and other hyperparameters are also updated from their conjugate full conditionals (Steps 5-10). Finally, we update ω_{hk} and based on the marginal likelihoods for $\mathbf{y} = \{\mathbf{y_i}\}_{\mathbf{i=1}}^{\mathbf{n}}$ and $\mathbf{S} = \{S_i\}_{i=1}^n$, respectively (Step 11).

5.3.2 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, $\mu_{\psi_k} = 0, \tau_{\psi_k} = 1$ for $j = 1, \ldots, p$. For P_0, $a_{\lambda_r} = b_{\lambda_r} = 0.5$, $a_\tau = b_\tau = 0.5$, $\nu_0 = 2, \Omega_0 = 0.1\mathbf{I}$. For H, we choose 50 equally spaced grid points for Γ^*_{mk} in (-2.5, 2.5) for all k. For others, $\kappa_{\omega_k} = 0.5$ for all k and $\mu_\mu = 0, \tau_\mu = 1$. For truncation, we let $T = 20$ which was shown to be large enough because N tends to converge to a small number (≤ 10).

5.4 Simulation Study

In order to illustrate the method, we conduct a simulation study. For $i = 1, \ldots, n$, the predictors $\mathbf{w_i} = (\mathbf{w_{i1}}, \ldots, \mathbf{w_{ip}})'$ are generated as

$$w_{ik} \overset{iid}{\sim} \text{Uniform}(w_{ik}; -2, 2)$$

We consider equally-spaced time points as $t_{ij} = j$ for $j = 1, \ldots, J$ and standardize t_{ij} prior to analysis. Then, the response is obtained for the following cases.

Case (1) $\quad y_{ij} = \mathbf{x'_{ij}}\boldsymbol{\beta}_{\mathbf{i}} + \epsilon_{\mathbf{ij}}, \quad \epsilon_{\mathbf{ij}} \sim \mathbf{N}(\mathbf{0}, \tau^{-\mathbf{1}}), \quad \boldsymbol{\beta}_{\mathbf{i}} \sim \mathbf{N_q}(\boldsymbol{\theta}, \boldsymbol{\Omega}),$

$\quad \mathbf{x_{ij}} = (\mathbf{t^3_{ij}}, \{\mathbf{t_{i3}} + (\mathbf{t_{ij}} - \mathbf{t_{i3}})_+\}^{\mathbf{3}})', \quad \tau = \mathbf{1}, \quad \boldsymbol{\theta} = (\mathbf{5}, -\mathbf{5})', \quad \boldsymbol{\Omega} = [\mathbf{0.5}, \mathbf{0.1}\,; \mathbf{0.1}, \mathbf{0.2}]$

Case (2) $\quad y_{ij} = \mathbf{x'_{ij}}\boldsymbol{\beta}_{\mathbf{i}} + \epsilon_{\mathbf{ij}}, \quad \epsilon_{\mathbf{ij}} \sim \mathbf{N}(\mathbf{0}, \tau^{*-\mathbf{1}}_{\mathbf{S_i}}), \quad \boldsymbol{\beta}_{\mathbf{i}} \sim \mathbf{N_q}(\boldsymbol{\theta}^*_{\mathbf{S_i}}, \boldsymbol{\Omega}^*_{\mathbf{S_i}}), \quad \Pr(\mathbf{S_i} = \mathbf{h}) = \sum_{\mathbf{h}=\mathbf{1}}^{\mathbf{4}} \frac{\mathbf{1}}{\mathbf{4}} \delta_{\mathbf{h}}$

If $S_i = 1, \quad \mathbf{x_{ij}} = \mathbf{t_{ij}}, \quad \tau = \mathbf{1}, \quad \boldsymbol{\theta} = -\mathbf{5}, \quad \boldsymbol{\Omega} = \mathbf{0.2}$

If $S_i = 2, \quad \mathbf{x_{ij}} = (\mathbf{t^3_{ij}}, (\mathbf{t_{ij}} - \mathbf{t_{i2}})^{\mathbf{3}})', \quad \tau = \mathbf{1}, \quad \boldsymbol{\theta} = (\mathbf{5}, -\mathbf{2})', \quad \boldsymbol{\Omega} = [\mathbf{0.1}, \mathbf{0.1}\,; \mathbf{0.1}, \mathbf{0.1}]$

If $S_i = 3, \quad \mathbf{x_{ij}} = (\mathbf{t^2_{ij}}, (\mathbf{t_{ij}} - \mathbf{t_{i4}})^{\mathbf{2}})', \quad \tau = \mathbf{1}, \quad \boldsymbol{\theta} = (\mathbf{0}, \mathbf{5})', \quad \boldsymbol{\Omega} = [\mathbf{0.2}, \mathbf{0.0}\,; \mathbf{0.0}, \mathbf{0.2}]$

If $S_i = 4, \quad \mathbf{x_{ij}} = (\mathbf{t^2_{ij}}, (\mathbf{t_{ij}} - \mathbf{t_{i2}})^{\mathbf{2}})', \quad \tau = \mathbf{1}, \quad \boldsymbol{\theta} = (\mathbf{0}, \mathbf{5})', \quad \boldsymbol{\Omega} = [\mathbf{0.2}, \mathbf{0.0}\,; \mathbf{0.0}, \mathbf{0.2}]$

Case (3) $\quad y_{ij} = \mathbf{x'_{ij}}\boldsymbol{\beta}_{\mathbf{i}} + \epsilon_{\mathbf{ij}}, \quad \epsilon_{\mathbf{ij}} \sim \mathbf{N}(\mathbf{0}, \tau^{*-\mathbf{1}}_{\mathbf{S_i}}), \quad \boldsymbol{\beta}_{\mathbf{i}} \sim \mathbf{N_q}(\boldsymbol{\theta}^*_{\mathbf{S_i}}, \boldsymbol{\Omega}^*_{\mathbf{S_i}})$

If $w_{i1} < -0.7$, $\quad$ same as $S_i = 1$ in case (2)

If $-0.7 < w_{i1} < 0.7$, $\quad$ same as $S_i = 2$ in case (2)

If $w_{i1} > 0.7$ and $w_{i2} < 0$, $\quad$ same as $S_i = 3$ in case (2)

If $w_{i1} > 0.7$ and $w_{i2} < 0$, $\quad$ same as $S_i = 4$ in case (2)

For all cases, $n = 500, p = 5, J = 8$ were commonly chosen. Case (1) is a null case where all subjects belong to one trajectory class with individual variability (Figure 5.1). In case (2) and (3), there exist 4 trajectory groups (linear and polynomial splines) and they are not related to predictors in case (2) while w_1 and w_2 are the predictors of trajectories.

After standardizing y_{ij}, we applied the model (5.2) with (5.4) using the choices of priors and hyperparameters discussed in section 2 and 3 and the following basis functions for $\mathbf{x_{ij}}$.

$$\mathbf{x_{ij}} = (\mathbf{t_{ij}}, \mathbf{t_{ij}^2}, \mathbf{t_{ij}^3}, (\mathbf{t_{ij}} - \mathbf{t_{i2}})^\mathbf{3}, (\mathbf{t_{ij}} - \mathbf{t_{i4}})^\mathbf{3})'$$

The MCMC algorithm described in section 3.1 was run for 1,000 iterations, with the first 500 iterations discarded as burn-ins. The MCMC chain appeared to converge rapidly and to mix efficiently.

In case (1), $\Pr(H_{0k}^N|\text{Data}) = 0.77, 0.79, 0.79, 0.78, 0.78$ for $k = 1, \ldots, 5$ showing that none of the predictors is related to the trajectories. Figure 5.1 shows the simulated (left) and estimated (right) individual trajectories (black) with population mean (red) of case (1). Figure 5.2 shows individual trajectories for 4 different groups in case (2) and (3). We obtained $\Pr(H_{0k}^N|Data) = 0.64, 0.64, 0.63, 0.67, 0.58$ in case (2) while $\Pr(H_{0k}^N|Data) = 0.00, 0.00, 0.58, 0.63, 0.70$ in case (3). This implies that the method detects the predictors w_1 and w_2 of trajectories well among other candidate predictors. Figure 5.3 shows the estimated trajectories classified by w_1 and w_2 in case (2) (left 4 panels) and case (3) (right 4 panels) suggesting that the method clusters the overtime trajectories well based on the important predictors w_1 and w_2.

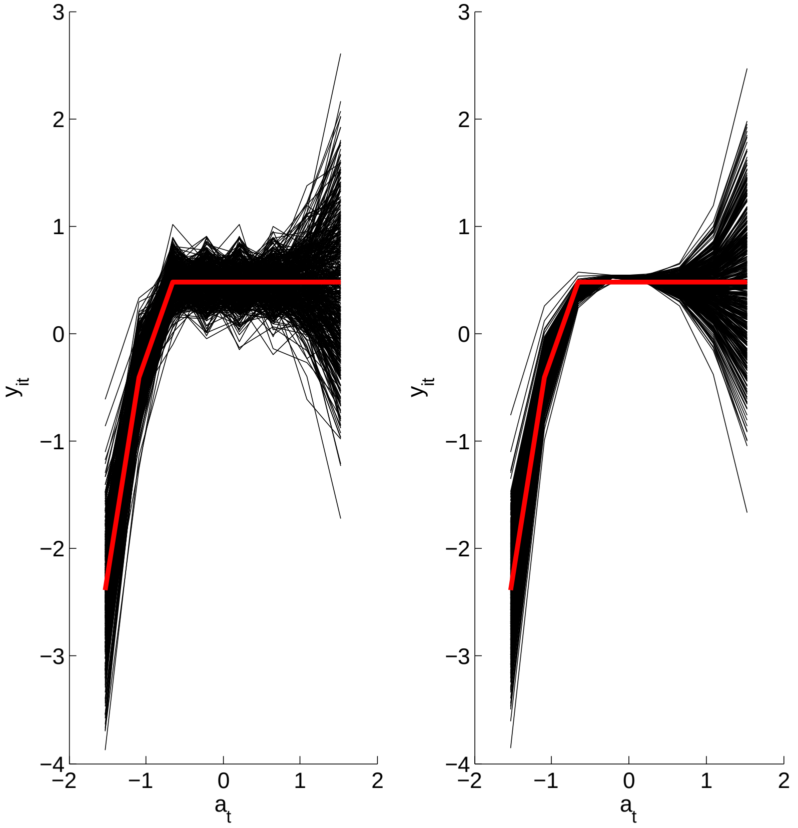

FIGURE 5.1: Case (1); simulated trajectory (left), estimated trajectory (right)

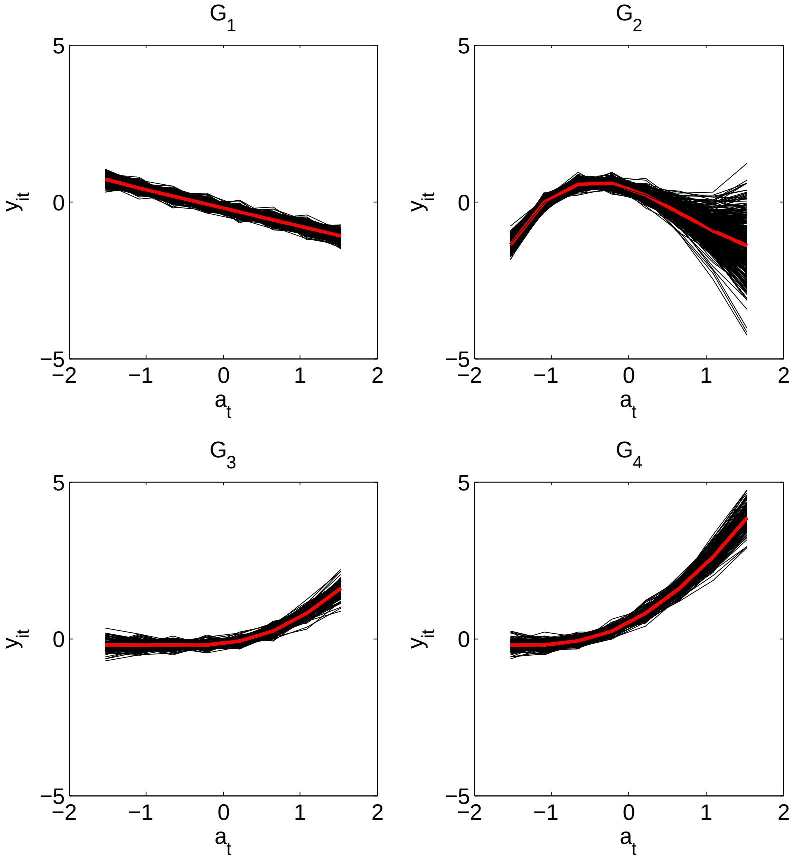

FIGURE 5.2: Case (2) and (3); simulated trajectories for 4 different groups. For case (3), $G_1 : w_1 < -0.7$, $G_2 : -0.7 < w_1 < 0.7$, $G_3 : w_1 > 0.7$ and $w_2 < 0$, $G_4 : w_1 > 0.7$ and $w_2 > 0$

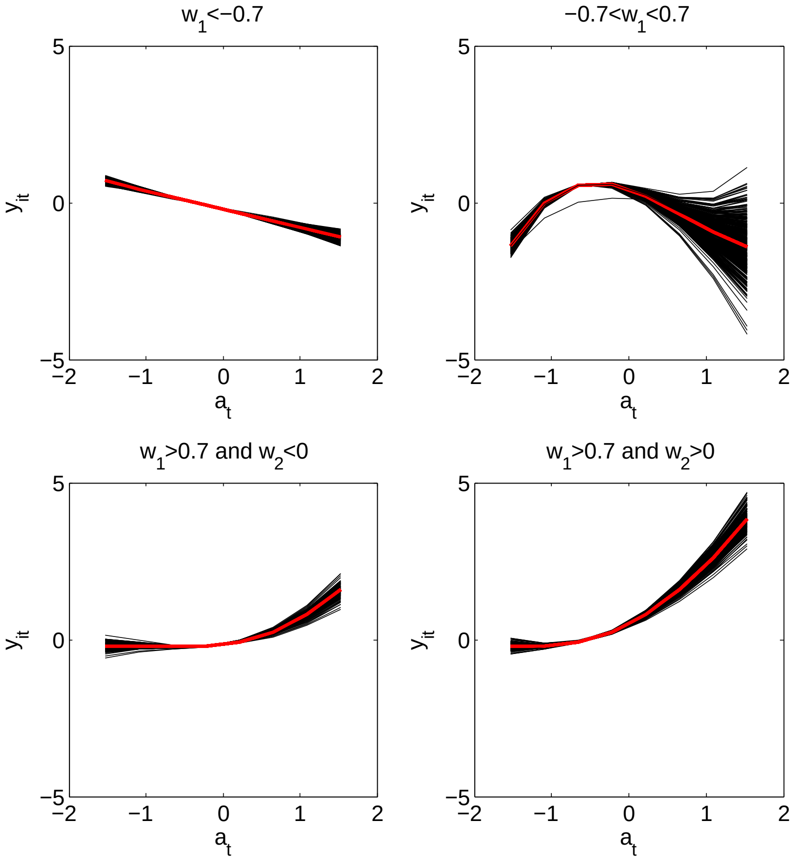

FIGURE 5.3: Case (3); estimated trajectories depending on w_1 and w_2

CHAPTER 6

CONCLUDING REMARKS AND FUTURE RESEARCH

The goal of this research is to develop nonparametric Bayes methodology for studying the relationship between a continuous response and predictors in a very flexible way. Primary focuses are on (1) estimating the conditional response distributions as flexibly changing across the predictor space, (2) testing hypothesis for identifying important predictors having effects on the response distribution both globally and within local regions of the predictor space. Although literature in nonparametric conditional distribution modeling is rich in frequentist framework and more recently in Bayesian framework, the contributions of this research are (1) to obtain a more intuitive and simpler approach so that the methods should be practically useful in many applications, (2) to introduce a formal hypothesis testing procedure for variable selection in conditional distribution modeling which has been addressed with limitations in scope by existing methods, (3) to extend the approach to a general setting where measurements are repeatedly obtained per subject and correlated within subject.

In chapter 3, we proposed a new stick-breaking prior for the collection of predictor-dependent random probability measures. The prior, called the lDP, was shown to be a useful alternative to recently developed prior models that induce predictor-dependence among distributions. The marginal DP structure of lDP should be useful in considering theoretical properties, such as posterior consistency and rates of convergence. The lDP is also appealing in that the construction is intuitive and leads to simple and interpretable expressions for the dependence in random measures at different locations, while also leading to straightforward posterior computation relying on truncation with a fair amount of accuracy. Such desirable properties of lDP were illustrated in conditional distribution modeling framework through lDP mixture of normal linear regressions.

In chapter 4, we proposed a more general approach for conditional distribution modeling with variable selection. The probit stick-breaking process (PSBP) was introduced as a new choice of prior for an uncountable collection of predictor-dependent random probability measures. The PSBP was shown to be a well-defined flexible prior for the dependent probability measures and particularly convenient for posterior computation and incorporating variable selection structure. We considered the PSBP mixture (PSBPM) of normal linear regressions for modeling the conditional distributions incorporating variable selection structure in both regression coefficients and mixing weights. Such structure allowed predictors to drop out of the model or to be included in the model so that the predictors' effects can be formally assessed both globally and locally using posterior inclusion probabilities. Using data augmentation technique, we developed an efficient stochastic search variable selection (SSVS) algorithm. Although we only illustrated continuous predictor cases in chapter 4, the method was shown to be generalized to incorporate categorical predictors.

In chapter 5, we extended the paper 2 method for longitudinal data setting where the response is measured over time per subject and considering within-subject dependence is desirable. Adding random effects in each mixture component, we considered the PSBPM of linear mixed effects (LME) model instead of the PSBPM of normal linear regressions. The PSBPM of LME model characterized the response distribution as predictor-dependent mixture of LME model which accounted for individual variability within each cluster. A variable selection structure was incorporated in the model allowing for formal testing of predictors' effects on the response distribution features such as mean or quantiles. In addition, using the fact that the model embeds a simple LME model as a special case, we proposed a formal testing of goodness of fit

(GOF) for a LME model.

We have shown that the proposed methods in chapters 3,4,5 performed well in various simulation cases and provided interesting results in epidemiological applications. However, there are a number of problems to be addressed for the methods to be more reliable and practical. Specific to each chapter, the issues are summarized as follows.

- Chapter 3

  - The hyperparameter ψ plays an important role and a fixed constant assumption made in paper 1 is restrictive. The method should be improved such that uncertainty for ψ is allowed particularly in an adaptive way that the neighborhood size can differ at different predictor regions depending on the data richness and sparsity (e.g. $\psi_{\mathbf{x}}$) or each mixture component can have its own neighborhood size depending on its location on the space $\mathcal{X}'$ (e.g. ψ_h).

  - Relying on truncation approach for posterior computation approximates the infinite probability measure into a finite one. This can be improved using other approaches which avoid both truncation and marginalization. Such methods include retrospective sampling (Papaspiliopoulos and Roberts, 2007) and slice sampling (Walker, 2007)

- Chapter 4

  - The method should be improved such that implementing with high-dimensional predictors is feasible and reliable results can be obtained although relatively small samples are available.

  - Efficient approaches should be developed for formal hypothesis testing of interactions among the predictors and for identifying local regions of high-dimensional predictor spaces across which the response distribution changes.

- Chapter 5

  - The method should be improved to incorporate random effect selection along with fixed effect selection. As the marginal likelihoods for model comparison are not available in closed forms with variable selection structure for random effects, approximation techniques may make the computation more straightforward without involving complicated MCMC sampling techniques. Allowing random effect selection overcomes the limitation of current approach where the predictors with mixed effects cannot be tested for their effects on other response distribution features than the marginal mean.

In addition, more general to nonparametric Bayes methodology, the following issues can be listed.

- Improvement for computational time is needed, in particular, for high-dimensional settings which become common in many applications with the development of technology to generate complex data.

- Prior specification is of concern given that nonparametric Bayes approach is infinitely parameterized and requires a number of hyperparameters to be specified. It has been shown that results can be sensitive to the choice of hyperparameters in many cases.

- Improvement for mixing of the MCMC chain is important, in particular, for variable selection or hypothesis testing problem where one often has multi-modal posterior distributions and it is hard to prevent the chain from staying in local modes.

- Developing formal methods for hypothesis testing is important for various research questions that can arise in highly flexible nonparametric model (e.g. model selection, assessing the goodness of fit for parametric models, testing for interactions among the predictors in conditional distribution modeling)

APPENDIX A

Proofs in Chapter 3

Proof of Lemma 1

An infinite number of locations $\mathbf{\Gamma} = \{\Gamma_h, h = 1, \ldots, \infty\}$ are generated from H on $\mathcal{X}'$. Any ψ-neighborhood of $\mathbf{x}$ defined as $\eta_{\mathbf{x}}^{\psi} = \{\mathbf{x}' : d(\mathbf{x}, \mathbf{x}') < \psi,\, \mathbf{x}' \in \mathcal{X}'\}$ with $\psi > 0$ is a subset of $\mathcal{X}'$. The regularity condition 1 for H ensures that there is a positive probability for a location Γ_h to be generated in any $\eta_{\mathbf{x}}^{\psi}$. Therefore, there are also an infinite number of locations in $\eta_{\mathbf{x}}^{\psi}$ for all $\mathbf{x} \in \mathcal{X}$ and $\psi > 0$, which implies $N(\mathbf{x}) = \infty$. Then, $\sum_{l=1}^{N(\mathbf{x})} p_l(\mathbf{x})$ almost surely by lemma 1 in Ishwaran and James (2001).

Proof of Theorem 1

Assume that $\mathcal{G}_{\mathcal{X}} \sim lDP(\alpha, G_0, H, \psi)$. Then, from the definition of the lDP in (5)-(7), we can reexpress (7) as $G_{\mathbf{x}} = \sum_{l=1}^{N(\mathbf{x})} V_l^{(\mathbf{x})} \prod_{j<l}(1 - V_j^{(\mathbf{x})})\delta_{\theta_l^{(\mathbf{x})}}$, where $V_l^{(\mathbf{x})}$ is the lth element of $\mathbf{V}(\mathbf{x})$ and $\theta_l^{(\mathbf{x})}$ is the lth element of $\mathbf{\Theta}(\mathbf{x})$. Note that it follows from the proof of Lemma 1 that $N(\mathbf{x}) = \infty$. Since the random weights and atoms are generated by iid sampling from Beta$(1, \alpha)$ and G_0, respectively, independently from the location, we have $V_l^{(\mathbf{x})} \overset{iid}{\sim}$ Beta$(1, \alpha)$ independently from $\Theta_l^{(\mathbf{x})} \overset{iid}{\sim} G_0$, for $l = 1, \ldots, \infty$. Hence, it follows directly from Sethuraman's (1994) representation of the DP, that $G_{\mathbf{x}} \sim DP(\alpha G_0)$, $\forall \mathbf{x} \in \mathcal{X}$.

Proof of Theorem 2

Given $\mathbf{\Gamma}$ and $\mathbf{V}$,

$$\begin{aligned} Pr(\phi_i = \phi_j | \mathbf{x}_i, \mathbf{x}_j, \mathbf{\Gamma}, \mathbf{V}, \psi) &= \sum_{\{(k,l):\pi_k(\mathbf{x}_i)=\pi_l(\mathbf{x}_j)\}} p_k(\mathbf{x}_i) p_l(\mathbf{x}_j) \\ &= \sum_{h \in \mathcal{L}_{\mathbf{x}_i} \cap \mathcal{L}_{\mathbf{x}_j}} V_h^2 \prod_{m \in \mathcal{S}_h} (1 - V_m)^2 \prod_{n \in \mathcal{S}'_h} (1 - V_n) \end{aligned}$$

For the definition of $\mathcal{S}_h$ and $\mathcal{S}'_h$, refer to the equation (3.9) in section 3.2. Marginalizing out $\mathbf{V}$ over the Beta distribution,

$$Pr(\phi_i = \phi_j | \mathbf{x}_i, \mathbf{x}_j, \mathbf{\Gamma}, \alpha, \psi) = \frac{2}{(\alpha+1)(\alpha+2)} \sum_{h \in \mathcal{L}_{\mathbf{x}_i} \cap \mathcal{L}_{\mathbf{x}_j}} \left(\frac{\alpha}{\alpha+2}\right)^{\#\mathcal{S}_h} \left(\frac{\alpha}{\alpha+1}\right)^{\#\mathcal{S}'_h}$$

In order to marginalize out $\mathcal{S}_h$ and $\mathcal{S}'_h$, we introduce $Z_{\gamma_j} \overset{iid}{\sim}$ Bernoulli $(P_{\mathbf{x}_i,\mathbf{x}_j})$ as described in the formulations from (3.9) through (3.10) in section 3.2. Then,

$$Pr(\phi_i = \phi_j | \mathbf{x}_i, \mathbf{x}_j, \mathbf{\Gamma}, \alpha, \psi) = \frac{2}{(\alpha+1)(\alpha+2)} \sum_{j=1}^{\infty} Z_{\gamma_j} \left(\frac{\alpha}{\alpha+2}\right)^{\sum_{k=1}^{j-1} Z_{\gamma_k}} \left(\frac{\alpha}{\alpha+1}\right)^{j-1-\sum_{k=1}^{j-1} Z_{\gamma_k}}$$

After marginalizing out the $\{Z_{\gamma_j}\}_{j=1}^{\infty}$ as in the proof of theorem 3, we obtain:

$$Pr(\phi_i = \phi_j | \mathbf{x}_i, \mathbf{x}_j, \alpha, \psi) = \left[\frac{2}{(\alpha+1)(\alpha+2)}\right] \left[\frac{P_{\mathbf{x}_i,\mathbf{x}_j}(\alpha+2)(\alpha+1)}{\alpha(1+P_{\mathbf{x}_i,\mathbf{x}_j})+2}\right] = \frac{2P_{\mathbf{x}_i,\mathbf{x}_j}}{(1+P_{\mathbf{x}_i,\mathbf{x}_j})\alpha+2}$$

Proof of Theorem 3

From (3.10),

$$Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B) | \mathbf{\Gamma}\} = \frac{2}{\alpha+2} \sum_{j=1}^{\infty} Z_{\gamma_j} \left(\frac{\alpha+1}{\alpha+2}\right)^{\sum_{k=1}^{j-1} Z_{\gamma_k}} \left(\frac{\alpha}{\alpha+1}\right)^{j-1},$$

where Z_{γ_j} are iid draws from Bernoulli$(P_{\mathbf{x}_1,\mathbf{x}_2})$. Taking expectation of $\{Z_{\gamma_j}\}_{j=1}^{\infty}$ with respect to

Bernoulli$(P_{\mathbf{x}_1,\mathbf{x}_2})$,

$$E[Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B)\}] \;=\; \frac{2}{\alpha+2} P_{\mathbf{x}_1,\mathbf{x}_2} \sum_{j=1}^{\infty} \left(\frac{\alpha}{\alpha+1}\right)^{j-1} E\left[\left(\frac{\alpha+1}{\alpha+2}\right)^{Y_j}\right],$$

where $Y_j \sim$ Binomial$(j-1, P_{\mathbf{x}_1,\mathbf{x}_2})$. Using the Binomial Theorem, the expectation on the right is marginalized out with respect to Binomial$(j-1, P_{\mathbf{x}_1,\mathbf{x}_2})$, which results in

$$Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B)\} = \frac{2}{\alpha+2} P_{\mathbf{x}_1,\mathbf{x}_2} \sum_{j=1}^{\infty} \left[\frac{-\alpha P_{\mathbf{x}_1,\mathbf{x}_2}}{(\alpha+2)(\alpha+1)} + \frac{\alpha}{\alpha+1}\right]^{j-1}$$

Since $|\frac{-\alpha P_{\mathbf{x}_1,\mathbf{x}_2}}{(\alpha+2)(\alpha+1)} + \frac{\alpha}{\alpha+1}| \leq 1$, the infinite sum on the right converges. Then,

$$Corr\{G_{\mathbf{x}_1}(B), G_{\mathbf{x}_2}(B)\} = \left(\frac{2P_{\mathbf{x}_1,\mathbf{x}_2}}{\alpha+2}\right)\left(\frac{(\alpha+2)(\alpha+1)}{\alpha(1+P_{\mathbf{x}_1,\mathbf{x}_2})+2}\right) = \frac{2P_{\mathbf{x}_1,\mathbf{x}_2}(\alpha+1)}{(1+P_{\mathbf{x}_1,\mathbf{x}_2})\alpha+2}$$

Proof of Theorem 4

Due to the marginal DP property and using the inequality on the left in (3.11) with n=1, we get $||\mu_N(\mathbf{x}) - \mu_\infty(\mathbf{x})|| \leq 4\Bigg(1 - E\Bigg[\Bigg(\sum_{h=1}^{N(\mathbf{x})-1} p_h\Bigg)\Bigg]\Bigg)$, where μ_N, μ_∞, N in (3.11) are replaced by $\mu_N(\mathbf{x})$, $\mu_\infty(\mathbf{x})$, $N(\mathbf{x})$, respectively, and n is substituted by 1. Here, $N(\mathbf{x})$ is random differently from N in (3.11). Conditioned on $N(\mathbf{x})$ but marginalizing out p_h, we get $||\mu_N(\mathbf{x}) - \mu_\infty(\mathbf{x})|| \leq 4E\Bigg[\Bigg(\frac{\alpha}{1+\alpha}\Bigg)^{N(\mathbf{x})-1}\Bigg]$. Note that $N(\mathbf{x}) \sim$ Binomial$(N, P_{\mathbf{x}})$ as discussed in section 3.3. Then, using the Binomial Theorem, we obtain $||\mu_N(\mathbf{x}) - \mu_\infty(\mathbf{x})|| \leq 4\Bigg(\frac{\alpha+1}{\alpha}\Bigg)\Bigg[1 - \Bigg(\frac{1}{\alpha+1}\Bigg)P_{\mathbf{x}}\Bigg]^N$.

APPENDIX B

Proofs in Chapter 4

Proof of Lemma 1

Following the proof of Lemma 1 for the KSBP (Dunson and Park, 2008), $\sum_{h=1}^{\infty}\pi_h(\mathbf{x}) = \mathbf{1}$ 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 $\log\{1-\Phi(\eta_h(\mathbf{x}))\} \leq \mathbf{0}$, 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:

$$l(\mathbf{y},\mathbf{S}|\boldsymbol{\Lambda}) = \int \prod_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{y_i}|\mathbf{x_i},\boldsymbol{\theta}_{\mathbf{S_i}}) \prod_{\mathbf{h=1}}^{\infty}(\boldsymbol{\theta}_{\mathbf{h}}|\boldsymbol{\gamma}_{\mathbf{h}})\mathbf{d}\boldsymbol{\Theta} \times \int \prod_{\mathbf{i=1}}^{\mathbf{n}}(\mathrm{S}_{\mathbf{i}}|\mathbf{x_i},\Xi) \prod_{\mathbf{h=1}}^{\infty}(\boldsymbol{\xi}_{\mathbf{h}}|\boldsymbol{\gamma}_{\mathbf{h}})\mathbf{d}\Xi \tag{B.1}$$

Because $S_i \leq N$, we reexpress $(B.1)$ as:

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

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

$\mathbf{\Lambda}_{+}^{N} = \{\boldsymbol{\gamma}_h\}_{h>N}$. Then,

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

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

MCMC algorithm

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

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

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

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

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

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

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

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

$$\boldsymbol{\beta}^*_{\gamma_h,h} \sim N(\boldsymbol{\beta}^*_{\gamma_h,h}; [\tau^*_h \mathbf{X}'_{\gamma_\mathbf{h},\mathbf{h}}\mathbf{X}_{\gamma_\mathbf{h},\mathbf{h}} + \boldsymbol{\Sigma}^{\mathbf{-1}}_{\gamma_\mathbf{h},\mathbf{h}}]^{\mathbf{-1}}[\tau^*_\mathbf{h}\mathbf{X}'_{\gamma_\mathbf{h},\mathbf{h}}\mathbf{y_h}], [\tau^*_\mathbf{h}\mathbf{X}'_{\gamma_\mathbf{h},\mathbf{h}}\mathbf{X}_{\gamma_\mathbf{h},\mathbf{h}} + \boldsymbol{\Sigma}^{\mathbf{-1}}_{\gamma_\mathbf{h},\mathbf{h}}]^{\mathbf{-1}}),$$

where $\mathbf{X}_{\gamma_\mathbf{h},\mathbf{h}}$ 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 $p_{\gamma_h} = \sum_{j=1}^p \gamma_{hj}$,

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

7. Update g :

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

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

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

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

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

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 $h = 1, \dots, T$:

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

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

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

APPENDIX C

Proofs in Chapter 5

MCMC algorithm

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

$$\Pr(S_i = h) = \frac{\pi_h(\mathbf{w_i})\mathbf{N}(\mathbf{y_i}; \mathbf{X_i}\boldsymbol{\beta}_\mathbf{i}, \tau_\mathbf{h}^{*-\mathbf{1}}\mathbf{I_{n_i}})\mathbf{N}(\boldsymbol{\beta}_\mathbf{i}; \boldsymbol{\theta}^*_\mathbf{h}, \boldsymbol{\Omega}^*_\mathbf{h})}{\sum_{h=1}^T \pi_h(\mathbf{w_i})\mathbf{N}(\mathbf{y_i}; \mathbf{X_i}\boldsymbol{\beta}_\mathbf{i}, \tau_\mathbf{h}^{*-\mathbf{1}}\mathbf{I_{n_i}})\mathbf{N}(\boldsymbol{\beta}_\mathbf{i}; \boldsymbol{\theta}^*_\mathbf{h}, \boldsymbol{\Omega}^*_\mathbf{h})}$$

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 $W^*_{ih} = Z^*_{ih} + \sum_{k=1}^p \psi_{hk}|w_{ik} - \Gamma_{hk}|$.

3. Update ψ_{hk} for $k = 1, \dots, p$ and $h = 1, \dots, T-1:$ If $\omega_{hk} = 0$, $\psi_{hk} = 0$. If $\omega_{hk} = 1$,

$$\psi_{hk} \sim N_+(\psi_{hk};\, [\tau_{\psi_k} + \sum_{i:S_i \geq h} |w_{ik} - \Gamma_{hk}|^2]^{-1}[\tau_{\psi_k}\mu_{\psi_k} + \sum_{i:S_i \geq h} |w_{ik} - \Gamma_{hk}|U^*_{ih}], [\tau_{\psi_k} + \sum_{i:S_i \geq h} |w_{ik} - \Gamma_{hk}|^2]^{-1}),$$

where $U^*_{ih} = \alpha_h - Z^*_{ih} - \sum_{s=1, s\neq k}^p \psi_{hs}|w_{is} - \Gamma_{hs}|$.

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

$$\Pr(\Gamma_{hk} = \Gamma^*_{mk}) = \frac{\frac{1}{M_k}\prod_{i:S_i \geq h} N(Z^*_{ih}; \alpha_h - \sum_{s=1,s\neq k} \psi_{hs}|w_{is} - \Gamma_{hs}| - \psi_{hk}|w_{ik} - \Gamma^*_{mk}|, 1)}{\sum_{m=1}^{M_k} \frac{1}{M_k}\prod_{i:S_i \geq h} N(Z^*_{ih}; \alpha_h - \sum_{s=1,s\neq k} \psi_{hs}|w_{is} - \Gamma_{hs}| - \psi_{hk}|w_{ik} - \Gamma^*_{mk}|, 1)}$$

5. Update $\boldsymbol{\theta}_h^*$ for $h = 1, \ldots, T:$ With $n_h = \sum_{i=1}^n 1(S_i = h)$,

$$\boldsymbol{\theta}_h^* \sim N(\boldsymbol{\theta}_h^* \quad ; \quad [n_h \Omega_h^{*-1} + \boldsymbol{\Sigma}_0^{-1}]^{-1} [\Omega_h^{*-1} \sum_{i:S_i=h} \boldsymbol{\beta}_i + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\theta}_0], [n_h \Omega_h^{*-1} + \boldsymbol{\Sigma}_0^{-1}]^{-1})$$

where $\boldsymbol{\theta}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0$ is a diagonal matrix with λ_r^{-1} diagonal elements.

6. Update τ_h^* for $h = 1, \ldots, T:$ With $k_h = \sum_{i:S_i=h} n_i$,

$$\tau_h^* \quad \sim \quad \text{Gamma}(\tau_h^*; a_\tau + \frac{k_h}{2}, b_\tau + \frac{1}{2} \sum_{i:S_i=h} (\mathbf{y_i} - \mathbf{X_i} \boldsymbol{\beta}_\mathbf{i})' (\mathbf{y_i} - \mathbf{X_i} \boldsymbol{\beta}_\mathbf{i})$$

7. Update Ω_h^* for $h = 1, \ldots, T:$ With $n_h = \sum_{i=1}^n 1(S_i = h)$,

$$\Omega_h^{*-1} \sim \text{Wishart}(\Omega_h^{*-1}; \nu_0 + n_h, [\Omega_0 + \sum_{i:S_i=h} (\boldsymbol{\beta}_i - \boldsymbol{\theta}_h^*)(\boldsymbol{\beta}_i - \boldsymbol{\theta}_h^*)']^{-1})$$

8. Update $\boldsymbol{\beta}_i$ for $i = 1, \ldots, n:$

$$\boldsymbol{\beta}_i \sim N_q(\boldsymbol{\beta}_i; [\Omega_{S_i}^{*-1} + \tau_{S_i}^* \mathbf{X_i' X_i}]^{-\mathbf{1}} [\boldsymbol{\Omega}_{\mathbf{S_i}}^{*-\mathbf{1}} \boldsymbol{\theta}_{\mathbf{S_i}}^* + \tau_{\mathbf{S_i}}^* \mathbf{X_i' y_i}], [\boldsymbol{\Omega}_{\mathbf{S_i}}^{*-\mathbf{1}} + \tau_{\mathbf{S_i}}^* \mathbf{X_i' X_i}]^{-\mathbf{1}})$$

9. Update λ_r for $r = 1, \ldots, q:$

$$\lambda_r \sim \text{Gamma}(\lambda_r; a_{\lambda_r} + \frac{1}{2} T, b_{\lambda_r} + \frac{1}{2} \sum_{h=1}^T \theta_{hr}^{*2})$$

10. Update $\mu:$

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

11. Update ω_{hk} for $k = 1, \ldots, p$ and $h = 1, \ldots, T:$

$$Pr(\omega_{hk} = 1) \quad = \quad \frac{a_{hk}}{a_{hk} + b_{hk}},$$

$$
\begin{aligned}
a_{hk} &= \kappa_{\omega_k} \times \int \prod_{i:S_i \geq h, S_i \neq T} N(Z^*_{ih}; \alpha_h - \sum_{k=1}^{p} \psi_{hk} |w_{ik} - \Gamma_{hk}|, 1) N_+(\psi_{hk}; \mu_{\psi_k}, \tau^{-1}_{\psi_k}) d\psi_{hk} \\
b_{hk} &= (1 - \kappa_{\omega_k}) \times \prod_{i:S_i \geq h, S_i \neq T} N(Z^*_{ih}; \alpha_h - \sum_{s=1, s\neq k}^{p} \psi_{hs} |w_{is} - \Gamma_{hs}|, 1)
\end{aligned}
$$

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.

Beal, M., Ghahramani, Z. and Rasmussen, C. (2001). The infinite hidden markov model. In *Neural Information Processing Systems (NIPS) Conference*, British Columbia, Canada.

Berger, J. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association* **453**, 174–184.

Bigelow, J. and Dunson, D. (2008). Bayesian semiparametric joint models for functional predictors. *Journal of the American Statistical Association* to appear.

Bishop, C. and Svensen, M. (2003). Bayesian hierarchical mixtures of experts. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, Sanfrancisco : Morgan Kaufmann.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via polya urn schemes. *The Annals of Statistics* **1**, 353–355.

Blei, D. and Jordan, M. (2006). Variational inference for dirichlet process mixtures. *Bayesian Analysis* **1**, 121–144.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric bayesian model for randomized block designs. *Biometrika* **88**, 275–285.

Chaudhuri, P. (1995). Nonparametric estimates of regression quantiles and their local bahadur reresentation. *The Annals of Statistics* **2**, 760–777.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.

Chib, S. and Hamilton, B. (2002). Semiparametric bayes analysis of longitudinal data treatment models. *Journal of Econometrics* **110**, 67–89.

Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Conigliani, C., Castro, J. and O'Hagan, A. (2000). Bayesian assessment of goodness of fit against nonparametric alternatives. *Canadian Journal of Statistics* **2**, 327–342.

Dahl, D. and Newton, M. (2007). Multiple hypothesis testing by clustering treatment effects. *Journal of the American Statistical Association* **102**, 517–526.

De Iorio, M., Muller, P., Rosner, G. and MacEachern, S. (2004). An anova model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

De la Cruz-Mesia, R., Quintana, F. and Marshall, G. (2008). Model-based clustering for longitudinal data. *Computational Statistics and Data Analysis* **52**, 1441–1457.

Dowse, K., Zimmet, P., Alberti, G., Bringham, L., Carlin, J., Tuomlehto, J., Knight, L. and Gareeboo, H. (1993). Serum insulin distributions and reproducibility of the relationship between 2-hour insulin and plasma glucose levels in asian indian, creole, and chinese mauritians. *Metabolism* **42**, 1232–1241.

Dunson, D. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association* **100**, 618–627.

Dunson, D. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics* **7**, 551–568.

Dunson, D., Herring, A. and Siega Riz, A. (2008). Bayesian inference on changes in response densities over predictor clusters. *Journal of the American Statistical Association* to appear.

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

Dunson, D. and Peddada, S. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrikca* **95(4)**, 859–874.

Dunson, D.B., H.-A. and Engel, S. (2008). Bayesian selection and clustering of polymorphism in functionally-related genes. *Journal of the American Statistical Association* to appear.

Dunson, D.B., P.-N. and Park, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society, Series B* **69**, 163–183.

Escobar, M. (1994). Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Fan, J., Yao, Q. and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83**, 189–206.

Fan, J. and Yim, T. (2004). A cross validation method for estimating conditional densities. *Biometrika* **91**, 819–834.

Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2**, 615–629.

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.

Gelfand, A., Kottas, A. and MacEachern, S. (2004). Bayesian nonparametric spatial modeling with dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

George, E. and Mcculloch, R. (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.

Ghidey, W., Lesaffre, E. and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics* **60**, 954–953.

Ghosal, S., Ghosh, J. and Ramamoorthi, R. (1999). Posterior consistency of dirichlet mixtures in density estimation. *The Annals of Statistics* **27**, 143–158.

Ghosal, S. and Van der Vaart, A. (2007). Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statistics* **35(2)**, 697–723.

Green, P. and Richardson, S. (2001). Modeling heterogeneity with and without dirichlet process. *Scandinavian Journal of Statistics* **28**, 355–375.

Griffin, J. and Steel, M. (2006). Order-based dependent dirichlet processes. *Journal of the American Statistical Association* **101**, 179–194.

Hansen, B. and Pitman, J. (2000). Prediction rules for exchangeable sequences related to species sampling. *Statistics and Probability letters* **46**, 251–256.

Hyndman, R., Bashtannyk, D. and Grunwald, G. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics* **5**, 315–336.

Ishwaran, H. and Zarepour, M. (2002a). Exact and approximate sum-representations for the dirichlet process. *Canadian Journal of Statistics* **30**, 269–283.

Ishwaran, H. and Zarepour, M. (2002b). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica* **12**, 941–963.

Jacobs, R., Peng, F. and Tanner, M. (1997). A bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks* **10**, 231–241.

Jain, S. and Neal, R. (2007). Splitting and merging components of a nonconjugate dirichlet process mixture model. *Bayesian Analysis* **5**, 1–38.

James, G. and Sugar, C. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* **98**, 397–408.

Jones, B. and Nagin, D. (2007). Advances in group-based trajector modeling and an sas procedure for estimating them. *Sociological Methods & Research* **35**, 542–571.

Kim, S., Tadesse, M. and Vannucci, M. (2006). Variable selection in clustering via dirichlet process mixture models. *Biometrika* **94**, 877–893.

Kleinman, K. P. and Ibrahim, J. (1998). A semiparametric bayesian approach to the random effects model. *Biometrics* **54**, 921–938.

Kottas, A., Branco, M. and Gelfand, A. (2002). A nonparametric bayesian modeling approach for cytogenetic dosimetry. *Biometrics* **58**, 593–600.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Leslie, D., Kohn, R. and Nott, D. (2007). A general approach to heteroscedastic linear regression. *Statistics and Computing* **17**, 131–146.

Lucas, J., Carvalho, C., Wang, Q., Bild, A., Nevins, J. and West, M. (2006). Sparse statistical modeling in gene expression genomics. *Bayesian Inference for Gene Expression and Proteomics* **100**, 155–176.

MacEachern, S. (1994). Estimating normal means with a conjugate style dirichlet process prior. *Communication in Statistics-Simulation and Computation* **23**, 727–741.

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

MacEachern, S., Clyde, M. and Liu, J. (1999). Sequential importance sampling for nonparametric bayes models: The next generation. *Canadian Journal of Statistics* **27**, 251–267.

MacEachern, S. and Müller, P. (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

MacLehose, R., Dunson, D., Herring, A. and Hoppin, J. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology* **18**, 199–207.

Neal, R. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 238–297.

Pennell, M. and Dunson, D. (2006). Bayesian semiparametric dynamic frailty models for multiple event time data. *Biometrics* **62**, 1044–1052.

Pennell, M. and Dunson, D. (2008). Nonparametric bayes testing of changes in a response distribution with an ordinal predictor. *Psychometrika* **64(2)**, 413–423.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158.

Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Annals of Probability* **25**, 855–900.

Quintana, F. (2006). A predictive view of bayesian clustering. *Journal of Statistical Planning* *and Inference* **136**, 2407–2429.

Quintana, F. and Iglesias, P. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society B* **65**, 557–574.

R., K. and Basset, G. (1978). Regression quantiles. *Econometrika* **46**, 35–50.

Ray, S. and Mallick, B. (2006). Funcational clustering by bayesian wavelet methods. *Journal of the Royal Statistical Society, Series B* **68**, 305–332.

Rigby, R. and Stasinopoulous, D. (2005). Generalized additive models for location, scale and shape. *Applied Statistics* **54(3)**, 507–554.

Rodriguez, A., D. D. and Gelfand, A. (2008). The nested dirichlet process (with discussion). *Journal of the American Statistical Association* **103**, 1131–1144.

Rosen, O., J. W. and Tanner, M. (2000). Mixtures of marginal models. *Biometrika* **87**, 391–404.

Rosen, O. and Cohen, A. (2003). Analaysis of growth curves via mixtures. *Statistics in Medicine* **22**, 3641–3654.

Smith, J., Everhart, J., Dickson, W., Knowler, W. and Johannes, R. (1998). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Symposium on Computer Applications and Medical Care*. IEEE Computer Society Press.

Walker, S. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation* **36**, 45–54.

West, M., Muller, P. and Escobar, M. (1994). Hierarchical priors and mixture models with application in regression and density estimation. Technical report, Duke University.

Xing, E., Sharan, R. and Jordan, M. (2004). Bayesian haplotype inference via the dirichlet process. In *Proceedings of the International Conference on Machine Learning (ICML).*

Yu, K. and Jones, M. (1998). Local linear quantile regression. *Journal of the American Statistical Association* **93**, 228–237.

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*. Notrh-Holland/Elsevier.

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **57**, 795–802.